Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de pós-graduação em química

Matheus Ferraz

Enhanced Sampling Simulations and Machine Learning to Address Viral Infections

Recife
2024

Matheus Ferraz

Enhanced Sampling Simulations and Machine Learning to Address Viral Infections

Tese de doutorado apresentada à coordenação do curso de Pós-graduação em Química, como um dos requisitos para a obtenção do título de Doutor em Química.

Área de Concentração: Química teórica e Computacional

Orientador: Dr. Roberto D. Lins

Coorientadora: Dra. Rebecca C. Wade

Recife

2024

Matheus Vitor Ferreira Ferraz

Enhanced sampling simulations and machine learning to address viral infection.

Tese apresentada ao Programa de Pós-Graduação no Departamento de Química Fundamental da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Química.

Aprovada em: 29 de junho de 2023

BANCA EXAMINADORA

Prof. Roberto Dias Lins Neto (Orientador)
Fundação Oswaldo Cruz - PE

Prof. Munir Salomão Skaf
Universidade Estadual de Campinas

Prof. Rafael de Cássio Bernardi
Auburn University

Profa. Ariane Ferreira Nunes Alves
Technische Universität Berlin

Dr. Carlos Henrique Bezerra da Cruz
Fundação Oswaldo Cruz - PE

Folha de aprovação: Inserir a folha de aprovação enviada pela Secretaria do curso de Pós-Graduação. A folha deve conter a **data de aprovação**, estar **sem assinaturas** e em formato **PDF**.

Dedico esta tese à minha família.

ACKNOWLEDGEMENTS

Agradeço ao CNPq e ao DAAD pelo apoio financeiro para realização desse trabalho.

Ao meu orientador, Roberto Lins, agradeço por todo o apoio, estímulo e orientação fornecidos. Sou grato pelo suporte, pela amizade cultivada e pelos conhecimentos compartilhados, que ultrapassaram os limites da ciência e moldaram minha visão de mundo. Agradeço imensamente por todas as discussões enriquecedoras, pelas correções de textos, apresentações e pelas propostas de projetos desafiadores. Lembro-me do primeiro encontro com Roberto onde perguntei se poderia tentar (trabalhar com biofísica computacional), e desde então, sou grato por toda a orientação profissional e trabalhos desenvolvidos. Aqui expresso toda minha admiração e inspiração por Roberto.

À minha orientadora, Rebecca Wade, que me acolheu em seu grupo de pesquisa durante o doutorado sanduíche no Heidelberg Institute for Theoretical Studies. Agradeço imensamente pela paciência, disponibilidade, orientação e valiosas críticas construtivas. Rebecca sempre foi extremamente solícita, contribuindo significativamente para o meu desenvolvimento acadêmico.

À Camilla Adan, minha dupla, por todas as infinitas discussões científicas, colaborações e amizade. Agradeço pelo empenho na obtenção dos dados experimentais desta tese.

A todos os colegas e ex-colegas do BIOMAT, em especial, Bruno, Carlos, Danilo, Emerson, Fernando e Isabelle por todo o apoio, amizade e ricas discussões científicas.

Aos colegas do HITS durante meu doutorado sanduíche, *vielen Dank für alles!* Abraham, Alexandros, Christina, Giulias (North and South) Jonathan, Manuel, Mislav, Tomassio e Marco, por terem me recebido tão bem, por todas as conversas pós-almoço, e conhecimentos trocados. Agradeço especialmente a Stefan por todo o suporte técnico e disponibilidade e a Giulia D'Arrigo por todo o auxílio e discussões nas simulações de τ-RAMD.

À minha família, em especial meu pai Alexandre (*in memoriam*), minha mãe, Carla e minha irmã Amanda, os quais sempre me apoiaram incondicionalmente e proporcionaram um ambiente estável. Agradeço por todo apoio fornecido e disposição frente a qualquer necessidade e terem sido meus maiores incentivadores.

A Rômulo, pelo suporte, incentivo, carinho e principalmente por muita paciência. Sou grato por ter você ao meu lado durante essa jornada.

Aos amigos, pelas risadas e momentos de descontração, Allan, Bia, Catarina, Fabrício,

Júlia, Leo e Marjorie. Apesar da distância física, se fizeram presentes nos últimos anos e sei que sempre posso contar com eles. Agradeço, também, a Bruno Chausse por sempre me salvar em Heidelberg.

A Erico Teixeira e José pelo auxílio e discussões no treinamento das redes neurais artificiais.

À Universidade Federal de Pernambuco, ao Instituto Aggeu Magalhães e Heidelberg Institute for Theoretical Studies pelo suporte técnico e infraestrutura.

Por fim, agradeço a todos que estiveram presentes, direta ou indiretamente, contribuíram para a realização deste trabalho e estiveram presente em minha formação acadêmica.

Ao longo do desenvolvimento desta tese, foram utilizadas as seguintes alocações computacionais: Supercomputador Santos Dummont lotado no Laboratório Nacional de Computação Científica (LNCC - MCTI) e arquitetura computacional do Heidelberg Institute for Theoretical Studies. Os trabalhos contidos nesta tese foram auxiliados pelas seguintes agências: CAPES, CNPq, FACEPE, FIOCRUZ, Klaus Tschira Foundation e DAAD.

"Reality is frequently inaccurate."

Douglas Adams

The restaurant at the end of the universe (1980)

RESUMO

Os surtos virais representam uma grande ameaça à humanidade, ressaltando a necessidade de desenvolvimento de medicamentos antivirais e vacinas. Compreender a estrutura e a dinâmica viral é crucial para avanços nessa área. Nesse sentido, nossa hipótese é que métodos computacionais possam auxiliar na compreensão da estrutura e dinâmica de proteínas virais para o combate de infecções virais. Assim, esta tese tem por objetivo apresentar o desenvolvimento, aplicação e validação de métodos baseados em simulações moleculares e aprendizado de máquina (AM) e está dividida em três partes principais. Na primeira parte, abordamos o cálculo da variação da energia livre de Gibbs (ΔG) de ligação para complexos proteína-proteína. Desenvolvemos uma rede neural artificial que apresentou um erro médio absoluto de 1.5 kcal/mol para a predição da ΔG absoluta em complexos proteína-proteína. Além disso, identificamos a correlação entre a ΔG da ligação da proteína E6 ao complexo E6AP/p53 e o risco de oncogênese mediado pelo Papilomavírus Humano (HPV). Descritores moleculares, como ligações de hidrogênio e energia livre de solvatação, foram correlacionados ao risco de oncogênese, permitindo o desenvolvimento de um protocolo baseado em AM para a predição do potencial oncogênico dos tipos não classificados de HPV. Na segunda parte, propusemos uma abordagem baseada em AM e dinâmica molecular acelerada para o desenho computacional de nanocorpos (Nbs). Como prova de conceito, aplicamos essa abordagem ao desenho de Nbs contra o domínio de ligação ao receptor (RBD) do SARS-CoV-2. Avaliações experimentais demonstraram que um dos Nbs desenhados apresentou alta afinidade de ligação (45 nM) a uma partícula pseudoviral do SARS-CoV-2, comparável à dos anticorpos neutralizantes. Isso demonstra o potencial da geração de biofarmacêuticos desta estratégia. Por fim, abordamos o cálculo da cinética e mecanismo de dissociação em complexos proteína-proteína e proteína-peptídeo, os quais apesar da importância no desenvolvimento de (bio)fármacos têm recebido menor atenção até o momento. Validamos o procedimento de simulação de dinâmica molecular com aceleração aleatória (RAMD) para complexos proteína-peptídeo, obtendo um valor R^2 de 0.86 para o tempo de residência ($1/k_{off}$) de um conjunto de sistemas peptídeo-MHC. Além disso, a RAMD forneceu insights sobre a seletividade de neutralização de um Nb em relação ao RBD do SARS-CoV-1, mas não em relação ao SARS-CoV-2.

Palavras-chaves: Doenças infecciosas. Engenharia de proteínas. Dinâmica molecular. Cinética de ligação. Termodinâmica de ligação.

ABSTRACT

Viral outbreaks represent a significant threat to humanity, highlighting the need for the development of antiviral drugs and vaccines. Understanding viral structure and dynamics is crucial for advancements in this field. In this regard, our hypothesis is that computational methods can assist in comprehending the structure and dynamics of viral proteins for combating viral infections. Thus, this thesis aims to present the development, application, and validation of methods based on molecular simulations and machine learning (ML), divided into three main parts. Firstly, we address the calculation of the Gibbs free energy change (ΔG) upon protein-protein binding. We have developed an artificial neural network that achieved a mean absolute error of 1.5 kcal/mol in predicting the absolute ΔG for protein-protein complexes. Additionally, we identified a correlation between the ΔG of the E6 protein binding to the E6AP/p53 complex and the risk of oncogenesis mediated by Human Papillomavirus (HPV). Molecular descriptors such as hydrogen bonds and solvation free energy have been correlated with oncogenesis risk, enabling the development of an ML-based protocol to predict the oncogenic potential of unclassified HPV types. In the second part, we proposed an approach based on accelerated molecular dynamics and machine learning for the computational design of nanobodies (Nbs). As a proof of concept, we applied this approach to design Nbs targeting the receptor-binding domain (RBD) of SARS-CoV-2. Experimental evaluations demonstrated that one of the designed Nbs exhibited high binding affinity (45 nM) to a SARS-CoV-2 virus-like particle, comparable to that of neutralizing antibodies. This demonstrates the potential of generating biotherapeutics using this strategy. Lastly, we addressed the computation of kinetics and dissociation mechanisms in protein-protein and protein-peptide complexes, which have received less attention despite their importance in (bio)drug development. We validated the random acceleration molecular dynamics simulation (RAMD) procedure for protein-peptide complexes, obtaining an R^2 value of 0.86 for a set of peptide-MHC systems. In addition, RAMD provided insights into the neutralization selectivity of an Nb towards the RBD of SARS-CoV-1 but not SARS-CoV-2.

Keywords: Infectious diseases. Protein engineering. Molecular Dynamics simulations. Binding kinetics. Binding thermodynamics.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

τ	Residence time
Δ G	Gibbs free energy change
k_D	Binding affinity (dissociation constant)
k_{on}	Association rate
k_{off}	Dissociation rate
Å	Angstrom
ϕ	Electrostatic potential
μs	Microsecond
ps	Picosecond
ns	Nanosecond
nm	Nanometer
AB	AdaBoost
Adam	Adaptive moment estimation
AMBER	Assisted Model Building with Energy Refinement
ANN	Artificial neural networks
APBS	Adaptive Poisson-Boltzmann Solver
AS	Alanine scanning
AUC	Area under the curve
BD	Brownian dynamics
BO	Born-Oppenheimer
cAbs	Conventional antibodies
CDR	Complementary determining region

CHARMM	Chemistry at Harvard Macromolecular Mechanics
CHIKV	Chikungunya virus
COM	Center-of-mass
CV	Collective variable
def2-TZVP	Karlsruhe basis set valence triple-zeta polarization
DL	Deep learning
DT	Decision Tree
EDA	Exploratory data analysis
GP	Gaussian Process
GROMOS	GROningen MOlecular Simulation
HPV	Human papillomavirus
HIV	Human immunodeficiency virus
HR	High Risk
KS	Kolmogorov-Smirnov
LDA	Linear Discriminant Analysis
LR	Low Risk
MD	Molecular dynamics
MD-IFP	Molecular dynamics interaction fingerprint
MHC	Major histocompatibility complex
ML	Machine learning
MLP	Multiple layer perceptron
MST	Microscale thermophoresis
NB	Naïve Bayes

NPT	Number of particles, pressure, and temperature
NVT	Number of particles, volume, and temperature
NN	Neural Networks
OPLS	Optimized Potentials for Liquid Simulations
PBC	Periodic boundary conditions
PB	Poisson-Boltzmann
PDB	Protein data bank
PME	Particle Mesh Ewald
PR	Parrinello-Rahman
PSSM	Position-specific score matrix
QDA	Quadratic Discriminant Analysis
RBD	Receptor-binding domain
RAMD	Random acceleration molecular dynamics
REF15	Rosetta energy function 2015
RF	Random Forest
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
ROC	Receiver operating characteristic
SARS-CoV	Severe acute respiratory syndrome virus
SASA	Surface accessible solvent area
SDA	Simulation of Diffusional Association
SGD	Stochastic gradient descent
SPR	Surface plasmon resonance

SVM	Support Vector Machine
vdW	van der Waals
VHH	Variable heavy chain domain of a heavy chain antibody
VLP	Virus-like particle
k-fold CV	k-fold cross validation
MD-IFP	Molecular dynamics interaction fingerprint
K-S	Kolmogorov-Smirnov
SDA	Simulation of Diffusional Association
RBD	Receptor-binding domain
SPR	Surface plasmon resonance
T	Absolute temperature
R	Universal gas constant

CONTENTS

1 INTRODUCTION

Viruses are tiny infectious agents that consist of genetic material (either DNA or RNA) surrounded by a protein coat called a capsid. They are the most common and widespread type of evolutionary entity (NORRBY, 2008), and can cause a wide range of diseases in humans, from mild and self-limiting to severe and potentially deadly. Viral infections have been responsible for an astounding number of deaths in human history. For example, human immunodeficiency virus (HIV) has infected 75 million people and killed approximately 32 million <https://www.who.int/gho/hiv/en/>, and as of June 7, 2023, the Severe acute respiratory syndrome virus (SARS-CoV-2) (the causative agent of COVID-19) has infected over 760 million people and caused almost 7 million deaths <https://covid19.who.int/>. These viral infections have had a significant impact on human history, caused millions of deaths and have led to economic, health, and educational disruptions (NAYAK et al., 2022; PIOT et al., 2001; WALMSLEY; ROSE; WEI, 2021).

Viral outbreaks still pose a serious threat to human and animal populations, driving the ongoing development of antiviral drugs, diagnostic, and vaccines. This development is further enhanced by a detailed understanding of viral structure and dynamics at a molecular level. Structural biology methods like cryo-electron microscopy and tomography allow researchers to observe atomic-level details within large molecular complexes and subcellular compartments (CALLAWAY, 2020b; JIANG; TANG, 2017), and have recently made it possible to observe complete virus replication cycles in detail (KLEIN et al., 2020; TUROŇOVÁ et al., 2020).However, these techniques offer limited insight into thermal fluctuations or mechanistic information, such as the distribution of intermediates along assembly/disassembly pathways (LYNCH et al., 2023). In parallel with advances in these experimental techniques, the development of accurate physical models, efficient algorithms, and high-performance computing infrastructure (PÁLL et al., 2020; LEE et al., 2018; PHILLIPS et al., 2020; MELO; BERNARDI, 2023) has enabled the use of computation, including molecular simulations and machine learning (ML), to study viral structures and enhance our understanding of viruses and their behavior at the molecular level (MACHADO; PANTANO, 2021; JEFFERYS; SANSOM, 2019; HADDEN; PERILLA, 2018; LAINE et al., 2018).

Among computational techniques, molecular dynamics (MD) simulations, which involve integrating Newton's equations of motion to predict the time evolution of molecular systems

(ADCOCK; MCCAMMON, 2006), are known as a "computational microscope" and are widely used in biomolecular simulations. MD can be used to study the dynamics and behavior of viral proteins and other biomolecules, providing insights into mechanisms of viral infections such as permeability, dynamics, and drug binding sites (MARZINEK; HUBER; BOND, 2020; SOÑORA et al., 2021). Particularly, during the COVID-19 pandemic, MD simulations have played a central role in revealing molecular details that are not accessible through experimental techniques (ARANTES; SAHA; PALERMO, 2020), including the dynamics of full-length model of the spike protein (CASALINO et al., 2020), rational drug design campaigns (GOSSEN et al., 2021), the observation of opening of the SARS-CoV-2 spike protein (SZTAIN et al., 2021), characterization of the heparin-induced inhibition of SARS-CoV-2 (PAIARDI et al., 2022), and even the simulation of SARS-CoV-2 in a respiratory aerosol (DOMMER et al., 2021).

Recent advances in MD simulations have enabled not only the understanding of phenomena but also predicting macroscopic properties from molecular systems, such as accurate characterization of biomolecular binding thermodynamics and kinetics. These properties are crucial for therapeutic design, particularly in the context of viral infections, where protein-target interactions are critical in their development, progression, and treatment. The binding affinity (k_D), which is a measure of the strength of the interaction between two molecules, is one of the most commonly used quantitative metrics to describe biomolecular interactions. Thus, several methods have been developed to calculate the binding affinity between proteins and their targets, many of them based on unbinding simulations (WOO; ROUX, 2005; JORGENSEN, 2010) or rigorous statistical-mechanics treatment. Recently, there has been a shift towards computing binding kinetics properties, as drug efficacy correlates better with binding kinetics (particularly dissociation rates, k_{off}) than binding affinity (COPELAND; POMPLIANO; MEEK, 2006). In addition, it has been shown that mutations at protein-protein interfaces associated with diseases significantly affect unbinding kinetics (DAVID et al., 2012). Despite this critical importance and advances in MD simulations, computing binding kinetics and thermodynamics remains challenging due to long biological timescales involved in binding and unbinding phenomena, and in the complex protein dynamics (HOLLINGSWORTH; DROR, 2018; WANG et al., 2022).

In this regards, enhanced sampling methods (KAMENIK; LINKER; RINIKER, 2022) have been developed to simulate biomolecular binding and dissociation processes, and predict the associated binding kinetic and thermodynamics parameters. These sampling methods all have a common characteristic of speeding up specific events (HUBER; TORDA; GUNSTEREN, 1994;

NEAL, 1996). Acceleration can be achieved through various means, such as adding an external potential to the original potential, using restraints to collect statistics at a particular point in the phase space, or changing the system Hamiltonian to match a reference Hamiltonian (BERNARDI; MELO; SCHULTEN, 2015; AHMAD et al., 2022; YANG et al., 2019).

In addition, in computational biochemistry, ML has had a significant impact due to the large amounts of data generated and available data sets. ML has been used to predict properties of interest, analyze and design molecules, and gain insights into biochemical processes (CERIOTTI; CLEMENTI; LILIENFELD, 2021; HUANG; LILIENFELD, 2021; RUDORFF; LILIENFELD, 2021; ANAND et al., 2022; WANG et al., 2022b). DeepMind's Alphafold technology (JUMPER et al., 2021) is an example of a revolutionary method for predicting protein structures with high accuracy. By combining molecular simulations with ML, it is possible to shed light on important biochemical processes, such as the binding kinetics and thermodynamics of protein-targets, and aid in the design of novel therapeutic molecules.

The main goal of this thesis is to develop, validate, and apply computational methodologies that utilize physics-based simulations and ML to understand the molecular basis of infectious diseases, particularly, to simulate the binding kinetics and thermodynamics of proteins and their target, to ultimately rationally design inhibitory molecules.

1.1 ORGANIZATION OF THE THESIS

In addition to the introduction, which also presents the biochemical processes of interest (section 1.2), this thesis contains five other chapters as follows: Chapter 2 provides an overview and the theoretical foundations of the computational methods employed in this work. Chapter 3 relates to the computation of protein-protein Gibbs free energy change (ΔG) of binding, a challenging task in computational biophysics. In this chapter, we have deployed an artificial neural networks (ANN) method to predict the absolute ΔG of binding and applied it to computer-designed aptamers against Chikungunya virus; in addition, we have correlated the ΔG of binding of E6 protein to the E6AP/p53 complex with Human Papillomavirus (HPV)-mediated oncogensis risk. In Chapter 4 we describe a data-centric and enhanced sampling-based approach to the design of artificial nanobodies (Nbs). In Chapter 5, computation of residence time is addressed. We validate the τ-RAMD protocol for computing protein-peptide complex residence time, an overlooked aspect in molecular simulations, and identify molecular factors that determines the structure-kinetics relationship for a Nb binding SARS-CoV-2 S

protein. Chapter 6 presents the final considerations of this thesis.

1.2 BIOCHEMICAL PROCESSES

1.2.1 Thermodynamics and kinetics of protein-target binding

1.2.1.1 Molecular binding thermodynamics

Protein-target binding kinetics and thermodynamics refer to the rates at which proteins bind to their targets and the energy changes associated with this process, respectively. The binding process is influenced by various intermolecular forces, such as electrostatic, van der Waals, and hydrophobic forces. Hydrophobic interactions, in particular, play a critical role in protein-protein binding. From a thermodynamic perspective, the process involves changes in entropy (S) and enthalpy (H) as the proteins interact and form the complex. The noncovalent association between a protein P and its target T in solution, to form the complex PT, can be seen as a one-step reaction:

$$P + T \leftrightharpoons PT \tag{1.1}$$

This reaction can be described by the equilibrium (or association) constant, k_a, defined by:

$$k_a = \frac{[PT]_{eq}}{[P]_{eq}[T]_{eq}} \tag{1.2}$$

Where the square brackets denote the molarity of the species. The reciprocal of k_a is the dissociation constant ($k_D = 1/k_a$) and is defined as the equilibrium constant for the opposite reaction. k_D corresponds to the binding affinity between the two species, and corresponds the target concentration for which an equal probability of bound and unbound protein is achieved. The connection between the equilibrium constants and thermodynamics is provided by the Gibbs free energy of binding (ΔG_B) at constant temperature and pressure (Equation 1.3):

$$\Delta G_B = -k_B T ln(\frac{k_D}{C_0}) \tag{1.3}$$

Where T is the absolute temperature, k_B is the Boltzmann constant, and C_0 is the standard state concentration of 1 mol/L. The binding affinity (k_D) of a protein for its target reflects

the strength of the interaction between the two molecules. Accurate determination of k_D is crucial as it serves as a vital parameter in drug discovery, design, and comprehension of the molecular foundation of protein-target interactions. There are several experimental techniques to determine the k_D, among them, surface plasmon resonance (SPR), isothermal titration calorimetry (ITC), fluorescence polarization, radioligand binding, microscale thermophoresis (MST), and bio-layer interferometry (BLI). For review, see (JARMOSKAITE et al., 2020).

1.2.1.2 How can molecular binding thermodynamics be computed?

Numerical simulations based on the fundamental principles of statistical mechanics have made it possible to reliably determine free energy changes and, therefore, to estimate k_D. Developments in methodology and computational power have made free energy calculations more robust and widely applicable (KING et al., 2021; ARMACOST; RINIKER; COURNIA, 2020). However, accurately estimating the free energy of a system requires overcoming several challenges. The free energy of a system, denoted by F,[1] is a thermodynamic potential whose natural independent variables are those of the canonical ensemble. It can be expressed in terms of the partition function Q and the inverse temperature β, which is divided by Boltzmann's constant k_B, as follows:

$$F = -\beta^{-1} ln Q(N, V, T) \tag{1.4}$$

This equation forms the fundamental connection between thermodynamics and statistical mechanics in the canonical ensemble. Estimating the value of F is equivalent to calculating Q (The state variables N (number of particles),V (volume), and T (temperature) were omitted for clarity). However, evaluating Q is a challenging task. For simplicity, we will consider a classical description of the system in Cartesian coordinates, assuming that the system is at thermodynamic equilibrium. In this case, Q can be expressed as:

$$Q = \frac{1}{h^{3N} N!} \int \int e^{-\beta H(p,r)} dp dr \tag{1.5}$$

[1] Here, F corresponds to the Helmholtz free energy. Unlike the Gibbs free energy (G), which is used to describe a system at constant temperature and pressure, F is used to describe a system at constant temperature and volume. The choice to derive the formalism using Helmholtz free energy is justified by the fact that the Helmholtz free energy is expressed as a function of the internal energy, which can be directly calculated from the positions and velocities of the particles in the system, unlike the enthalpy (a component of G).

Since the integral in Equation 1.5 is $6N$-dimensional and the integrand is positive definite, the calculation of the absolute free energy is only possible in certain cases where an analytical expression for the partition function can be derived. Typically, this applies to small and straightforward systems governed by a simple Hamiltonian, such as the ideal gas or harmonic oscillator (MCQUARRIE, 2000). However, in larger systems with more significant particle interactions, obtaining an analytical formulation of the partition function is generally not feasible.

In practical applications, it is usually sufficient to calculate relative free energies (Equation 1.6):

$$\Delta F_{BA} = F_B - F_A = -\beta^{-1} ln \frac{Q_B}{Q_A} \tag{1.6}$$

Where A and B denote two different states represented by two different Hamiltonians, H_A and H_B, respectively. For example, A and B can refer to distinct conformations of the same molecular system in which the attainable conformational space is confined within the desired regions for a given set of restraints.

When performing molecular simulations to determine differences in free energy, the main challenges revolve around selecting an appropriate Hamiltonian and sampling scheme. The calculation of free energy differences involves three basic components (CHRIST; MARK; GUNSTEREN, 2010):

- Sampling (section 4.2.2.4 and 2.1.2): A method to generate configurations that have a probability density consistent with the experimental conditions to estimate the ensemble average;
- A model Hamiltonian (section 2.1.1): A molecular model used to calculate the energy and forces so that all configurations have the correct relative probability;
- A method to estimate the ΔF (sections 2.1.3) and 2.3: Various approaches can be used to estimate the difference in F, ranging from free energy perturbation and thermodynamic integration to metadynamics and equilibrium free energy differences using non-equilibrium methods (For example, Jarzynski's identity or Crooks fluctuation theorem)(JARZYNSKI, 1997).

In this thesis, the sampling procedure was mostly based on MD simulations using classical force field potentials (Hamiltonian). To estimate free energy differences, enhanced sampling MD and neural networks were utilized.

1.2.1.3 Molecular binding kinetics

High affinity is an important consideration in the design of inhibitors, such as drugs or neutralizers. However, the thermodynamics of protein-target binding alone does not provide a complete understanding of the binding mechanism. The association and dissociation rates of a protein with its target depend on transient interactions with the surrounding environment, which cannot be fully described by a state function like the ΔG_B (VIVO et al., 2016). To this end, binding kinetics can afford an additional understanding of binding for providing out-of-equilibrium parameters.

For the reaction in Equation 1.1, the rate of forward reaction is of second order in reactant concentration and is characterized by the association rate constant (k_{on}), measured in $M^{-1}s^{-1}$. The reverse process is of first order and is characterized by the dissociation rate constant (k_{off}), measured in s^{-1} (Figure 1). The rates of these processes can be combined to provide a phenomenological rate equation for the protein-target binding:

$$\frac{d[PT]}{dt} = k_{on}[P][T] - k_{off}[PT] \tag{1.7}$$

The equation 1.7 is written as the law of mass action. At the equilibrium, when $\frac{d[PT]}{dt} = 0$, the k_D can be expressed as ratio of k_{off} over k_{on}:

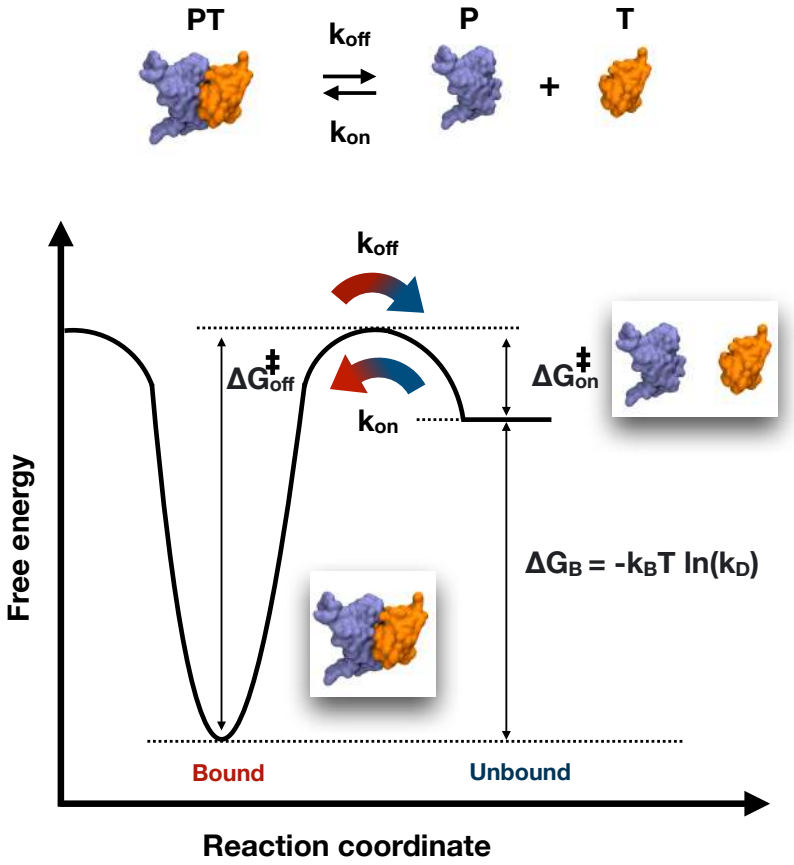
$$k_D = \frac{k_{off}}{k_{on}} \tag{1.8}$$

k_{off} is directly related with the lifetime of the protein-target binding, and as in any other first-order reaction, its reciprocal, referred to as residence time ($\tau = 1/k_{off}$), denotes the average life span of a protein-target complex.

The process of (un)binding can be represented by a double-welled one-dimensional potential mean force (Figure 1). This potential has two minima separated by a high barrier, such that transitions between the two minima (i.e., from a bound state to an unbound state) are rare events compared to the dynamics within each minimum (intrabasin dynamics). The rate constant for these transitions can be calculated using transition state theory by the Eyring equation (Equation 1.9) (ZHOU, 2010):

$$k = k_0 e^{-\Delta G^{\ddagger}/(k_B T)} \tag{1.9}$$

Figure 1 – Simplified one-dimensional potential mean force represented by the free energy as a function of the reaction coordinate for the unbinding process. The one-step reaction for unbinding is shown with the protein, target, and complex structures represented by their molecular surface. The protein-target complex is depicted as a deep free-energy minimum on the left side (the bound state), while the dissociated complex is depicted as a higher-energy minimum on the right side (the unbound state). The free-energy difference between these minima (ΔG_B) is a measure of the thermodynamics of binding, while the dissociation and association rate constants, k_{off} and k_{on}, respectively, determine the kinetics of (un)binding. These rate constants are related to the free-energy barriers between the minima and the transition state, $\Delta G^{\ddagger}_{off}$ and $\Delta G^{\ddagger}_{on}$, respectively.



Source: Reproduced and adapted from Decherchi and Cavalli (DECHERCHI; CAVALLI, 2020)

In the expression for the rate constant for a chemical reaction, $\Delta G^{\ddagger}$ is the activation free energy, which is also known as the free-energy barrier. It represents the energy required to overcome the barrier separating the reactants and products. The constant k_0 is a constant that takes into account the frequency of transition attempts and the probability of recrossing events from the transition state (such as the Arrhenius constant).

1.2.1.4 How can molecular binding kinetics be computed?

When it comes to simulating binding kinetics, encounter events can be used to calculate k_{on} in a simulation of a period box of the relevant solutes. However, this approach is frequently computationally inefficient. Thus, the Northrup Allison McCammon (NAM) algorithm (NORTHRUP; ALLISON; MCCAMMON, 1984) was developed by Northrup et al. for estimating bimolecular association rate constants using Brownian Dynamics simulation (GABDOULLINE; WADE, 1997). The algorithm involves defining two intermolecular distances at the "b-surface" and "q-surface" where the forces between the molecules are centrosymmetric and simulating only the two associating solutes. The Smoluchowski equation is used to calculate the rate constant at which the two molecules at a defined bulk concentration approach the intermolecular distance defined by the b-surface ($k_d(b)$).

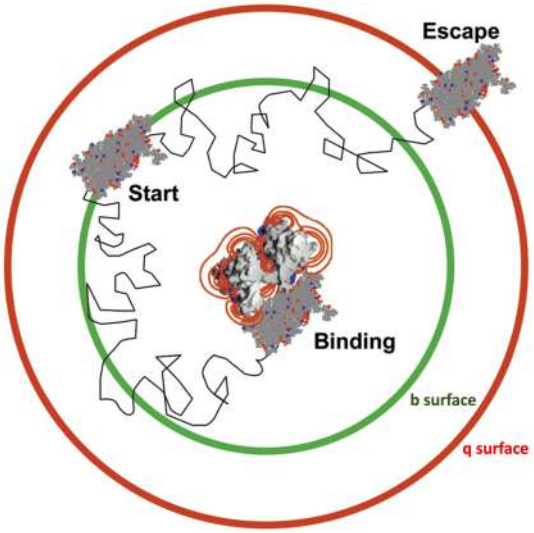
$$k_d(b) = \frac{4\pi D}{\int_0^\infty \frac{e^{\frac{E(r)}{k_B T}}}{r^2} dr} \tag{1.10}$$

This equation equation can be utilized for calculating the rate constant of bimolecular association:

$$k_{on} = k_d(b) \frac{\beta}{1 - (1 - \beta)\frac{k_D(b)}{k_D(q)}} \tag{1.11}$$

The NAM algorithm uses Brownian Dynamics simulation to estimate bimolecular association rate constants. The simulation generates trajectories of two molecules, and if they approach each other based on predefined reaction criteria, the simulation is stopped and counted as a "reaction" trajectory. If the intermolecular distances exceed the q-surface, the simulation may be restarted at the b-surface or stopped based on an analytically calculated probability of escape. The NAM algorithm involves defining two intermolecular distances, the b-surface and q-surface, where forces between molecules are centrosymmetric.

Figure 2 – Schematic illustration of Brownian dynamics simulations in the determination of association rate constants. In each simulation, two proteins are initially positioned at a distance b from each other, represented by a green surface. The simulation focuses on the dynamic diffusion process of one protein, monitoring its movement until it either binds with the other protein or moves away to a greater distance q, depicted by a red surface.



Source: Reproduced from Elcock, Sept, and McCammon (ELCOCK; SEPT; MCCAMMON, 2001).

Despite the efficacy of BD simulations, conventional MD simulations are able to capture spontaneous ligand binding to target proteins and predict corresponding k_{on}, at the cost of very long simulations. Recent studies have successfully captured ligand binding events using tens-to-hundreds of microseconds of conventional MD simulations. For instance, MD simulations have predicted the association rate of Dasatinib drug to its target Src kinase in a classical atomistic simulation of ca 35 μs (SHAN et al., 2011), and also the binding of benzene to the L99A mutant of T4 lysozyme via 0.48 μs via weighted ensemble simulations (RAY; STONE; ANDRICIOAEI, 2021). In addition, microsecond MD simulations have enabled accurate prediction of host-guest binding thermodynamics and kinetics. However, typical small-molecule ligands dissociation events from proteins are hard to be captured in MD simulations (SHAN et al., 2011). Usually, external forced have been used to simulate biomolecular dissociation (BERNARDI et al., 2019).

This highlights the difficulties associated to computing unbinding kinetics. The dissociation rate, which determines the residence time, is more difficult to compute than the binding affinity. Accurately calculating the dissociation rate requires extensive sampling of the transition state, which often involves multiple pathways in the protein-ligand configurational space, unlike the binding affinity that can be estimated using a simpler two-state endpoint approach

(BRUCE et al., 2018). In addition, protein-peptide and protein-protein complexes residence time calculations have received comparatively less attention than protein-small molecule binding. For protein-peptide complexes, large conformational changes of peptides pose an extra challenge in modelling (ROBUSTELLI; PIANA; SHAW, 2020; ZOU et al., 2020). For protein-protein, these are known to have stronger binding affinity compared to protein-small molecule and protein-peptide interactions (WANG et al., 2023), and the binding and unbinding processes occur over significantly longer timescales. Despite these challanges, microsecond simulations have been used to derive unbinding kinetics for protein-peptide and protein-protein complexes (ZWIER et al., 2016; PAN et al., 2019; PAUL et al., 2017).

1.3 SYSTEMS STUDIED

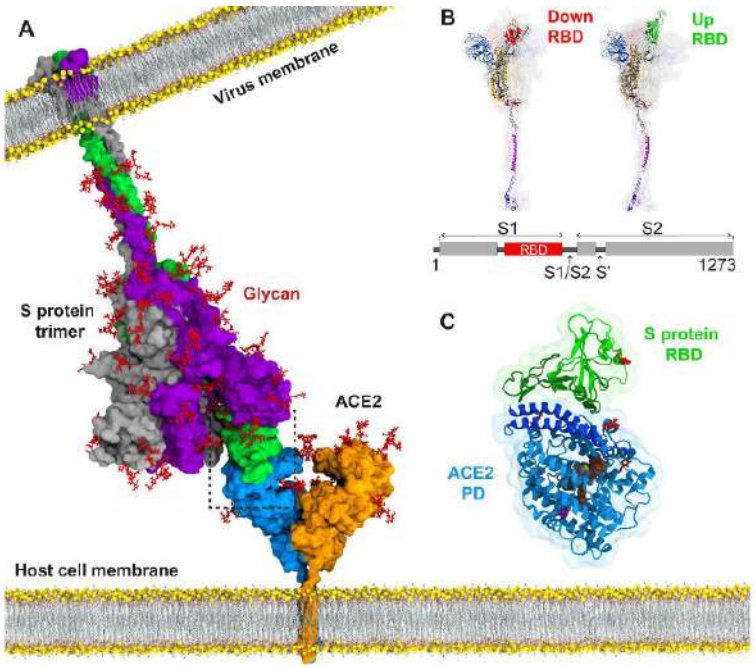
1.3.1 Nanobodies targeting SARS-CoV-2 Receptor Binding Domain

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) belongs to the betacoronavirus family and has caused several waves of coronavirus-induced disease (COVID-19) with profound impacts on the economy and health systems worldwide. One of the most critical targets for vaccines, therapeutics, and diagnostic development against SARS-CoV-2 is the homotrimeric spike (S) glycoprotein. The S protein is involved in receptor recognition, virus attachment, and entry through binding the human angiotensin-converting enzyme (hACE2) in the host cell. It is found exposed on the virion surface and can trigger a succession of adaptive immune responses (MURIN; WILSON; WARD, 2019).

The S protein is comprised of two subunits: S1 and S2. The S1 subunit at the N-terminal region is responsible for virus attachment and contains the receptor-binding domain (RBD) (BAUER et al., 2022), which directly binds to hACE2 and is one of the main targets of nAbs produced from the human immunological response. During the fusion, the S protein undergoes considerable conformational changes, especially around the RBD, in which two main conformations are observed from cryogenic electron microscopy results: the down-state (shielded from receptor binding) and the up-state (receptor accessible) (WRAPP et al., 2020b; WALLS et al., 2020). Upon receptor binding, the unstable 3-up state is induced, shedding the S1 sub-unit, and refolding the S2 to present the fusion peptide towards the opposing membrane (WALLS et al., 2019). To preclude viral entry, high-affinity binding molecules, such as nanobodies and antibodies, bind to, or close to, the receptor binding motif of the RBD. Also

other neutralization mechanisms have been observed, such as locking the RBDs in the down conformation or hindering the fusion (SCHOOF et al., 2020; KOENIG et al., 2021), high-affinity binders are required.

Figure 3 – Atomistic model for the binding between SARS-CoV-2 S protein and ACE2. (A) Full-length S protein complexed with ACE2. S protein is a homotrimer (green, purple, gray), incorporated to the viral membrane. The RBD, in green, interacts with ACE2. (B) Two conformations for the RBD protomer: up and down (C) Interaction between RBD and ACE2



Source: Reproduced from Taka et al. (TAKA et al., 2021)

While monoclonal antibodies have long been crucial in combating viral diseases, the COVID-19 pandemic has exposed limitations in its global production. To overcome this, nanobodies (Nbs) have emerged as an alternative (GÜTTLER et al., 2021). Derived from single-domain antibodies in the Camelidae family, Nbs possess advantages: small size (15 KDa), exceptional refolding capability, and limitless modifications due to their recombinant nature (MUYLDERMANS et al., 2013). These characteristics make Nbs promising for overcoming limitations of traditional antibodies and opening new avenues in the fight against COVID-19 (VALENZUELA-NIETO et al., 2022).

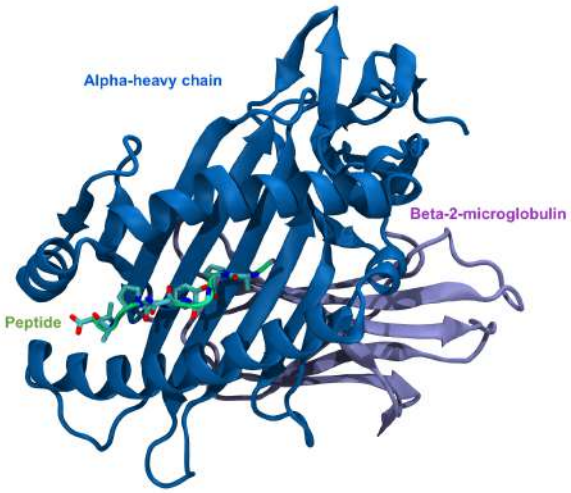
1.3.2 Peptides-MHC Class I

MHC class I molecules are a type of cell surface receptor that present peptides to the immune system (KULSKI et al., 2002; WIECZOREK et al., 2017). Specifically, MHC class I

molecules present peptides derived from intracellular proteins to CD8+ T cells, which are a type of white blood cell that can recognize and eliminate cells that are infected with viruses or other pathogens (ALBERTS et al., 2002). The binding of a peptide to an MHC class I molecule is a crucial step in the process of activating CD8+ T cells, and it allows the immune system to identify and target cells that are infected or otherwise abnormal (HEWITT, 2003; MATSUMURA et al., 1992).

The MHC-I is a heterodimeric protein, composed of two polypeptide chains (Figure 4): an alpha chain, referred to as the heavy chain, and a beta-2 microglobulin chain (JONES, 1997), responsible for stabilizing the complex . The alpha chain is subdivided into alpha 1, alpha 2, and alpha 3 domains, where alpha 1 and alpha 2 form a groove-like structure called the peptide-binding groove, which anchors peptides and interacts with the T-cell receptor for recognition (HATEREN et al., 2010). The alpha chain is encoded by the MHC gene, which is one of the most polymorphic genes in the human genome. This means that there are many different versions of the alpha chain, and each individual has a unique combination of MHC class I molecules on their cells. In the alpha 3 region, besides the non-covalent binding with beta-2 microglobulin, there is the coupling of the accessory molecule of the cytotoxic T cell, CD8. The beta-2m chain is a non-polymorphic protein that is present in all cells, and it helps to stabilize the MHC class I molecule (ANTONIOU; POWIS; ELLIOTT, 2003).

Figure 4 – Representation of the peptide-MHC complex topology. The alpha-heavy chain is depicted in blue, the beta-2-microglobulin in purple, and the peptide in green. The licorice representation of the peptide highlights its atom types, with carbon atoms shown in cyan, nitrogen atoms in blue, and oxygen atoms in red



The space between the two helices forms a groove that accommodates peptides through two main mechanisms: (i) the establishment of conserved hydrogen bonds between the side chains of the MHC molecule and the peptide backbone, and (ii) the positioning of specific peptide side chains, known as anchor residues, within defined pockets. These anchor residues are typically located at positions P2 or P5/6 and PΩ (HUNT et al., 1992; FALK et al., 1991).

1.3.3 E6 and E6AP from the human papillomavirus binding p53

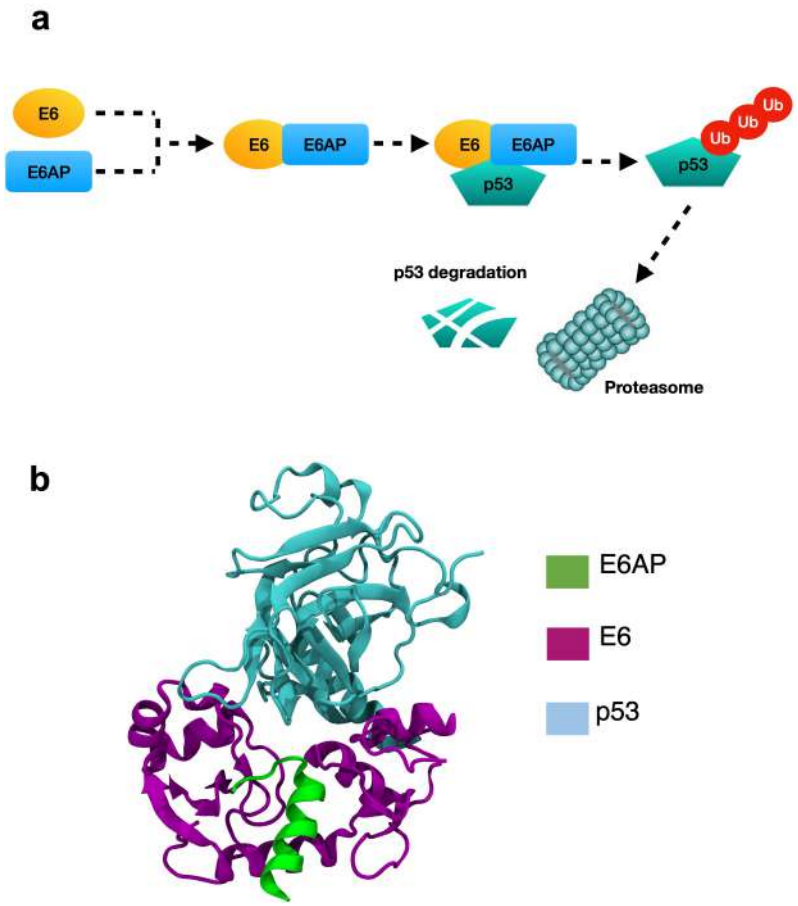
Human papillomaviruses (HPV) are small DNA viruses that infect the mucosal and cutaneous epithelia of vertebrates. They are etiologically associated to to several human cancers and are the second leading cause of cancer-related death among women (MARTEL et al., 2017). While over 120 HPV types have been identified based on sequence similarities, only a small number have been associated with cancer development. Thus, the HPVs can be classified into high and low oncogenic potential risk. Among them, HPV16 and HPV18, are the most common high-risk HPV and are associated with 70% of cervical carcinomas and the majority of HPV-positive head-and-neck cancers (YU; MAJERCIAK; ZHENG, 2022).

HPVs possess oncogenes called E6 and E7, which are highly expressed in HPV-positive

cancers and have been shown to transform cells and induce oncogenic effects in both tissue culture and animal models (HOPPE-SEYLER et al., 2018). These proteins target tumor suppressors p53 and retinoblastoma, respectively, resulting in their degradation via ubiquitin-mediated pathways (Figure 5a). This process leads to cell proliferation, survival, genome instability, and evasion of innate immune responses (HUIBREGTSE; SCHEFFNER; HOWLEY, 1991; HUIBREGTSE; SCHEFFNER; HOWLEY, 1993; DYSON et al., 1989).

HPV E6 oncoproteins have the ability to interfere with the transcriptional activity of p53 and facilitate its degradation. This interaction is mediated by E6-associated protein (E6AP), a member of the HECT E3 ubiquitin ligase family, which serves as a link between E6 and p53 (SCHEFFNER et al., 1993; HUIBREGTSE; SCHEFFNER; HOWLEY, 1993), forming the ternary complex E6/E6AP/p53 (Figure 5b) (MARTINEZ-ZAPIEN et al., 2016). Extensive *in vitro* and *in vivo* studies have elucidated the role of E6AP in the degradation process. The complex formed by E6, E6AP, and p53 enhances the ligase activity of E6AP towards p53, leading to its degradation via the intracellular ubiquitin-proteasome system (MARTINEZ-ZAPIEN et al., 2016; DREWS; BRIMER; POL, 2020; ANSARI; BRIMER; POL, 2012; POL; KLINGELHUTZ, 2013; ZANIER et al., 2013).

Figure 5 – (a) Schematic representation of the mechanism by which the HPV16 E6 protein targets the p53 protein for degradation through the ubiquitin/proteasome pathway. The process begins with the association of the E6 oncoprotein with the ubiquitin-protein ligase E6AP, forming a dimeric complex. This complex then binds to the p53 protein, and E6AP catalyzes multi-ubiquitination of p53 in the presence of ubiquitin and other enzymes of the ubiquitin pathway (b) Representative structure of the proteins in the E6/E6AP/p53 complex in cartoon model (α-helices are shown as spirals; β-strands as arrows and unstructured regions as coils); proteins are color-coded according to labels in the figure (i.e., E6 in magenta, E6AP in green and p53 in cyan).



Source:(a) Adapted from (GHITTONI et al., 2010)

1.4 HYPOTHESIS

Enhanced sampling methods, in combination with physics-based simulations and ML, can provide a powerful tool for predicting binding kinetics and thermodynamics of viral proteins, which is crucial for developing effective therapies and treatments against viral infections.

2 THEORETICAL METHODS

2.1 MOLECULAR DYNAMICS SIMULATIONS

MD is a widely used computational approach in which classical equations of motion are solved numerically to generate trajectories that can be used to estimate macroscopic observables (KARPLUS; MCCAMMON, 2002). The trajectories generated using MD comprise a set of microstates consistent with the canonical distribution, i.e., they produce a sampling of the canonical phase space distribution (TUCKERMAN, 2010).

Quantitatively, one of the most basic goal of MD is to obtain equilibrium expectations for a given property A. A system in a state **x** presents the average value of an observable A given by equation 2.1:

$$\mathbb{E}[A] = \int A(\mathbf{x})\mu(\mathbf{x})d\mathbf{x} \tag{2.1}$$

Where $\mu(\mathbf{x})$ corresponds to the equilibrium distribution (the probability to find a molecule in state **x** at equilibrium conditions). A typical distribution is the Boltzmann distribution in the canonical ensemble at a temperature T (equation 2.2), which calculates the probability of a system being in the state **x** with energy $U(\mathbf{x})$ (MCQUARRIE; SIMON, 1997).

$$\mu(\mathbf{x}) \propto e^{-\frac{U(\mathbf{x})}{k_B T}} \tag{2.2}$$

In MD simulations, the initial state **x** of ensemble is propagated through the phase-space using a physically-motivated propagator via Newton's equations of motion (equation 2.3) (ALDER; WAINWRIGHT, 1959), generating a collection of classical microscopic configurations in a particular equilibrium ensemble.

$$m\frac{d^2\vec{r}}{dt^2} = f = -\nabla E^{Pot}(\vec{r}) \tag{2.3}$$

Where m is the particle mass, and f is the force acting upon the particle and it is independent of the particles' momenta. For any arrangement of the atoms in the system, the force is computed by differentiating a potential energy ($E^{Pot}(\vec{r})$) function that describes the interactions between the atoms (2.1.1). As from the obtained forces, the acceleration can be determined. The integration of equations of motion is carried out numerically using proper

integrators (2.1.1.1) and provides trajectories that describe how the positions, velocities and accelerations of the particles vary with the time, which can be used to compute time averages.

For a long enough MD trajectory, all relevant configurations in the phase-space should be generated (sampled), especially configurations with a well-defined probability distribution (according to equation 2.2, the lower the energy of a state, which can be evaluated as $E^{pot}(\vec{r})$, the higher the probability of the system to be found). According to the ergodic hypothesis, for a system with this property, the expectation value from equation 2.1, which represents experimental averages, can be computed from the time averages obtained via MD simulations.

In practice, the equation 2.3 is commonly modified by introduction of a thermostatting (reviewed in (HÜNENBERGER, 2005)) or/and barostatting (reviewed in the introduction of (HÜNENBERGER, 2002)) to simulate the systems in a given ensemble. The success of configurations generation consistent with the ensemble of choice will heavily depend on the correct description of energies within the system (2.1.1) (LIANG; FOX; BOWEN, 1996) and in the integration of motion's equation.

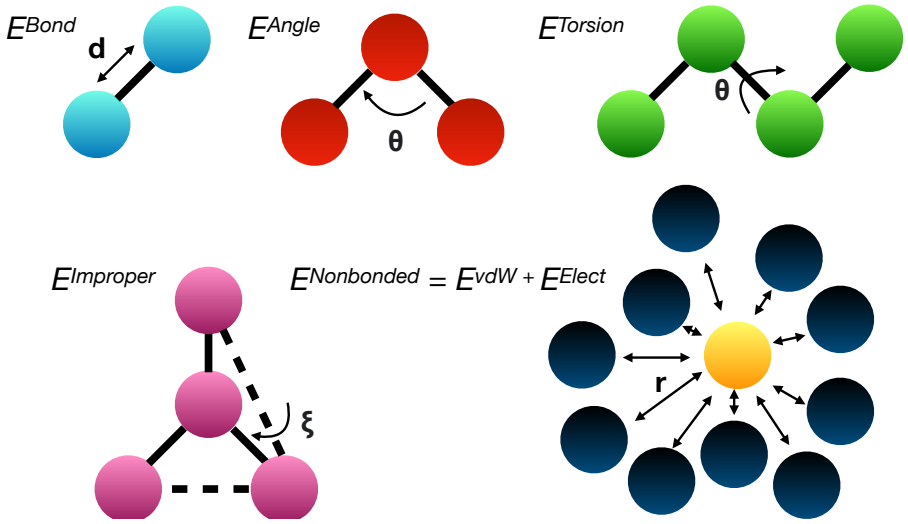
2.1.1 Potential energy

The interactions between the atoms in the system are defined via a mathematical expression that describes the potential energy with respect to the the atomic coordinates referred to as force field. The mathematical formulation for a typical biomolecular force field still uses a similar functional as the one proposed by Levitt and Lifson (LEVITT; LIFSON, 1969), shown in Equation 2.4.

$$E^{pot}(\vec{r}) = E^{Bond}(\vec{r}) + E^{Angle}(\vec{r}) + E^{Torsion}(\vec{r}) + E^{Improper}(\vec{r}) + E^{Elect}(\vec{r}) + E^{vdW}(\vec{r}) \quad (2.4)$$

Where E^{pot} corresponds to the potential energy function of the system with the coordinates defined by $\vec{r}$. The potential energy function is comprised by two group of terms, where the first four terms describe the bonded interactions (potentials for bond stretching, bond-angle bending, dihedral-angle torsion, and improper dihedral-angle bending (or out-of-plane distorions), respectively), and the two latter for the nonbonded ones (terms to describe the Coulomb (elecrostatic) and van der Waals (vdW) interactions). Figure 6 shows a schematic representation of the types of interactions between atoms which are present in almost all force field implementations.

Figure 6 – Schematic illustration of the individual contributions in a classical force field containing the following terms: interactions along the covalent bonds due to the bond stretching (E^{Bond}), the angle flexion and bending (E^{Angle}), dihedral angle torsions ($E^{Torsion}$), improper dihedral angle bending ($E^{Improper}$), and the nonbonded terms ($E^{Nonbonded}$): van der Waals (E^{vdW}), and electrostatic interactions (E^{Elect}).



Source: Adapted from Riniker (RINIKER, 2018)

Since the 1980s, four major families of force fields have been established, Assisted Model Building with Energy Refinement (AMBER) (WEINER et al., 1984), Chemistry at Harvard Macromolecular Mechanics (CHARMM) (BROOKS et al., 1983), GROningen MOlecular Simulation (GROMOS) (GUNSTEREN; KARPLUS, 1982), and Optimized Potentials for Liquid Simulations (OPLS) (JORGENSEN; TIRADO-RIVES, 1988), which are being continuously improved. The basic functional form of these force field families is very similar, however, their technical implementation, parametrization strategies, and philosophies differ among them (RINIKER, 2018). To illustrate the individual contribution to a typical force field, the functional form of the GROMOS 54A7 force field (SCHMID et al., 2011) is presented (Equation 2.5). For a thorough description of each term, see (OOSTENBRINK et al., 2004):

$$E^{Pot}(\vec{r}) = \sum_{n=1}^{N} \frac{1}{4} K_b [b_n^2 - b_{0n}^2]^2 + \sum_{n=1}^{N_\theta} \frac{1}{2} K_{\theta n} [cos\theta_n - cos\theta_0 n]^2 + \sum_{n=1}^{N_\varrho} \frac{1}{2} K_{\varrho n} [\varrho_n - \varrho_{0n}]^2 + \sum_{n=1}^{N_\varphi} \frac{1}{2} K_{\varphi n} [1 + cos(\varphi_n) cos(m_n \varphi_n)] + \sum_{pairsi,j} \left(\frac{C_{12ij}}{r_{ij}^{12}} - \frac{C_{6ij}}{r_{ij}^{6}} \right) + \sum_{pairsi,j} \frac{q_i q_j}{4\pi \ \epsilon_0 \epsilon_1} \frac{1}{r_{ij}} \quad (2.5)$$

The terms describe, successively, bond stretching, bond angle, improper and proper rotations, nonbonded van der Waals interactions, and electrostatic interactions (described by

the Coulomb potential). In practice, the Coulomb term is not used as such, and treatment of electrostatic interactions requires special care (HÜNENBERGER, 1999; DARDEN; YORK; PEDERSEN, 1993b; TIRONI et al., 1995), such as the use of the Ewald, Particle Mesh techniques and the Fast Multipole methods (POLLOCK; GLOSLI, 1996). A comprehensive overview on the electrostatic description for biomolecular systems was published by Cisneros and collaborators (CISNEROS et al., 2014). The parameters of the force field are normally obtained by fitting quantum mechanical or experimental values.

Several approximations enable the successful application of force fields to describe the interactions within a system. First, the Born-Oppenheimer (BO) approximation (BORN; HEISENBERG, 1985), which separates the motion of molecules into two levels: electronic and nuclear, and considers the electrons to be moving in a field of fixed nuclei (SZABO; OSTLUND, 2012). This enables the wave functions of atomic nuclei and electrons in a molecule to be treated separately, allowing the energy of a molecule in the ground electronic state to be calculated as a function of the nuclear coordinates only. The nuclear Hamiltonian is assumed to be time-independent and its potential component can be described by force field energies (equation 2.4).

The second approximation is the use of fractional point charges to represent the charge distribution within a molecule (LEACH, 2001). These charges are intended to reproduce the electrostatic properties of the molecule and are located at the nuclear centers, referred to as partial atomic charges (2.1.1.1). In most MD simulations, the partial atomic charges are kept fixed, which reduces computational cost compared to polarizable force fields, in which the partial charges can change in response to the conformation and environment (JING et al., 2019). Nevertheless, fixed charges can provide reliable results.

2.1.1.1 Deriving partial atomic charges

Partial atomic charges (q) cannot be directly calculated from quantum mechanics, as there are no wavefunctions associated to charges. Thus, there are several ways to calculate these charges. Schemes that calculate charges that are consistent with the electrostatic potential (ϕ) are widely used (SIGFRIDSSON; RYDE, 1998; WANG; CIEPLAK; KOLLMAN, 2000).

The electrostatic potential at a given position corresponds to the interaction energy acting on a unit positive charge located at that position, and it can be calculated from a wavefunction using the equation 2.6:

$$\phi(r) = \phi_{nucl}(r) + \phi_{elec}(r) = \sum_{A=1}^{M} \frac{Z_A}{r - R_A} - \int \frac{\rho(r)}{|r' - r|} dr' \tag{2.6}$$

Where $\rho(r)$ is the electronic charge density at the point r, Z_A is the nuclear charge of the atom A, and R_A is the position vector the atom A. The summation runs over all atoms and the integral runs over all space.

The idea is to derive the set of partial atomic charges that best produces the electrostatic potential at points in the space around the molecule. To this end, Cox and Williams (COX; WILLIAMS, 1981) have proposed to minimize the least-square fitting difference between the quantum mechanical electrostatic potential of equation 2.6 and the potential from the point charges (equation 2.7), complying to the condition that the sum of the charges should be equal to the net charge of the molecule.

$$\chi^2_{esp} = \sum_i (\phi_i^0 - \phi_i^{calc})^2 \tag{2.7}$$

In the works of this thesis, when necessary to calculate partial atomic charges, a modified version of the electrostatic potential method, termed restrained electrostatic potential (RESP) fit (BAYLY et al., 1993), was used. In RESP, restraint hyperbolic functions (equation 2.8) are used on non-hydrogen atoms during the fitting of the partial charges to the electrostatic potentials. This leads to the attenuation of the charges on some atoms, alleviating an issue with the conventional electrostatic potential method which overestimates bond polarities in the gas-phase (CORNELL et al., 2002) and lowers the charge variation as a response to the molecular conformations (REYNOLDS; ESSEX; RICHARDS, 1992).

$$\chi^2_{resp} = \chi^2_{esp} + \chi^2_{restr} \tag{2.8}$$

Where

$$\chi^2_{restr} = k_{restr} \sum_j ((q_j^2 + b^2)^{1/2} - b) \tag{2.9}$$

The tightness of the hyperbola around its minimum is determined by a constant b, whereas the strength of the restraint is determined by a weight k_{restr}.

2.1.2 Stochastic alternative to Newtonian Dynamics

The Newtonian equation of motion, represented by equation 2.3, can be altered to account for the impact of solvent particles in the solute by introducing an extra frictional term and

adding random force to the systematic forces. This introduces a stochastic dynamic process that can be modeled using the Langevin equation (equation 2.10) (MCQUARRIE, 2000). In its basic form, the friction kernel is assumed to be independent of space and time for every particle. This approach represents the influence of the environment on the systematic, internal force in an averaged manner, neglecting explicit hydrodynamic interactions. The internal force is therefore supplemented with a frictional term that is proportional to the velocity and a random force $\vec{R}$ that crudely mimics molecular collisions and viscosity in a realistic cellular environment:

$$m\frac{d^2\vec{r}}{dt^2} = f = -\nabla E^{Pot}(\vec{r}) - m\gamma\frac{d\vec{r}}{dt} + \vec{R}(t) \tag{2.10}$$

The parameter γ in the Langevin equation is known as the damping constant or collision parameter, and it is expressed in reciprocal units of time. The random-force vector $\vec{R}$ is a stationary Gaussian process that has specific statistical properties, which are given by:

$$\left\langle \vec{R}(t) = 0 \right\rangle, \left\langle \vec{R}(t)R(t')^T = 2\gamma k_B T m\delta(t-t') \right\rangle \tag{2.11}$$

The value of γ determines the relative dominance of inertial forces compared to random external forces. The Langevin forces in the stochastic dynamics model mimic the effects of collisions between the solute (biomolecule) and solvent molecules. In the simulations of this thesis, the Langevin dynamics have been used to control the temperature, keeping it constant. In practice, specifically, during the integration of the equations of motion, a fraction of the kinetic energy of each atom is regularly extracted from the system at every time step. This energy is then replaced by a fraction of a random velocity, which follows a known velocity distribution such as the Maxwell-Boltzmann distribution, based on a defined temperature. The relaxation time is a crucial parameter of this dynamic process and dictates how quickly the system reaches the target temperature T.

In Langevin dynamics, the characteristic vibrational frequencies of a molecule in vacuum are dampened, particularly the low-frequency vibrational modes which become overdamped, and various correlation functions are smoothed. The magnitude of these disturbances relative to Newtonian behavior is dependent on the value of γ (BARTH; SCHLICK, 1998). For each particle, a suitable physical value of γ can be determined using Stokes' law for a hydrodynamic particle with a radius a and mass m in a solvent of viscosity η.

$$\gamma = 6\pi\eta a/m \tag{2.12}$$

Increasing γ causes the system to shift from the inertial regime to the diffusive (Brownian) regime. At zero inertia, the resulting dynamics are commonly known as Brownian dynamics (BD). It is noteworthy that Brownian motion, which describes the random movement of solutes in a solvent, was named after Robert Brown, who first observed it in pollen particles. Later, Albert Einstein and Marian von Smoluchowski developed a mathematical theory to describe the diffusive movement of solute particles over time, which showed that the average displacement is proportional to the square root of time.

2.1.2.1 Brownian Dynamics

In BD, since the inertial relaxation times are short compared to the timescale of interest, it is often possible to ignore inertia in the governing equation 2.10, and discards the momentum variables, assuming $m\frac{d^2\vec{r}}{dt^2} = 0$. Thus, from equation 2.10:

$$0 = -\nabla E^{Pot}(\vec{r}) - m\gamma\frac{d\vec{r}}{dt} + \vec{R}(t) \tag{2.13}$$

And using Einstein relatio, $D = k_B T/\gamma$, it is convenient to write the BD equation as:

$$\frac{d\vec{r}}{dt} = -\frac{-D}{k_B T}\nabla E^{Pot}(\vec{r}) + \vec{R}(t) \tag{2.14}$$

BD simulations use larger time steps, typically on the order of picoseconds or longer, compared to molecular and Langevin dynamics simulations, which use timesteps on the order of femtoseconds. These simulations are computationally efficient because molecules are often simulated as rigid bodies, and an implicit solvent treatment is applied.

Equation 2.15 is fundamental to Brownian dynamics BD, as it establishes a correlation between the variation of particle position in one dimension (Δx) and a time interval (Δt) in Brownian motion.

$$\overline{\Delta x^2} = 2D\Delta t \tag{2.15}$$

BD also incorporates intermolecular forces to calculate particle drift. As such, an expanded version of Equation 2.15 is employed to describe the position change ($\Delta\vec{r}$) of a particle during an interval Δt. This extension accounts for external and solute-solute forces, in addition to random "kicks" that simulate stochastic collisions between the solute and solvent particles, represented by a random displacement vector, and hydrodynamic interactions. Ermak and

McCammon (ERMAK; MCCAMMON, 1978) derived a basic BD propagation scheme from the generalized Langevin equation in the high-friction limit (Equation 2.16), while also accounting for hydrodynamic interactions, intermolecular forces, and random displacements at each time step:

$$\Delta\vec{r} = \frac{\Delta t}{k_B T} F(\vec{r}) D + \vec{R} \tag{2.16}$$

In BD simulations, the force, $F(\vec{r})$ is calculated including several types of interactions, such as electrostatic, van der Waals, and desolvation. Electrostatic forces are typically the most significant due to their long-range characteristics.

2.1.2.2 Continuous electrostatics

The treatment of biomolecules in a continuous medium, such as for BD simulations, is based on models developed by Born (BORN, 1920), Onsager (ONSAGER, 1936), and Tanford-Kirkwood (TANFORD; KIRKWOOD, 1957). In these models, the solute is described as a low dielectric constant cavity immersed in a medium with dielectric constant ϵ=78. In this approximation, the protein and the model compound are represented as a rigid object with a dielectric constant ϵ_p=4-6. Implicit solvent models are commonly described using continuous electrostatics treatment. This treatment is based on the numerical solutions to the Poisson-Boltzmann (PB) equation (GILSON et al., 1993; GILSON; SHARP; HONIG, 1988), which combines principles of statistical mechanics (Boltzmann distribution for charge density, ρ) with electrostatic equations (Gauss's law relating electrostatic potential, Φ, to charge density). The PB equation is given by Equation 2.17:

$$\nabla \cdot [\epsilon(x)\nabla\Phi(x)] = -4\pi\rho_{soluto}(x) - 4\pi \sum_{i=1}^{n_i} q_i c_i exp[-q_i\Phi(x)/k_B T)] \tag{2.17}$$

In this equation, ∇ represents the gradient vector, ϵ is the position-dependent dielectric function. This position-dependent description takes into account the difference in polarizability between the macromolecule and the solvent. For a dielectric solution occupying a volume V and containing N_i ions with charge q_i for n_i species i, let $c_i = N_i/V$ be the concentration of ionic species i. The product $q_i\Phi(x)$ is an approximation of the effective potential of mean force for type-i electrolytes at position x, given a specific solute configuration, which is equivalent to the energy required to bring the ion from infinity to position x.

The Debye-Hückel theory addresses the effect of ionic strength in the medium, allowing for the application of the Poisson-Boltzmann equation to systems that can be represented as charges in a medium with a uniform dielectric constant and low ionic strength. To represent the ionic atmosphere of a solute immersed in an aqueous solution and counterions, a linearized approximation of the Poisson-Boltzmann (PB) equation can be employed.

The linearized PB equation (Equation 2.17) is obtained by performing a Taylor series expansion of the Boltzmann factor and truncating it beyond the first-order terms, resulting in Equation 2.18:

$$\exp\left(-\frac{q_i\Phi(x)}{k_BT}\right) \approx 1 - \frac{q_i\Phi(x)}{k_BT} \tag{2.18}$$

The linearized version is valid when the energies involved are much smaller than the thermal energy, which is often the case for monovalent electrolytes in dilute solutions.

The Poisson-Boltzmann equation PB equation is a non-linear elliptical partial differential equation that has closed-form solutions only for very simple geometries. For biomolecular applications, numerical solutions, such as finite difference techniques, are necessary due to the complex shapes involved (WANG et al., 2009). The solution of PB equation has been extensively utilized to investigate various biological processes. These include predicting pKa values (ALEKSANDROV; ROUX; JR, 2020), computing solvation and binding free energies (NGUYEN; WANG; WEI, 2017; SWANSON; HENCHMAN; MCCAMMON, 2004; JEAN-CHARLES et al., 1991), and studying protein folding and design (MARSHALL; VIZCARRA; MAYO, 2005).

2.1.3 Enhanced sampling simulations

For many phenomena of interest, if high free energy barriers or kinetic traps are present, it may not be possible to observe rare events or slow processes within the typical simulation time. This is because generating long enough molecular dynamics (MD) trajectories to compute the expectation value in equation 2.1 through direct averaging becomes infeasible, and enhanced sampling techniques may be required. An example of such a phenomenon is the simulation of the dissociation of proteins and their binding targets to obtain kinetics and thermodynamics parameters (as described in section 1.2.1) as the experimental timescales necessary to observe unbinding can reach minutes or even days.

Enhanced sampling techniques are used to increase the rate of conformational sampling

and escape from local minima in a system. These methods can be divided into two categories: those that require a predefined reaction coordinate (collective variable (CV)-based methods) (VIVO et al., 2016), and those that do not (Hamiltonian-based methods). CVs (2.1.3.1) are used as reaction coordinates to summarize the behavior of the entire system and a bias is applied to these coordinates during the simulation. This reduces the phase space to the space of the CVs and results in a dimensional reduction of the free energy surface. On the other hand, Hamiltonian-based enhanced sampling techniques involve modifying the energy function of the system to allow for more efficient sampling of the phase space (TORRIE; VALLEAU, 1977; SUGITA; OKAMOTO, 1999).

2.1.3.1 Collective variables

A CV is a differentiable function of the atomic coordinates, and is used to describe the collective behavior of molecular systems, such as its overall conformation, orientation, or shape. It should accurately reflect the state of the simulated system, including any metastable states (LAIO; GERVASIO, 2008).

The most general definition of a CV, ξ, is as a differential function of a vector of 3N atomic Cartesian coordinates, $\mathbf{X}$:

$$\xi(\mathbf{X}) = \xi(x_1, x_2, ..., x_N) \tag{2.19}$$

Depending on the structure of the system, $\xi(\mathbf{X})$ is often a function of significantly fewer arguments than 3N, or it can be expressed as a combination of such functions:

$$\xi(\mathbf{X}) = \xi(z^1(\mathbf{X}), z^2(X0, ..., z^\alpha(\mathbf{X})...) \tag{2.20}$$

With a much smaller number of base functions $z^\alpha(\mathbf{X})$ than the number of atoms. We refer to $z^\alpha(\mathbf{X})$ as a component of the CV: in the simplest and most common scenarios, a single component z sufficient to represent the CV (FIORIN; KLEIN; HÉNIN, 2013). Typically, the CVs are a function of geometric terms such as distances, angles, dihedrals, or purely chemical terms as coordination number or the total potential energy of the simulation box (TIANA, 2008).

The equilibrium behavior of these variables is completely defined by the probability distribution

$$P(s) \propto \int e^{-U(\mathbf{X})/k_B T} \delta(s - s(\mathbf{X})) \tag{2.21}$$

Where s are the CVs to be biased, $s(\mathbf{X})$ is the value of the CV at a time t, and $U(\mathbf{X})$ is the internal energy. The probability can be expressed in energy units as a free energy landscape $F(s)$:

$$F(s) = -k_B T log(P(s)) \tag{2.22}$$

2.1.3.2 Metadynamics

In Metadynamics, the system's evolution is influenced by a potential that is dependent on its history and is constructed by summing Gaussian functions (HUBER; TORDA; GUNSTEREN, 1994) along the trajectory in the collective variable (CV) space. After an initial adjustment period, the bias potential balances the free energy landscape and gives an estimate of its dependence on the CVs. The concept behind metadynamics is to strategically fill the free energy minima of a metastable state with bias potentials (as described in equation 2.26) in order to facilitate exploration of other states within the energy landscape.

The system's Hamiltonian, H, can be augmented with a bias potential, V_{Bias}, in order to enhance the system's sampling by discouraging revisiting of already sampled. The bias potential,V_{Bias}, is continuously updated by adding a bias at rate ω and depends on collective variables. The update process is mathematically represented as:

$$\frac{\partial V_{Bias(s)}}{\partial t} = \omega \delta(s - s(\mathbf{X})) \tag{2.23}$$

As the simulation time, t_{sim}, increases, the accumulated bias potential converges to the free energy with opposite sign:

$$V_{Bias}(s) = \int_0^{t_{sim}} \omega \delta(|s - s(\mathbf{X})|) dt \Rightarrow F(s) = \lim_{t_{sim} \to \infty} V_{Bias}(s) + C \tag{2.24}$$

For computational efficiency, the update process is discretized into time intervals, λ, and the delta function is replaced with a localized positive kernel function, K. The bias potential then becomes a sum of kernel functions centered at the instantaneous collective variable values at the time λ_j:

$$V_{Bias}(s) \approx \lambda \sum_{j=0}^{\lfloor \frac{t_{sim}}{\lambda} \rfloor} \omega K(|s - s(\mathbf{X})|) \tag{2.25}$$

Typically, the kernel used is a multi-dimensional Gaussian function, where the covariance matrix has diagonal non-zero elements only:

$$V_{bias}(s) \approx \lambda \sum_{j=0}^{\lfloor \frac{t_{sim}}{\lambda} \rfloor} \omega exp\left(-\frac{1}{2}\left|\frac{s - s(\mathbf{X})}{\sigma}\right|^2\right) \tag{2.26}$$

Where λ, ω, and σ are determined *a priori* and kept constant during the simulation.

Despite its success, there is a need for improvement in several aspects of Metadynamics. Firstly, determining when to end a Metadynamics simulation can be challenging. In a single run, the free energy estimate fluctuates around the correct value and the average error is proportional to the square root of the rate at which the bias potential is deposited (BUSSI; LAIO; PARRINELLO, 2006; LAIO et al., 2005). Slowing down this rate reduces the error but increases the time needed to fill the free energy landscape. Additionally, continuing the simulation for too long carries the risk of pushing the system into regions of configuration space that are not physically meaningful (MICHELETTI; LAIO; PARRINELLO, 2004; GERVASIO; LAIO; PARRINELLO, 2005; ENSING; KLEIN, 2005; WU; SCHMITT; CAR, 2004; BABIN et al., 2006).

To address these issues, a variant of metadynamics referred to as well-tempered metadynamics (BARDUCCI; BUSSI; PARRINELLO, 2008) has been devised. Well-tempered metadynamics has been inspired by the self-healing umbrella sampling method (MARSILI et al., 2006), In conventional Metadynamics, Gaussian functions with unchanging heights are continually added throughout the simulation. This leads the system to investigate high-free energy areas, and the estimate of free energy obtained from the bias potential fluctuates around the actual value. In contrast, Well-Tempered Metadynamics adjusts the height of the Gaussian functions over time according to the equation 2.27:

$$W = \omega_0 exp\left(-\frac{V(s(\mathbf{X}))}{k_B \Delta T}\right) \tag{2.27}$$

Where ω_0 is an initial Gaussian height, ΔT is an input parameter with the dimension of a temperature. With the adjustment of the Gaussian height, the bias potential gradually converges in the long run, but it does not fully balance the underlying free energy. The rate at which $V(s(\mathbf{X}), t)$ changes is given according to equation 2.28:

$$V(s(X), t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(s) + C \tag{2.28}$$

In this equation, T represents the temperature of the system. Over a long period of time, the CVs sample an ensemble at a temperature $T + \Delta T$, which is higher than the system

temperature T. The parameter ΔT can be adjusted to control the extent of free energy exploration. A ΔT value of 0 corresponds to standard MD, while ΔT approaching infinity corresponds to standard metadynamics (SCHAFER; SETTANNI, 2020). An important concept in well-temperature metadynamics is the bias factor (γ), which represents the ratio between the temperature of the CVs ($T + \Delta T$) and the system temperature (T) (Equation 2.29).

$$\gamma = \frac{T + \Delta T}{T} \tag{2.29}$$

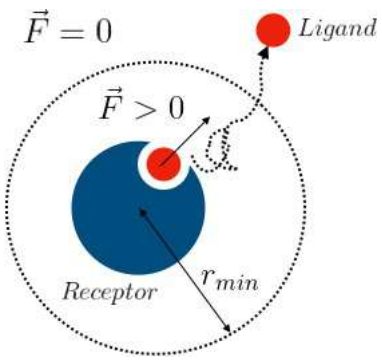
The bias factor parameter enables the adjustment of the region within the free energy landscape to be explored. A low bias factor may not sufficiently enhance sampling, while a γ of 1 indicates no metadynamics bias at all. Conversely, a high bias factor expands the region to explore, potentially making convergence more challenging.

2.1.3.3 τ-random acceleration MD (τ-RAMD)

RAMD is a method to study how ligands dissociate from the binding pockets of receptors. (LÜDEMANN; CARUGO; WADE, 1997; LÜDEMANN; LOUNNAS; WADE, 2000). It assumes generating many trajectories results in a representative set of motions for the true unbinding process. It works by applying a randomly oriented force of constant magnitude on the center of mass of a ligand to simulate its exit (Figure 7). The movement of the ligand is then monitored over N simulation steps, and if the change in its position is less than a set threshold distance (r_{min}), the direction of the force is reassigned. If the change is greater, the simulation continues with the same force direction for another N steps.

Figure 7 – Schematic visualization of the RAMD protocol. It shows the application of a randomly directed constant force on the ligand's center of mass. The dissociation is defined as the minimum distance between the center of mass of the ligand and the receptor

.



An extension of the RAMD method, referred to as τ-RAMD, has shown to efficiently

estimate relative dissociation kinetic rates (k_{off}) (KOKH et al., 2018). It is based on the assumption that relative residence times (τ, i.e., the reciprocal of the k_{off}) for a group of compounds can be estimated by directly determining the simulation time needed for ligand dissociation during enhanced molecular dynamics simulations (MOLLICA et al., 2015). Molecules that dissociate slowly will need more time to exit the binding pocket, or require a higher force to be applied within a designated simulation time. Conversely, molecules that dissociate more rapidly will leave the binding pocket more swiftly and require a comparatively weaker force for exit within a set simulation time (KOKH; WADE, 2021; NUNES-ALVES; KOKH; WADE, 2021; BERGER et al., 2021).

For sampling of the bound state in RAMD, multiple starting points usually need to be prepared, and for each of these starting points (replica), several trajectories are generated (typically 15). The effective relative τ is calculated as the simulation time at which dissociation was observed for 50% of the simulation runs averaged by bootstrapping (KOKH et al., 2018).

2.2 FUNDAMENTALS OF THE ROSETTA PACKAGE

Rosetta is a widely-used software suite that has multiple applications in protein structure prediction, functional design, and protein engineering (DAS; BAKER, 2008). It is designed to predict protein structures accurately by modeling their energy landscape and determining their stable, native conformation based on physical principles. To achieve its goals, Rosetta performs two primary tasks: sampling relevant conformational or sequence spaces and accurately evaluating and ranking the energies of structural models. The suite has various applications, including predicting protein structures from sequence information, docking proteins with other molecules, designing new proteins, and modifying the function of existing proteins (KAUFMANN et al., 2010).

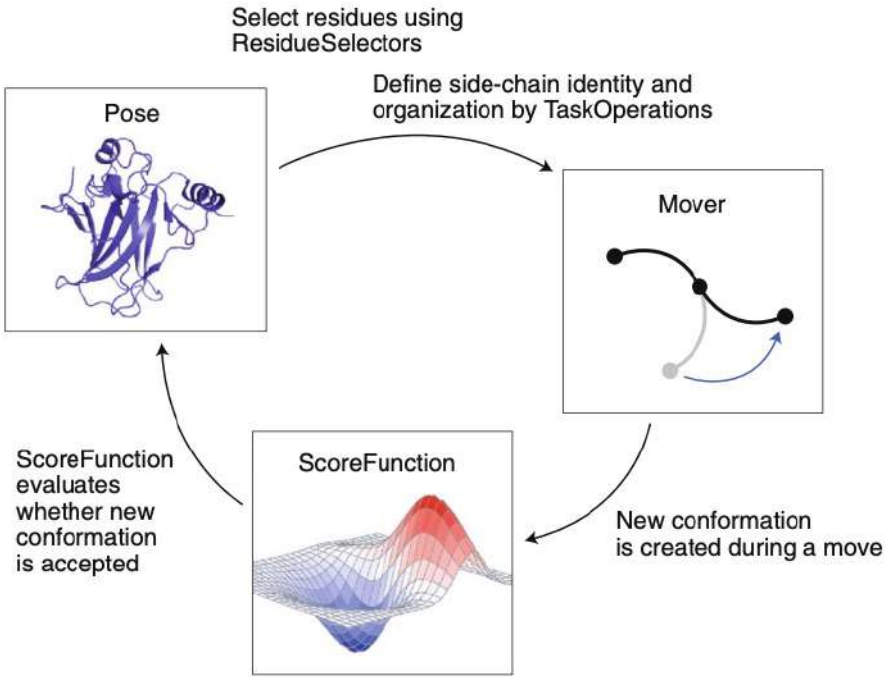
Figure 8 illustrates the general process used by the Rosetta program. It typically begins with a biomolecular conformation called a *Pose*. Specific residues within the *Pose* can be selected using *ResidueSelectors*, and *TaskOperations* can be applied to specify how side-chain optimization or mutation should be handled. The *Movers* control the changes made to the *Pose* conformation, while the *ScoreFunction* is used to evaluate the resulting conformation. The Metropolis criterion and the energy difference between the original (E_{orig}) and new (E_{new}) conformations determine whether the conformational changes are accepted or rejected.

$$\text{If } E_{new} < E_{orig}\text{, accept}$$

$$\text{If } E_{new} \geq E_{orig}\text{, accept with probability } P = e^{-(E_{new}-E_{orig})/k_BT}$$

The above principle of evaluating changes in conformation based on energy difference and the Metropolis criterion is commonly applied to tasks such as protein structure prediction, modeling, and design. To achieve the desired results, multiple independent trajectories are generated, and the final models are evaluated based on the specific objective.

Figure 8 – In a Rosetta protocol, essential elements include the Pose representing the biomolecule in a specific conformation. ResidueSelectors select residues, TaskOperations define behavior for optimization or mutation, and Movers control conformational changes. Evaluation is done using a ScoreFunction, and acceptance is determined by the Metropolis criterion. Multiple sampling trajectories explore the conformational space, with final models evaluated based on protocol objectives.



Source: Adapted from (LEMAN et al., 2020)

In recent years, significant progress has been made in the field of protein structure prediction and design with the advent of deep learning techniques (discussed in section 2.3.4) (MOULT et al., 2018; CALLAWAY, 2020a; KRYSHTAFOVYCH et al., 2019)). This has resulted in a shift in focus within the Rosetta framework towards the implementation of deep learning-based methods. One such example is the development of trRosetta (YANG et al., 2020) for protein structure prediction and design. This method utilizes a deep neural network to predict inter-residue distances and orientations based on a protein's amino acid sequence. These predictions are then used as restraints to direct structure prediction through energy minimization using the

Rosetta package (DU et al., 2021). Similar approaches can be used for protein design using the inverse approach. Many other methods for computational protein design using deep learning within the Rosetta framework have also been developed (WANG et al., 2022b; DAUPARAS et al., 2022).

2.2.1 Potential energy

The Rosetta energy function is an all-atom mathematical model used to approximate the energy of a biomolecule conformation. It is based on physical principles and and is designed to reproduce the behavior of these molecules in aqueous solutions. The Rosetta energy function has been continuously improved over many years (LEAVER-FAY et al., 2013), and consists of a linear combination of weighted score terms that balances physics-based and statistically derived potentials. The current version of the rosetta energy function is the Rosetta Energy Function 15 (REF15)[1] (ALFORD et al., 2017). Its functional form (Equation 2.30) includes, respectively, terms for describing van der Waals energies (6-12 Lennard-Jones potential), hydrogen bonds (orientation-dependent hydrogen bonding potential), electrostatics (based on Coulombic potential), disulfide bonds (orientation-dependent model), solvation (Lazarid-Karplus), backbone torsion angles (knowledge-based conformation-dependent amino acid internal free energy), sidechain rotamer energies, and an average unfolded state reference energy (Figure 9) (LEMAN et al., 2020). These terms are combined to produce an overall score that reflects the stability of the biomolecule conformation.

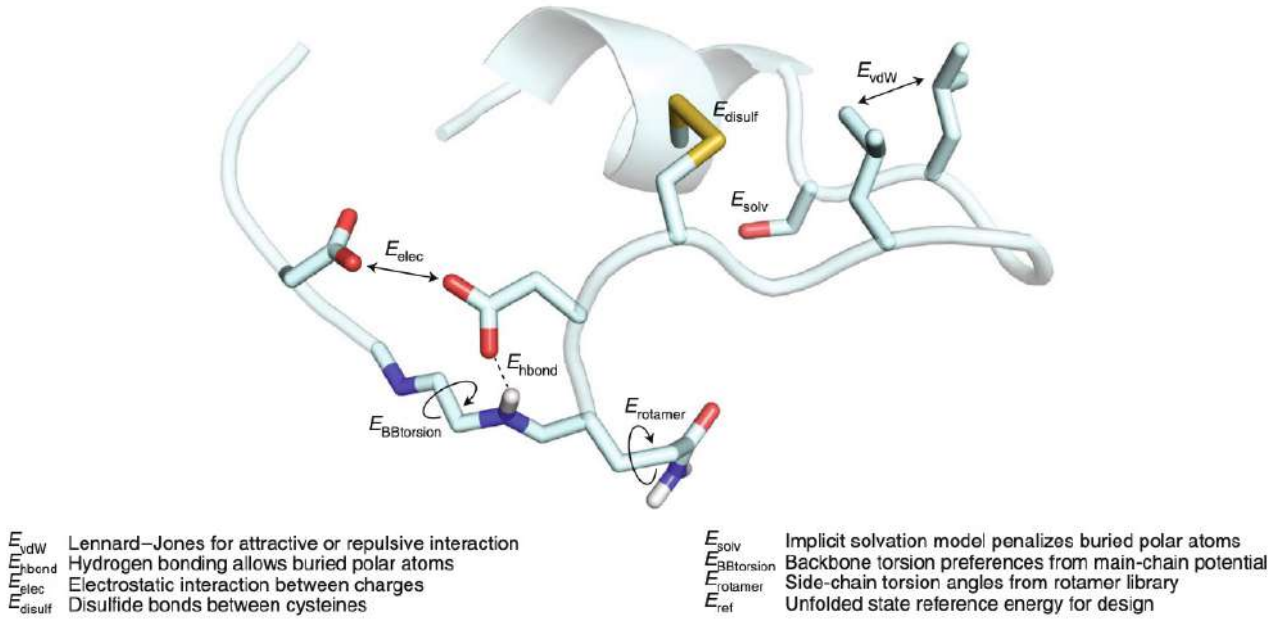
$$E = E^{vdW} + E^{HBond} + E^{Elect} + E^{Disulf} + E^{Solv} + E^{BBTorsion} + E^{Rotamer} + E^{Ref} \quad (2.30)$$

To accurately model the energy landscape of proteins, some energy terms in the Rosetta energy function are decomposed into multiple components. For example, the van der Waals energy term is split into attractive and repulsive components that act between different residues. This allows each component to be separately parameterized for better accuracy. A detailed description of the all-atom score function can be found in the reference (ALFORD et al., 2017).

The Rosetta energy function has been demonstrated to accurately predict the relative stability of proteins (CUNHA et al., 2015) and reproduce thermodynamics observables such as liquid-phase properties and liquid-to-vapor transfer free energies (LEMAN et al., 2020). It also

[1] **N.B**. REF15 stands for Rosetta Energy Function 2015.

Figure 9 – Schematic representation of the terms of the Rosetta score function



Source: Adapted from (LEMAN et al., 2020)

closely resembles the hydrogen bond geometries observed in high-resolution crystal structures (O'MEARA et al., 2015). However, it has some limitations: it does not directly estimate entropy, which can be roughly taken into account by the solvation energy term; knowledge-based terms are often derived from crystal structures, which only represent a single state and do not account for flexibility; and the solvation model is implicit, which is fast but does not accurately model interactions with solvent, such as relevant hydrogen bonds.

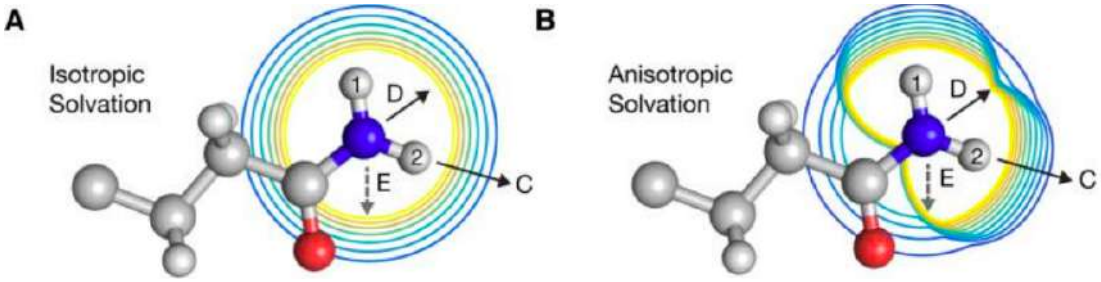
Since solvation plays a critical role in protein structure, being regarded as the dominant driving force for protein folding and association (LEVY; ONUCHIC, 2006), this energy term will be discussed further.

2.2.1.1 Solvation model

The solvation component of the Rosetta force field represents the solvent as bulk water and is based on the Lazaridis-Karplus implicit solvation model (Gaussian-exclusion model) (LAZARIDIS; KARPLUS, 1999). It applies an energetic penalty for the desolvation of side-chains, with charged side chains having a higher penalty and hydrophobic side chains having a smaller penalty. The solvation model consists of two parts: an isotropic solvation energy (Figure 10A) that assumes evenly distributed water around the atoms and an anisotropic solvation energy

(Figure 10B) that considers specific water molecules near polar atoms forming a solvation shell.

Figure 10 – The two components of the Lazaridis-Karplus solvation model used in Rosetta: (A) an isotropic term, and (B) an anisotropic term. The figures A and B show the contrast between isotropic and anisotropic solvation of the NH_2 group by CH_3 on the asparagine side chain. The potential is computed as the necessary energy to remove the water molecules around NH_2 when this is approached by the group CH_3. The contour lines indicate the variation in energy from low (blue) to high (yellow)



Source: Reproduced from Alford *et al.*, 2017 (ALFORD et al., 2017)

The solvation of individual groups in a molecule can change when the molecule undergoes conformational change or when two molecules come into close proximity with each other. This is due to two factors: 1) the exclusion of solvent from the space occupied by other groups in the solute, and 2) the modification of solvent density and the orientational distribution of solvent. In the Lazaridis-Karplus model, the second effect is not considered for nonpolar groups, as it is believed to have a small effect, but it is partially taken into account for polar groups through the use of a distance-dependent dielectric constant. The free energy of solvation ($\Delta G^{Solvation}$) proposed is written as a sum over all the atomic contributions (Equation 2.31).

$$\Delta G^{Solvation} = \Delta G^{Reference} - \sum_{j} \int_{V_j} f_i(r_{ij}) d^3r \tag{2.31}$$

Where $\Delta G_i^{Reference}$ (reference solvation energy) corresponds to the solvation free energy of i in a small molecule in which the group i is completely solvent-exposed. The second term in the equation 2.31 is an integral over the solvation free energy density of the group i in a position r. This term also contains contributions from the solute-solvent energy, solvent-reorganization energy, and the solvent-reorganization entropy. The integral is calculated over the volume V_j, of the group j, which displaces the water molecules around i; the summation is over all the groups j around i, where r_{ij} is the distance between the groups i and j. To make the calculations more numerically treatable, the integral is replaced by the product between the solvation free energy density of the atom i and the volume V_j, where the latter is approximated

by the volume of a sphere of a given radius. Thus, the equation 2.31 is written as equation 2.32:

$$\Delta G_i^{Solvation} = \Delta G_i^{Reference} - \sum_{ji} f_i(r_{ij}) V_j \tag{2.32}$$

The functional form of f_i can be determined using statistical mechanical equations and computer simulations (LAZARIDIS, 1998), and it is defined by a Gaussian function. The function that Rosetta uses is a modified version of the proposal by Lazaridis-Kaplus, called the desolvation function, (f_{desolv}), which describes the energy required to remove water molecules from an atom i, when it is approached by an adjacent atom j. The interaction energy of pairs varies with the separation distance, $d_{i,j}$, the experimentally determined free energy of the water-vapor transition, ΔG_i^{Free}, the sum of atomic radii, $\sigma_{i,j}$, the correlation length, $\lambda_{i,j}$, and the atomic volume of the desolvating atom, V_j. It is calculated using the following expression (equation 2.33):

$$f_{dessolv} = -V_j \frac{\Delta G_i^{Free}}{2\pi^{3/2}\lambda_i \sigma_i^2} exp\left[-\left(\frac{d_{i,j} - \sigma_{i,j}}{\lambda_i} \right)^2 \right] \tag{2.33}$$

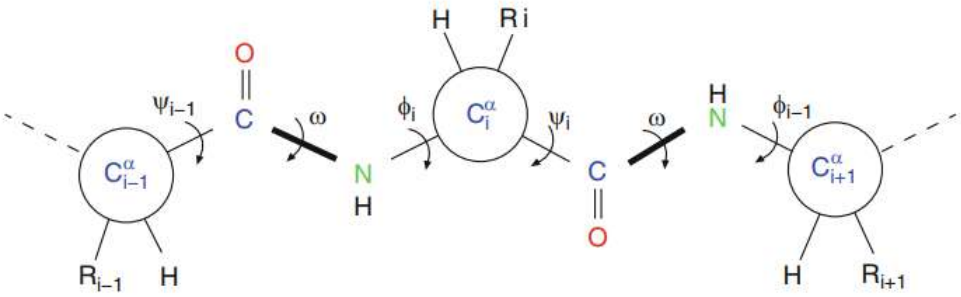
2.2.2 Backbone conformational sampling

The package Rosetta uses a torsion space representation in which the backbone conformations are specified as a list of torsion angles ϕ , ψ and ω (Figure 11). The Rosetta program distinguishes between global conformational sampling of the main chain and local refinement. Global conformations of the main chain are modeled based on a library of peptides containing nine or three amino acids. For the local refinement,the package Rosetta utilizes a MC Metropolis sampling to optimize the ϕ and ψ angles, which are calculated in such a way to not perturb the global folding of the protein (KAUFMANN et al., 2010).

The package Rosetta calculates the probability of placing a side chain of a given backbone conformation in terms of ϕ and ψ angles. The propensity, $P(aa|\phi,\psi)$, i.e, to probability to occur a given amino acid given the occurrence of ϕ and ψ angles (Bayes' theorem), is derived using a kernel density estimate using the following expression (equation 2.34) (ALFORD et al., 2017):

$$P(aa|\phi,\psi) = \frac{P(\phi\psi|aa)P(aa)}{\sum_{aa'} P(\phi,\psi|aa')} \tag{2.34}$$

Figure 11 – Rotational flexibility in polypeptides



Torsion angles (dihedrals) ϕ , ψ and ω of the backbone
Source: Reproduced from Schlick (SCHLICK, 2010)

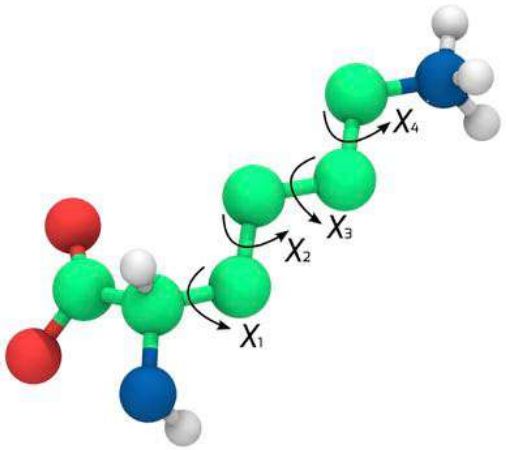
The basic operation used by the Rosetta program to explore conformational changes is called fragment insertion. In this process, a consecutive window of three and nine residues is selected and the torsion angles of these residues are replaced with the torsion angles from a known protein fragment. This method is commonly used in protein structure prediction. For more precise sampling, such as in loop modeling and protein design, fragment insertion is combined with additional operations. These include random perturbation of torsion angles, selection of fragments that do not significantly perturb the global structure, and fast torsion angle optimization to balance global perturbations in the backbone. Essentially, these operations involve randomly perturbing the angles (ROHL et al., 2004).

2.2.3 Side chain conformational sampling

In the side chains sampling, the package Rosetta reduces the number of sampled conformations by using discrete conformations of the side chains observed in the PDB (KUHLMAN; BAKER, 2000; JR; KARPLUS, 1993). In most cases, these conformations correspond to rotamers or local minimum in the potential energy surface with frequencies calculated from conformational analysis of organic molecules. These rotamers represented by χ_n (Figure 12) capture allowed combinations between the side chain, as well as the backbone angles, in such a way to reduce the conformational space. The *Simulated Annealing* MC algorithm (KIRKPATRICK; GELATT; VECCHI, 1983) is employed to search for the rotamer combinations occupying the global minima in the energy function.

From the rotamers library, the Rosetta package derives the conformation probabilities by means of k χ angles using the equation 2.35, where T is the number of rotamer angles plus

Figure 12 – Rotational flexibility in polypeptides



Torsion angles (dihedrals) of the side chain
Source: Reproduced from Verli (VERLI, 2014)

1 (ALFORD et al., 2017).

$$P(\chi|\phi,\psi,aa) = P(rot|\phi,\psi,aa)[\prod_{k<T} P(\chi_k|\phi,\psi,rot,aa)]P(\chi_T,\phi,\psi,rot,aa) \tag{2.35}$$

2.3 MACHINE LEARNING

Machine learning (ML) is a field of artificial intelligence concerned with the design and development of algorithms that can learn from and make predictions on data (MITCHELL, 1997). Fundamentally, ML models are committed to recognizing and learn data patterns. The term was first introduced in 1959 by Arthur Samuel, an IBM employee who was a pioneer in computer gaming and artificial intelligence (SAMUEL, 1959). It is also attributed to Arthur Samuel the following definition for ML:

> "*The field of study that gives computers the ability to learn without being explicitly programmed*" (Arthur Samuel, 1959)

Since then, ML has experienced significant growth and development in recent decades, leading to its widespread use in a variety of fields and applications (SCHMIDHUBER, 2015).

These include technology fields like web search and natural language processing, as well as scientific fields like medical diagnostics and bioinformatics (BINDER et al., 2021; ARDILA et al., 2019; SENIOR et al., 2020). ML has also been used to improve discovery and design in areas such as materials, chemicals, and chemical processes (KADURIN et al., 2017; SANCHEZ-LENGELING; ASPURU-GUZIK, 2018; RUDORFF; LILIENFELD, 2021; SOARES et al., 2022).

A typical workflow for machine learning consists of gathering and preparing the data, choosing a representation for the system under study to build a data set, training the model (train model candidates, evaluate model accuracy, and tune hyperparameters), and finally testing the model out of sample (KEITH et al., 2021). The goal of the training process is to find a mathematical representation, or model, that accurately maps inputs to their corresponding outputs. Among other classifications, ML can be divided into two main types: supervised and unsupervised learning. Supervised ML involves training a model on labeled data to make predictions and can tackle two main types of problems: regression (predicting a continuous numerical value) and classification (predicting a categorical label). On the other hand, unsupervised ML involves training a model on unlabeled data to discover patterns or structure, and techniques such as clustering, data visualization, and dimensionality reduction are utilized.(GÉRON, 2022).

2.3.1 Training models

The choice of model architecture defines a set of possible solutions, which is optimized for the training data set by selecting the best parameters. The optimization is guided by a loss function ($L(y, \hat{y})$), which is a measure of how well a model is able to predict the true output given a set of inputs (HASTIE et al., 2009). It is a function of the model's predictions and the true values, and is used to optimize the model's parameters during training. In mathematical terms, a loss function can be defined as follows:

$$L(y, \hat{y}) = f(y, \hat{y}) \tag{2.36}$$

Where y is the true output and $\hat{y}$ is the model's prediction. The function f measures the discrepancy between the true and predicted outputs. The goal of training a model is to find the parameters of the model that minimize the loss function. In other words, the training process is an optimization problem that aims to find the values of the model parameters that produce the smallest difference between the model's predictions and the actual target values.

This is usually achieved using the gradient descent algorithm (BEACH, 1953; WEDDERBURN, 1960), in which it iteratively updates the parameters of a model in the direction of steepest decrease in the loss function until a minimum is found.

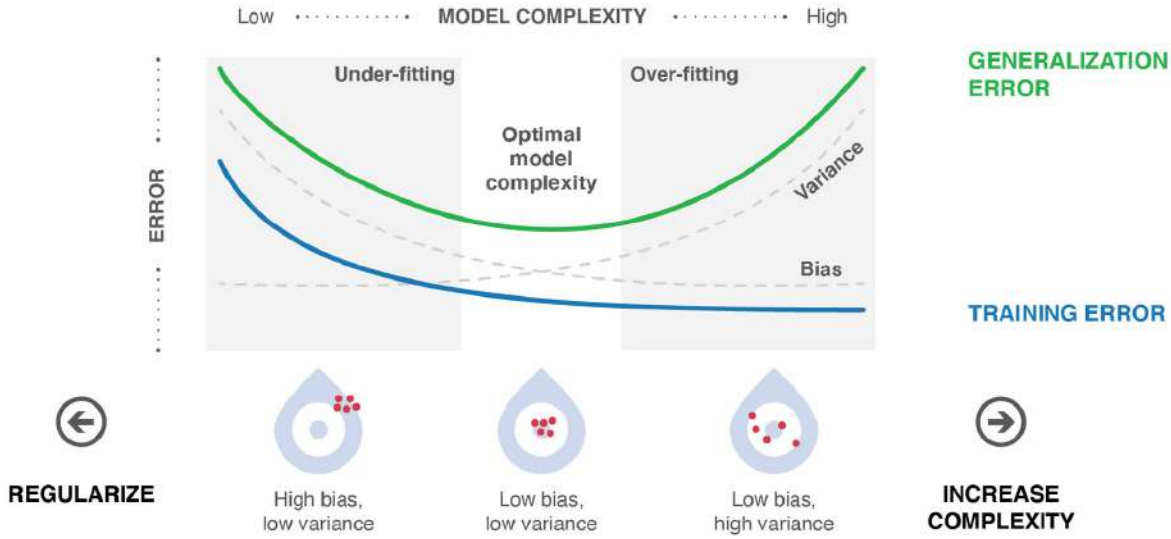
ML models contain hyperparameters, which are are parameters that are set before training the ML model and control the behavior and performance of the model (CLAESEN; MOOR, 2015). They are not learned during training, unlike model parameters, which are learned from the data. These hyperparameters influence the model or the optimization algorithm and play a crucial role in the model's effectiveness. Complex hyperparameter spaces are usually optimized using grid or random searches (HUTTER; HOOS; LEYTON-BROWN, 2014; YANG; SHAMI, 2020) and their performance is measured by evaluating the model on the validation data set, a process referred to as model selection.

2.3.2 Overfitting and underfitting

Machine learning aims to create a model that can generalize well to new data, but this is often challenging due to overfitting and underfitting. Overfitting happens when a model is too complex and fits training data too closely, leading to poor performance on new data. Underfitting happens when a model is too simple and misses the underlying patterns in the data. To alleviate these problems, regularization of the model addresses overfitting by adding a penalty term to the objective function, reducing model complexity. Techniques like L1 and L2 regularization can help.

Finding the appropriate amount of regularization to manage under- and overfitting is known as attaining a good bias-variance trade-off (GEMAN; BIENENSTOCK; DOURSAT, 1992) (Figure 13). Bias refers to the error that is introduced by approximating a real-world problem with a simplified model. Variance refers to the error that is introduced by the model's sensitivity to fluctuations in the training data. The bias-variance trade-off refers to the trade-off between these two sources of error. To find the right balance between bias and variance, it is important to choose a model that is complex enough to capture the underlying patterns in the data but not too complex that it overfits. This can be achieved by using techniques such as cross-validation, discussed below.

Figure 13 – During the training process, supervised learning algorithms must find a balance between two types of errors: bias and variance. When a model is highly biased, it is based on incorrect assumptions about the problem being addressed, resulting in under-fitting. Conversely, a model with high variance is too sensitive to small fluctuations in the data and may pick up random noise, resulting in overfitting.



Source: Adapted from (KEITH et al., 2021)

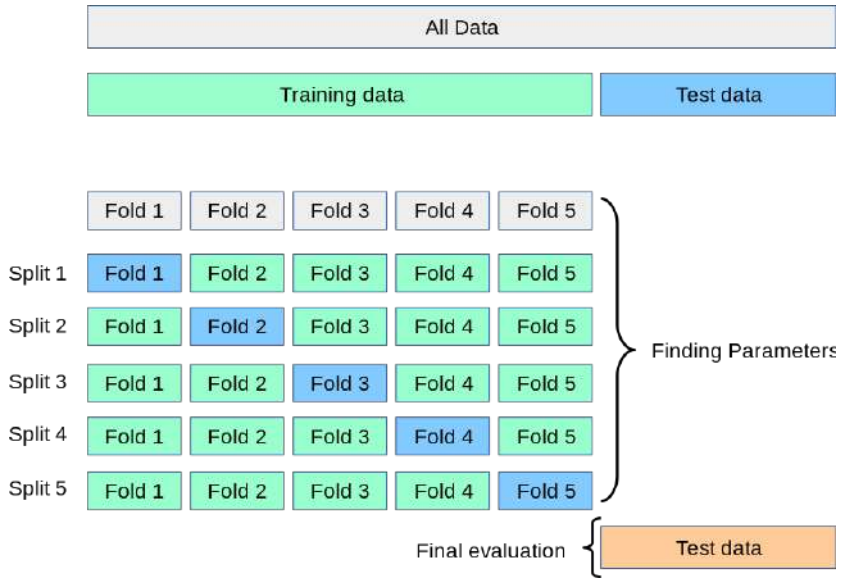
2.3.3 Evaluating models

The validation process involves testing the ML model on a separate dataset to evaluate its performance and generalizability. This process involves testing the model on new data that was not used in training and comparing its predictions to the actual outcomes. The metrics to evaluate a ML model is highly dependent on the nature of the problem. For example, for regression R-squared and root mean squared error are commonly used, whereas for classification accuracy and precision are good predictors of performance. The definition of these metrics will be presented in the methodological section of each chapter as necessary.

To validate a ML model, the data is split into two parts: training and testing data. The training data is used to fit the model and adjust its parameters, while the testing data is used to evaluate the model's performance on unseen data. The most common technique is the train-test split, where the data is randomly divided into training and testing sets. However, this approach can lead to high variance if the split is not representative of the data distribution. Thus, k-fold cross-validation is an alternative (Figure 14), where the data is divided into k equal-sized folds and use each fold once as testing data while the remaining folds are used for

training.

Figure 14 – Example of a k-fold cross validation protocol. To perform k-fold cross-validation, the training set is partitioned into k smaller sets, with other approaches following similar principles. The procedure involves training a model using one of the k folds as training data while validating it on the remaining data (i.e., a test set) to calculate a performance measure like accuracy. This process is repeated for each of the k folds. The reported performance measure for k-fold cross-validation is the average of the computed values during the loop.



Source: Scikit-learn

2.3.4 Artificial neural network

Artificial Neural Networks (ANN) (MCCULLOCH; PITTS, 1943; ROSENBLATT, 1958) are computational models inspired by the structure and function of the human brain. They are composed of layers of interconnected "neurons", which process and transmit information. The input layer receives the input data, and each subsequent layer applies transformations to the data through a set of weights and biases, until the output layer produces the desired output. The weights and biases are adjusted during the training process in order to minimize the error between the predicted output and the target output.

Mathematically, an ANN can be represented as a function $f(\mathbf{x};\theta)$, where $\mathbf{x}$ is the input and θ represents the network parameters (weights and biases). Each layer of the network applies a linear transformation to the input data, followed by a non-linear activation function:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \tag{2.37}$$

$$\mathbf{a}^{(l)} = g^{(l)}(\mathbf{z}^{(l)}) \tag{2.38}$$

Where $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weights and biases at layer l, and $g^{(l)}$ is the activation function at layer l. The output of the network is then given by:

$$\mathbf{y} = f(\mathbf{x}; \theta) = \mathbf{a}^{(L)} \tag{2.39}$$

In which L is the total number of layers in the network. The training process involves adjusting the parameters θ in order to minimize the error between the predicted output $\mathbf{y}$ and the target output $\mathbf{y}_{\text{true}}$. This is typically done using a loss function, which quantifies the discrepancy between the predicted and target outputs.

One type of ANN is the multi-layer perceptron (MLP) (MCCULLOCH; PITTS, 1943) (Figure 15), which consists of an input layer, one or more hidden layers, and an output layer. The input layer receives the input data, and the output layer produces the model's predictions. The hidden layers process the input data and transmit it to the output layer. An MLP can be represented mathematically as follows:

$$\mathbf{y} = f_L(\mathbf{W}LfL - 1(\dots f_2(\mathbf{W}_2 f_1(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2)\dots) + \mathbf{b}_L) \tag{2.40}$$

In this equation $\mathbf{x}$ is the input data, $\mathbf{y}$ is the output, f_i is the activation function for layer i, $\mathbf{W}_i$ is the weight matrix for layer i, and $\mathbf{b}_i$ is the bias vector for layer i. The input $\mathbf{x}$ is transformed and processed by the layers of the MLP, and the final output $\mathbf{y}$ is produced by the output layer of the network. The activation function f determines the output of each neuron in the network and is typically a non-linear function, such as the sigmoid function ($f(x) = \frac{1}{1+e^{-x}}$) or the rectified linear unit (ReLU) function ($f(x) = \max(0, x)$).

The weights and biases are learned during the training process, which involves adjusting their values in order to minimize the error between the model's predictions and the true labels. This can be done using a variety of optimization algorithms, such as stochastic gradient descent (SGD) (PLATT, 1998) or Adaptive moment estimation (Adam) (KINGMA; BA, 2014).

SGD updates the weights and biases at each training iteration according to the following equation:

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha \frac{\partial L}{\partial \mathbf{W}} \tag{2.41}$$

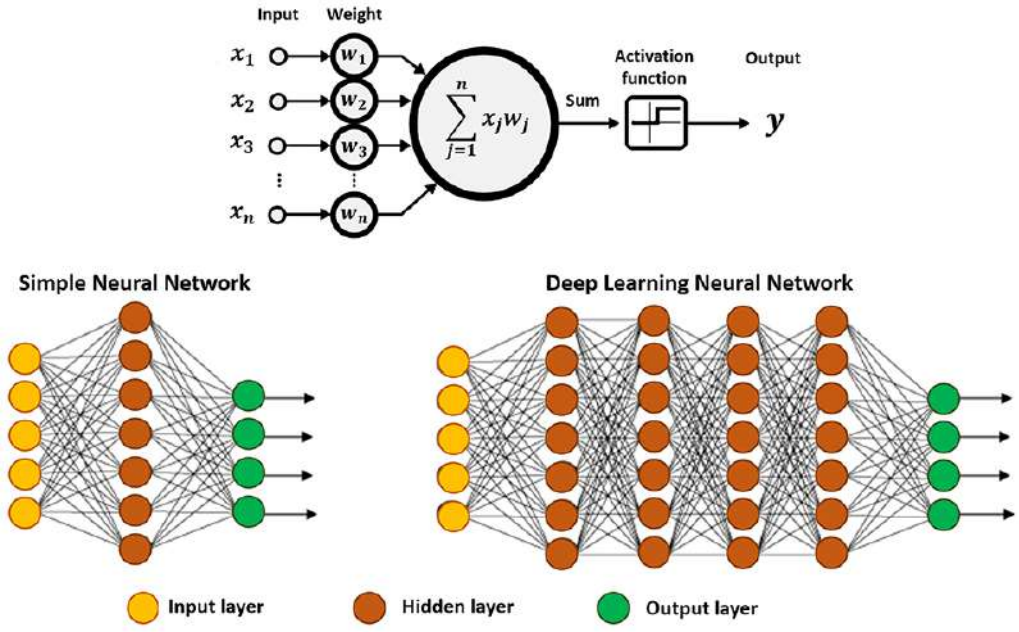
$$\mathbf{b} \leftarrow \mathbf{b} - \alpha \frac{\partial L}{\partial \mathbf{b}} \tag{2.42}$$

Where α is the learning rate and $\frac{\partial L}{\partial \mathbf{W}}$ and $\frac{\partial L}{\partial \mathbf{b}}$ are the gradients of the loss function with respect to the weights and biases, respectively. On the other hand, Adam is an optimization algorithm that utilizes the gradient of the loss function and exponentially decaying average of past squared gradient values to update the weights of the model, with a learning rate that decreases over time. This is done using the following update rule:

$$w_t = w_{t-1} - \alpha * m_t / (\sqrt{v_t} + \epsilon) \tag{2.43}$$

Where w_t is the weight at time step t, α is the learning rate, m_t is the exponential moving average of the gradient, v_t is the exponential moving average of the squared gradient, and ϵ is a small value added for numerical stability.

Figure 15 – Schematic representation of an artificial neuron (top) and a simple neural network with input, hidden, and output layers (bottom-left), along with a deep neural network featuring at least two hidden layers or nodes (bottom-right). The calculations are carried out through the connections, comprising input data, pre-assigned weights, and defined paths through the activation function. In case the outcome deviates significantly from the expected, the connection weights are adjusted, ensuring the analysis continues until achieving optimal accuracy



Source: Reproduced from Cova and Pais (COVA; PAIS, 2019)

Recurrent Neural Networks (RNN) (RUMELHART; HINTON; WILLIAMS, 1985) and Convolutional Neural Networks (CNN) (LECUN et al., 1998) stand out as two popular deep learning models widely utilized across diverse applications. RNNs, specifically designed for sequential

data like text or speech, possess the ability to process such data by incorporating feedback connections. This unique characteristic enables RNNs to capture intricate dependencies among elements in the sequence. On the other hand, CNNs specialize in image recognition and computer vision tasks, leveraging convolutional layers to scan images and detect specific features and patterns. Due to their proficiency in handling high-dimensional data, CNNs excel in tasks such as image classification, object detection, and segmentation.

3 MACHINE LEARNING BINDING THERMODYNAMICS

This chapter was based on the following publications:

An artificial Neural Network Model to Predict Structure-based Protein-protein Free Energy of Binding from Rosetta calculated Properties, Matheus Ferraz, Jose Neto, Roberto Lins, Erico Teixeira, *Physical Chemistry Chemical Physics* **2023**, 25, 7257-7267, DOI:10.1039/D2CP05644E

Association strength of E6 to E6AP/p53 complex correlates with HPV-mediated oncogenesis risk, Matheus Ferraz, Isabelle Viana, Danilo Coêlho, Carlos Cruz, Maíra Lima, Madson Aragão, Roberto Lins, *Biopolymers* **2022**, 113, 10, e23524 DOI:doi.org/10.1002/bip.23524

3.1 BACKGROUND

To understand many chemical and biochemical processes, it's necessary to examine their free energy behavior. Examples include molecular association, immiscible liquid partitioning, and protein stability, among others. As the free energy can be related to binding affinities, it is crucial in fields like computer-aided protein design against viral targets, where high affinity binders are often required. By computationally predicting binding affinities, the need for experimental measurements of thermodynamic properties can be reduced, making free energy estimation an essential factor in this area of research.

Calculating binding free energies for protein-protein and protein-ligand interactions is difficult due to the complex nature of the Hamiltonian in biological systems. Throughout the doctorate period, various techniques were employed to estimate these binding free energies without relying on directly estimating the partition function. Although not extensively covered in this thesis, references are provided for further reading. These references include studies that calculated the ΔG of binding caused by point mutations in the variants of concern and their impact on antibody binding in the receptor-binding domain (RBD) of SARS-CoV-2, using scoring functions (FERRAZ et al., 2021). Additionally, enhanced sampling simulations (metadynamics) were employed to investigate how SARS-CoV-2 persisted in a human population with

a high immunity barrier in the Amazonas population, one of the hardest-hit Brazilian states by the COVID-19 epidemic (NAVECA et al., 2022). Furthermore, metadynamics simulations were utilized in drug design projects targeting Staphylococcus aureus dihydrofolate reductase (MATOS et al., 2022), 3-Hydroxykynurenine Transaminase from *Aedes aegypti* and *Anopheles gambiae* (MACIEL et al., 2023), and the main protease of SARS-CoV-2 (GODOY et al., 2022).

Although these techniques have been successfully applied for the aforementioned works, they also have presented certain limitations. For instance, the Rosetta score function proved capable of computing the $\Delta\Delta G$ of binding for protein-protein complexes. However, its effectiveness in predicting the relative ΔG of binding, especially for large systems that are now routinely studied, was limited. The calculation of binding energy differences could be prone to inaccuracies due to noise in the energy function. Consequently, when it came to predicting the impact of point mutations in significant systems, we resorted to enhanced sampling simulations. Nevertheless, these simulations are generally not practical for high-throughput applications. In this chapter, we propose the utilization of ML techniques to accurately predict the absolute ΔG of binding for protein-protein interactions based on the three-dimensional structure of the complex. This approach has shown great promise in effectively estimating the affinity of computer-designed arbovirus-binding proteins. Furthermore, we have employed ML to explore the molecular factors influencing affinities in protein-protein systems, specifically in studying the correlations between the oncogenic potential of HPV-induced carcinogenesis and the interactions within the ternary complex E6/E6AP/p53.

These discoveries provide valuable insights into the molecular mechanisms underlying the battle against viral infections, with direct implications for the development of novel diagnostics, therapies, and vaccine formulations. Unlike methods designed for calculating ΔG of binding for protein-small molecule interactions, the field of protein-protein binding prediction is relatively underdeveloped, despite its crucial significance in protein computational design, particularly in the identification of high-affinity binders.

3.2 AN ARTIFICIAL NEURAL NETWORK MODEL TO PREDICT STRUCTURE-BASED PROTEIN-PROTEIN FREE ENERGY OF BINDING FROM ROSETTA-CALCULATED PROPERTIES

3.2.1 Introduction

Protein-protein association is a pervasive phenomenon in physiological and biotechnological processes, ranging from mechanisms with cell interactions and disease modulation to industrial applications of metabolic control.(JUBB et al., 2017; SINGH et al., 2013; CUTCLIFFE et al., 2011; JONES; THORNTON, 1996; FERRAZ et al., 2021) To characterize this association, Gibbs free energy (ΔG) of binding is one of the most fundamental thermodynamic quantities. In addition to the comprehension of biomolecular processes, the development of novel biomaterials (e.g., towards biomedical,(VIANA et al., 2013) vaccinal,(MARCANDALLI et al., 2019; BOYOGLU-BARNUM et al., 2021) catalytic,(JIANG et al., 2008) development of biosensors,(FENG et al., 2015) and industrial applications(BENKOULOUCHE et al., 2021)) relies substantially on the binding affinity (k_D), which is directly related to the ΔG between the involved binding partners ($\Delta G{=}{-}RTlnk_D$, where T is the absolute temperature and R is the universal gas constant).

Although the ΔG of binding is a central concern for protein modeling and design, and the calculations by computational means have been an active area of development since the beginning of the 1980s,(BARROS et al., 2022; GILSON et al., 1997) obtaining it accurately at a low computational cost remains a challenge for the molecular simulation community.(FLEISHMAN et al., 2011b) This difficulty is even more pronounced when trying to obtain the ΔG for liquids and flexible macromolecules, mainly due to insufficient sampling and inaccuracies in the description of the potential energy surface.(GUNSTEREN; DAURA; MARK, 2002)

However, despite the challenge, methods like molecular dynamics and Monte Carlo simulations afford a range of rigorous approaches such as thermodynamic integration or free energy perturbation,(ZWANZIG, 1954) which provide accurate results. In these techniques, nonphysical pathways via alchemical methods (e.g., thermodynamic integration, free energy perturbation)(POHORILLE; JARZYNSKI; CHIPOT, 2010) are simulated by connecting two end states by a coupling parameter. Even though these methods offer accurate values, for example, achieving an error of ca. 1.0 kcal.mol^{-1},(WANG et al., 2015; CLARK et al., 2019) a major hurdle is the amount of sampling needed to simulate the regions of the phase space that

make important contributions to the ΔG. On the other hand, end-point state methods, such as molecular mechanics generalized Born surface area (MM-GBSA) and molecular mechanics Poisson Boltzmann surface area (MM-PBSA), (SRINIVASAN et al., 1998b; KOLLMAN et al., 2000; SRINIVASAN et al., 1998a) have been used extensively to compute the ΔG of binding, (XUE et al., 2018; XUE et al., 2022) as they exempt the need for simulating intermediate states as in alchemical methods and make use of implicit solvation, reducing the computational cost. Despite that, to account for entropic changes in MM-PBSA and MM-GBSA, typically, standard normal mode is used, which is time-consuming and approximate.(GŌ; SCHERAGA, 1969; BROOKS; JANEŽIČ; KARPLUS, 1995) Thus, in many studies, the entropic contribution is often neglected, leading to inconsistent results. Aiming improvement of computational costs maintaining accuracy, enhanced sampling methods (e.g., metadynamics,(LAIO; PARRINELLO, 2002) umbrella sampling(TORRIE; VALLEAU, 1977)) simulate a physical pathway between two predefined states as a function of collective variables (CV). However, the proper choice of CV is a non-trivial task for most complex systems,(SITTEL; STOCK, 2018) and the methods often suffer from a lack of convergence, leading to multiple simulations runs. Thus, atomistic simulations are bound to the size of the system, high-performance computing resources available and adequate sampling. As a result, these methods are routinely computationally prohibitive for high-throughput applications, for example, for protein design, where thousands or even millions of protein complexes must be evaluated.(CHEVALIER et al., 2017; CAO et al., 2020)

Alternatively, scoring functions are used as a fast approach to predict the protein-protein ΔG of binding,(KASTRITIS; BONVIN, 2010) and can be physics-, empirical-, knowledge-based, or a combination of those.(LIU; WANG, 2015) Despite their swiftness in calculating the ΔG, thus far, the performance of scoring functions is uneven across different chemical systems, presenting a substandard correlation with the experimental and even relative ΔG.(KASTRITIS; BONVIN, 2010; MARILLET et al., 2017; GROMIHA et al., 2017)

Against this backdrop, supervised machine learning (ML) techniques have witnessed an unprecedented evolution in structure-based computational biology, gaining impetus in a wide range of applications.(TANG et al., 2019; DAS; CHAKRABARTI, 2021; JUMPER et al., 2021) Concerning the prediction of binding protein-ligand ΔG of bidding for various complexes, ML models reliably outperformed classical scoring functions,(ASHTAWY; MAHAPATRA, 2014; DURRANT et al., 2013) and several methods have been devised. However, only low to moderate correlation with experimental measurements was achieved with success conditioned to the chosen data set.(MOAL; AGIUS; BATES, 2011) For example, the PROtein binDIng enerGY prediction

(PRODIGY), a robust state-of-art web server descriptor of protein-protein ΔG of binding,(XUE et al., 2016) an empirical linear equation composed of terms based on interfacial contacts and properties from the non-interface surface that make a statistically significant contribution to the binding affinity. The test set for this protocol presents an unprecedented root mean-square-error (RMSE) of 1.89 kcal.mol^{-1}, demonstrating that there is still room for improvement in this field.

To this end, we devised a neural network model trained with the interface and folding properties for a set of 81 protein-protein complexes,(VANGONE; BONVIN, 2015) calculated by Rosetta, a state-of-art software package for protein modeling and design.(LEAVER-FAY et al., 2011a) The Rosetta contains an empirical energy function,(ALFORD et al., 2017) which is based on the parametrization and transferability of small-molecule potentials, and X-ray crystal structural data to calculate protein geometries. While its functional form, which includes atomic packing, pairwise-additive implicit solvation model, and physical- and statistical-based potentials, can satisfactorily account for protein stability(CUNHA et al., 2015) and reproduction of some thermodynamic observables (e.g., liquid-phase properties and liquid-to-vapor transfer free energies(PARK et al., 2016)), it was not specifically designed for the calculation of ΔG of binding. Nevertheless, Rosetta contains a suite of tools to analyze protein-protein interfaces, thereby, calculating useful metrics to evaluate the interface. Since most of the physical factors that govern the ΔG of binding are related to their interface properties and chemical interactions,(REICHMANN et al., 2005) machine-learning methods can learn functional relationships from the features calculated using the Rosetta package to build a predictor model.(FERRAZ; ADAN; LINS, 2020)

Harnessing ML methodologies, metrics derived from the Rosetta package have already been previously used along to calculate $\Delta\Delta$G of binding for mutations,(SHRINGARI et al., 2020) to the identification of false positive leads in structure-based virtual screening,(ADESHINA; DEEDS; KARANICOLAS, 2020) to predict sites of tolerability to acridonylalanine,(GIANNAKOULIAS et al., 2021) and to discriminate between foldable or aggregating nanobodies.(FERRAZ; ADAN; LINS, 2020) Here, we showcase an ML-based model to compute the absolute ΔG of binding between protein pairs utilizing Rosetta-derived parameters. Our model presents a reduction in the RMSE of ca. 1 kcal.mol^{-1} if compared to the current state-of-the-art tool PRODIGY, representing a novel mean to the computation of protein-protein affinity. Furthermore, our devised approach can enable interoperability with the Rosetta-package applications to aid the design of functional proteins with engineered properties and functions.

3.2.2 Computational procedures

3.2.2.1 Libraries

In this work, the versions of the following Python libraries were used: matplotlib(HUNTER, 2007) 3.2.2, numpy(HARRIS et al., 2020) 1.21.5, pandas(MCKINNEY et al., 2010) 1.3.5, plotly(INC., 2015) 5.5.0, scikit-learn(PEDREGOSA et al., 2011a) 1.0.2, scipy(VIRTANEN et al., 2020) 1.4.1, seaborn(WASKOM, 2021) 0.11.2, tensorflow(ABADI et al., 2015) 2.8.0, and umap-learn(MCINNES et al., 2018) 0.5.2.

3.2.2.2 Data sets

3.2.2.2.1 Input data

The data set used in this study to train the model was retrieved from Vangone and Bonvin (2015)(VANGONE; BONVIN, 2015) and downloaded from the PRODIGY webpage (<https://bianca.science.uu.nl/prodigy/dataset>).(XUE et al., 2016) It comprises 81 bound protein-protein PDB crystallographic structures with known experimentally measured ΔG of binding (ranging from -4.3 to -18.6 kcal.mol^{-1}). Despite the relative data set small size, it encompasses a well-thought variety of structure complexes, constituted by a wide range of protein–protein interfaces, including antibody-antigen, enzyme-containing, and other complexes (for example, G-proteins). This data set was built from the published protein-protein binding affinity structure-based benchmark by Kastritis et al. (2011),(KASTRITIS et al., 2011) which originally contains 144 non-redundant protein-protein complexes. We highlight that the reason to choose this small data set was reasoned by its capability in providing a consistent model to compute protein-protein affinity, such as the Prodigy.

3.2.2.2.2 Test data

To test the trained model, two external data sets were used. For the first one, a data set containing 50 general protein–protein complexes with known ΔG of binding was built. This data set was curated from the PDBbind database (<http://www.pdbbind.org.cn/>),(WANG et al., 2004) by filtering crystal structures released in the PDB from 2010 to 2021 and with a high electron density resolution, i.e., lower than 1.5 Å (Table S1). Given that the values

in the PDBbind data set are given as a function of the k_D, it was used $\Delta G = -RTlnk_D$, to compute the ΔG of binding. The systems were estimated at a temperature of 298 K, with an R of 1.987 x 10^{-3} kcal.K^{-1}.mol^{-1}. Structures presenting k_I (inhibition constant) or IC_{50} (half-maximum inhibitory concentration) instead of the k_D were disregarded from our analysis. The second external data set consisted of a non-redundant set of 47 nanobody–antigen crystal structures with experimentally determined ΔG of binding.(ZAVRTANIK; HADŽI, 2019) Nbs are highly stable proteins capable of binding to a diversity of antigens with high affinity and specificity.(MUYLDERMANS et al., 2013) Therefore, these have been deemed one of the most promising tools in nanomedicine and nanoengineering for diagnostics and therapeutic applications.(ZIMMERMANN et al., 2020)

3.2.2.3 Structure preparation and generation of molecular descriptors

The structures within the PRODIGY data set were already treated to not contain crystallographic observed water structures, ions, cofactors, or ligands. To prepare the structures, initially, these were geometry-optimized in a stepwise manner with the minimizer routine of the Rosetta package v.3.12 consisting of three steps: 1) minimization of the sidechain χ angles; 2) full romateric packing; 3) minimization of the sidechain, backbone, and rigid body orientation. A tolerance of 0.0001 for 50,000 steps was used, solved by the Broyden-Fletcher-Goldfarb-Shanno algorithm with the Armijo-Goldstein criterion.(HEAD; ZERNER, 1985) Extra χ_1 and χ_2 rotamer angles were used for all residues passing a cutoff of 18 neighbors. Hydrogens were added using the Rosetta, and pair terms for histidine–histidine interactions were set to zero.

Starting from the minimized structures, the *InterfaceAnalyzerMover* of Rosetta was used in a *RosettaScripts*(FLEISHMAN et al., 2011a) workflow to calculate 44 folding and interface-related properties, of which some of the terms are described in the Supporting Information (Table S2), given that most knowledge-based Rosetta terms are not interpretable.(LEMAN et al., 2020) The *InterfaceAnalyzerMover* calculates the interface's properties by separating the chains, then packing the side chains, followed by the binding partners' re-association. The properties are taken as the difference between the bound and unbound states. To calculate the solvent-accessible surface area (SASA), the Le Grand and Merz method(LEAVER-FAY et al., 2007) is used by the Rosetta package. A description of polar burial atoms definitions used in this study is provided in the Supporting Information. The parsed command lines and XML script files are made available in the Supporting Information. All the calculations within the

Rosetta package were performed using the all-atom REF15(ALFORD et al., 2017) potentials unless explicitly stated otherwise.

3.2.2.4 Exploratory data analysis (EDA)

Initially, it was performed an exploration of the property values range to identify which ones should be discarded or which protein-protein structures represent an outlier. This was accomplished by visualizing the data distributions through histograms. In addition, the iForest(LIU; TING; ZHOU, 2008) anomaly detection algorithm was used. As a result of this process, only 81 bound protein-protein complexes from the input data were kept.

3.2.2.5 Machine learning

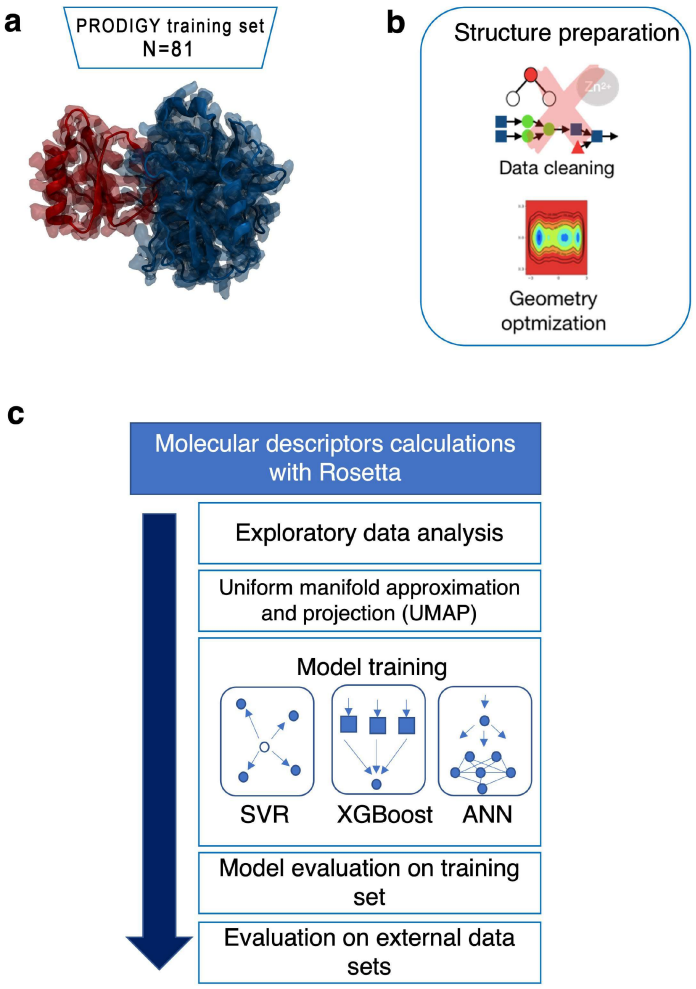
To develop a machine learning model, it is necessary to apply a set of steps: to split the data, pre-process the information, select the most important features, choose the model, validate the results, and evaluate with a test set. This process is described in the following sections and a general pipeline of the protocol devised in this work is shown in Figure 3.3.4.

3.2.2.5.1 Split, Pre-processing, and Dimensionality Reduction

To help the model avoid over-learning and presenting poor performance, it is necessary to validate it with a different set applied in the training process. Thus, the input data is divided into training and validation subsets, following a general rule of thumb: 70% of the input was associated with the training set (57 structures) and 30% with validation (24 structures). Moreover, a machine learning model presents a better performance if all properties share the same or similar order of magnitude, something achieved when the characteristic values are standardized (Figure S1). Therefore, the training and validation sets were pre-processed: each feature was individually translated to be within a range from 0 (minimum) to 1 (maximum value).

To reduce the computational cost, overfitting, and spurious correlation of the features while preserving most of the relevant information, the machine learning pipeline can decrease the number of random variables under consideration. For this purpose, this work applied UMAP,(MCINNES et al., 2018). As a consequence, the features were reduced from 41 to 14.

Figure 16 – The computational workflow for the machine learning protocol used in this work comprises (a and b) data set building and structure preparation, and (c) model training and evaluation. Data from the PRODIGY set is cleaned and geometry-optimized. Molecular descriptors for each instance are calculated with the Rosetta package and served as the training data set. Support vector regression (SVR), extremely generalized boost (XGBoost), and artificial neural networks (ANN) are tested and evaluated using the training set and two different external data set (a PDBBind-derived and a set of antigen-antibody structures).

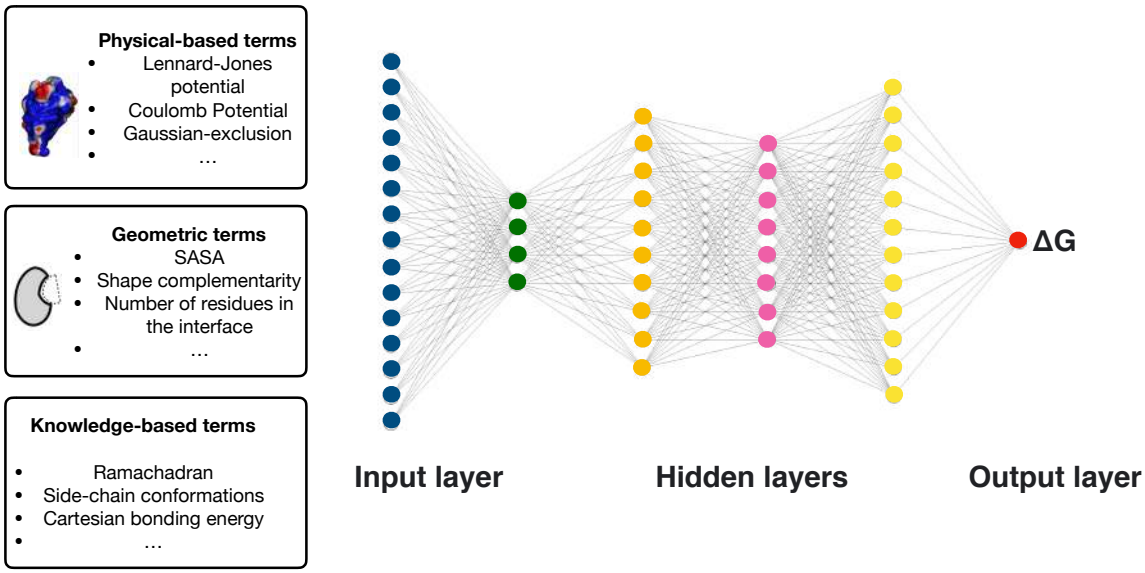


3.2.2.5.2 Artificial Neural Networks model training

We have used different configurations of ANN(MCCULLOCH; PITTS, 1943) regressor to build a relationship between the Rosetta-calculated features and the experimentally measured ΔG of binding. The ANN in this work has 14 features as input (the characteristics originated from the reduction), output with dimension one (ΔG prediction), and different input and hidden layers combinations were tested to assure the best fit possible. The input layer was tested using 1 to 4 neurons with the ReLU activation function. Besides that, we fitted a permutation of 1 to 3 hidden layers with 4 to 12 neurons for each layer, and the ReLU activation function. After testing the ANN permutations, we ended up with a neural network composed of a 4-neurons

input layer and 3 hidden layers with 10, 8, and 12 neurons, respectively (Figure 17).

Figure 17 – Schematic representation of the ANN model applied in the training. The insets demonstrate some of the calculated features grouped according to the type of terms.



In addition, each training used 500 epochs with a scheduler to reduce the learning rate when the epochs were equal to 5, 50, and 150. Besides, due to the stochastic nature of the algorithm, evaluation procedure, and numerical precision, the prediction is an average of 10-fold cross-validation (Figure S2), assuring convergence of the final model.

3.2.2.5.3 Classical machine learning model training

In addition to the ANN, two different algorithms based on different learning philosophies were trained: 1) XGBoost and 2) SVR. Hyperparameters were tuned based on an exhaustive grid search. Details of the implementation are found in the code deposited on GitHub (<https://github.com/jcsn13/PPSUS>).

3.2.2.5.4 Models' evaluation metrics

To assess the distance between the predicted value and the actual one, the RMSE (Equation 3.1) was calculated between the predicted and experimental value, i.e., the difference between

the predicted ΔG of binding ($predicted_i$) and the actual ΔG of binding ($actual_i$) for each i-th sample from the N structures within a given data set.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(predicted_i - actual_i)^2}{N}} \tag{3.1}$$

In addition to the RMSE, to evaluate the correlation between the predicted and experimental values, the $R_{Pearson}$ was also calculated (Equation 3.2):

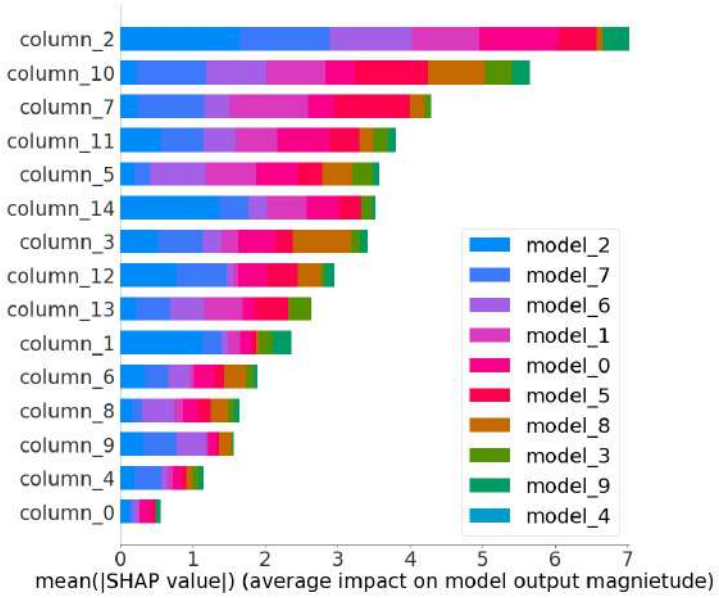
$$R_{Pearson} = \frac{\sum_{i=1}^{N}(predicted_i - \overline{predicted})(actual_i - \overline{actual})}{\sqrt{\sum_{i=1}^{N}(predicted_i - \overline{predicted})^2}\sqrt{\sum_{i=1}^{N}(actual_i - \overline{actual})^2}} \tag{3.2}$$

Where $\overline{predicted}$ and $\overline{actual}$ are the mean values of the predicted and actual ΔG values respectively. These metrics provide a comprehensive evaluation of the models' ability to predict experimental binding free energies. The RMSE measures the average difference between the predicted and experimental values, while the Pearson correlation coefficient measures the strength of the linear relationship between the predicted and experimental values.

3.2.2.5.5 Feature Importance

The explainer methodology involves identifying how each original feature is correlated with the UMAP components, and combining this information with the importance of each UMAP in the models. First, we applied the Python SHAP (SHapley Additive exPlanations)(LUNDBERG; LEE, 2017) library to determine the importance of each of the 15 UMAP components for each model using 10-fold cross-validation (Figure 18).

Figure 18 – Impact of SHAP components (denoted as the 15 columns). The models are color-coded according to the 10-fold model training.



Then, the importance of each original feature (FI_j) was estimated by a double sum, for each model (m) and UMAP component (c), as the product of the correlation between the original feature and the UMAP component and the SHAP score of the UMAP component in the model ($shap_{c,m}$) (Equation 3.3).

$$FI_j = \sum_{m=0}^{9} \sum_{c=0}^{14} umap_{c,j} \cdot shap_{c,m} \tag{3.3}$$

Where $1 \leq j \leq 41$ represents all original features.

3.2.3 Results and discussion

3.2.4 Rosetta force field accuracy

To evaluate the quality of Rosetta's ΔG calculation using the Rosetta Energy Function 15 (REF15), the latest Rosetta functional available, we have assessed how the computed values compare to the experimental measurements. To this end, the data set utilized to train the model was utilized. The performance of the calculated ΔG of binding was estimated based on Pearson's correlation coefficient $R_{Pearson}$, outputted was 0.45 for the geometry-optimized structures and 0.35 for the crystallographic structures as they are in the PDB. The RMSE was not considered for this evaluation since the Rosetta provides non-realistic values for the ΔG.

The $R_{Pearson}$ demonstrates that geometry-optimization for these structures is necessary, and that the Rosetta energy function per *se* yields a moderate correlation, i.e., here defined as a $R_{pearson} \geq 0.45$, to the experimental data.

3.2.5 Exploratory data analysis (EDA)

In this work, Rosetta interface parameters were calculated for a set of 81 relaxed protein–protein complexes to build functional relationships targeting the ΔG of binding between protein–protein complexes. Initially, an exploratory analysis was carried out to analyze how the features correlate with the experimental ΔG. A moderate correlation was observed for Sc_int_area (shape complementarity divided by the interface area), dSASA_int (total solvent-accessible surface area (SASA)), and nres_int (number of resides in the interface (Table S3). As it can be seen, this moderate negative correlation is observed between properties related to the interface area and the number of residues in the interface. It suggests that for our training data set, greater interfaces (relative to extension size and/or the number of contacts) tend to have higher affinities.

Two features (packstat and yhh_planarity) were identified to present constant values for all the complexes, being disregarded because no information can be extracted from them. These two quantities refer, respectively, to a measure of packing-goodness and a torsional potential to maintain tyrosine hydroxyl group in the plane of the aromatic ring. In addition, it was observed that three structures from the training set presented unusual values for some of the properties. The outlier value of the total_score (the total sum of the individual residue scores terms) was associated with the 1E4K structure on the data set, fa_sol (Lazaridis-Karplus solvation energy)(LAZARIDIS; KARPLUS, 1999) associated with the 1DE4 structure, and dG_cross (interaction energy) with the 1GXD structure. Furthermore, the isolation forest (iForest), a machine learning algorithm to identify anomalies, was applied. In iForest,(LIU; TING; ZHOU, 2008) a score is assigned to each sample, and the lowest scores represent outliers. In Table 1, we observe that the three structures identified as outliers based on their property values also have lower anomaly scores. Consequently, these three structures were removed from the data set.

Table 1 – The six lowest anomaly scores from the Isolation Forest algorithm

Anomaly score from iForest	PDB ID
-0.301443	1GXD
-0.256578	1DE4
-0.081814	1E4K
-0.046498	1EWY
-0.038241	2OOB
-0.027419	1HE8

3.2.6 Models' evaluation

We have trained three ML-based methods based on different learning philosophies: extreme gradient boost (XGBoost),(CHEN; GUESTRIN, 2016) an ensemble method; support vector regression (SVR),(SMOLA; SCHÖLKOPF, 2004) a class of instance-based algorithms; and artificial neural networks (ANN), based on a multilayer perceptron.The training procedure followed the same strategy of PRODIGY (to use the whole training set)(XUE et al., 2016), and it is described in section **??**. To evaluate the performance of the models, two external data sets were built as described in the Test Set section. The first one consists of a data set presenting high resolution ($\leq$ 1.5 Å) crystallography-determined structures of general protein-protein interface. The second one comprises a data set containing structures for nanobodies (Nbs) targeting their antigens.

Initially, we have evaluated the training set metrics for our devised models and the PRODIGY web server (Table 2). The main evaluation metric utilized in this work was the RMSE to assess the accuracy of the predictions. The main reason for this, is that we are interested in developing a model that minimizes the error compared to experimental data. Therefore, since the RMSE provides a direct assessment of the model's error, it was utilized. In contrast, metrics such as the $R_{Pearson}$ can be misleading to interpret the accuracy of ML algorithms as it does not take into account the error or uncertainty associated with the predicted values. However, the $R_{Pearson}$ was calculated for the sake of comparison between different models. As it can be seen, for the training set, XGBoost presents the best performance with an $R_{Pearson}$ of 0.94, followed by the PRODIGY. The RMSE also follows the same trend. It is worth noting that our presented RMSE differ fairly from that of the PRODIGY publication,(XUE et al., 2016) as their reported RMSE is of 1.89 kcal.mol^{-1}, this happens because for our study, we have removed three outliers of the training data set. Despite a better performance for the training

set, for both test sets, ANN presents a reduction in the RMSE as compared to PRODIGY (0.77 kcal.mol^{-1} for the Nbs data set, 1.18 kcal.mol^{-1}for the data set of general interfaces, and 1.00 kcal.mol^{-1}for the junction of both data sets) and the other trained models, SVR and XGBoost. One possible explanation for this, is that the PRODIGY may have achieved lower error on the training set by overfitting to the data, resulting in poor generalization to the test set. Our model, on the other hand, may have achieved better generalization and a smaller bias to the test set at the expense of slightly higher error on the training set, which suggests that the ANN may be better able to generalize to unseen data. Thus, the ANN was the chosen model for predicting ΔG of binding.

Table 2 – Evaluation metrics (RMSE and $R_{Pearson}$ for the training and test sets.

Model	Train - R	Train - RMSE	Test - R	Test - RMSE
XGBoost	0.94	1.15	0.01	2.59
SVR	0.49	2.43	0.00	2.65
ANN	0.52	2.43	0.11	2.14
PRODIGY	0.74	1.9	-0.04	3.13
Model	Nbs - R	Nbs - RMSE	PDBBind - R	PDBBind - RMSE
XGBoost	0.15	2.61	0.04	2.55
SVR	0.10	2.77	-0.08	2.55
ANN	0.45	1.66	-0.12	2.45
PRODIGY	0.18	2.43	-0.08	3.60

From Table 2, it can be observed that even though our ANN model presents a consistently better $R_{Pearson}$ across the separated and joined test sets when compared to PRODIGY, it stills presents very low values, indicating lack of correlation between the experimental and predicted data. To investigate this weak correlation, we have identified and removed from the test sets the "challenging cases". These challenging cases were defined as targets presenting an error higher than 3 kcal.mol^{-1} for our ANN regressor, resulting in the removal of 13 structures (2BSE, 4WEM, 5XKH, 5KY4, 5KY5, 6DDM, 6FG8, 4LHJ, 5B78, 4M0W, 4UYP, 4YI8, and 5NT7). For these 13 structures, the error obtained from PRODIGY predictions were equivalently high (Data in GitHub), being for some cases even higher such as 13 kcal.mol^{-1}. Upon removal of the challenging cases, the $R_{Pearson}$ for our ANN model rises to 0.53, and for PRODIGY to 0.31 (Figure 19), along with a decrease in the RMSE. After removing challenging cases, the RMSE of our model for the PDBBind and Nbs datasets was 1.46 and 1.44 kcal.mol^{-1}, respectively, with an $R_{Pearson}$ of 0.41 and 0.59 (Figure S3). For the PRODIGY, the RMSE values were 2.48 kcal.mol^{-1} (with an $R_{Pearson}$ of 0.32) and 2.25 kcal.mol^{-1} (with

an $R_{Pearson}$ of 0.27) for the PDBBind and Nbs data sets, respectively. It is important to note that, regardless of the presence of challenging cases, the RMSE of our model remained lower to PRODIGY and comparable to other existing models, as will be discussed further in the paper.

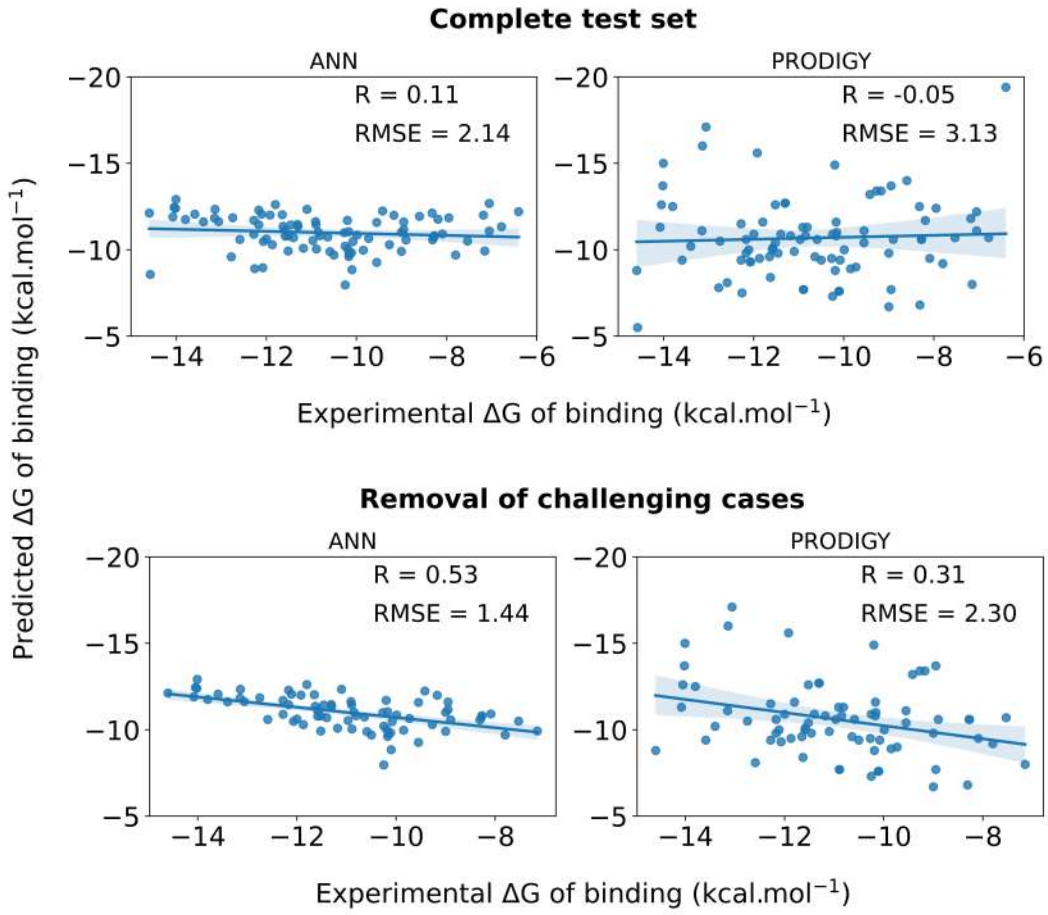
Among the challenging cases that we have identified, we have found that the PDBs ID 4WEM, 5XKH, 5KY4, 5KY5, 6DDM, and 6FG8 correspond to glycosylated structures. These structures contain glycans, which are complex carbohydrates that can modulate the dynamics and structure of proteins. In our calculations, we do not take glycans into account, which may be a factor in the poor performance of our model and the PRODIGY model for these cases. Structures such as 4LHJ were obtained at pH 4.5, which can affect the conformations due to differences in protonation states, in this work, considered at a pH 7.0. In 5B78, it presents a zinc-finger motif, characterized by a conserved pattern of cysteine and histidine residues that coordinate zinc ions, and has effect in the atomic parameterers of the coordinating residues, specifically the charge distribution.(FERRAZ et al., 2022)

However, for the other cases, it was not clear why the predictions were not accurate. One of the putative reasons could be related to the inability in the crystal structure to consider conformational rearrangements upon binding. This is supported by the difficulties in obtaining accurate ΔG of binding for PRODIGY, which also utilizes bound-state parameters.

While there are numerous methods for computing the binding free energy between protein-ligand complexes, the options for calculating the binding free energy between protein-protein interactions along with systematic comparisons are relatively scarce. In a previous study,(KASTRITIS; BONVIN, 2010) various computational methods were evaluated for their ability to predict the ΔG of binding. In this study, the authors have compared the performance of different methods using the same data set as PRODIGY and our work have used for training. Their results revealed a $R_{Pearson}$, ranging from -0.29 to 0.18, emphasizing the complexity of this task. It is worth noting that in this work, the authors have mainly employed docking score functions for protein-protein interactions as a mean of assessing the relative ΔG of binding for different targets. However, such methods and respective score functions are primarily used for ranking different docking poses, and may not be optimal for accurately determining the relative ΔG of binding for different targets. As such, it is hard to compare the performance of these different methods with ours, which was validated using a different test set and is specifically designed for the calculation of binding free energy rather than for ranking native-like poses.

Panday and Alexov,(PANDAY; ALEXOV, 2022) have reported on a Gaussian-based method,

Figure 19 – Linear relationship between the experimental and predicted ΔG of binding for the full validation set and for the training set without challenging cases



termed f5-MM/PBSA/E, for calculating binding entropy in protein-protein interactions and incorporating it in MM-PBSA calculations. The model was validated using a data set consisting of 46 protein–protein binding cases, based on the same training set of this work (consisting of 81). The model takes into account the binding entropy using independently energy-minimized unbound and bound structures and yields an RMSE of 2.6 to 3.2 kcal.mol^{-1}. Even though they are not directly comparable, the obtained RMSE for our ANN model presents a lower error and falls within the same range for the f5-MM/PBSA/E method. Another method to compute protein-protein ΔG of binding uses co-alchemical water approach to study the efficacy of free energy perturbation calculations for charge-changing mutations at the protein-protein interface.(CLARK et al., 2019) The authors achieved an overall RMSE of 1.2 kcal.mol^{-1} on a set of 106 cases involving a change in net charge. It is important to highlight that despite the remarkable accuracy, the method was restricted to assess the impact of net charge alterations,

being an efficient protocol to compute affinities between related complexes.

Wang and collaborators (WANG et al., 2022a) have developed a general calculation method of ΔG of binding for protein-proteins using metadynamics simulations. The method was validated using 19 non-congeneric protein-protein complexes, resulting in an RMSE of 2.0 kcal.mol^{-1}. Metadynamics is a simulation method that uses bias potentials added to the Hamiltonian of the system to enhance the sampling of rare events along the pathway of CVs. Thus, one of the main strengths of metadynamics simulations consist of exploring multiple metastable states allowing the calculation of the free energy landscape of the system in multiple dimensions, providing a detailed view of the underlying molecular mechanisms involved in the (un)binding process. Therefore, metadynamics can capture the mechanism from the bound to unbound state transition in full atomistic detail.(GERVASIO; LAIO; PARRINELLO, 2005; BUSSI; LAIO, 2020) In addition, it was used with explicit solvation, providing a more realistic contribution of the solvent.

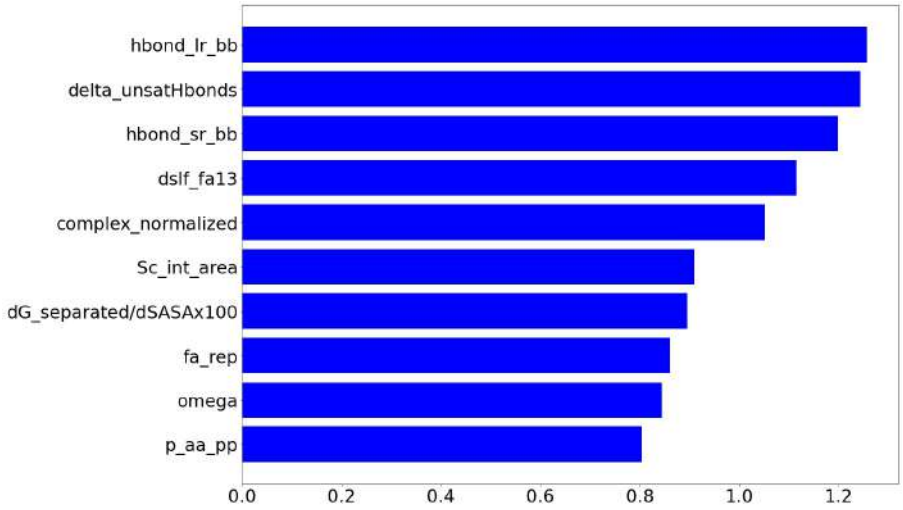
To compare the robust metadynamics-based method with our protocol, we have applied our pipeline to the same set of 19 protein-protein complexes of their study. While the metadynamics method obtained an RMSE of 2.0 kcal.mol^{-1}, our ANN presented an RMSE of 2.7 kcal.mol^{-1} (Table S4). It must be noted that it is not unexpected that our model was shown to be less accurate than metadynamics, as metadynamics is a very powerful method based on rigorous statistical thermodynamics concepts. One additional reason for discrepancy is that in their model, they have converted the dissociation constant (k_D) to binding ΔG considering a temperature of 310 K, while for our model, we have validated it at a temperature of 298 K. However, even though our model presents a higher RMSE, it is more computationally efficient, especially for large systems or systems with multiple metastable states mainly due to convergence issues of metadynamics. In some cases for the metadynamics-based method more than one simulation had to be carried out. Another advantage of our method is that it does not require complex set-up or choice of collective variables. This makes our method an attractive alternative for certain applications where computational efficiency is a key concern, such as for screening tasks or high throughput applications.

3.2.7 Features importance analysis

We have developed a model explainer to better understand the factors that influence the predictions made by our model as described in Section 3.2.2.5.5. From this process, it was

possible to estimate the ANNs' feature importance from all the original features generated by Rosetta (Figure S4), and therefore, it was possible to make the model interpretable. The highest ten importance (Figure 20) (from highest to lowest: hbond_lr_bb, delta_unsatHbonds, hbond_sr_bb, dslf_fa13, complex_normalized, Sc_int_area, dG_separated /dSASAx100 , fa_rep, omega, and p_aa_pp) were deemed the most relevant to show how the model calculates the ΔG of binding and to allow for the identification of relations between the features and output. These features can be divided into three categories based on their nature referred to as: 1. physical-based (hbond_lr_bb,hbond_sr_bb,dslf_fa13,fa_rep); 2. geometric-based (delta_unsatHbonds, Sc_int_area, complex_normalized, dG_separated/dSASA x100); and 3. knowledge-based (omega, p_aa_pp).

Figure 20 – Highest ten feature importance scores calculated.



The physical-based terms refer to the low and short range backbone energetic contributions to the hydrogen bons (hbon_lr_bb and hbond_sr_bb, respectively). The dslf_fa13 is the disulfide geometry potential, and fa_rep is the Pauli repulsive component of the Lennard-Jones potential. For the geometric terms, delta_unsatHbonds is the number of unsatisfied hydrogen bonds in the interface, i.e., buried hydrogen bonds. Sc_int_area denotes the shape complementarity of the interfaces divided by the interface area. Complex_normalized is the average energy of the residues in the complex, dG_separated/dSASAx100 is the binding free energy divided by the total SASA multiplied by 100. The omega and p_aa_pp refer respectively to the Omega dihedral in the backbone potential and the probability of amino acid to be at a given ϕ/ψ.

The geometric-based terms are not surprising, since it displays a higher weight to features

related to the extension/complementarity of the binding interface, and we have demonstrated a moderate correlation of these features to the binding affinity. Insofar the physical-based are mostly related to the patterns of hydrogen and intrachain disulfide bonds and the Lennard-Jones potential. In general, van der Waals forces are determinants of the formation of a protein-protein complex. The relation between hydrogen-bonding in protein interfaces and binding affinity is well described,(CHEN et al., 2016) in which hydrogen bonds are generally considered to be facilitators of protein binding since they can modulate the binding affinity by displacing protein-bound water molecules into the bulk solvent and reducing competitive interference for hydrogen bonds with water.(ABEL et al., 2008; BARILLARI et al., 2007) In addition, disulfide bonds have already been correlated to the binding affinity for reducing the entropic penalty due to the immobilization of mobile regions upon binding,(MENDOZA et al., 2020) especially for antibodies and nanobodies. Knowledge-based terms are less interpretable, but it shows that structural features are important for the outcomes of our model. In a previous study by Erijman et al (2016),(ERIJMAN; ROSENTHAL; SHIFMAN, 2014) the hydrogen bonds pattern and geometric complementarity were two of the features correlating with binding affinity, corroborating our findings.

3.2.8 Model's consideration

Our model demonstrates good accuracy using simple physical-chemical descriptors as input. Despite the complexity of the problem and limited training data, the model was able to achieve promising results, being comparable to highly accurate methods based on rigorous physics models, such as metadynamics.

One of the advantages of our model is its computational efficiency. It uses ANNs and is therefore able to compute the ΔG of binding very quickly, rivaling the speed of PRODIGY which takes a few seconds to complete the same calculation. In structural biology, it is often a good practice to geometry-optimize structures in order to alleviate atomic clashes and ensure experimental consistency. This is why our model includes an extra step for geometry optimization, which takes around 3 minutes on average, including the calculation of interface parameters, and can be done on a personal computer. While this extra step adds to the overall computation time, our model is still faster than other techniques such as MM-PB/GBSA or enhanced sampling methods, which can take significantly longer to complete. These methods are time-intensive, making our model a more computationally efficient option.

Our model performed very well for predicting the ΔG of binding for Nbs-antigens, being a promising tool for industrial and biomedical research. Experimental and computational efforts have been focused on optimizing Nbs as high-affinity binders, making the prediction of their ΔG an important inquiry. Soler et al. (2018)(SOLER et al., 2018), have used a combination of molecular dynamics simulations and scoring functions to predict the ΔG of binding for Nbs-antigen complexes, but only using a limited number of structures (seven in this case). In contrast, our approach consistently predicts the Nbs-antigen ΔG of binding and has been benchmarked using ca. 40 structures. The remarkable performance of the nanobodies targeting their antigens is likely related to the physical-chemical characteristic of their interfaces. Typically, nanobodies–antigens interfaces share common structural patterns related to general protein–protein interactions. The nanobody paratopes are enriched with aromatic residues, bearing a more hydrophobic character. Besides, nanobodies usually bind their antigens to more rigid and structured epitopes enriched with aromatic residues.(ZAVRTANIK et al., 2018) These general features can alleviate the patterns identified by the algorithms. Adversely, the data set of general protein interfaces contains uneven classes of protein–protein interactions, and hence, different nature of interactions.

Despite its good prediction capability, our model presents some limitations. Such as the lack of parameters for some molecular classes, e.g. as cofactors and carbohydrates. Not including carbohydrates in the analyses has shown to negatively impact the accuracy of predictions (even though the Rosetta software is capable of handling glycans,(FRENZ et al., 2019) it is currently not possible to use Rosetta in conjunction with InterfaceAnalyzer to carbohydrates). Another potential limitation of our model may be the variability in the conditions and methods used to collect the experimental data. This heterogeneity could affect the accuracy of the model's predictions. It is also important to mention that protein-protein binding is a thermodynamically driven process involving entropy changes that leads to conformational rearrangement of the binding partners. While the entropic term is implicitly modeled through the Gaussian exclusion solvent model (which at the considered temperature and solvation density it represents more than half of the entropic term) within the Rosetta functional, flexibility is not taken into account, since it utilizes a static structure representing a single state of the configurational energy landscape. Techniques like MM-PBSA or MM-GBSA, which are based on molecular dynamics simulations, may have an advantage in cases where entropy differences are small across the systems and can be canceled out in the calculations. However, these techniques are far more computationally expensive (requiring from hours to day and usage of

GPU resources), and our devised model still presents a good accuracy in predicting ΔG of binding. While out of the scope of this work, we hypothesize that the issue with dynamism may be alleviated through the inclusion of protein dynamism by generating a thermodynamic-consistent ensemble of conformational states using MD, MC simulations, or even generative neural network models.(ZHAO et al., 2008)

3.2.9 Application to engineered proteins against the domain B of Chikungunya virus envelope protein

The infection caused by the Chikungunya virus (CHIKV) is a highly concerning illness due to its severe febrile symptoms and debilitating joint pain, which can significantly impact quality of life. Currently, there are no effective treatments or vaccines available for this disease. Among the various proteins found in CHIKV proteome, the E2 envelope protein is especially critical to the infection process, as it interacts with cellular receptors and is the primary target of neutralizing antibodies.

In an effort to develop new protein-based therapies that target the E2 protein, Purificacao (2022) (thesis unpublished) employed computational techniques to engineer artificial proteins. Using molecular dynamics (MD) simulations and Rosetta parameters, three proteins were identified as promising candidates based on their structural and dynamic properties. These proteins were then subjected to microscale thermophoresis (MST) to measure their binding affinity. This value was expressed in units of mol.L^{-1} and then converted to energy in units of kcal.mol1 using the equation $\Delta G = -RTlnk_D$, with a temperature of 298 K assumed. Applying our ANN-devised protocol to the predicted three-dimensional structures for the designed proteins bound to the E2 domain, we obtain a close agreement with the experimentally measured binding affinity (Table 3). This highlights that the developed tool can be an important for protein engineering projects.

Table 3 – Comparison between experimental and predicted binding free energy for computer-designed proteins against the E2B domain of CHIKV.

System	**Experimental ΔG [kcal.mol^{-1}]**	**ANN ΔG [kcal.mol^{-1}]**
1	-9.5	-8.9
2	-8.9	-9.1
3	-9.4	-8.5

3.2.10 Conclusions

We present our newly deployed model for the prediction of the ΔG of binding between protein pairs. This model is based on simple descriptors of protein–protein interface and protein folding properties along with machine learning tools. Despite its simplicity it has been demonstrated to be consistent, providing consistent protein–protein ΔG of binding in a matter of seconds to minutes. Our model presents a reduction in the RMSE of nearly 1 kcal.mol^{-1} if compared to the PRODIGY web server. The application of the model presented in this work is of great interest to calculating ΔG, especially for the binding of nanobody-antigen complexes, opening venues for applications in the biomedical and biotechnological industries.

Given the simplicity of its implementation, the users can add extra sequence-based or structure-based terms that can boost the algorithm performance. While the results of our model show promise for the field of computational biology, it is important to note that all computer-based prediction methods, including ours, have a major limitation: the inhomogeneity of the experimental techniques exploited in the data sets. It is well-known that the measured ΔG of binding relies, among other factors, on the technique and physical-chemical conditions, and that changing one or both aspects will lead to different ΔG of binding. Therefore, the model deployed in this work is not tied to any specific technique, but it has demonstrated good accuracy to predict ΔG of binding regardless of the technique. This limitation, however, is inherent to any developed computer-based model up to date since the developed computer models do not commit to any technique in individual.

3.3 ASSOCIATION STRENGTH OF E6 TO E6AP/P53 CORRELATES WITH HPV-MEDIATED ONCOGENESIS RISK

3.3.1 Introduction

Papillomaviruses belong to the *Papillomaviridae* family and are classified into five genera (alpha-, beta- gamma, mu- and nu-papillomavirus) encompassing 49 species and more than 200 viral types (BZHALAVA; EKLUND; DILLNER, 2015). Human papillomaviruses (HPVs) are human parasites with tropism for squamous epithelium (HAUSEN, 2009; HAUSEN, 1996). Over 120 types of HPVs have been identified and approximately one-third of those infect the squamous epithelia of the genital tract (VILLIERS et al., 2004). Among those, 15 are categorized as high

risk and are considered the major cause of cervical cancer among women all over the world, with over 99% of cervical lesions containing viral sequences (MOODY; LAIMINS, 2010). Within the high risk group, the HPV16 and 18 types are the most prevalent ones, followed by the 31, 33, 35, 42, 52 and 58 types (CLIFFORD et al., 2006). High-risk HPVs are also associated with many penile, vulvar and anal carcinomas and contribute to 40 of oral cancers (MOODY; LAIMINS, 2010). The remaining viral types are usually associated with benign lesions, such as warts and condylomas, and therefore are categorized as low risk types (EGAWA; DOORBAR, 2017).

HPVs are non-enveloped, double-stranded DNA viruses with approximately 8 kb in size. Transcription is initiated from more than one promoter region and is polycistronic, yielding multiple mRNAs with several open reading frames (ORFs) divided into: long control region (LCR), late (L) and early (E) ORFs (BURD, 2003; FAVRE, 1975). The LCR region encompasses the gene regulation elements, such as promoters and transcription regulatory elements. The HPV genomes do not encode polymerases nor other enzymes required for proliferation and therefore depend on the host cell machinery to mediate viral DNA synthesis (MOODY; LAIMINS, 2010). Late ORFs (L1 and L2) encode the viral capsid structural proteins, which are expressed in the final stages of cellular differentiation, allowing for virus particle assembling and release in the extracellular medium (SCHIFFMAN et al., 2007; DOORBAR, 2005; HAUSEN, 2002; ZHENG; BAKER et al., 2006; GRAHAM, 2010). The early ORFs are named E1, E2, E4, E5, E6 and E7 and encode the viral cycle regulatory proteins at the early stages of cellular differentiation. These proteins modulate a series of processes in the host cell leading to the establishment of abnormal cellular growth and cancer. Among these proteins, E6 and E7 are considered the key components in the cancer development pathway and malignancy by stimulating cell proliferation, inhibiting apoptosis and inducing cell transformation and immortalization.

Once HPV infects cells, the viral genome is established as an extrachromosomal element or episome and is utterly dependent on the host cellular replication proteins to mediate viral DNA synthesis (MITTAL; BANKS, 2017). The proliferation capacity of the HPV-infected cells is uncoupled from differentiation through the inactivation of key cell cycle regulators. Among these, the most prominent ones are members of the retinoblastoma (Rb) family. The HPV E7 protein binds to underphosphorylated forms of Rb family members and targets them for degradation, which results in cells re-entering the S phase (MOODY; LAIMINS, 2010; MITTAL; BANKS, 2017; GAGE; MEYERS; WETTSTEIN, 1990; POL; KLINGELHUTZ, 2013). The efficient binding of Rb by E7 can lead to inhibited cell growth and apoptosis through a p53-dependent

pathway. As a consequence, high-risk mucosal (hrm) HPV E6 proteins have evolved to overcome the oncosuppressor functions of p53 by targeting this protein for degradation through the ubiquitin-proteasome pathway via interaction with the human ubiquitin-ligase E6AP (MITTAL; BANKS, 2017; HUIBREGTSE; SCHEFFNER; HOWLEY, 1991). This process culminates in the formation of the E6/E6AP/p53 complex, which results in prevention of cell growth inhibition in both undifferentiated and differentiated cells and is the key point in the HPV-mediated oncogenesis process (MARTINEZ-ZAPIEN et al., 2016). In the E6-mediated degradation of p53, hrm-HPV E6 proteins interact with the acidic leucine (L)-rich LxxLL motifs of E6AP, leading to recruitment and polyubiquitination of p53 (POL; KLINGELHUTZ, 2013; ZANIER et al., 2013; POIRSON et al., 2017; CONRADY et al., 2020). Neither E6 nor E6AP alone are capable of recruiting p53 (MARTINEZ-ZAPIEN et al., 2016), however the isolated LxxLL peptide of E6AP is sufficient to render E6 liable to interact with p53 (ANSARI; BRIMER; POL, 2012). This interaction is the first step in the formation of the stable complex (MITTAL; BANKS, 2017; POIRSON et al., 2017; TOMAIĆ; PIM; BANKS, 2009), and is essential to allow the recruitment of other ubiquitin-ligases, leading to p53 degradation (NOMINÉ et al., 2006; DREWS; BRIMER; POL, 2020).

The crystallographic structure of the E6/E6AP/p53 complex reveals structural details on the complex formation (MARTINEZ-ZAPIEN et al., 2016). In this structure, the E6 protein is bound to an E6AP-derived 12-mer peptide containing the LQELL motif and to the DNA binding domain of p53. Although the described structure provides some insights on the role of the E6-binding region of E6AP in the E6-p53 interaction, it provides only limited information since the entire E6AP binding domain is comprised by 18 residues, whereas the crystallized peptide is a 12-mer, and a maltose-binding protein was fused to the E6AP-derived 12-mer peptide to allow efficient protein production. Therefore, missing interactions associated with the addition of an extra component to the complex may result in changes to the native conformation as well as to the stability of the complex. To address these issues, we have refined the E6/E6AP/p53 available structure by modeling and molecular dynamics simulations. The resulting data was subsequently used to model and assess the biophysical binding properties of 40 HPV E6 variants to the E6AP/p53 complex aiming to better understand the relationship between the formation of the trimeric complex and HPV-mediated oncogenesis at the molecular level. To this end, machine learning (ML) models were used to classify and identify biophysical-based features (geometric and statistical-mechanic energetic features) into the different HPV risk type. In summary, by harnessing ML algorithms associated with protein engineering methods,

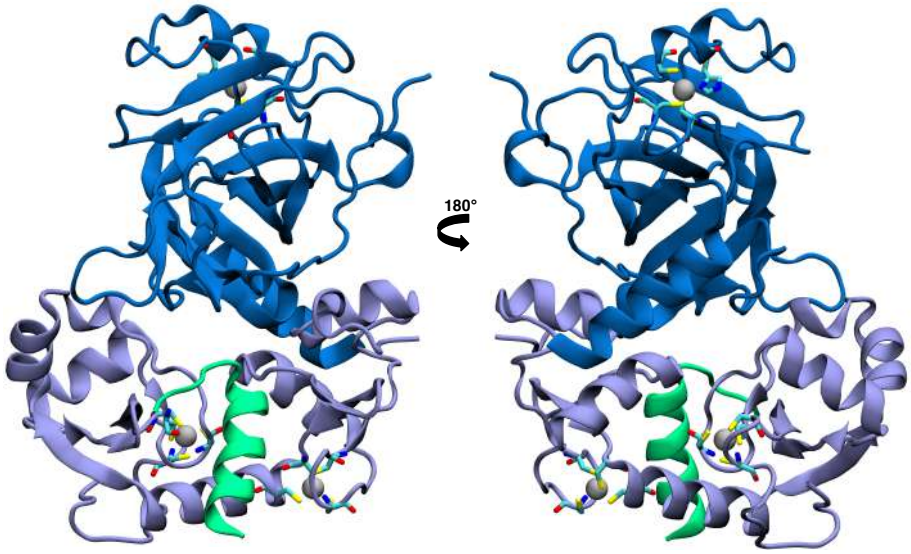
we seek to propose a consistent structure-based model that can classify between high and low risk HPV oncogenesis potential, that can be used as a fast-screening approach for novel and unclassified HPV variants. Moreover, our data shows an association between thermodynamics and HPV oncogenesis risk, and sheds light into the potential of p53 degradation by HPV types currently not classified as high-risk.

3.3.2 Computational procedures

3.3.2.1 Atomic charges' parameterization for the zinc finger motifs

E6 and p53 proteins have a total of three zinc-finger motifs (Figure 21), where the zinc ions are coordinated by deprotonated cysteine and histidine residues. Point charges for all zinc finger motifs were obtained by unrestricted Hartree-Fock self-consistent field calculations at theory level 6-31G* followed by RESP adjustment (BAYLY et al., 1993) using the NWChem software (VALIEV et al., 2010). A group charge approach was subsequently manually adjusted to ensure that each cysteine and histidine residue had the expected net charge of -0.5e, while the zinc ion charge was kept as $+2.0e$.

Figure 21 – Cartoon depiction of the E6/E6AP/p53 complex. The proteins E6, E6AP, and p53 are represented in purple, green, and blue, respectively. The zinc fingers are visualized with grey zinc atoms and the coordinating residues (His and Cys) depicted as spheres and cylinders, respectively.



3.3.2.2 Model building of the HPV E6 variants in complex with E6AP/p53.

An MD simulation was carried out to equilibrate the structure of the system after the inclusion of the E6AP peptide missing residues and removal of the fused maltose binding protein (FERRAZ et al., 2022). The equilibrated structure was used as the starting point to model the variants, which in its turn, the mutants' global minima was searched using a Monte Carlo-based protocol. The energy optimized most representative structure for the E6/E6AP/p53 complex retrieved from a cluster analysis of the MD trajectory was used as a starting point to generate molecular models for the additional 39 E6 HPV variants. All 40 variants share the same molecular pathogenesis involving the formation of the E6/E6AP/p53 complex (genus Alphapillomavirus); amino acids sequences were retrieved from the Papillomavirus Episteme Database – PaVE (<https://pave.niaid.nih.gov> (DOORSLAER et al., 2012). A list of the studied E6 HPV types and the classifications according their clinical oncogenic risk is shown in Table 4, where 16 are classified as low-risk (LR), 15 as high-risk (HR) and 9 as unclassified-risk (UR). HPV oncogenesis risk classification is presented here as a consensus compiled from literature data (SO et al., 2016; MUÑOZ, 2003; MEISAL et al., 2017; ANANDHARAJ et al., 2015; VARNAI et al., 2007) according to clinical data.

Table 4 – List of HPV E6 sequences used in this study and corresponding squamous cervical cancer risk classification.

Oncogenic potential	**HPV Type**
Low risk	2, 3, 6, 10, 11, 30, 40, 42, 44, 61, 62, 72, 81, 84, 89, 90
High risk	16, 18, 26, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73
Unclassified risk	7, 32, 34, 67, 71, 70, 82, 83, 87

The sequences of each viral type were pairwise aligned against the reference sequence of the E6 protein of HPV16 to identify the mutation sites. The online tool EMBOSS Stretcher (<https://www.ebi.ac.uk/Tools/psa/emboss_stretcher>) 52-54 was used with the following parameters: substitution matrix BLOSUM62, penalty for gap opening = 12 and penalty for gap extension = 2. After alignment, few N- and C-terminal unaligned residues were removed so that all HPV E6 sequences had the same number of residues in equivalent positions to each other for structural comparison purposes. It is worth noting that none of the few N- and C-terminal removed residues were neither interacting nor in contact with E6AP and/or p53. All modelled 40 E6 proteins have their amino acid sequences in positions ranging from 5 to

141, taken E6 from HPV16 as reference. To build the additional 39 E6/E6AP/p53 complexes, *in silico* mutations on E6 were performed using the Rosetta v.3.8 software suite (LEAVER-FAY et al., 2011b) on the MD-obtained structure of the HPV16 trimeric complex. All zinc finger motif topologies were included into the Rosetta's force field (Figure S1). The generated structures (including the reference HPV16 complex) were refined using the Rosetta package of software v. 3.10 (LEAVER-FAY et al., 2011b) aiming to achieve the closest state representing a native-like conformation. Starting from the MD-equilibrated structure, the modeling of the variants was subjected to a Rosetta-based Monte Carlo sampling, which includes sampling of the side and main chains, and which is amenable to capture conformational changes due to the mutations, exempting the use of time-consuming protocols such as MD simulations for all the variants. The refinement was carried out by considering the conformational fluctuations in response to thermal energy and the mutations. To this end, side chain and backbone sampling were performed to consider the subtle structural shift due to the mutations in the native structure. Initially, backbone sampling was accomplished using backrub-like backbone simulation (SMITH; KORTEMME, 2008) as implemented in Rosetta. These simulations consist of Monte Carlo (MC) moves to approximate the backrub motion, i.e., alternate hinge-like conformational shift of the backbone as observed in the crystal lattice of proteins.(DAVIS et al., 2006) A benchmarking and validation of the backrub Rosetta method on a set of 2000 point mutations can be found in Bordner et al. (BORDNER; ABAGYAN, 2004). In addition, it has also shown to improve the agreement between modelled structured and relaxation order parameters as obtained from NMR data (FRIEDLAND et al., 2008). Thus, these simulations are an efficient sampling algorithm to predict the structures of mutations. The backrub protocol was used with 10,000 MC trials generating 1,000 models for each variant. The temperature value in k_BT was set as 0.6 and the C-α atoms were defined as the main chain atoms pivot. Minor side chain sampling was carried out by defining the probability of making a side chain move as 25. To post-process the generated data, the selection criterion was the lowest energy structure as calculated with the Rosetta force field. A second step of refinement consisting of a high-resolution, full-atom MC search was used to refine the models at atomic level resolution through the RosettaDock's algorithm (GRAY et al., 2003) and energy function. To this end, prior to a local refinement docking calculation, a pre-packing step was performed to optimize the monomer's side chains separately. Docking Local refine was used to refine the models using only the high-resolution step of the docking protocol, which consists of side chain optimization via rotamer packing and continuous minimization. To increase fine grained rotamer selection,

extra side chain rotamers were added. It was also incorporated the side chain refinement technique proposed by Wang et al. (WANG; CIEPLAK; KOLLMAN, 2000), which enhances the recovery of correct rotameric side chain conformations. A total of 1,000 decoys were generated for each variant, in which the one containing simultaneously the lowest total score and lowest interface score was considered. For all the calculations and procedure described using the Rosetta package, the REF15 potentials (ALFORD et al., 2017) were employed, unless stated otherwise.

3.3.2.3 Binding free energy calculation

Binding free energies were calculated using the *InterfaceAnalyzer* protocol of the Rosetta package. The Rosetta package utilizes its energy function, which its functional consists of atomic packing and pairwise-additive approximation. Implicit solvation is included based on Lazaridis-Karplus Gaussian solvent-exclusion model for the solvation free energy (LAZARIDIS; KARPLUS, 1999). Therefore, the Rosetta energy function accounts for entropy. The Rosetta energy function is able to reproduce thermodynamics observables, such as liquid-phase properties and liquid-to-vapor transfer free energies (PARK et al., 2016), and has been used to calculate free energies of binding and differences in the free energy of binding presenting correlation with experimental observations (KORTEMME; BAKER, 2002; FERRAZ et al., 2021; BARLOW et al., 2018). The binding free energy was estimated by taking the difference between the energy of the complex and of the separated binding partners subsequently to repacking of the bound and unbound (bound human complex E6AP/p53 is separated from the HPV E6 subtypes) state. The measure of the binding free energy was averaged over 5 repeats.

3.3.2.4 Machine learning

To gain insight into properties that would possibly discriminate between high and low risk oncogenic potential, should they exist, ML algorithms emerge as convenient methods since they aim to extract complex patterns and relationship from the data sets and are able to predict properties of interest about the system under study. Given that the three-dimensional arrangement of protein structures dictates their chemical properties, ML can associate the latter with the former by using specific mathematical models and enough training data. HPV E6 variants complexed with E6AP/p53 were labeled into two groups: 1) high oncogenic

potential risk; and 2) low oncogenic potential risk, resulting into a binary classification task. To obtain features that describe the E6 - E6AP/p53 interaction interface and can be numerically translated to train and test our ML models as the input features, physical properties of the complexes' interfaces were calculated using the *InterfaceAnalyzer* in conjunction with the REF15 (ALFORD et al., 2017) and RosettaScript (FLEISHMAN et al., 2011a; BENDER et al., 2016). Through this protocol, 53 interface-related features were calculated for each of the modeled system. The considered features for the subsequent ML models' constructions are depicted on the Supplemental Information. These features encode geometrical and physical-chemical properties of the interfaces and were calculated by separating the chains of the complexes and packing them when separated. The maximum value for the surface area solvent exposed (SASA) to be considered as buried was of 0.01 Å, and the probe radius to calculate the SASA was 1.2 Å. Initially, the data was pre-processed by standardizing features, i.e., removing the mean and scaling to unit variance. The Boruta algorithm (KURSA; JANKOWSKI; RUDNICKI, 2010) was used to identify a subset of optimal variables and reduce the dimensionality of the data. Then, from the selected features, the linear discriminant analysis (LDA) (VAPNIK; VAPNIK, 1998) was solved by singular value decomposition, and no shrinkage was employed. The data was firstly fit, and then transformed. LD loadings were used to weight the importance of the features for binary classification. We considered a total of nine classification algorithms as implemented in the Python v. 3, along with Sci-Kit learn (PEDREGOSA et al., 2011b) library and Keras (GULLI; PAL, 2017). The used methods were k-Nearest Neighbors (kNN) (VAPNIK; VAPNIK, 1998), Support Vector Machine (SVM), Gaussian Process (GP), Decision Tree (DT73), Random Forest (RF74), Neural Networks (NN75), AdaBoost (AB76), Naïve Bayes (NB) (VAPNIK; VAPNIK, 1998), and Quadratic Discriminant Analysis (QDA) (VAPNIK; VAPNIK, 1998). Data manipulation was performed with Pandas library (MCKINNEY et al., 2011) and numerical calculations were carried out with the NumPy package (HARRIS et al., 2020). The parameters used for each method is depicted in the Supplemental Information, the Jupyter-notebooks, PDBs and data sets are accessible from the Supplemental Information and Github (<https://github.com/mvfferraz/HPVOncogenicPotential>). To assure reproducibility of the data, the random state seed was defined as 100.

3.3.2.5 Sequence logo

Residue conservation on the interface of E6 with E6AP/p53 was graphically represented by sequence logos generated for the three HPV risk groups (HR, LR, UR) using the Seq2Logo 2.0 server (available at: <http://www.cbs.dtu.dk/biotools/Seq2Logo/>) (THOMSEN; NIELSEN, 2012) on the P-Weighted Kullback-Leibler type. The Heuristic clustering method was used to reduce sequence redundancy and the weight on prior value was assigned to 200, since the alignment files contained less than 50 sequences. All graphical layout parameters were set to the default values.

3.3.3 Results and discussion

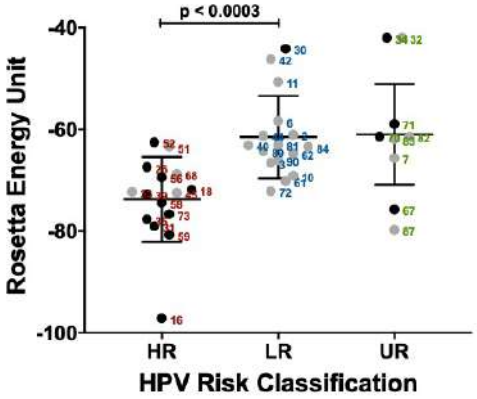
3.3.3.1 Relative binding free energies of HPV E6 variants to the E6AP/p53 complex

E6-mediated degradation of p53 is considered a crucial step in the development of HPV-associated cancers. Binding free energies of the 40 different HPV E6 variants with the E6AP/p53 complex was calculated to assess the influence of E6 association strength on HPV oncogenic potential (Figure 22).The E6 protein from HPV 16 exhibited the lowest binding free energy value from the assessed dataset. This variant is known to account for most cervical cancer cases and therefore it would be expected, should our hypothesis be held valid. Statistical significance of the calculated binding free energy average and the corresponding standard deviations for both groups was assessed by an unpaired t-test. A p-value $= 0.0003$ was obtained, suggesting that thermodynamics of association between HPV E6 to E6AP/p53 complex play a pivotal role in the oncogenic potential of HPV types. Analogously, differences in affinity of high- (HPV16 and 18) and low-risk (HPV6 and 11) E7 proteins to p105-RB have been previously studied. The authors have shown that the apparent affinity of association to p105-RB was lower for the low-risk types, suggesting a similar mechanism to what is herein presented for E6. However, association strength of HPV E7 to p105-RB did not correlate well with the ability to transform keratinocyte cell lines, which led the authors to state that such differences in biological behavior could not be only attributed to the potential of the E7 proteins to associate to p105-RB (MÜNGER et al., 1989).

Unclassified-risk E6 proteins have shown binding free energy values compatible with both, low- and high-risk HPV types. It is important noting that risk classification of HPV types

regarding cancer mediation is mainly based on epidemiologic data and that carcinogenicity risk has been established mainly based on cervical cancer cases. HPV DNA sequence variation and evolutionary relatedness have long been used to determine viral types and phenotypic characteristics. Similarly, sequence variations at certain sites of the viral genome have been associated to HPV's ability of modulating its carcinogenic properties and host immune responses (CHAN et al., 2013). A lack of consensus on some HPV types has been a matter of debate on the literature and some HPV types have been categorized as unclassified regarding to their epidemiologic oncogenic risk. Nevertheless, studies have consistently showed that high-risk HPV types induce the E6/E6AP/p53 complex to bind and ubiquitinate p53 oncosuppressor protein, leading to subsequently degradation of p53 (MESPLÈDE et al., 2012). In addition, E6 of some viral types categorized as low- and unclassified-risk have been shown to mediate p53 degradation with the same efficiency as the high-risk types (SCHIFFMAN; CLIFFORD; BUONAGURO, 2009; MESPLÈDE et al., 2012; FU et al., 2010; GALLOWAY; LAIMINS, 2015; KRANJEC; DOORBAR, 2016). This trait conferred the term of p53 degrader to these E6 variants. We have compiled all available literature data (up to our knowledge) for all known E6 variants classified as p53 degraders, which are represented by a filled black dot in Figure 22. It is noticeable that the majority of p53 degraders fall, epidemiologically, into the high-risk category. Only one low-risk type has been described as p53 degrader; its association energy falls far outside of the high-risk type interval. Most interestingly, four unclassified-risk viral serotypes (34, 67, 70 and 71) are listed as p53 degraders.Among them, E6 type 67 has a calculated binding free energy value to the E6AP/p53 complex comparable to the high-risk group.

Figure 22 – Binding free energy values for the association of E6 to the E6AP/p53 complex of 40 HPV viral types, as a function of oncogenic risk. (Energy values are listed as REU, which stands for Rosetta Energy Unit. Filled black dots correspond to HPV types where E6 has been reported as p53-degrader. HR: high-risk; LR: low-risk; UR: unclassified; types are listed sided by their corresponding dots and color-coded as a function of risk as red, blue and green, respectively, for further highlight.



3.3.3.2 Discriminating between high and low risk for oncogenic potential

If binding strength of E6 to the E6AP/p53 complex is related to HPV type oncogenic potential, via degradation of p53, it is expected that evolution has shaped high-risk E6 sequence/structure towards optimal interactions with the E6AP/p53 complex. In order to test this hypothesis, we have used statistical learning tools to identify physical-chemical components that are responsible for the higher affinity of oncogenic HR-type complexes. Having developed the dataset based on the Rosetta package structural features, we have applied linear discriminant analysis to weight which of the features are responsible to discriminate between HR and LR complexes. These features consist of a linear combination of traditional molecular mechanics physics-based terms, and empirical terms to quantitatively mimic geometric patterns observed in the PDB.

Initially, we have utilized the Boruta feature selection 69 approach to map the minimal-optimal set of features to the classification task. The Boruta algorithm applies a random forest 74 classifier to determine the relevant features subset, through a wrapper approach. Through the application of Boruta, seven features were single out: i. number of buried unsatisfied hydrogen bonds in the interface; ii. Lennard-Jones attractive potential of the complex; iii. Coulombic electrostatic potential of the complex; iv. Lazaridis-Karplus solvation62 free energy of the complex; v. Energy of backbone-side chain hydrogen bonds; vi. energy of short-range

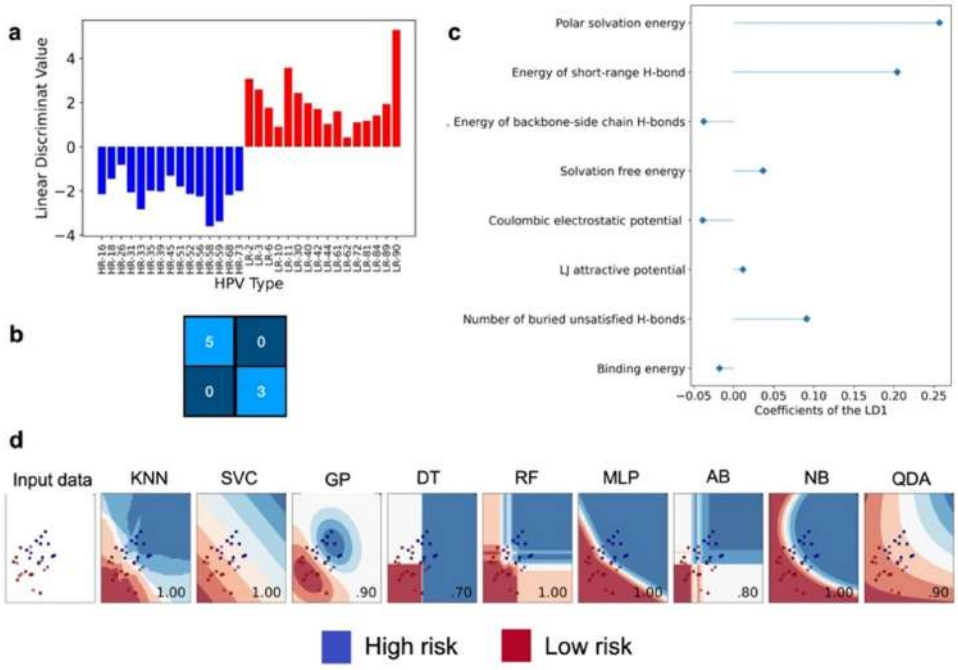
hydrogen bonds; and vii. polar solvation energy of the complex. In addition, given the importance of the binding energy, as we have previously demonstrated, this feature has also been included. The distribution of these variables is represented as a density plot found in the Supplemental Information (Figure S2).

Based on the filtered features, we have used classificatory ML-based algorithms to identify patterns capable to discriminate between low and high-risk type. Initially, the linear discriminant analysis (LDA) was employed. The LDA methodology consists of a projection of the input data to a linear subspace. This subspace is constituted of directions able to maximize the separation between the classes and minimize the separation among a class. This projection is based on Bayes' statistics, and it is a generalization of Fisher's linear discriminant. Therefore, the LDA allows for the separation of different classes by means of linear combination.

Using LDA-based classification for this dataset, the classes low and high-risk type are linearly separable. Since we have only two classes, a single LD function was specified. The LD values calculated for the first LD, i.e, the value for the linear combination of each value of the feature weighted by the LD coefficient, are positive for low-risk type, and negative for high risk-type (Figure 23A). This pattern was observed for all the samples in the dataset, corroborating with the 100% of explained variance considering an only LD component. A confusion matrix was built to diagnose the performance of the LDA. The data set was split into training set (75%) and test set (25%), and the confusion matrix is built based on the test set predictions. Each line of the confusion matrix represents a real class, whereas each column represents a predicted class. As it can be seen (Figure 23B), the LDA model was able to accurately predict the classes for the samples in the test set, suggesting the accuracy of the classification.

It is worth noting that an advantage of LDA over other classical ML methodologies, is that the LDA allows for computing the weight of each feature for the classification, rather than acting as a "black-box" these are expressed by means of the LD loading (also referred to as the slopes, coefficients, or weights). The loadings represent the scaling value of the LDA object for each feature on each discriminant function. The magnitude of the loading indicates which features contribute most to that discriminant function. Largest loadings, positive or negative, are associated to larger influence on the classification between the classes. The data show that the features with the highest modulus of the loadings are the polar solvation energy of the complex, and the energy of the short-range hydrogen bonds (Figure 23C). This result indicates that these two features present the highest contribution to linearly separate between

Figure 23 – Machine learning models for the binary classification between high and low risk groups. (a) First Linear discriminant value calculated for each sample of the dataset demonstrates that the explained variance considering only one LD is of 100%; (b) Confusion matrix for the LDA model, light blue indicates the number of correctly predicted instances, and dark blue is the number of instances that are mislabeled by the classifier (c) LD loadings for the Boruta-selected features; (d) Decision boundaries computed for nine different machine learning models and their mean accuracy. The decision boundaries are drawn considering the two features with highest LD loadings: polar solvation energy and energy of short-range H-bonds. The numerical value within the decision boundaries plot represents the mean accuracy. Classification is color-coded in (a) and (c), in which blue denotes LR and red denotes HR.



the classes by means of the LDA. We have previously demonstrated that the association free energy of E6 to the E6AP/p53 complex is, in average, statistically different between the low and high-risk type classes. Based on the LDA model, it suggested that the physical-chemical components within the association free energy that are responsible for the separation between the low and high-risk type are the solvation free energy of the complex and interface-hydrogen bonds energy.

Aiming to visualize how these two single features (polar solvation free energy of the complex and short-range hydrogen bond energy) impact the classification between low and high-risk types, we have plotted the decision boundaries for nine different ML/DL models considering these two features (Figure 23D). The decision boundary (or decision surface) consists of the region of a problem space in which a hypersurface partitions the vector space into a set for each class. Thus, the decision boundary is a graphical representation of the separation of

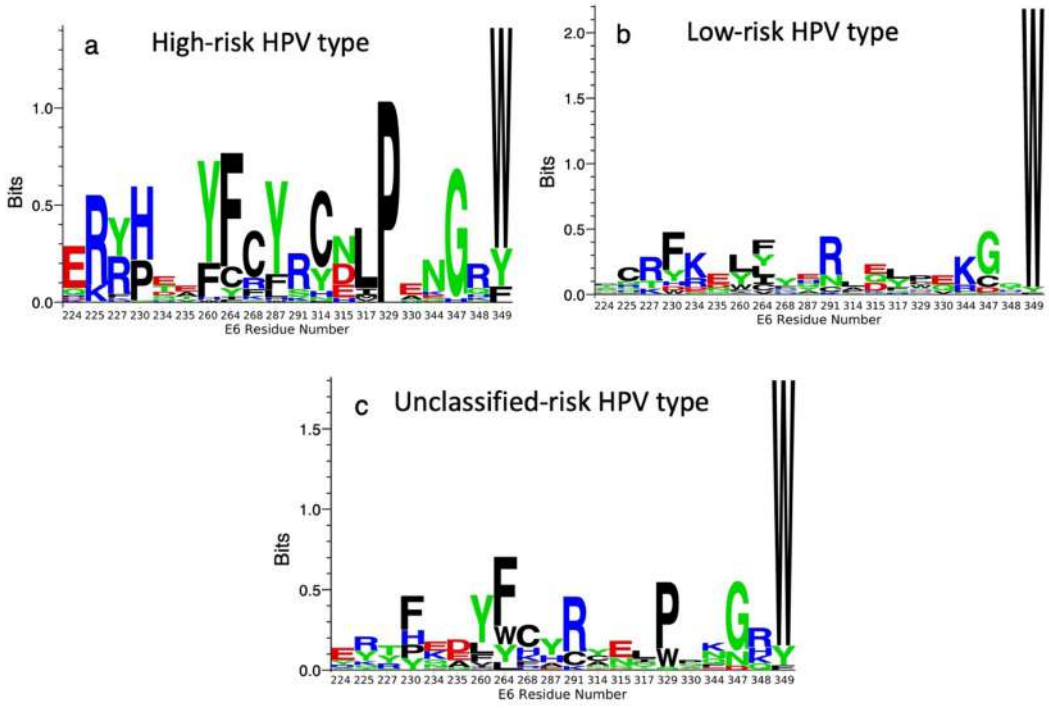
classes by a geodesic element. Figure 23D shows the mapping of the combination for these two features onto a plane (input data), and how different algorithms fit the data. Red dots on the input data represent high risk type, and blue dots represent the low-risk type. The different algorithms separate the sample differently, and in all of them, it can be seen disparities between both classes. To assess how well these decision boundaries can draw separation lines between the classes, we have calculated the training accuracy for each model. The accuracy was calculated based on a test set containing 30% of the original dataset. As it can be seen, all the models presented high accuracy, with values $> 70\%$. Accordingly, these results corroborate the relevance of these two features for the classification between low and high-risk type.

3.3.3.3 E6 interacting interface to E6AP/p53 pattern differs according to HPV risk type

The ML models suggest that the interface between E6 and E6AP/p53 complex present a different physical-chemical profile between high- and low-risk HPV types. Therefore, we have investigated the sequence diversity of the E6 interacting residues. An E6 residue was considered as interacting if it was within 4.0 Å from the interface between the E6 and E6AP/p53 dimer. The relative frequency of all interacting residues in E6 for all 40 HPV types was calculated and plotted as logo per clinical classification of the oncogenic potential of the serotypes (Figure **??**). While W349 is a highly conserved residue in E6 across all HPV types, remarkable differences can be seen when comparing the sequence pattern of the E6 interacting residues to the complex in high- and low-risk HPV types. High-risk HPV types have a highly conserved glutamic acid in position 224 (Figure **??**a), while low-risk HPV types do not display any preferred residue type (Figure **??**b). In addition, positions 234 and 235 in high-risk HPV types are occupied by negatively charged residues, while a high occurrence of positively-charged residues is present at position 234 in low-risk types. Other conserved residues among high-risk types that are not found in low-risk types include an aromatic residue at position 260 and 287 (Y or F), a proline at position 329 and an excess of positively charged residues at positions 225, 230 and 348. Interestingly, sequence profiling for clinically unclassified-risk HPV types shows a higher similarity pattern to high-risk types (Figure **??**). This finding corroborates with the fact that the calculated binding free energy values for most of the unclassified-risk types fall in the high-risk type interval (Figure 22). The fact that these HPV types have not been classified as high-risk suggests that other biological factors may act orchestrated with binding strength to influence HPV virulence. Nevertheless, our data show that a high affinity between an E6

protein and the E6AP/p53 complex is required by an HPV type to mediate oncogenesis.

Figure 24 – Relative frequency of E6 residues interacting with E6AP/p53 complex for (a) high-risk, (b) low-risk and (c) unclassified-risk HPV types.



3.3.3.4 ML-driven prediction of the oncogenic potential for the unclassified risk type

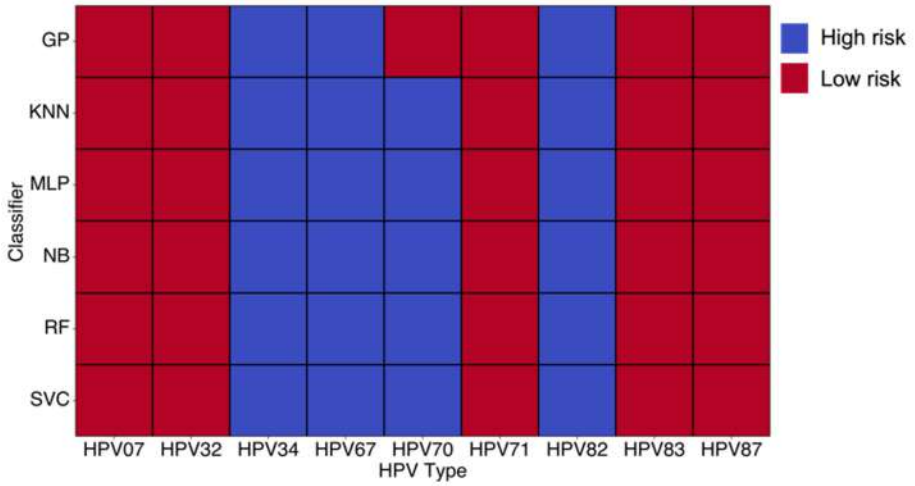
So far, the classification of HPV types into low-risk and high-risk lacks consensus for a clear-cut definition. Since the classification of the oncogenic potential is based on epidemiological studies, HPV types with low prevalence are mostly unclassified regarding to their risk. Of great interest is a consistent proposal of the risk for a given unclassified HPV type. This information is crucial for patient care and screening strategies by planning vaccinal strategies and large number of people testing program. One of the abilities of ML is to predict a class label based on patter recognition for a given input of data. Harnessing ML potential, we have trained nine different classifiers to predict the unclassified risk HPV types based on the physical-chemical properties, according to the Boruta-selected features including the binding energy. The performance of the classifiers was assessed through a combination of evaluation metrics. Initially, a 10-fold evaluation was employed to compute the accuracy of the model. A k-fold approach consists of splitting the training set in k-folds (here k=10) and calculating

the accuracy in a loop for each different dataset. The final accuracy is taken as the average for each separated set. To avoid bias from the dataset nature, a confusion matrix was built for the models. The dataset was split into a training set (70% of the total dataset), and a test set (30% of the total dataset). To quantify the information contained within the confusion matrix, threshold metrics were calculated. The threshold metrics are denoted by the precision (the rate of positive instances that are correctly detected by the classifier) and recall (or specificity, is the rate of positive instances that are mismatch predicted). The formulation for the threshold metrics is found at the Supplemental Information. Lastly, the areas under the curve (AUC) of the receiver operating characteristic (ROC) curves (Figure S3) were plotted, to provide further support to the previously calculated diagnostic metrics. In addition, a 10-fold approach was used to compute the mean ROC-AUC for each model.

Figure S4 presents the diagnostic metrics. Except for DT, AB, and QDA classifiers, the remaining presented a mean accuracy higher than 0.95. A striking feature of the ML models is that it can accurately classify between high and low risk, even though the alignment shows that for the unclassified-type risk does not present major sequences variation as compared to high and low risk type. These results demonstrate the robustness of our model and suggest a consistent classificatory pipeline to predict the risk type of unclassified and novel HPV types. Therefore, it is a promising tool to track unclassified-risk p53-degraders in HPV-related cancer surveillance and prevention strategies.

Based on the label-assignment proposed by the trained classifiers, we have compared the predictions for each unclassified risk HPV type (Figure 25) and proposed the classification as high or low risk taking into consideration the consensus of the classifiers. We have disregarded for our predictions the DT, AB, and QDA classifiers, since these presented the lowest assessment evaluations. The predictions performed with these algorithms are shown in the Supplemental Information (Figure S4). Ensuing the proposed pipeline, the classification for the unclassified HPVs type was defined as following: HPV-7 (low-risk), HPV-32 (low-risk), HPV-34 (high-risk), HPV-67 (high-risk), HPV-70 (high-risk), HPV-71 (low), HPV-82 (high-risk), HPV-83 (low-risk), and HPV-87 (low-risk). These results have implications in the development of strategies towards comprehensive HPV-mediated oncogenesis, including primary (vaccination against HPV) and secondary (screening and treatment of pre-cancerous lesions) preventions care.

Figure 25 – Color-coded representation for the predicted categorization for each unclassified risk type HPV according to the trained algorithm. Blue squares refer to a predicted high-risk HPV type, and red squares denote the predicted low-risk HPV type. Each row corresponds to the prediction for a given classifier. K-NN: k-Nearest Neighbor; SVC: Support Vector Classification; GP: Gaussian Process; DT: Decision Tree; MLP: Multilayer perceptron; AB: AdaBoost; QDA: Quadratic Discriminant Analysis; and NB: Naïve-Bayes.



3.3.4 Conclusions

As HPV adapts to the changing human population, selective pressure among different serotypes of HPV can confer biological and adaptive advantages, possibly changing the phenotype associated to a specific viral type. Although the ability of viral HPV E6 protein to bind and degrade p53 is the basis of HPV-induced cell transformation (ANNUNZIATA et al., 2018), the unveiling of the molecular basis of the E6/E6AP/p53 complex interactions is ultimately necessary to fully estimate the oncogenic potential of low risk and unclassified HPV types. Our findings show that binding free energies are possibly one of the main driving forces leading to the establishment of HPV-associated cancer. Using machine learning models in conjunction with classical force field, our data suggest that the features that most contribute to the binding energy are the energetic contribution of short-range hydrogen bonds, and polar solvation free energy. Taken together, these findings allow us to devise a fast-screening general pipeline to predict the oncogenic potential risk of unclassified HPV types, which can be used to track unclassified-risk p53-degraders in HPV-related cancer surveillance and prevention strategies.

SUPPLEMENTAL INFORMATION - CHAPTER 3

SI: AN ARTIFICIAL NEURAL NETWORK MODEL TO PREDICT STRUCTURE-BASED PROTEIN-PROTEIN FREE ENERGY OF BINDING FROM ROSETTA CALCULATED PROPERTIES

Rosetta parsed command lines

Energy minimization:

$ ./minimize.macosclangrelease -l [list-of-pdbs] -min_all_jumps true -run::min_type lbfgs_armijo_nonmonotone -use_input_sc true -ex1 -ex2 -extrachi_cutoff 1 -no_his_his_pairE true -no_optH false -ignore_unrecognized_res -ndruns 5

Properties calculations:

$ ./rosetta_scripts.macosclangrelease -l [list-of-minimized-pdb] -parser:protocol ifa.xml -ignore_unrecognized_res -no_his_his_pairE -out:file:score_only ifa.sc -no_optH false -ex1 -ex2 -use_input_sc -run::min_type lbfgs_armijo_nonmonotone -extrachi_cutoff 1 -linmem_ig 10 -atomic_burial_cutoff 0.01 -sasa_calculator_probe_radius 1.2

Rosetta scripts in XML format

XML to calculate interface properties (ifa.xml)

<ROSETTASCRIPTS>

<SCOREFXNS>

<ScoreFunction name="ref2015" weights="ref2015"/>

</SCOREFXNS> <FILTERS>

<ShapeComplementarity name="Sc" min_sc="2.0" write_int_area="1" jump="1" confidence="0" />

<Ddg name="ddg" scorefxn="ref2015" threshold="0" jump="1" repeats="5" repack="1" repack_bound="0" confidence="0" />

</FILTERS>

<MOVERS>

<InterfaceAnalyzerMover name="ifa" scorefxn="ref2015" pack_separated="1" pack_input="1" tracer="0" interface_sc="1" interface="A_B" />

</MOVERS>
<PROTOCOLS>
<Add mover="ifa" />
<Add filter="Sc" />
<Add filter="ddg" />
</PROTOCOLS>
</ROSETTASCRIPTS>

Polar atom definition

The SASA for a polar atom is calculated as the sum of the SASA for that specific atom and the SASA for any bound hydrogen. Polar atoms presenting SASA smaller than 0.1 Å^2 are considered buried. Hydrogen bonds between the donor and acceptors atoms with a SASA smaller than 3.0 Å^2 are considered buried. Atomic radii from the Reduce software (WORD et al., 1999) and a water probe radius of 1.2 Å^2 were employed to map buried polar atoms and hydrogen bonds. These values were reasoned by probability distributions of hydration water molecules around polar atoms from data collection of high-resolution PDB structures (MATSUOKA; NAKASAKO, 2009).

Table S1. Codes of the PDB used for the test set along with its binding affinity in kcal.mol^{-1}. Binding affinities were retrieved from the PDBind data set in form of k_D and converted using thermodynamic relationships.

PDB ID	k_D **(kcal.mol^{-1}))**	**PDB ID**	k_D **(kcal.mol^{-1}))**
2WH6	-10.5	5H3J	-8.95
2WP3	-8.31	5INB	-9.42
3V1C	-10.25	5MA4	-14.02
3VFN	-9.17	5NT7	-6.78
3WQB	-11.92	5TZP	-10.25
4B1Y	-8.95	5V5H	-9.27
4CJ0	-9.55	5XCO	-10.97
4CJ2	-10.85	5YWR	-10.1
4K5A	-10.89	6B6U	-6.4
4KT3	-13.06	6E3I	-11.58
4LZX	-11.31	6E3J	-12.07
4M0W	-7.05	6HER	-10.08
4NL9	-9	6JB2	-8.09
4PJ2	-14.08	6FU9	-9.99
4QLP	-13.14	6FUB	-10.27
4UYP	-14.58	6FUD	-9.73
4WND	-10.2	6J14	-11.46
4X33	-9	5IMK	-8.27
4YL8	-7.18	5IMM	-11.52
4Z99K	-11.8	5KXH	-8.6
5B78	-7.8	5KY4	-7.95
5DC4	-10.16	5KY5	-8.32
5DJT	-10.59	6DDM	-12.78
5E95	-10.64	6FG8	-8.19
5EP6	-8.37	6NE2	-12.11

Table S2. Calculated Rosetta folding and interface properties. Short description of the features based on the Rosetta package energy function. Only features representing energetic and/or geometric terms were considered.

Feature	**Description**
dslf_fa13	Disulfide geometry potential
hbond_bb_sc	Energy of backbone-side chain hydrogen bonding
hbond_lr_bb	Energy of long-range hydrogen bonding
p_aa_pp	Probability of amino acid at ϕ/ψ
lk_ball_wtd	Orientation-dependent solvation of polar atoms
hbond_sc	Energy of side chain to side chain hydrogen bonding
fa_atr	Attractive energy between two atoms on different residues separated by a given distance
fa_elec	Coulombic potential energy for two atoms separated by a given distance
complex_normalized	Average energy of a residue in the entire complex
total_score	Relative folding free energy
side1_score	Folding energy of the first interface
omega	Omega dihedral in the backbone
side2_score	Folding energy of the second interface
fa_rep	Lennard-Jones repulsive between atoms in different residues
Sc	Shape complementarity
dG_cross	Interaction energy
pro_close	Proline ring closure energy
dG_separated	Binding free energy
side2_normalized	Average per-residue energy on the second interface
fa_dun	Probability of a chosen rotamer is native-like conformation given backbone ϕ, ψ angles
ddg	Change in the binding free energy

Feature	Description
per_residue_energy_int	Average energy of each residue at the interface
fa_intra_rep	Intra-residue repulsive component
hbond_sr_bb	Energy of short-range hydrogen bonding
side1_normalized	Average per-residue energy on the first interface
dG_separated/dSASAx100	Binding free energy divided by the total solvent accessible surface area multiplied by 100
dG_cross/dSASAx100	Interaction energy divided by the total solvent accessible surface area multiplied by 100
fa_intra_sol_xover4	Gaussian exclusion implicit solvation energy
hbond_E_fraction	Contribution of the hydrogen bonding potentials to the binding energy
rama_prepro	Backbone torsion preference term
fa_sol	Gaussian exclusion implicit solvation energy
nres_all	Total number of residues
ref	Reference energy for each amino acid relatively to unfolding.
hbonds_int	Number of hydrogen bonds in the interface
delta_unsatHbonds	Number of buried hydrogen bonds in the interface
dSASA_polar	Polar solvent accessible surface area
dSASA_hphobic	Hydrophobic solvent accessible surface area
nres_int	Number of residues in the interface
dSASA_int	Total solvent accessible surface area
Sc_int_area	Shape complementarity divided by interface area

Table S3. Calculated $R_{Pearson}$ in ascending order for the correlation between the features value and the experimental ΔG.

Feature	$R_{Pearson}$
dslf_fa13	0.336668
hbond_bb_sc	0.252102
hbond_lr_bb	0.215789
p_aa_pp	0.210008
lk_ball_wtd	0.113917
hbond_sc	0.111662
fa_atr	0.109970
fa_elec	0.101131
complex_normalized	0.055793
total_score	0.047824
side1_score	0.046541
omega	0.031989
side2_score	0.029075
fa_rep	0.022766
sc_value	0.012860
Sc	0.004894
dG_cross	-0.033080
pro_close	-0.037478
dG_separated	-0.037859
side2_normalized	-0.048827
fa_dun	-0.053386
ddg	-0.053916

Feature	$R_{Pearson}$
per_residue_energy_int	-0.056136
fa_intra_rep	-0.056630
hbond_sr_bb	-0.056821
side1_normalized	-0.061778
dG_separated/dSASAx100	-0.064651
dG_cross/dSASAx100	-0.065108
fa_intra_sol_xover4	-0.068651
hbond_E_fraction	-0.070499
rama_prepro	-0.078543
fa_sol	-0.096505
nres_all	-0.121934
ref	-0.270164
hbonds_int	-0.346316
delta_unsatHbonds	-0.378498
dSASA_polar	-0.397664
dSASA_hphobic	-0.439358
nres_int	-0.451539
dSASA_int	-0.458725
Sc_int_area	-0.532643

Table S4. Comparison of the predicted ΔG of binding using the ANN and experimental ΔG of binding for the 19 cases of the metadynamics-validation set.

PDB ID	Experimental ΔG (kcal.mol^{-1})	ANN ΔG (kcal.mol^{-1})
1ACB	13.76	-11.254782
1AY7	13.76	-11.054798
1BVN	15.65	-11.545321
1EMV	19.32	-14.301220
1FFW	8.33	-8.465515
1KAC	11.11	-9.067882
1KTZ	9.27	-10.862952
1QA9	7.16	-8.139755
1R0R	14.94	-12.371928
1US7	8.28	-10.642823
2C0L	9.88	-12.066045
2OOB	5.99	-8.733976
2PTC	18.75	-13.219584
2UUY	11.7	-11.982295
3A4S	7.87	-8.636804
3BZD	9.95	-9.275232
3F1P	8.3	-9.549908
3LVK	9.25	-10.150698
3SGB	15.24	-11.496317

Figure S1. Histogram containing all the standardized range value for all features.

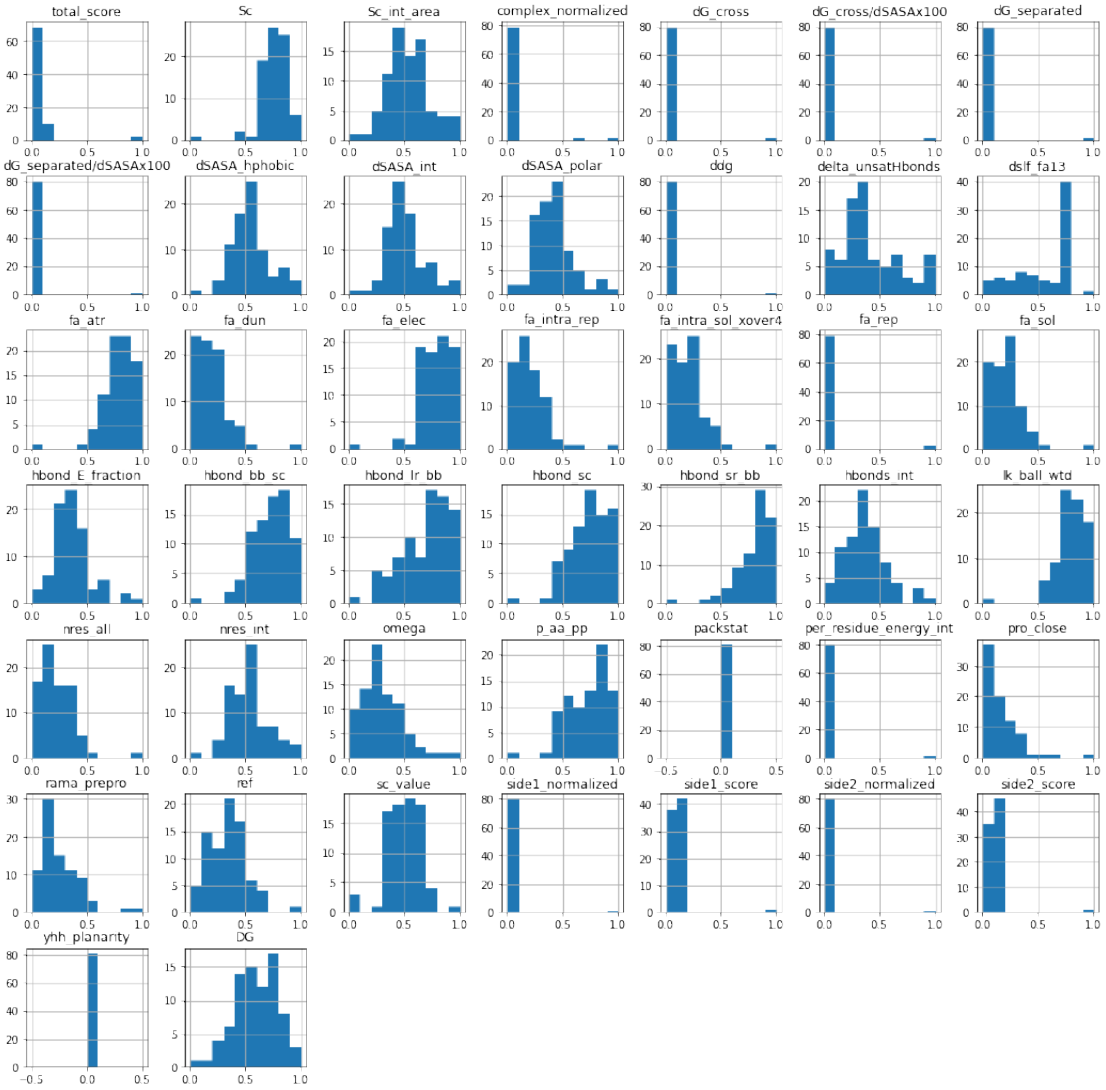
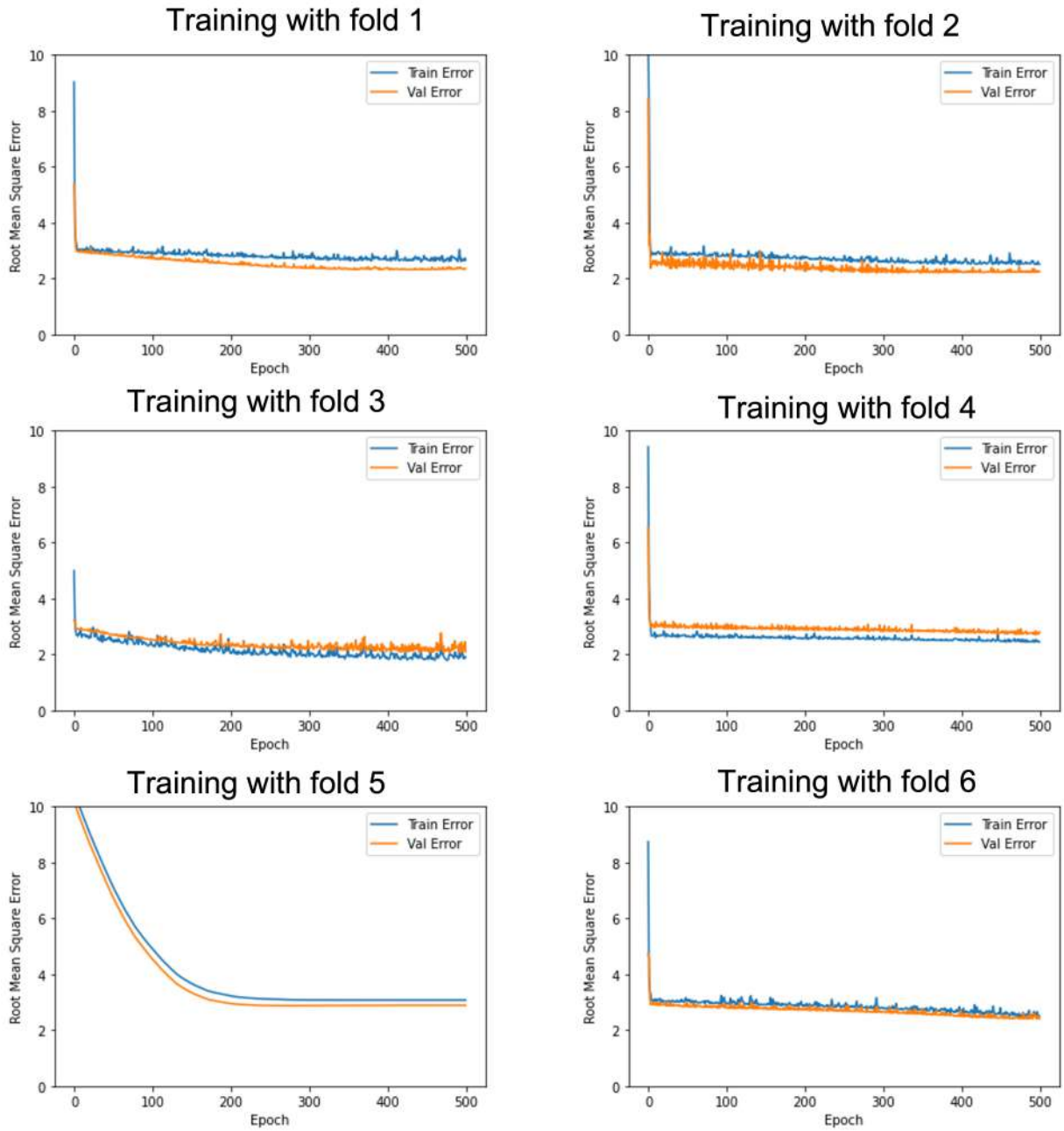
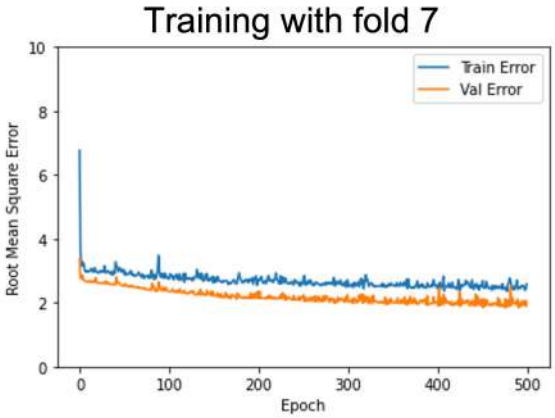
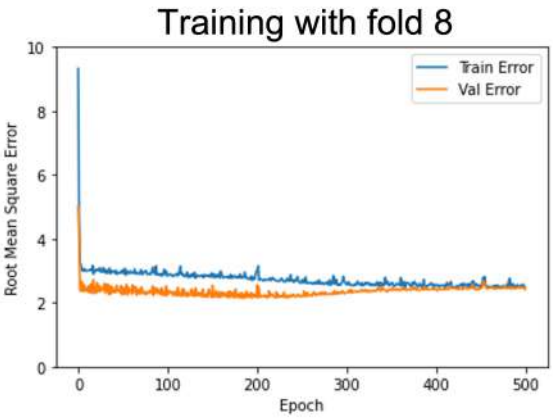
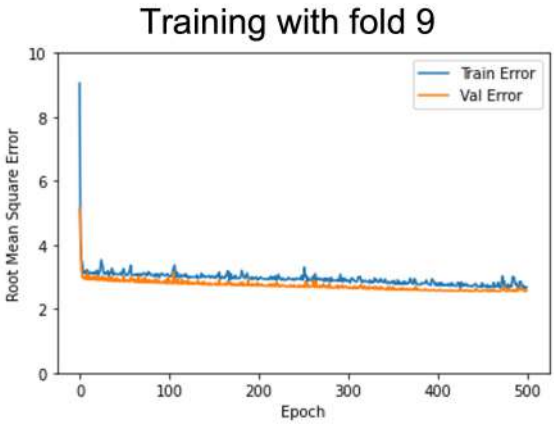


Figure S2. Evaluation of the number of epochs as a function of the root mean square error for a k-fold training where $k \in \{1, .., 10\}$.




Training with fold 7
Train Error
Val Error
Root Mean Square Error
Epoch
0
2
4
6
8
10
0
100
200
300
400
500


Training with fold 8
Train Error
Val Error
Root Mean Square Error
Epoch
0
2
4
6
8
10
0
100
200
300
400
500


Training with fold 9
Train Error
Val Error
Root Mean Square Error
Epoch
0
2
4
6
8
10
0
100
200
300
400
500

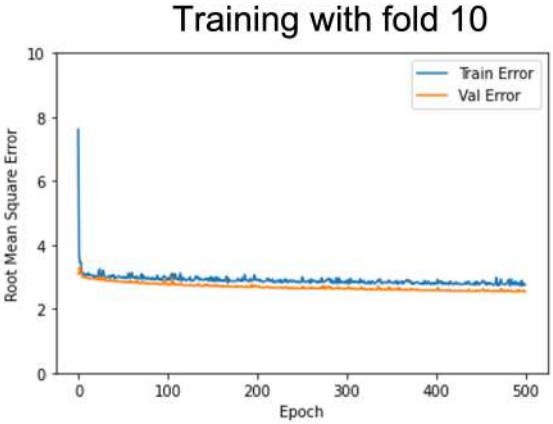

Training with fold 10
Train Error
Val Error
Root Mean Square Error
Epoch
0
2
4
6
8
10
0
100
200
300
400
500

Figure S3. Correlation between the predicted and experimental ΔG of binding for the separated training sets using the ANN and PRODIGY methods.

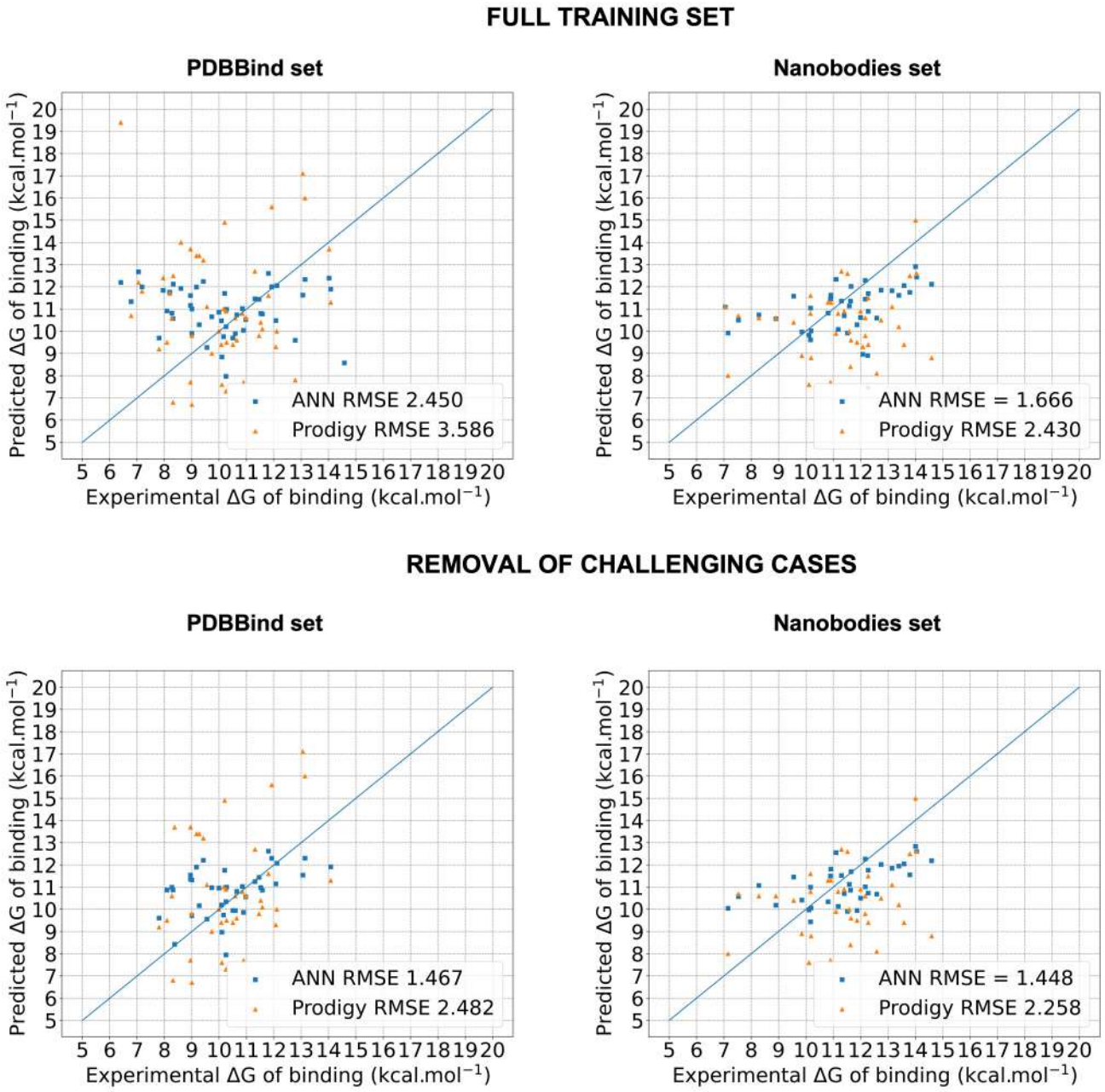
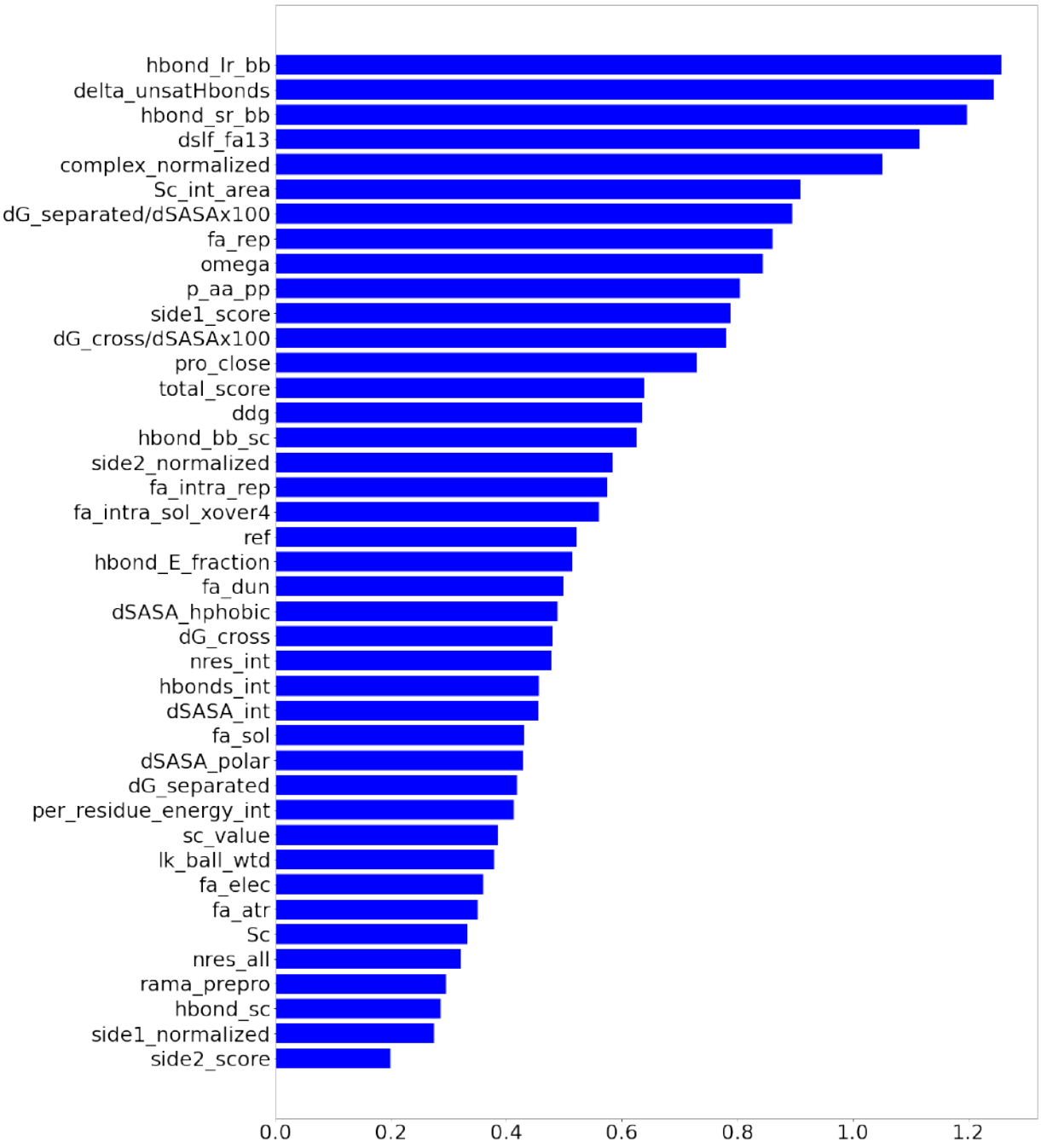


Figure S4. Feature importance score for all the features.



SI: ASSOCIATION STRENGTH OF E6 TO E6AP/P53 CORRELATES WITH HPV-MEDIATED ONCOGENESIS RISK

Machine learning models training

Ten different machine learning models were trained to plot the decision boundaries. For the AdaBoost, Gaussian Naïve Bayes, and Quadratic Discriminant Analysis classifiers, the default implementation of the algorithm was used. For the remaining models, a grid-search over pre-defined values was used. The grid-search method performs an exhaustive search over the specified values to optimize the input parameters of an estimator. The search is reasoned by a cross-validation scheme. The following parameters were identified as the best for each estimator:

- k-Nearest Neighbor: k = 1 (from k=1,2,3,4,5,6,7,9,10);
- Linear Support Vector Machine: C = 0.05 (from C= 0.025,0.05, 0.1,1,10,100,1000);

When not mentioned, parameters were taken as the default implementation since it yielded high-accuracy values. The grid-based search was carried out upon 30% of the total dataset (representing a test/train split), using a random seed of 100.

Threshold metrics to diagnose the performance of the classifiers For a binary classification, the threshold metrics are referred to as precision,recall, and f1-score, and these are calculated according to the assigned prediction: true positive (tp), true negative (tn), false positive (fp), and false negative (fn). Their formulas are calculated as:

$$Precision = \frac{tp}{tp + fp} \tag{3.4}$$

$$Precision = \frac{tp}{R}ecall = \frac{tp}{tp + fn} \tag{3.5}$$

$$f1 - score = \frac{tp}{tp + \frac{1}{2}(fp + fn)} \tag{3.6}$$

Figure S1. Point charge set for zinc finger clusters in E6 (upper) and p53 (lower) proteins, as used in the calculations performed by the Rosetta package. Residues are identified by one-letter code followed by their residue number. Zinc ion is highlighted by a yellow sphere. Atom names follow the standard PDB nomenclature.

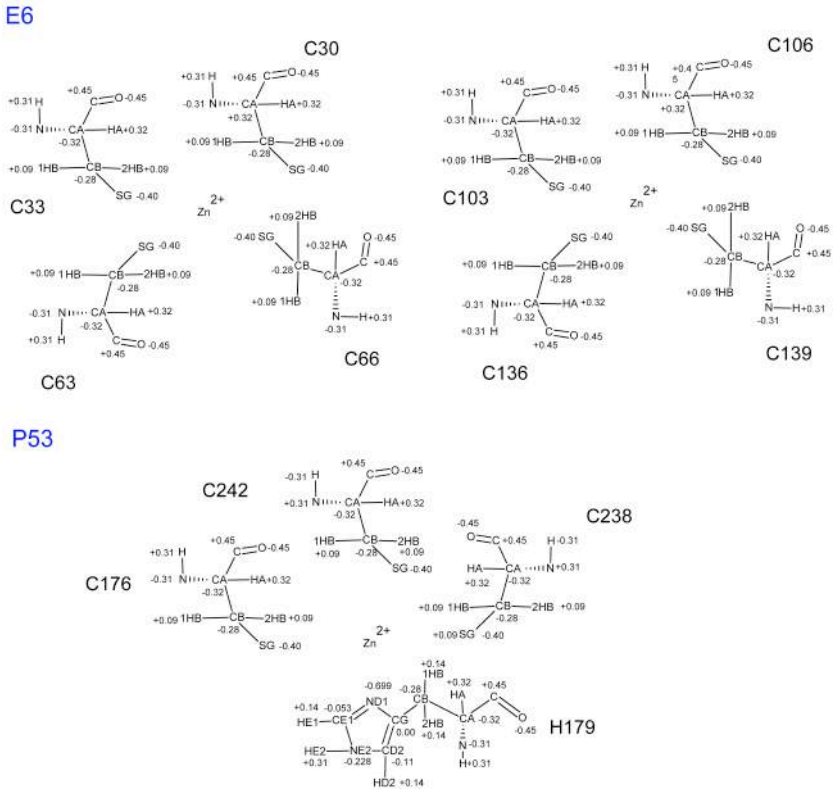


Figure S2. Representation of the distribution of the numerical Boruta-selected features through a kernel density estimate. The plots show the probability density function of the features. A smooth version of the histogram is represented. The probability densities are shown for (A) Binding energy; (B) LJ attractive potential; (C) Unsatisfied H-Bond; (D) Energy of short-rang H-bonds; (E) Coulombic electrostatic potential; (F) Energy of backbone-side chain H-bonds; (G) Solvation free energy; (H) Polar solvation energy. Orange and blue histograms show the distribution for high and low-risk types HPV, respectively.

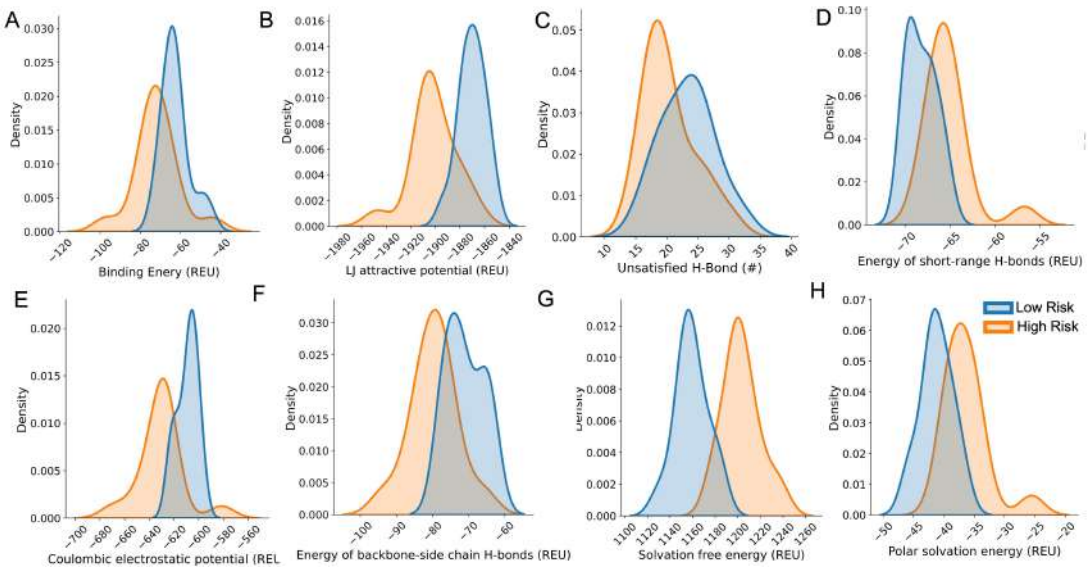


Figure S3. . ROC Curve analysis for the nine different classifiers (A) ADA; (B) GP; (C) k-NN; (D) DT; (E) MLP; (F) QDA; (G) GNB; (H) RF; (I) SVC. The paired results for the true/false positive rate of each classifier are plotted as points in a ROC space, and different discrimination cutoffs are graphically represented. The orange curve represents the classifier ROC curve, while the dashed blue lines denote a non-skilled classifier that only predicts a single class for all examples. The acronym for each classifier is depicted as follows: K-NN: k-Nearest Neighbor; SVC: Support Vector Classification; GP: Gaussian Process; DT: Decision Tree; MLP: Multilayer perceptron; AB: AdaBoost; QDA: Quadratic Discriminant Analysis; and NB: Naïve-Bayes.

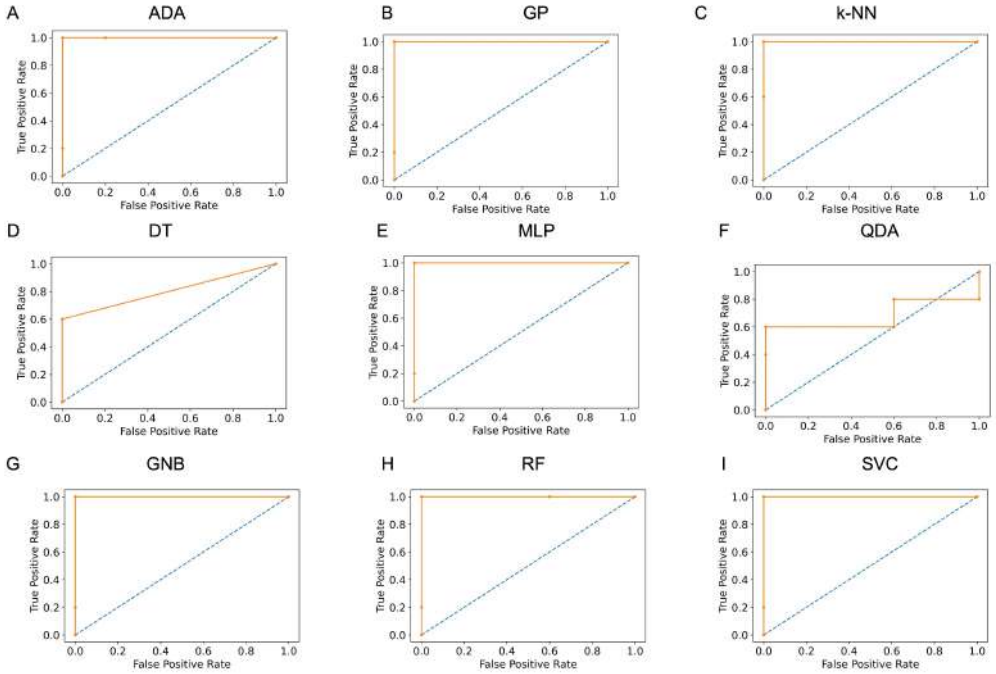


Figure S4. Diagnostic metrics to evaluate the performance of the nine different employed classifiers.

	k-NN	SVC	GP	DT	RF	MLP	AB	QDA	NB
10-fold Accuracy	0.97±0.03	0.99±0.02	0.99±0.02	0.85±0.08	0.95±0.05	1.00 ±0.00	0.83±0.13	0.60±0.12	0.96±0.07
Precision (Low risk)	1.000	1.00	1.00	0.83	0.83	1.00	0.83	0.62	1.000
Recall (Low risk)	1.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.000
f1- Score (Low risk)	1.000	1.00	1.00	0.91	0.91	1.00	0.91	0.77	1.000
Precision (High risk)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall (High risk)	1.00	1.00	1.00	0.80	0.80	1.00	0.80	0.40	1.00
F1- Score (High risk)	1.00	1.00	1.00	0.89	0.89	1.00	0.89	0.57	1.00
ROC-AUC	1.000	1.00	1.00	0.80	1.00	1.00	1.00	0.68	1.000
10-fold ROC-AUC	1.00±0.02	1.00±0.00	0.99±0.02	0.82±0.10	0.99±0.01	1.00±0.00	0.86±0.13	0.70±0.16	0.98±0.05

4 COMPUTATIONAL DESIGN OF NANOBODIES

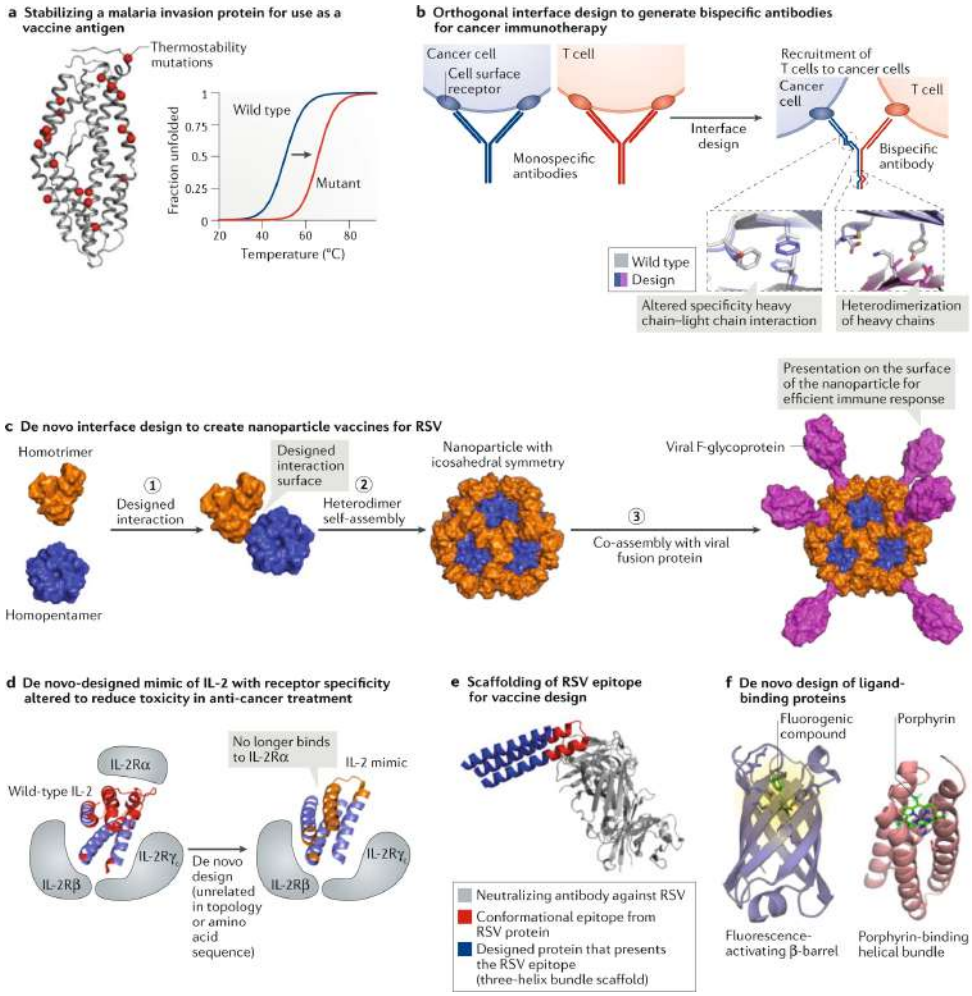
4.1 BACKGROUND

In the past two decades, protein engineering has witnessed significant progress through two distinct approaches: directed evolution (ROMERO; ARNOLD, 2009) and knowledge-based force-field modeling (HUANG; BOYKEN; BAKER, 2016; BAKER, 2010). These approaches have led to the development of engineered proteins with wide-ranging applications in industry, health-care, medicinal sciences, and nanobiotechnology (PAN; KORTEMME, 2021; GAINZA-CIRAUQUI; CORREIA, 2018; AUSTIN et al., 2018; VIANA et al., 2023). Particularly, the creation of proteins capable of effectively binding to specific protein targets holds immense importance in the development of therapies (ZHANG; ANDERSEN; GERONA-NAVARRO, 2018; ZAJC et al., 2020), diagnostic (QUIJANO-RUBIO et al., 2021), and vaccine candidates for viral infections (CORREIA et al., 2014). Currently, common experimental methods involve immunizing animals with the target protein to stimulate antibody production or screening extensive libraries of antibodies or scaffold proteins for their binding abilities. While these methods are robust, they demand substantial experimental endeavors and offer limited control over the properties of the resulting binding molecules.

To overcome these issues, there is a growing interest in the computational design of protein binders, offering a faster and more controlled pathway to affinity reagents with desired biophysical properties (KORTEMME; BAKER, 2004; MARCHAND; HALL-BEAUVAIS; CORREIA, 2022). This approach relies on computational techniques to precisely design proteins that target specific regions on the surface of other proteins. While protein structure prediction aims to forecast the spatial arrangement of amino acid atoms in a given sequence (DILL; MACCALLUM, 2012), protein design focuses on determining an amino acid sequence capable of folding into a particular protein structure or carrying out a specific function known as the "inverse protein folding problem" (YUE; DILL, 1992). Thus, the protein design conundrum involves finding a sequence that can adopt a desired tertiary structure that will exert a specific function. Improved computational methods and advancements in DNA synthesis and sequencing (WRENBECK; FABER; WHITEHEAD, 2017) have made it affordable to test large sets of computationally designed sequences. This affordability enables protein designers to engineer proteins with functions useful in biotechnology and medicine. Figure 26 illustrates a selection of protein-engineering projects from recent years that demonstrate the application of computational protein design,

with molecular modeling playing a vital role in the design process.

Figure 26 – Curated selection of instances where computational design has been employed to create proteins with significant applications in research and medicine.



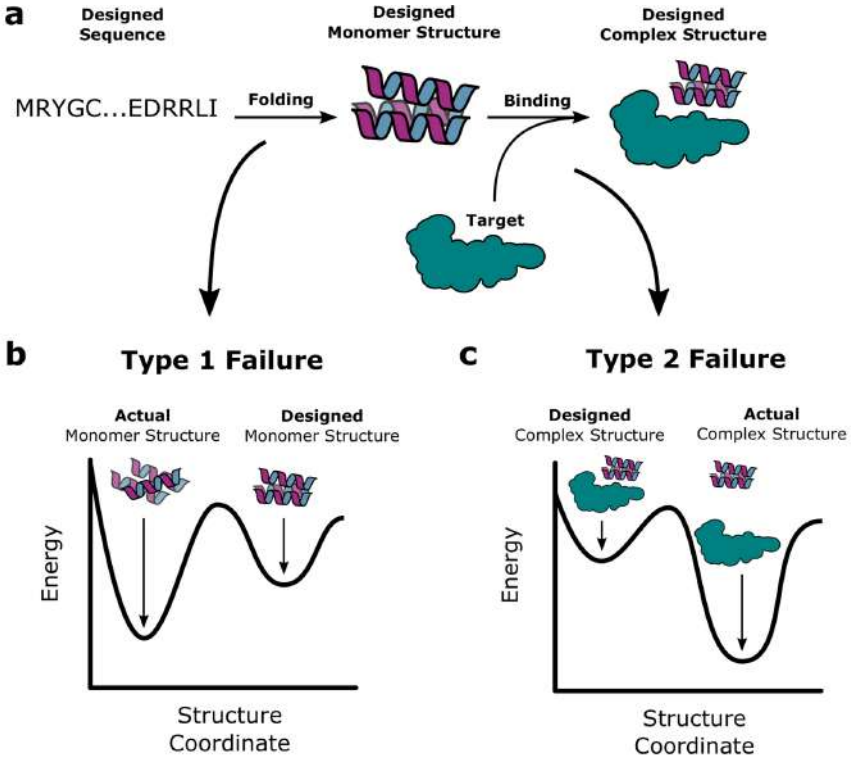
Source: Reproduced from Kuhlman and Bradley (KUHLMAN; BRADLEY, 2019)

The Rosetta package can optimize protein structure and sequence through an iterative process, considering energy factors. It employs fixed backbone sequence optimization and flexible backbone energy minimization in multiple rounds (KUHLMAN et al., 2003). Sequence optimization uses Monte Carlo simulated annealing with the Dunbrack library (JR; KARPLUS, 1993) for conformational sampling. The backbone relaxes after each round to accommodate the designed amino acids. Alongside physically-based protein design, data-driven methods have been developed by leveraging abundant protein data. Among these methods, deep learning has significantly impacted protein design (DING; NAKAI; GONG, 2022), using iterative transformations to train neural networks approximating complex functions in high-dimensional spaces. However, the overall success rate of designing binders for diverse protein targets using these

methods remains low (BENNETT et al., 2023).

There are two primary types of failure in computational protein design(BENNETT et al., 2023): type 1 - the designed sequence may not fold into the intended monomer structure (Figure 27A), and type 2 - the designed monomer structure may not effectively bind the target (Figure 27B). Success requires the designed sequence to have the lowest energy state in isolation and for the complex formed between the designed monomer structure and the target to have sufficiently low energy. Challenges arise from inaccuracies in the energy function, represented as pairwise decomposable terms (STRANGES; KUHLMAN, 2013), and the extensive sampling space.

Figure 27 – (a) Key requirements for successful binder design: (Left) The designed sequence must fold to the designed binder monomer structure. (Right) The structure must then form the designed interface with the target protein. Failure modes include (b) Type-1 failures, where the sequence fails to fold to the monomer structure, and (c) Type-2 failures, where the sequence folds correctly but does not form the desired interface.

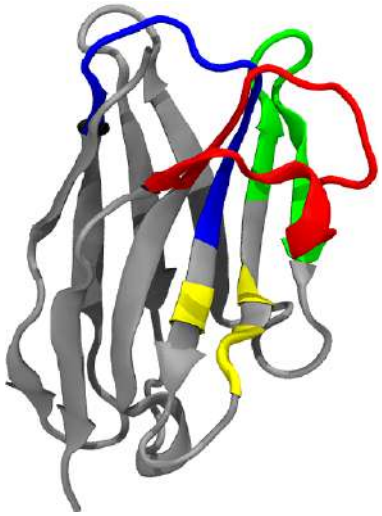


Source: Adapted from Bennett et al. (BENNETT et al., 2023)

In this chapter, we propose a pipeline for the computational design of nanobodies (Nbs), an important antigen-binding class of the recombinant variable domains of heavy-chain-only antibodies (MUYLDERMANS et al., 2013). The general structural topology of Nbs is depicted in Figure 28. Nbs have a core structure composed of a pair of β-sheets, formed by, re-

spectively, 4 and 5 antiparallel β-strands connected by loops and a disulfide bridge. Unlike conventional antibodies (cAbs), Nbs have three highly variable loops (H1, H2, and H3), which correspond to the Complementary Determining Region (CDR) responsible for antigenic binding and specificity. The overall structure is maintained by conserved framework regions. In Nbs, the absence of the variable light chain and light-heavy domain interface is compensated by the Nb Tetrad residues (Y/P37, E44, R/C45, and G47), replacing nonpolar side chains with polar ones (REVETS; BAETSELIER; MUYLDERMANS, 2005; MUYLDERMANS, 2001). These substitutions increase hydrophilicity and solubility, crucial for Nb stability (BARTHELEMY et al., 2008). The high conservation of these residues, demonstrated in sequence alignment studies (MITCHELL; COLWELL, 2018; KUNZ et al., 2017), highlights their evolutionary-driven role in Nb structure.

Figure 28 – Cartoon representation of the overall topology of an Nb (PDB ID: 3DWT)(VINCKE et al., 2009). The Nb domain consists of 9 β-strands linked by loop regions, 3 of these constitute the CDR region and are colored in green, blue, and red. The framework region separated by the hypervariable loops are colored in silver. The Nb tetrad residues are highlighted in yellow.



Given the Nb's tetrad importance in maintaining its folded structure and stability, these residues can be considered key to engineer novel Nbs. To ascertain that changes in the Nb tetrad would negatively impact the Nb folding, we have previously designed a Nb by altering the tetrad residues. The obtained chimera presented low expression yields and the absence of a well-defined globular three-dimensional structure due to aggregation (unpublished data). On the contrary, attempts to "camelize" human/murine Abs by grafting the Nb tetrad to the Ab heavy chain's corresponding position has resulted in structural deformations of the framework

β-sheet, leading to poor stability and aggregation (ROUET et al., 2015) Thus, to investigate the role of the Nb's tetrad to engineer stable Nbs and reducing the chance of type 1 failure, during the PhD period, we have used ML to predict the effect of the removal of the Nb's tetrad contribution by training ML algorithms using a data set of stable Nbs' structures, and Nbs devoid of the tetrad, by replacing the tetrad by alanine (FERRAZ; ADAN; LINS, 2020). We have used structural physical chemical descriptors as features. The results suggested that the loss of stability due to the tetrad's absence is chiefly driven by the entropic contribution. In addition, we propose a support vector classifier algorithm that discriminates between stable and non-stable Nbs, achieving an accuracy of 0.8 for a 10-fold cross validation.

Thus, in this chapter we propose an approach to bypass the type 2 failure by incorporating machine learning techniques. Additionally, to improve sampling, we incorporate dynamism by employing techniques to enhance the sampling of MD simulations.

4.2 AN INTEGRATED DATA-CENTRIC AND ENHANCED SAMPLING-BASED APPROACH TO THE DESIGN OF NANOBODIES

4.2.1 Introduction

Ever since their discovery, single-domain binding fragment of heavy-chain camelid antibodies (MUYLDERMANS, 2013), referred to as nanobodies (Nbs), have gained considerable attention in translational research as therapeutic and diagnostic tools against human diseases and pathogens (MIR et al., 2020). Along with their small size (15 kDa) and favorable physical-chemical properties (e.g., thermal and environmental stabilities), Nbs display binding affinities equivalent to conventional antibodies (cAbs) (MUYLDERMANS, 2013; VINCKE; MUYLDERMANS, 2012). Moreover, their heterologous expression in bacteria allows overcoming cAbs production pitfalls, such as high production cost and need of animal facility (JOVČEVSKA; MUYLDERMANS, 2020; MORRISON, 2019). Hence, Nbs are considered as a promising tool against numerous diseases. A variety of Nbs is currently being investigated under pre-clinical and clinical stages against a wide range of viral infections (BEGHEIN; GETTEMANS, 2017; KONWARH, 2020).

Given the versatility of Nbs, they have shown a great range of neutralization against many viruses, such as influenza (OLGA et al., 2021), HIV-1 (STROKAPPE et al., 2019) and ebola (OLGA; YU et al., 2021). However, the discovery of novel Nbs is a challenging task, as extensive laboratory screenings are necessary to identify Nbs that specifically bind to the desired target.

This process can be both time-consuming and expensive. Certain challenging targets still pose difficulties, such as receptors and channels, proteins belonging to closely related families, peptides prone to aggregation, and short-lived protein aggregates associated with diseases (APRILE et al., 2020; HUTCHINGS; COLUSSI; CLARK, 2019).

In this context, computational Nb design offers a promising solution for overcoming these limitations by significantly reducing the time and costs associated with Nb discovery (KALITA; TRIPATHI; PADHI, 2023). It allows for a highly controlled screening of desired biophysical properties and enables the precise targeting of specific epitopes of interest. However, these calculations are usually based on molecular modeling, and the challenge of achieving exhaustive sampling makes design simulations resource-intensive and imprecise (ALVIZO; MAYO, 2008; CHILDERS; DAGGETT, 2017; LIPPOW; TIDOR, 2007; GAINZA-CIRAUQUI; CORREIA, 2018). As a result, the design of binding proteins, including Nbs, has generally yielded low success rates, necessitating recursive experimental affinity maturation (CHEVALIER et al., 2017; BARAN et al., 2017; TILLER et al., 2017).

Typically, multiple rounds of experimental affinity maturation are necessary due to the challenging nature of predicting affinities. This difficulty arises partly from sampling limitations and the failure of energy functions to capture all interface details, particularly the delicate balance between polar interactions and solvation in the binding interface (STRANGES; KUHLMAN, 2013). To address these issues, we present a novel design method for Nbs in this study. Our approach utilizes extensive sampling via metadynamics and incorporates ML to predict binding affinities with high accuracy, comparable to experimental results (FERRAZ et al., 2023), presented in Chapter 3.

As a proof of concept of our approach, we have utilized our protocol to design Nbs that specifically target SARS-CoV-2, the virus responsible for the global COVID-19 pandemic (PHELAN; KATZ; GOSTIN, 2020; WU et al., 2020a; ZHU et al., 2020). This disease has had a significant impact worldwide, causing widespread morbidity and mortality (CHAN et al., 2020). Despite substantial progress in vaccination and disease control, the emergence of SARS-CoV-2 variants poses a threat to the efficacy of neutralizing antibodies and vaccines (KHOURY et al., 2023). In addition, while neutralizing antibodies against SARS-CoV-2 have shown protective effects, they may have a short duration and not be present in all infected individuals (ROBBIANI et al., 2020). Therefore, it remains crucial to develop therapies for COVID-19 that can be easily adapted to different variants of concern. Our Nbs were designed to specifically target the receptor binding domain (RBD) of the virus, aiming to preventing it from binding to the human

receptor angiotensin-converting enzyme-2 (hACE2) and inhibiting the mechanism of cell entry.

4.2.2 Computational procedures

4.2.2.1 Computational design

Two different computational approaches targeting distinctive regions of the S protein were used to design specific VHHs, totalling 1 million designed structures. All the computational design steps were carried out using the Rosetta package of software v. 3.12 (LEAVER-FAY et al., 2011b), by means of RosettaScripts (FLEISHMAN et al., 2011a) protocols written in the XML language. The Rosetta energy function 2015, REF15 (ALFORD et al., 2017), was employed throughout all this work. XML scripts are made available at the Supplemental Information.

The computational design was based on the native interactions within the interface of SARS-CoV-2 RBD bound to a cAb (Ab CR3022) (YUAN et al., 2020) and a Nb (VHH-72, which originally binds to SARS-CoV-1 RBD) (WRAPP et al., 2020a). Initially, the three-dimensional structures of the complexes (PDB IDs: 6W41 and 6WAQ, respectively) were geometry-optimized using the Rosetta FastRelax protocol with coordinate constraints[1] on the heavy atoms. Since there is no available crystal structure for SARS-CoV-2 RBD complexed to Nb VHH-72, it was modelled based on the structure for SARS-CoV-1 RBD bound to Nb VHH-72 (section 4.2.2.1.1). Starting from the optimized structures, 0.5 μs of unrestrained MD simulations were carried out (details described in the section 4.2.2.4) using the Gromacs engine and parameters of the AMBER force field ff14SB. The last snapshot of these simulations were used as the starting point of metadynamics simulations, in order to sample a broad conformational space according to the protocol defined in (FERNÁNDEZ-QUINTERO et al., 2019; FERNÁNDEZ-QUINTERO et al., 2019; FERNÁNDEZ-QUINTERO et al., 2021; SEIDLER et al., 2023).

Metadynamics simulations were carried out in GROMACS utilizing the PLUMED 2.5 implementation (TRIBELLO et al., 2014). To focus on the key binding interactions, a collective variable based on the ψ torsion angles of CDR loops 2 and 3 was chosen (as these are the most important ones for binding). To this end, alphabeta (PIETRUCCI; LAIO, 2009), a CV that measures the similarity of each ψ angle of the CDRs 2 and 3 to a reference value of π rad was used. This CV was computed using PLUMED 2's MATHEVAL and COMBINE features, and has been used to capture essential conformational movements (SEIDLER et al., 2023). The

[1] This corresponds to "restraints" in other programs.

metadynamics parameters included a Gaussian height of 1.2 kJ/mol, a width of 1.5, deposition every 500 steps, and a bias factor of 24. The choice of alphabeta CV and parameters was based on (LÖHR; SORMANNI; VENDRUSCOLO, 2022). Simulations were carried out under NPT ensemble conditions for 12 μs for Nb VHH-72 complexed to SARS-CoV-2 RBD, and 5.5 μs for cAb CR3022 complexed to SARS-CoV-2 RBD following the same set up as the unrestricted MD simulations (section 4.2.2.4). The convergence of the metadynamics simulations is shown in the supplemental information.

From the metadynamics simulations, 25 different conformations, associated to the lowest energy states in the reconstructed free energy surface were extracted. These structures were used to map hot spots using computational alanine scanning (section 4.2.2.1.2). For the case of the Nb VHH-72, the hot spots in the Nb were mapped and kept fixed for the computational design described as follows. However, for the CR3022-RBD system, the hot spots identified in the cAb were grafted onto the surface of an artificially engineered Nb based on the framework of an "universal" humanized Nb (VINCKE et al., 2009) following a seeded interface design approach referred to as motif graft (SILVA; CORREIA; PROCKO, 2016). The set up for engineering this artificial Nb is described in section 4.2.2.1.3.

Subsequently to the hot spot definition and system building, 500,000 novel sequences were proposed through random and combinatorial mutations through the Monte Carlo search by Rosetta for each system. Mutations were introduced only in the CDR loops' regions in the presence of the antigen. This step was conducted to optimize the binding affinity towards the SARS-CoV-2 RBD, and to stabilize the surrounding region of the grafted loops. The remodeling of the CDR loops was carried out with ProteinInterfaceDesign subroutine, in which it repacks and designs residues close to the specified protein – protein interface considering a cut-off radius of 8 Å from the opposite interface. However the grafted hot spots were not allowed to be altered. 4 cycles of design followed by energy minimization were carried out. In each cycle, two different energy functions were used alternately: the "hard" energy function and the "soft" energy function. The cycles were performed in the following order: cycle 1 (hard), cycle 2 (soft), cycle 3 (hard), and cycle 4 (soft). The "hard" energy function refers to the default all-atom Rosetta energy function, which includes strong repulsion forces. On the other hand, the "soft" energy function is a modified version of the default energy function, known as "soft-repulsive" (soft-rep). This modified version reduces the emphasis on van der Waals overlaps and strain caused by residue conformational changes.

For each cycle, two different energy functions were used alternated: the hard energy

function and soft energy function (i.e, cycle 1 (hard), cycle 2 (soft), cycle 3 (hard), cycle 4 (soft). The first one is the all-atom "hard-repulsive" Rosetta energy function, which is the default energy function, and the second one is a modified version of this energy function known as "soft-repulsive" (soft-rep), where van der Waals overlaps and strain caused by residue conformational changes are given less weight.

The computational design process involved using phylogenetic information from a position-specific-scoring matrix (PSSM) constructed through multiple sequence alignment (MSA) of 10,000 homologous sequences to the parental Nb. This protocol was proposed by Goldenzweig and collaborators (GOLDENZWEIG et al., 2016). The BLASTp server (ALTSCHUL et al., 1990; ALTSCHUL et al., 1997) was utilized with non-redundant databank and maximum number of hits and e-value equals to 10^{-4}, resulting in the PSSM that represents the likelihood of observing each amino acid at different positions in the sequence. Positive values in the PSSM indicate frequent occurrence of a particular amino acid at a specific position (ALTSCHUL et al., 1990). This information guides the design of novel proteins by focusing on residues that are commonly observed, enhancing protein stability. Conversely, mutations leading to amino acids not found in natural sequences tend to decrease stability, a phenomenon referred to as consensus effect (STEIPE et al., 1994; LEHMANN et al., 2000; MAGLIERY, 2015).

4.2.2.1.1 Modelling of SARS-CoV-2 RBD bound to VHH-72

The coordinates for the atomic positions of the SARS-CoV-2 RBD were extracted from the chain A of the PDB ID 6M0J (LAN et al., 2020) and it was refined by 5 rounds of relaxation using the Rosetta package v. 3.12. The structures of the SARS-CoV-1 complexed with Nb VHH-72 was superimposed by structural alignment with the structure of SARS-CoV-2 RBD. To approach the native structure of Nb VHH-72 bound to SARS-CoV-2 RBD, local docking calculations were carried out using the RosettaDock protocol (GRAY et al., 2003) and Rosetta package v. 3.12. Initially, a docking prepack procedure was carried out by separating the binding partners until they are out of contact, packing the side of the unbound complex, and bringing the unbound partners back to the original orientation. Starting from the prepacked complex, a total of 50,000 conformations were generated by keeping atomic positions of the SARS-CoV-2 fixed, while docking the Nb VHH-72 locally in the binding site. Docking is carried out by perturbing the structure of Nb VHH-72 by a 2 Å translation and 6° rotation. Rotamers for the input structure was used between Monte Carlo and Minimization cycles. The

technique described by Wang and collaborators (WANG; BRADLEY; BAKER, 2007) was employed to recover the correct rotameric side chain. This technique utilizes discrete rotamer libraries, representing pre-defined side-chain conformations observed in protein structures, as a starting point. Additionally, it incorporates side-chain conformations from the unbound structures of the proteins involved in the docking process. To sample off-rotamer conformations, torsion space minimization is employed.

To assess the quality of our predictions, the complex presenting the lowest interface score value was used as the reference structure, to which, the remaining structures in the output decoy were compared by RMSD. Then, a docking funnel was built plotting interface score versus the calculated RMSD. If the plot shows that the structures presenting the lowest RMSD also display the lowest interface scores, it is likely that the reference structure is likely to be a near-native conformation (CHAUDHURY et al., 2011).

4.2.2.1.2 Hot spots mapping

Alanine scanning (AS) mutagenesis using the Robetta webserver (<http://old.robetta.org>) (KORTEMME; BAKER, 2002; KORTEMME; BAKER, 2004) was conducted to identify hot spots for grafting onto a VHH surface. By replacing a residue's side chain in the interface with a methyl group (-CH3), the AS technique calculates the change in Gibbs free energy of binding ($\Delta\Delta$G) to assess its contribution to affinity and stability of the individual proteins. Residues within the interaction interface were considered if they met either of the following criteria: 1) having at least one atom within a 4.0 Ådistance from an atom in the opposite interface, or 2) being buried in the interface upon complex formation. The latter is measure by an increase in the number of C_β atoms within a sphere with a radius of 8 Å around the C_β atom of the residue o interest. Hot spots were determined based on alanine mutations that resulted in a $\Delta\Delta$G exceeding 1.0 kcal/mol, indicating destabilization of the complex. This threshold was established with a success rate of 79% in predicting hot spots from a test set of 233 mutations in protein-protein complexes (KORTEMME; BAKER, 2004).

4.2.2.1.3 Hot spots grafting

The initial structure of the antigen – antibody complex was retrieved from the protein data bank (PDB) file with the accession code 6W41 (YUAN et al., 2020). This structure corresponds to the SARS-CoV-2 RBD protein complexed to the human antibody CR3022. Once the hot

spots were mapped, antigen-binding loops regions (CDR 1, CDR 2, and CDR3 of the VHHs) capable of accommodating these hot spots residues in their native conformation of the initial complex was scanned. To this end, a library of ca. 80 crystal structures of Nbs – Antigens was used (ZAVRTANIK et al., 2018), in which only the monomeric Nb was considered after geometry-optimization with the FastRelax protocol. A virtual screening was performed to identify loops containing backbones that can be superimposed with the predicted hot spots with a root mean square deviation lower than 1.0 Å. After identifying loops capable to receive the hot spots in their native conformation, the loops were grafted onto the framework of a humanized Nb referred to as universal VHH framework (VINCKE et al., 2009), belonging to the CabBCII-10 Nb. The etymology for this Nb arises from the competence of this framework in tolerating loops grafting without compromising its stability, solubility, and functionality of the original loops. The PDB code for the CabBCII-10 Nb is 3DWT (VINCKE et al., 2009).

4.2.2.2 Filtering of molecules via interface parameters

Initially, Nbs with potential to bind the S protein were selected based on interface metrics calculated using the Rosetta package. These metrics compare the structures in their bound and unbound states. The following interface parameters were used for filtering: (1) Shape complementarity (Sc) $>$ 0.7; (2) Binding and interaction energies $<$ -30 REU (Rosetta energy units); (3) Binding energy density $<$ -2.0 REU/Å2; (4) Number of unsatisfied hydrogen bonds (H-unsat) $<$ 10; (5) Evaluation of packing (packstat) $>$ 0.6; (6) Hydrophobic quantification in the interface proportional to that of the native complex; (7) Monomeric VHH folding free energy (total score) lower than the reference native structure; (8) RMSD $<$ 1.0 Å between the initial and relaxed VHH structure using the FastRelax protocol. These parameters were based on the properties' values for these metrics within a data set of ca. 80 Nbs-Antigen natural interface (ZAVRTANIK; HADŽI, 2019). Visual inspection was manually performed to ensure no buried charged residues or excessive alanine residues were present in the interface.

4.2.2.3 Filtering of molecules via machine learning

ML was employed as a post-processing step following the selection of potential candidates using interface parameters. Specifically, we utilized an Artificial Neural Network (ANN) model developed and discussed in Chapter 3 of this thesis (FERRAZ et al., 2023).

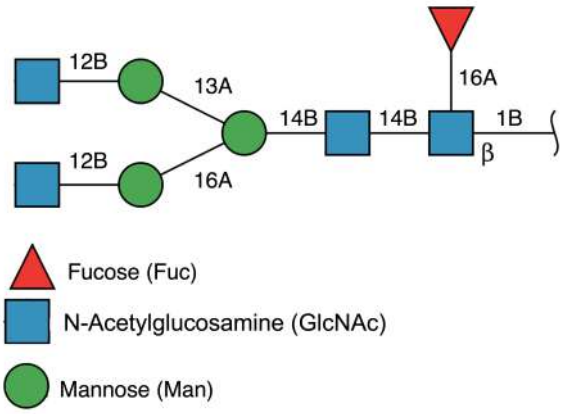
To address the limitations of our ANN model in predicting the ΔG of binding for glycosylated proteins, such as the RBD of SARS-CoV-2, we incorporated the glycan moiety into the Rosetta force field. The setup of the glycosylated system is elaborated in Section 4.2.2.4. For the Rosetta force field, the glycan was treated as a non-canonical residue. To generate backbone-dependent amino acid rotamer libraries, we employed the MakeRotLib approach (RENFREW et al., 2012). The partial atomic charges were obtained from the CHARMM36 force field (HUANG; JR, 2013) and adjusted using a group optimization scheme to ensure consistency with the Rosetta partial charges (PARK et al., 2016).

For the calculation of binding free energies, we followed the same protocol outlined in Chapter 3. Beginning with the initial structure, a geometry optimization step was performed. Subsequently, interface parameters were calculated and employed as input for the ANN. The binding free energies were expressed in kcal.mol^{-1} units.

4.2.2.4 Molecular dynamics set up and simulations

MD simulations were used to assess the internal dynamics of the filtered molecules and to generate a set of configurations consistent with a thermodynamic ensemble for the starting complexes structures. The Gromacs 2020.5 engine was used. SARS-CoV-2 RBD-bound glycans were added using the Glycan Reader & Modeler (PARK et al., 2019) input generator from the CHARMM-GUI (LEE et al., 2016) web server (<http://www.charmm-gui.org/>). An N-glycan was added to the residue N343 of the RBD, based on the mapping for the SARS-CoV-2 spike protein O- and N-glycosilation profile (WATANABE et al., 2020; SHAJAHAN et al., 2020). The chemical structure of the added glycan is shown in Figure 29. Histidine protonation states at pH 7.4 were predicted based on the theoretical value of their side-chains' pKa using the PROPKA code (OLSSON et al., 2011).

Figure 29 – This diagram illustrates the structure of the SARS-CoV-2 RBD-bound N-glycan attached to N-343. The N-glycan is composed of three components: N-Acetylgucosamine (depicted by a blue square), mannose (depicted by a green circle), and fucose (depicted by a red triangle). The α and β linkage types are represented by A and B, respectively. An α glycosidic linkage occurs between carbons with the same stereochemistry, while a β glycosidic linkage occurs between carbons with different stereochemistry. The numbering between the glycosidic linkages indicates the carbon involved in the bond, where the first number is the carbon number of the first monosaccharide and the second number is the carbon number of the second monosaccharide . For example, 12B signifies that carbon 1 from monosaccharide 1 is linked to carbon 4 of monosaccharide 2 in a β-type linkage



The systems were centered in cubic boxes treated for periodic boundaries conditions with edge dimensions of 100 Å and 200 Å from the center of the solute for the monomeric and complexed systems, respectively. A saline physiological concentration (NaCl) was added to assure an ionic strength of 150 mM. The AMBER ff14SB (MAIER et al., 2015) all atom parameters set was used to describe the interatomic interactions of the protein and ions, the GLYCAM06 (KIRSCHNER et al., 2008) for the carbohydrates, and the TIP3P water model (JORGENSEN et al., 1983) was employed to explicitly solvate the systems. A cut-off distance of 12 Å was used for all short-range interactions, and the Particle Mesh Ewald (PME) (DARDEN; YORK; PEDERSEN, 1993b) method was employed to treat electrostatic interactions using a 12 Å cut-off. Force switch was used with a switch distance of 10 Å. Initially, the systems were subjected to 5,000 steps of the steepest descent energy minimization restraining the positions of the protein atoms, followed by 5,000 steps of the same minimization protocol without positional restraints in the solute. Equilibration of each system was carried out in a stepwise fashion. Firstly, the system was heated by generating velocities according to a Maxwell-Boltzmann distribution at 5 K and increasing the temperature until reaching 303.15 K for 500 ps in the canonical ensemble (maintaining the number of particles, volume, and temperature fixed, NVT). Positional re-

straints were added to the heavy backbone atoms of the protein with a force constant of 1000 kJ.mol^{-1}.nm^{-2} for a period of 500 ps. Then, 500 ps was carried out in the NPT ensemble having a reduced force constant of 20 kJ.mol^{-1}.nm^{-2}. Subsequently, 1 ns of equilibration was run without restraints. The temperature was held constant through the Langevin thermostat (ANDERSEN, 1980) with a collision frequency $\gamma = 0.5$ ps^{-1}, and the pressure was kept at 1 atm by means of the Berendsen barostat (BERENDSEN et al., 1984) with a relaxation time of 0.4 ps. The production runs were carried out in the NPT ensemble maintaining the pressure constant using the Parrinello-Rahman barostat (PARRINELLO; RAHMAN, 1981), with a relaxation time constant of 2 ps and the Langevin temperature control to maintain the systems at 1 atm and 303.3 K. All the covalent bonds involving a hydrogen were holonomically constrained using the LINCS method (HESS et al., 1997), allowing an integration time-step of 2 fs.

4.2.2.5 Molecular protein – protein docking

Molecular docking calculations were performed to elucidate the higher affinity of the designed Nb VHH-72.1 for the full S protein within the VLP than for the SARS-CoV-2 RBD. The modelled structure of the designed Nb VHH-72.1 bound to the SARS-CoV-2 was used as the template to be superimposed on the full S protein structure. To this end, the following S protein structures were considered: no RBD up (PDB ID: 6VXX (WALLS et al., 2020)), one RBD in the up position (PDB ID: 6VSB (WRAPP et al., 2020b)), two RBDs in the up position (PDB ID: 6X2B (HENDERSON et al., 2020)). The S proteins were retrieved from the SwissModel repository for SARS-CoV-2-related proteins (<swissmodel.expasy.org/repository/species/2697049>), in which the missing residues for the flexible loop regions have been modelled using the Generalized Kinematic Closure (KIC) algorithm. The structure of the designed Nb was docked using local docking protocol from Rosetta as from the superimposed structures towards the S proteins. The local docking was carried out in two steps: initially the structure of the designed Nb was allowed to translate 1.5 Å and rotate 6° in arbitary directions before the start of every individual simulation in the centroid mode. Then, the high-resolution full atom mode was carried out, in which the backbone of the proteins is not moved. A total of 10,000 structures for each system was generated. The lowest energy structure for each system was selected and refined through the FastRelax protocol. The refined structures were used to calculate the binding free energies by means of the InterfaceAnalyzer, an application from the package Rosetta that combines a set of tools to analyze protein-protein interfaces, including

calculating the bindign free energy.

4.2.2.6 Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA)

We employed MM-PBSA calculations to estimate the binding free energy between the designed Nb-72-1 and the RBD and fully glycosylated S protein (using two RBD ups, built as described previously). MM-PBSA energies were computed over a 100 ns MD simulation using the same set up as the unrestricted MD simulations, except for the inclusion of a hydrogen mass repartitioning scheme (HOPKINS et al., 2015), allowing an integration time step of 4 fs. The calculations were performed using the g_mmpbsa tool (KUMARI et al., 2014), which is based on GROMACS and APBS (BAKER et al., 2001). Snapshots were taken every 50 ps per complex during the last 50 ns of the production trajectories, resulting in a total of 300 frames per complex. The solvation free energy's electrostatic contribution was determined by solving the linearized Poisson-Boltzmann equation (HONIG; NICHOLLS, 1995), while the non-electrostatic term was considered using the SASA nonpolar model. The APBS calculations utilized the force field from the MD simulations and a grid spacing of 0.5 Å to numerically solve the linearized Poisson-Boltzmann equation. The solvent was characterized by a dielectric constant of 80 and an ionic strength of 0.15 M NaCl, with sodium and chloride ions having atomic radii of 0.97 Å and 1.81 Å, respectively. A low-dielectric value of 4 was assumed.

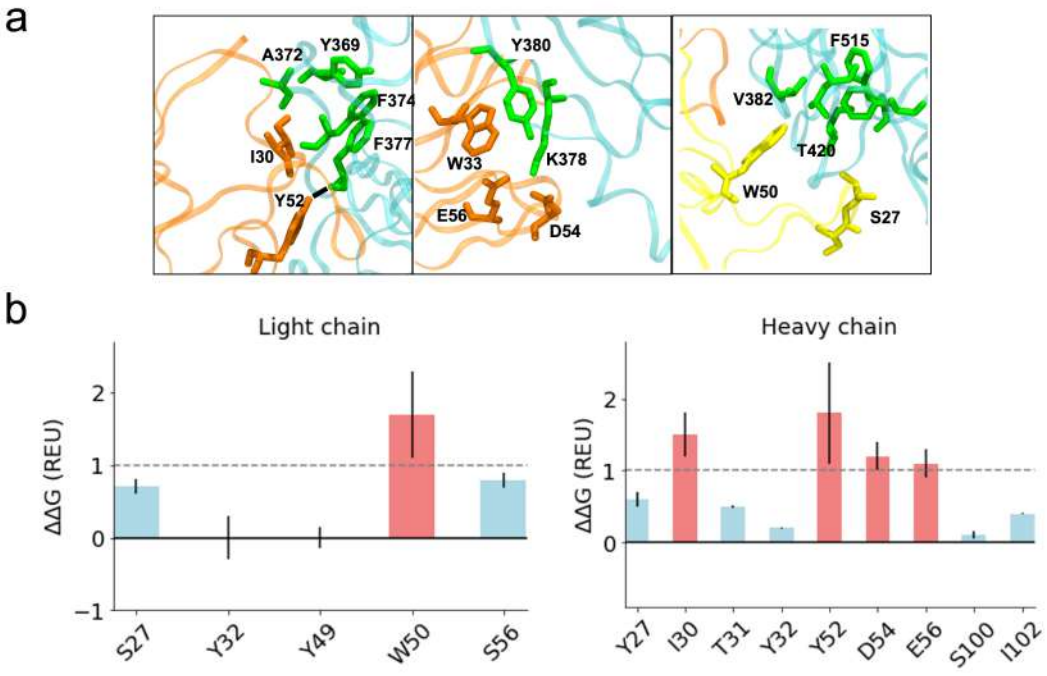
4.2.3 Results and discussion

4.2.3.1 Design of CR3022-derived Nb

To identify the key residues that are the most relevant for the interaction between the human Ab CR3022 and the SARS-CoV-2 RBD (Figure 30a), a computational mutagenesis by AS was carried out. The energetic contribution of each sidechain in the protein – protein interface was estimated by the binding $\Delta\Delta$G, and the important residues for the interaction in the Ab were grafted onto the scaffold of an artificial Nb. As it can be seen, the antibodies' residues that most contribute to the binding to SARS-CoV-2 RBD are located within the heavy-chain domain of the Ab Figure 30b. These residues are placed towards a structured epitope, comprising β-sheets and α-helixes, which is a good starting point to direct the Nb against, since Nbs bind preferentially to structured epitopes (ZAVRTANIK et al., 2018).

The AS result reveals a major role of the heavy chain for binding, and it also highlights the importance of the residues I30 and Y52. I30 is involved in a hydrophobic core, along with the W33, which has not been predicted to be a hot spot, notwithstanding, W33 has an important structural role involved in an aromatic interaction, as it can be observed in the crystallographic structure. According to the AS calculations, Y52 is the residue that presents the greatest contribution to the stabilization of the complex. Y52 residue interacts with F377 from the SARS-CoV-2 RBD through the formation of a hydrogen bond between the hydroxyl hydrogen side chain of Y52 (donor) and the oxygen of F333 main chain (acceptor). It is important to highlight that the computational design of precise hydrogen bonds in the interface is still a challenge to the design of interfaces (MAGUIRE et al., 2018), which in part due to deficiencies in the energy function (STRANGES; KUHLMAN, 2013). Therefore, the importance of this residue is not just crucial from a binding thermodynamics perspective, but also to maintain a hydrogen bond network already formed with the epitope. One of the advantages of this seeded approach to the computational design is due to beginning the design with some residues with contacts present in the native protein.

Figure 30 – (a) Important interactions between CR3022 and the receptor-binding domain (RBD) of SARS-CoV-2. The heavy chain of CR3022 is depicted in orange, the light chain in yellow, and the SARS-CoV-2 RBD in cyan. Dashed lines indicate the presence of hydrogen bonds(b) Computational mutagenesis by alanine scanning for the residues in the interaction interface. Each bar denotes the predicted binding $\Delta\Delta G$ for a given residue upon alanine mutation. A threshold of 1 REU, shown as a dashed line, was chosen as cut-off to predict destabilizing effect (red bars), suggesting importance for binding. Blue bars represent stabilizing or no effect in the $\Delta\Delta G$.



Moreover, the residues D54 and E57 present a lower contribution to the binding energetic of the interface, even though the crystallographic structure shows that these residues are involved in a salt bridge within K378. When it comes to the light chain of CR3022, W50 shows a great contribution to the stability of the complex, being involved in a hydrophobic core, and S27 makes a hydrogen bond with T430 from the RBD. Therefore, these residues were also taken into consideration to be grafted. In addition to identifying residues to be grafted, these residues also profile a pharmacophore model to map the nature of relevant interactions for the binding of the CR3022 Ab against the SARS-CoV-2 RBD. As it can be seen, hydrophobic (dispersion) and electrostatic interactions are critical to assure the binding of the Ab to its target. Since there was no possible conformation to simultaneously graft the selected residues from the heavy and light chain, we have opted to graft the residues from the heavy chain, as these present a higher contribution. Thus, the following residues from the heavy chain of the Ab were considered as important residues to be grafted onto a VHH binding loop: I30, W33, Y52, D54, and E56. These residues were scanned through a library containing 800 different CDRs. The screening resulted in one CDR capable of accommodating the five residues in their native conformation with a backbone atom RMSD of 1 Å. The trialed loop corresponds to the H3 loop of the PDB code 5M13 (ZIMMERMANN et al., 2018), a synthetic VHH complexed to the maltose-binding periplasmic protein.

The loop H3 was grafted onto the H3 region of the CabBCII-10 Nb while preserving the loops H1 and H2 of the scaffold. This scaffold was chosen due to its acceptance of grafted binding-loops from other VHHs without compromising its stability (VINCKE et al., 2009). It is important to highlight that the graft of residues being carried out in the loop H3 presents a higher tolerance for receiving grafted residues since contrary to the other binding loops, the loop H3 does not fit into any canonical structure because of its sequence hypervariability (MITCHELL; COLWELL, 2018). Furthermore, typically, the loop H3 is the one that most contributes to the binding affinity against the antigen given its usual location in the center of the binding site, making a greater number of contacts with the antigen (MACCALLUM; MARTIN; THORNTON, 1996).

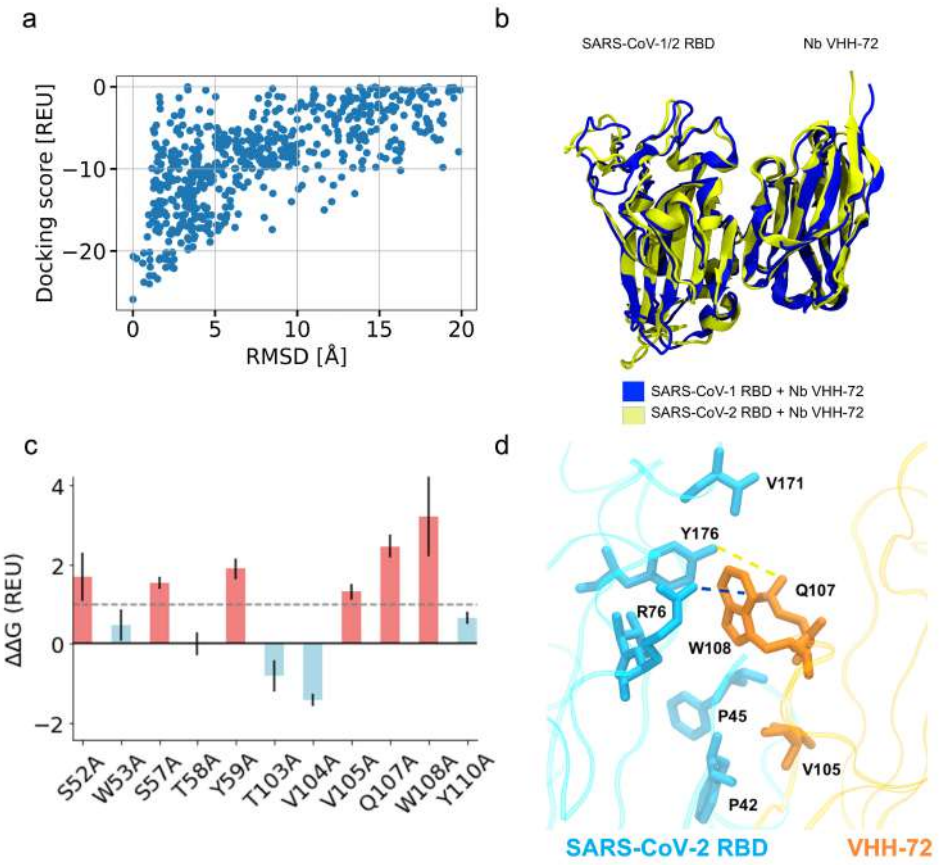
4.2.3.2 Design of VHH-72-derived Nb

The docking funnel for the docking calculations between the SARS-CoV-2 RBD and the Nb VHH-72 (Figure 31a) indicates that structures with low RMSD relative to the chosen

reference structure (the one with the lowest docking score) also exhibit lower docking scores. This suggests that there is a significant degree of convergence, and the reference structure is a representative model for the native structure of the bound state between the Nb VHH-72 and the RBD. Figure 31b shows the structural superimposition of the backbone of the crystal structure of SARS CoV-1 RBD bound to Nb VHH-72 (blue) and the modeled structure of SARS CoV-2 RBD bound to Nb VHH-72 (yellow). The superimposition renders an RMSD for the backbone atoms of ca 1.5 Å.

The 25 lowest-energy structures from the metadynamics simulations were used for AS calculations. As it can be seen in Figure 31c, residues S52, S57, Y59, V105, Q107 and W108 presented a significant contribution to the stability of the complex and these were considered to be hot spots and were kept fixed while all the remaining residues from the CDRs were randomly altered via a Monte Carlo scheme. From the AS calculations, residues Q107 and W108 contribute the most to the binding affinity. W108 makes a cation-π bond with R76, stabilizing electrostatic interaction between a positively charged cation and the polarizable π electron cloud of an aromatic ring. It is considered one of the strongest driving forces in biological association processes (SALONEN; ELLERMANN; DIEDERICH, 2011). Moreover, W108 engages in hydrophobic interactions within a hydrophobic patch, involving Y176, and also with V171 through their aromatic rings. Q107 forms a hydrogen bond with Y176, while V105 plays a crucial role by interacting with P42 and P45 in a hydrophobic pocket.

Figure 31 – (a)Scatter plots depicting the relationship between Rosetta energy scores and RMSD, illustrating funnel-like distributions that emphasize the accuracy of the prediction of the SARS-CoV-2 and Nb VHH-72 complex. (b) Structural alignment of the complexes SARS-CoV 1 RBD + Nb VHH-72 (crystallography, blue) and SARS-CoV 2 RBD + Nb VHH-72 (docking, yellow). Structures are shown in cartoon representation. (c) Computational mutagenesis by alanine scanning for the residues of the Nb in the interaction interface. Each bar denotes the predicted binding $\Delta\Delta G$ for a given residue upon alanine mutation. A threshold of 1 REU, shown as a dashed line, was chosen as cut-off to predict destabilizing effect (red bars), suggesting importance for binding. Blue bars represent stabilizing or no effect in the $\Delta\Delta G$. (d) Schematic representation of relevant interactions in the binding interface of the VHH-72 (residues in orange) and SARS-CoV-2 RBD (blue). Residues are shown in licorice representation. Yellow dashed lines represent hydrogen bonds, and blue dashed lines represent cation-π interaction.

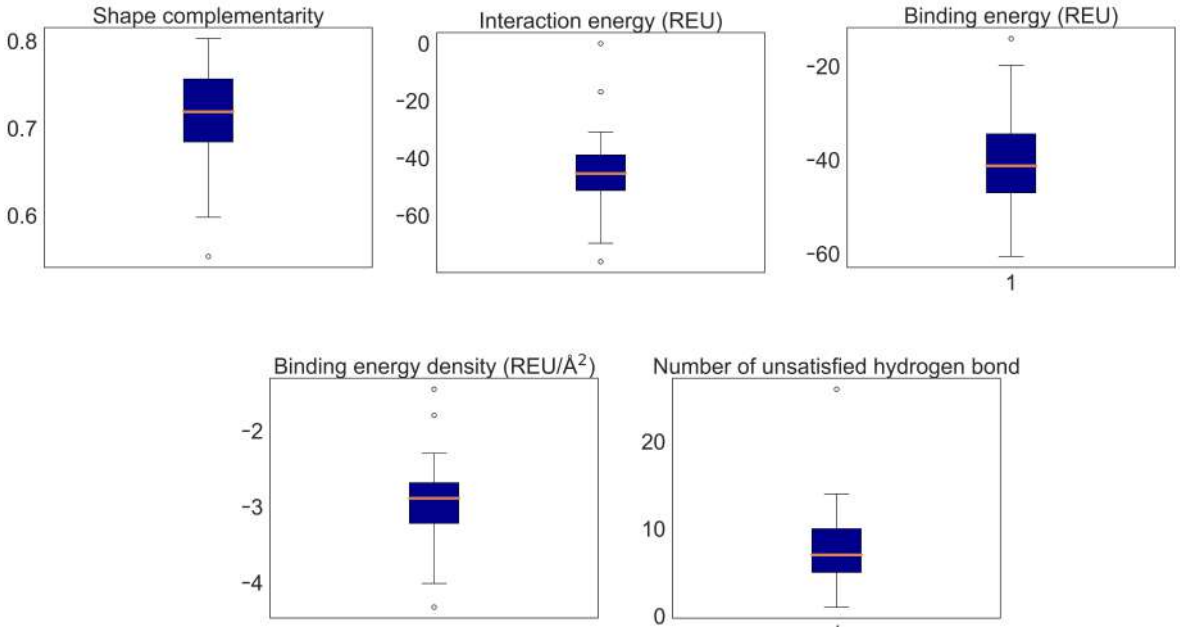


4.2.3.3 Filtering of molecules via interface parameters

A total of one million molecules were designed. These were initially filtered based on calculated interface properties. The interface parameters, calculated via Rosetta, for a data set of natural Nb-antigens interface are shown in Figure 32. The range of these five values was used as the first filtering step. Out of the one million designed proteins, only seven remained after this filtering step, by considering simultaneously the whole range: five based on the structure of the Nb VHH-72 and two based on the cAb CR3022. These seven protein complexes were selected for the next step, which consisted of predicting their binding ΔG

using ML.

Figure 32 – Box plot for visually displaying the distribution and skewness of the interface parameters in 80 natural Nbs-antigen interfaces showing the data quartiles and averages. The interquantile range is shown as a solid blue box, where the top and bottom of the box denotes the upper and lower quantile, respectively. The median is depicted as an orange line. Outliers are represented by circles.



In order to measure their binding affinity, microscale thermophoresis (MST) was carried out by M.Sc. Camilla Adan. The aim of the assessment was to determine if the designed molecules bind to the targets and compare with the predicted affinities of our protocol. To this end, binding experiments were carried out between the designed Nbs and SARS-CoV-2 RBD and the whole SARS-CoV-2 virion using SARS-CoV-2 virus like particles (VLPs), which are virus structures resembling the actual virus but lacking viral genetic material, rendering them non-infectious.

4.2.3.4 Filtering of molecules via machine learning

We utilized the ANN developed in Chapter 3 to predict the binding affinity between the designed Nbs and the RBD of SARS-CoV-2. Our initial investigation focused on examining the impact of including glycosylation on the predictions. To assess the effect of glycosylation in our ANN, we calculated the ΔG of binding for the crystal structures of cAb CR3022 and Nb VHH-72 both bound to SARS-CoV-2 and -1, respectively. Additionally, we considered the

docking structure of Nb VHH-72 bound to SARS-CoV-2. These calculations were performed both with and without glycans attached to the RBD.

The affinity values, both experimental and predicted (with and without the RBD-attached glycans), are presented in Table 5. It is evident that the inclusion of glycans leads to a better agreement with the experimental data, except for the case of CR3022 bound RBD CoV-2, which is overestimated. However, in order to thoroughly evaluate the impact of including glycans on for the calculations using our ANN model, a consistent benchmarking analysis should be conducted. Nonetheless, these findings are encouraging for the screening of the filtered designed Nbs, as they closely resemble the glycosylated structures we have tested in this section.

Table 5 – Comparison of the predicted binding affinities constant (k_D) using artificial neural networks for the Nb VHH-72 and cAb CR3022 and their targets SARS-CoV-1/2 RBDs. k_{DExp} is the experimentally measured binding affinity; k_{DGlyc} is the ANN-predicted binding affinity considering glycosylated structures; $k_{DNoGlyc}$ is the ANN-predicted binding affinity considering non-glycosylated structures;

System	**PDB ID**	k_{DExp}	k_{DGlyc}	$k_{DNoGlyc}$	Experiment
CR3022 + RBD CoV-1	7JN5	1.00 x 10^{-9}	3.45 x 10^{-9}	7.20 x 10^{-9}	BLI[a]
CR3022 + RBD CoV-2	6W41	1.50 x 10^{-7}	1.81 x 10^{-9}	3.3 x 10^{-10}	BLI[a]
VHH 72 + RBD CoV-1	6WAQ	1.20 x 10^{-9}	1.40 x10^{-9}	1.80 x 10^{-9}	SPR[b]
VHH 72 + RBD CoV-2	Modelled	3.86 x 10^{-8}	9.10 x 10^{-8}	9.01 x 10^{-7}	SPR[b]

All the binding affinities are shown in M. Binding affinities (k_D) were obtained via the predicted ΔG using $\Delta G = -RTlnK_D$

[a] (YUAN et al., 2020)

[b] (WRAPP et al., 2020a)

From the seven selected proteins, one protein from each group was selected based on the highest predicted affinity, and followed to experimental binding characterization. The selected proteins are here refereed to as Nb VHH-72.1 (k_D= 2.2 x $10^{-8}M$)and Nb Ab.2 (k_D= 1.5 x 10 ^{-9}M), derived from the Nb VHH-72 and the engineered CabBCII-10, respectively.

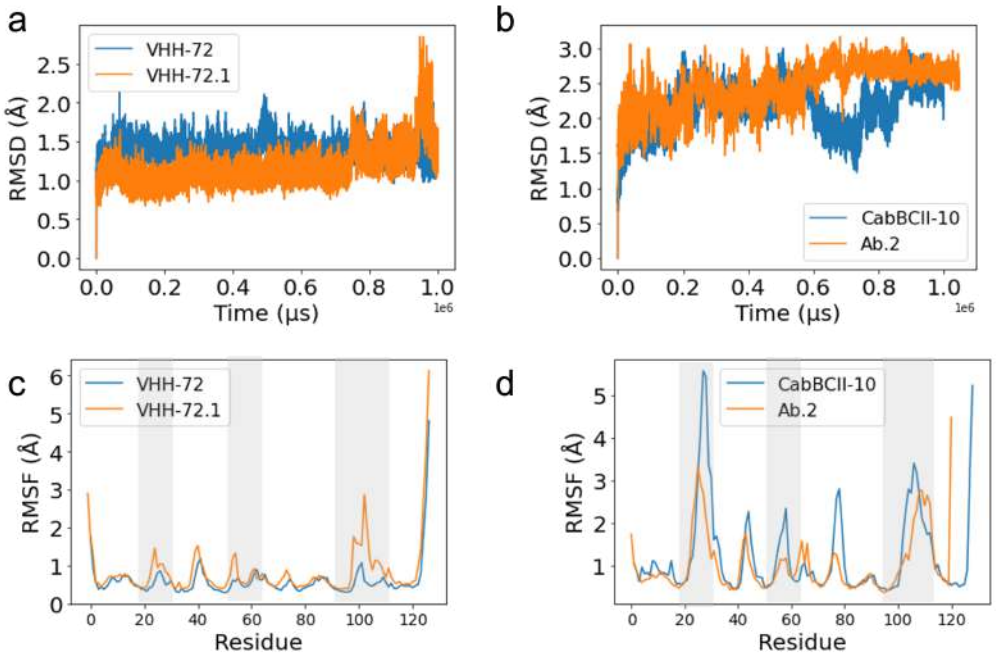
4.2.3.5 Conformational stability of the designed molecules

Prior to the experimental evaluation, MD simulations were used to assess the internal dynamics of the designed Nbs and to compare them with their native counterpart, i.e, Nb VHH-72 for the Nb VHH-72.1, and CabBCII-10 for Nb Ab.2. The average structural prop-

erties obtained from the simulations were captured and recorded as a time series. Figure 33 shows that the simulated structural ensembles exhibited root-mean-square deviation values for the α-carbon atoms in the corresponding X-ray structures of 2.5 Å or below. For the designed structures, the RMSD has a similar behavior. These low RMSD values indicate a significant structural conservation for the native and designed Nbs, with no notable conformational changes during the simulated time. The root-mean-square fluctuation (RMSF) values, which reflect atomic displacements, shows that the designed Nb VHH-72.1 has a higher flexibility in the CDR-3 compared to the Nb VHH-72, and apart from that it does not present differences in the RMSF pattern. For the Nb Ab.2, we observe a stabilization notable around the CDR-1 and -2, whereas CDR-3 presents a similar flexibility, even though the screened loop was grafted onto this region. Overall, it suggests that the designed Nbs maintain the folding structure with increased or decreased flexibility in the binding loops, which can be associated to better adaptation and conformational adjustments to the target molecule.

In addition, the foldability, i.e., the ability of a designed sequence to fold into the predicted 3D structure, was evaluated using the RoseTTaFold (BAEK et al., 2021) via Google Colab (<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/RoseTTAFold.ipynb>) using default parameters. It has been shown that structure prediction serves as a valuable filter in the protein design process. When the designed amino acid sequences are subjected to structure prediction calculations, the expected outcome is the generation of similar structures to the designed models (PROCKO et al., 2014). If the structure prediction results in an alternative conformation or fails to converge to the energy minimum within the conformational landscape, it suggests that the designed sequence may not fold correctly.

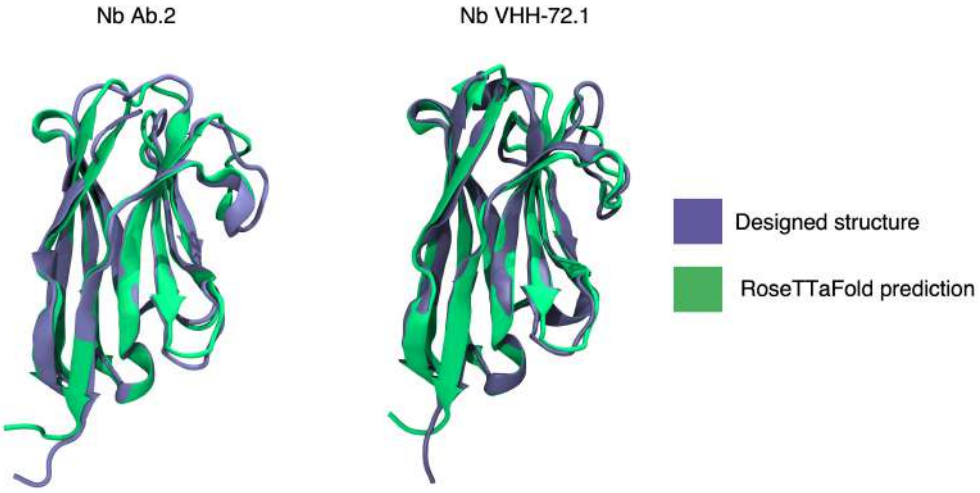
Figure 33 – Time-series properties obtained from MD simulations. (a-b) RMSD as a function of the time between the α carbons from the simulated structural ensemble and the crystallographic or modelled structures. (c-d) Per-residue RMSF for the α carbons calculated for the last 800 ns of simulation. Shaded gray area represents the CDRs 1-3. Blue line is used for the native Nb, and orange line for the designed.



The predicted structures obtained from RoseTTaFold exhibited a local distance difference test (lDDT) score of 0.80 (Figure S2). In the absence of a reference structure, the lDDT metric evaluates the internal consistency within the predicted structures themselves. It assesses the agreement and consistency between different regions or local environments within each predicted structure. A higher lDDT score indicates stronger agreement and consistency among different parts of the predicted structure. Therefore, based on the lDDT scores, the predictions generated by RoseTTaFold appear to be representative of the native state.

Figure 34 depicts the structural alignment of the native and designed structures. The comparison reveals that the predicted structure generated by Rosetta closely matches the computer-designed prediction, exhibiting an RMSD of approximately 1 Å. This finding suggests that both designed proteins are highly likely to fold according to the intended design.

Figure 34 – Structural alignment between the designed structure (iceblue) and the RoseTTaFold prediction (green). The high match between the structures provides evidence of the designed sequences' foldability. Structures are represented in cartoon.



4.2.3.6 Experimental binding affinity measurements

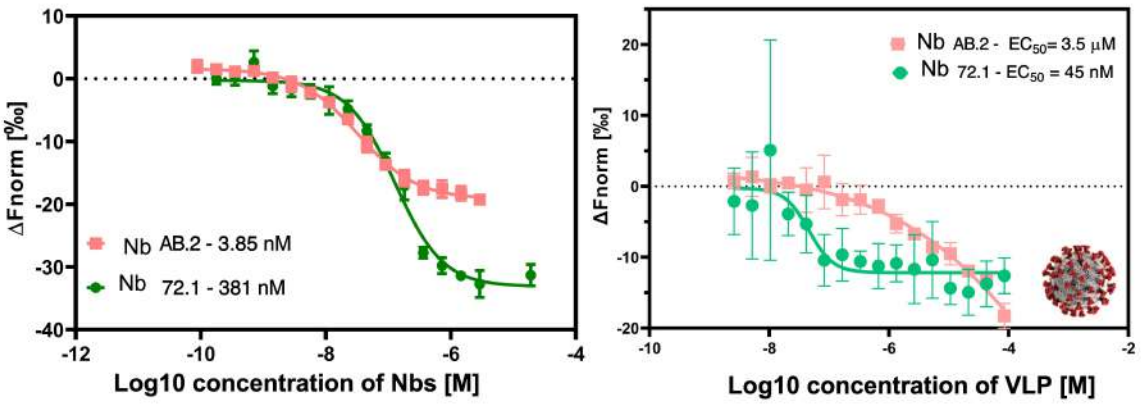
To obtain the experimental values for the binding affinities, the Nbs were heterologously expressed, purified, and their affinities were measured against isolated SARS-CoV-2 RBD and virus-like particles (VLP). All experiments were conducted by the Ph.D. candidate Camilla Adan (UFPE), supervised by Dr. Isabelle Viana (FIOCRUZ). Binding affinities were obtained by microscale thermophoresis (MST). Thermophoresis is a phenomenon where molecules move along a microscopic temperature gradient, influenced by changes in charge, mass, or hydration layer resulting from intermolecular interactions. To analyze binding events, a fluorophore-labeled molecular species is maintained at a constant concentration, and an unlabeled species is gradually titrated until all binding sites become saturated. This approach allows the acquisition of a dose-response curve, enabling the determination of the dissociation constant (k_D).

First, the affinity of Nbs Ab.2 and 72.1 towards the purified recombinant RBD protein was determined. The concentration of labeled RBD was kept constant while the unlabeled Nbs were serially diluted. The MST analysis revealed a k_D of 381 $\pm$ 57 nM (3.8 x $10^{-7}M$) for the interaction between labeled RBD and Nb 72.1. In contrast, the interaction between labeled RBD and Nb Ab.2 exhibited a k_D of 7.1 nM $\pm$ 6.4 (7.1 x $10^{-9}M$). Moving on to the interaction between Nbs and the VLP, the Nbs were maintained at a fixed concentration of 100 nM, while the VLP was titrated across 16 points. It is important to note that the binding between Nbs and the VLP is not 1:1, as observed with the RBD. Hence, the Hills

equation (HILL, 1913; TSO et al., 2018) was employed to estimate the affinity expressed by the EC50. Surprisingly, the affinity of Nbs targeting the VLP exhibited an opposite behavior compared to the RBD. The EC50 for Nb 72.1 binding the VLP was 45 $\pm$ 16 nM (4.5 x $10^{-8}M$), indicating an increase compared to the isolated RBD. Conversely, Nb Ab.2 displayed a significantly reduced affinity of 3.5 $\pm$ 2.8 μM (3.5 x 10^{-6} M). The wide range of error bars observed in the VLP binding affinity can be attributed to the inherent variability in RBD exposure. This variability arises due to the transient nature of RBDs in the S protein, as each S protein consists of three RBDs that can adopt different configurations—either with 0, 1, or 2 RBDs in the "up" conformation. This dynamic nature of RBD exposure introduces significant fluctuations, resulting in a broader distribution of binding affinities and, consequently, larger error bars in the VLP binding measurements.

These findings highlight the promising potential of Nb 72.1 as a high-affinity neutralizing biotherapeutic against SARS-CoV-2, as it presents a typical affinity of neutralizing cAbs. On the other hand, even though Nb Ab.2 had a high affinity for the RBD, it may not be suited for therapeutic applications. However, it must be pointed that Nb Ab.2 was designed towards the RBD, and presented a k_D similar to that predicted by the ANNs.

Figure 35 – Nonlinear adjustment through microscale analysis and thermophoresis. (left) MST experiment conducted with varying concentrations of Nb Ab.2 and Nb 72.1 relative to the labeled RBD (100nM). (right) MST investigation involving the modulation of SARS-CoV-2 VLP concentration in relation to the labeled Nb Ab.2 and Nb 72.1 (100nM).



Results and figure by M.Sc Camilla Adan

To analyze the experimental data, the fluorescence corresponding to the unbound state (Fnorm) was subtracted as a baseline from the fluorescence value corresponding to the bound state, resulting in ΔFnorm (SEIDEL et al., 2013). Figure 35 presents the ΔFnorm values

obtained from independent repetitions of the MST experiment, with error bars representing the standard deviation between repetitions.

4.2.3.7 Elucidating the differential binding between the Nbs and the S protein and the RBD

In this section, we employ molecular modeling tools to investigate the peculiar behavior observed in the cross differential binding affinities. Our analysis highlights the significance of considering the complete structure of the S protein in computational protein design endeavors, as various factors can contribute to distinct affinities between the isolated RBD and the entire S protein. Notably, a recent study by Baker and collaborators (CAO et al., 2020) successfully designed mini-proteins targeting the RBD of SARS-CoV-2, leading to effective neutralization of the virus. However, the authors did not compare the binding affinities of these designed mini-proteins with both the RBD and S proteins. However, our results suggest that it is not always the case that proteins directed against the RBD will be as effective for the S protein.

4.2.3.7.1 Nb - AB.2

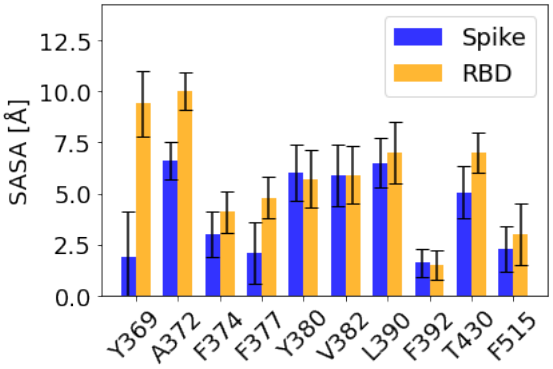
As demonstrated by the MST assays, the Nb Ab.2 presented a high affinity against the isolated SARS-CoV-2 RBD. However, the affinity has drastically decreased when compared to the S protein within the VLP. A putative explanation is that the epitope in which this Nb was designed against, and that CR3022 binds, is buried within the S protein, and it only becomes exposed, and therefore effective for binding, when two RBDs are in the up position, and the targeted RBD presents a rotation of ca. 30° (YUAN et al., 2020). Currently, it is assumed that the lack of neutralization of the CR3022 antibody against the SARS-CoV-2 is attributed to its lower affinity to the SARS-CoV-2 RBD in comparison with the SARS-CoV-1 RBD. However, our assays demonstrate that despite the high affinity of the designed VHH for the isolated RBD, it does not bind tightly enough to the S protein within the VLP, presumably because of the conformational arrangement of the epitope in the S protein.

In this regard, one of the putative mechanisms that preclude the neutralization of the SARS-CoV-2, is that the necessary conformation prone to binding, i.e., in which the epitope is exposed, is not easily achieved. A study demonstrated that a single mutation in this epitope, P384A, leads to neutralization of the SARS-CoV-2 by the CR3022 antibody (WU et al., 2020b). The replacement of the proline, which assures rigidity for the binding site, by an alanine could

turn the region more flexible and allows the S protein to sample more conformational states. Thus, we hypothesize that efficient binding towards the S protein is dependent on its dynamism. To test our hypothesis, 200 ns of MD simulations were carried out for the isolated RBD and the whole SARS-CoV-2 S protein with 2 RBDs up. Both systems were fully glycosylated.

Figure 36 shows that specifically for the residues Y369, A372, and F377, a higher solvent exposure is observed, suggesting that these residues are not available for binding in the S protein. It is worth noting that F377 makes a hydrogen bond interaction with Y52, one of the grafted residues from CR3022. In addition, the residues Y369, A372, and F377 are involved in a hydrophobic core that is solvent-exposed in the RBD, suggesting a structural role. In this case, for binding.

Figure 36 – The solvent accessible surface area (SASA) calculated for the epitope residues in the SARS-CoV-2 RBD from simulations of the isolated RBD and the S protein with 2 RBD ups. The SASA values were averaged over the last 150 ns of the simulation. The blue bars represent the SASA values for the residues in the entire S protein, while the orange bars represent the SASA values for the residues specifically in the RBD. The error bars indicate the standard deviation.

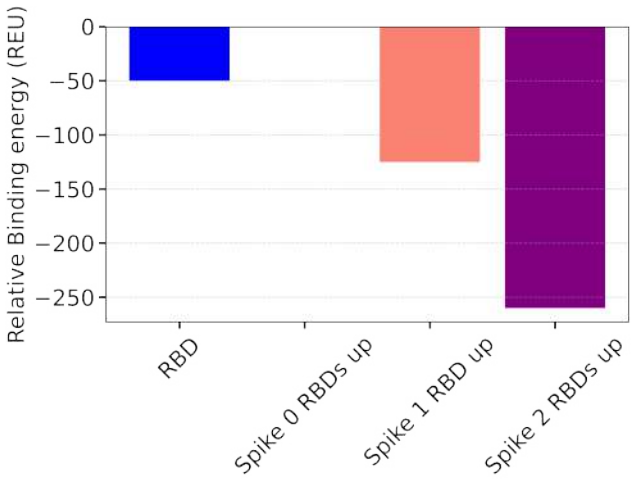


4.2.3.7.2 Nb - 72.1

Regarding the difference in affinity between the Nb VHH-72.1 and the RBD and S proteins, we utilized molecular docking techniques to examine the interaction of the Nb VHH-72.1 with the different conformations of the S protein as well as with the RBD. Since the binding of the native Nb VHH-72 requires at least one RBD in the up position for binding with the S protein (WRAPP et al., 2020a), we did not consider the S protein with the three RBDs in the down position. Conversely, when a single RBD is in the up position, there exists a flexible loop region that hinders the interaction between the Nb VHH-72.1 and the up RBD. However, upon loop remodeling, it is possible to dock the Nb VHH-72.1 in the S protein in a similar fashion as it binds the isolated RBD.

As it can be seen in Figure 37, the ΔG of binding as calculated using Rosetta is more favorable between the S protein with both 1 and 2 RBDs up than in the binding to the isolated RBD. One of the putative reasons for this, is the higher number of contacts and greater interaction interface when bound to the S protein.

Figure 37 – Relative binding free energy calculated using the Rosetta package potentials of the binding between the designed Nb VHH-72.1 and the RBD, and different conformations of SARS-CoV-2 S protein.



While the observed results successfully reproduce the experimental trend, they fall short in providing a comprehensive explanation for the underlying physicochemical factors that contribute to a higher affinity for the S protein as opposed to the RBD. In order to delve deeper into this aspect, we have carried out 100 ns of MD simulations for the Nb VHH-72.1 bound to the RBD and S protein with 2 RBDs up of SARS-CoV-2. Fully glycosylated structures were used. As it can be seen in the schematic representation in Figure 38, the Nb VHH-72.1 makes a significant number of contacts with the glycan moieties from the opposite RBD, suggesting that these glycans can have a buffering effect. For the sake of representation Figure 38 shows a representative frame of the S protein with 2 RBDs up bound to the Nb VHH-72.1.

End-state free energy calculations were performed using MM-PBSA. As shown in Table 6, the binding between the S protein and Nb VHH 72.1 is found to be more favorable than that with the RBD, as confirmed by experimental data and Rosetta calculations. While these calculations do not consider the entropic contribution, which cannot be directly compared across the systems due to their structural differences, they provide valuable insights into the individual components. Specifically, the binding of the Nb exhibits more favorable short-range interactions (van der Waals and Coulomb) within the S protein. In contrast, the polar solvation energy for the S protein is less favorable compared to the RBD, which is expected as the Nb,

when bound to the RBD, has increased contact with the solvent.

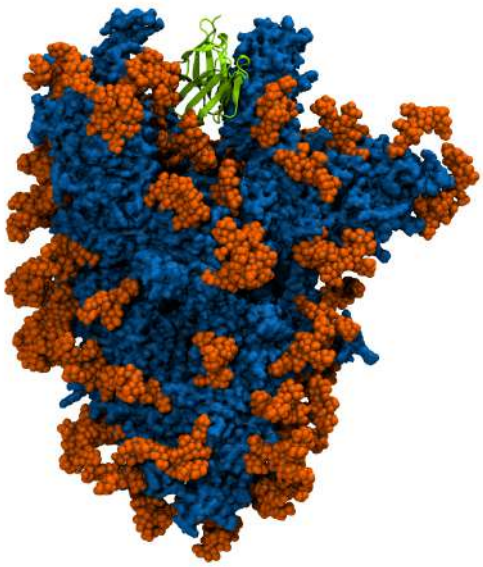
Table 6 – Decomposed free energy terms and total free energies obtained from MM-PBSA calculations for the Nb VHH-72.1 and RBD and S protein of SARS-CoV-2.

System	van der Waal	Electrostatics	Polar solvation	SASA	ΔG
Nb + RBD	-79.2 ± 7.7	-140.0 ± 12.0	129.0 ± 21.2	-8.1 ± 0.7	-98.7 ± 21.1
Nb + S	-135.2 ± 6.1	-157.4 ± 10.5	179.3 ± 8.9	-17.4 ± 1.0	-130.5 ± 12.6

All the energies are presented in kcal/mol.

Furthermore, we conducted an analysis of hydrogen bonds between Nb VHH-72.1 and the RBD and S protein of SARS-CoV-2 throughout the simulation period (Figure S3). On average, the Nb VHH-72.1 forms 7.41 ± 1.42 hydrogen bonds with the RBD, while with the S protein, it forms 14.38 ± 3.33 hydrogen bonds, which can be attributed to the larger binding interface area between Nb VHH-72.1 and the S protein.

Figure 38 – Representative structure of the SARS-Cov-2 S protein with 2 RBDs up bound to the designed Nb VHH-72.1. The S protein is shown in blue surface representation, while the Nb is show in a green cartoon representation. Glycans are shown in orange van der Waals representation.



4.2.4 Conclusions

In this work, we have proposed a pipeline for designing antigen-binding Nbs inspired by the native cAbs/Nbs-antigen interface. We utilized enhanced sampling simulations with metadynamics to harness the power of computational methods. As a proof-of-concept, we applied the pipeline to design two Nbs against SARS-CoV-2 RBD, employing two different design approaches: hot spots grafting and interface redesign. The results demonstrated that both designed proteins effectively bound to the RBD of SARS-CoV-2 with comparable affinities as predicted by ANN. However, an interesting observation emerged when analyzing the behavior of the S protein. It exhibited a contrasting pattern compared to the RBD, indicating that the efficient molecular recognition capability of protein S relies on more complex factors beyond the mere affinity for the RBD.

Furthermore, our findings suggest that the binding affinity is influenced by the selectivity towards specific conformational states of the spike protein. This highlights the importance of considering the spike's conformational variability in understanding the binding dynamics. Such information regarding the spike's differential binding preference could be invaluable for guiding future engineering strategies. Although further studies are required to assess the efficacy of these designed Nbs in terms of *in vitro* and *in vivo* virus neutralization, our data indicates promising potential for their application as prophylactic or therapeutic molecules. Specifically, the Nb VHH-72.1 displayed a binding affinity consistent with neutralizing molecules against SARS-CoV-2 in the nanomolar range (DORMESHKIN et al., 2022; HUO et al., 2020; TZOU et al., 2020), positioning it as a promising candidate in the fight against the pandemic.

Additionally, the low production cost associated with the Nbs makes them suitable for large-scale implementation. Combined with the scalability and cost-effectiveness, a protocol that enables tailoring Nbs for specific targets, are crucial factors in addressing the current and upcoming variants of concern, which have demonstrated the ability to evade immunity.

SUPPLEMENTAL INFORMATION - CHAPTER 4

SI: AN INTEGRATED DATA-CENTRIC AND ENHANCED SAMPLING-BASED APPROACH TO THE DESIGN OF NANOBODIES

Docking command lines using Rosetta

- *Docking prepack*

> $Rosetta/main/source/bin/docking_prepack_protocol.linuxgccrelease -use_input_sc -s [input.pdb] -ex1 -ex2

- *Local docking*

> $Rosetta/main/source/bin/docking_protocol.macosclangrelease -partners A_B -dock_pert 3 8 -use_input_sc -s [input.pdb]

Example of the structures relaxation XML using Rosetta

<ROSETTASCRIPTS>
<MOVERS>
<FastRelax name="fstrlx" repeats="4"/>
</MOVERS>
<FILTERS>
<Geometry name="omega" omega="150" cart_bonded="100" confidence="0"/>
<Rmsd name="rmsd" confidence="0" superimpose="1"/>
</FILTERS>
<PROTOCOLS>
<Add filter_name="omega"/>
<Add mover_name="fstrlx"/>
<Add filter_name="rmsd"/>
</PROTOCOLS>
</ROSETTASCRIPTS>

Command line for hot spots grafting using Rosetta

>$Rosetta/main/source/bin/rosetta_scripts.macosclang- release -database $Rosetta/main/-database/ -l scaf- folds.list -use_input_sc -nstruct 1 -parser:protocol MotifGraft_bb.xml

Example of the hot spots grafting XML file using Rosetta

<ROSETTASCRIPTS>

<SCOREFXNS>

<ScoreFunction name="ref15" weights="ref2015.wts" />

</SCOREFXNS>

<RESIDUE$_SELECTORS$ >

</RESIDUE$_SELECTORS$ >

<TASKOPERATIONS>

</TASKOPERATIONS>

<FILTERS>

</FILTERS>

<MOVERS>

<MotifGraft name="motif_grafting" context_structure="context.pdb" motif_structure="motif.pdb"

RMSD_tolerance="5" NC_points_RMSD_tolerance="5" clash_score_cutoff="5" clash_test_residue
hotspots="1:4,1:4:6,1:2" combinatory_fragment_size_delta="0:0,0:0,0:0" max_fragment_replacement_
8:8,-8:8,-8:8" full_motif_bb_alignment="0" graft_only_hotspots_by_replacement="0"/>

</MOVERS>

<APPLY_TO_POSE>

</APPLY_TO_POSE>

<PROTOCOLS>

<Add mover_name="motif_grafting"/>

</PROTOCOLS>

<OUTPUT />

</ROSETTASCRIPTS>

Example of the general design XML file using Rosetta

<ROSETTASCRIPTS>

<SCOREFXNS>

<ScoreFunction name="ref2015_cst" weights="ref2015_cst.wts" />

<ScoreFunction name="interchain_cen" weights="interchain_cen.wts" />

<ScoreFunction name="dockingfxn" weights="docking.wts" />

<ScoreFunction name="ref2015_full" weights="ref2015">

<Reweight scoretype="coordinate_constraint" weight="0.4"/>
<Reweight scoretype="res_type_constraint" weight="0.4"/>
</ScoreFunction>
<ScoreFunction name="soft_rep_full" weights="soft_rep">
<Reweight scoretype="coordinate_constraint" weight="0.4"/>
<Reweight scoretype="res_type_constraint" weight="0.4"/>
</ScoreFunction>
<ScoreFunction name="ref_pssm" weights="ref2015">
<Reweight scoretype="coordinate_constraint" weight="0.4"/>
</ScoreFunction>
<ScoreFunction name="ref2015" weights="ref2015"/>
</SCOREFXNS>
<RESIDUE_SELECTORS>
<Index name="CDRs" resnums="218-230,245-246,249-252,292-296,300,303-305,307-309"
/>
<Not name="noCDR" selector="CDRs" />
</RESIDUE_SELECTORS>
<TASKOPERATIONS>
<InitializeFromCommandline name="ifcl"/>
<OperateOnResidueSubset name="rpkonly" selector="noCDR" >
<RestrictToRepackingRLT/>
</OperateOnResidueSubset>
<ProteinInterfaceDesign name="InterfaceDesign" repack_chain1="1" repack_chain2="1"
design_chain1="0" design_chain2="1" interface_distance_cutoff="8" jump="1"/>
<ExtraRotamersGeneric name="extra" ex1="1" ex2="1" ex2aro="1" />
<ProhibitSpecifiedBaseResidueTypes name="ala" base_types="ALA" />
</TASKOPERATIONS>
<FILTERS>
<ShapeComplementarity name="Sc" min_sc="2.0" write_int_area="1" jump="1" con-
fidence="0" />
<Ddg name="ddg" scorefxn="ref2015" threshold="0" jump="1" repeats="5" repack="1"
repack_bound="0" confidence="0" />
<BuriedUnsatHbonds name="unsat" confidence="0"/>

```
<IRmsd name="rmsd" jump="1" threshold="3" scorefxn="ref2015" />
</FILTERS>
<MOVERS>
<CoupledMovesProtocol name="coupled_moves" task_operations="rpkonly,InterfaceDesign,ala"/>
<PackRotamersMover name="soft_design" scorefxn="soft_rep_full" task_operations="rpkonly,Inter
<PackRotamersMover name="hard_design" scorefxn="ref2015" task_operations="rpkonly,InterfaceD
<MinMover name="em" chi="1" bb="1" jump="ALL" cartesian="0" type="lbfgs_armijo_nonmonot
tolerance="0.01" max_iter="200" />
<MinMover name="soft_min" scorefxn="soft_rep_full" chi="1" bb="1" jump="ALL"
cartesian="0" type="lbfgs_armijo_nonmonotone" tolerance="0.01" max_iter="200"/>
<MinMover name="hard_min" scorefxn="ref2015_full" chi="1" bb="1" jump="ALL"
cartesian="0" type="lbfgs_armijo_nonmonotone" tolerance="0.01" max_iter="200" />
<Backrub name="backrub_motion" pivot_residues="108-238"/>
<GenericMonteCarlo name="backrub" recover_low="1" temperature="1.0" trials="500"
scorefxn_name="REF2015" mover_name="backrub_motion"/>
<RotamerTrialsMinMover name="RTmin" scorefxn="ref2015" task_operations="rpkonly"/>
<InterfaceAnalyzerMover name="ifa" scorefxn="ref2015" pack_separated="1" pack_input="0"
tracer="0" interface_sc="1" interface="A_B" />
<FavorSequenceProfile name="FPS" pssm="vhh.pssm" scaling="none"
scorefxns="ref_pssm" weight="1" chain="2" />
<ParsedProtocol name="design">
<Add mover="backrub" />
<Add mover="coupled_moves" />
<Add mover="soft_design" />
<Add mover="soft_min" />
<Add mover="backrub" />
<Add mover="hard_design" />
<Add mover="hard_min" />
</ParsedProtocol>
<LoopOver iterations="1" mover_name="design" name="design_protocol" />
</MOVERS>
<PROTOCOLS>
<Add mover="FPS" />
```

<Add mover="design_protocol" />

Add mover="RTmin" />

Add mover="hard_min" />

<Add mover="ifa" />

<Add filter="Sc" />

<Add filter="ddg" />

</PROTOCOLS>

<OUTPUT scorefxn="ref2015" />

</ROSETTASCRIPTS>

Plumed input file example for metadynamics calculation

RESTART

MOLINFO MOLTYPE=protein STRUCTURE=template.pdb

WHOLEMOLECULES ENTITY0=1-1911

ALPHABETA ...

ATOMS1=@phi-100 REFERENCE=pi

ATOMS2=@phi-101

ATOMS3=@phi-102

ATOMS4=@phi-103

... Extensive list of the residues LABEL=abbb

... ALPHABETA

PBMETAD ...

LABEL=pbmetad

ARG=abbb

PACE=500 HEIGHT=1.2 BIASFACTOR=24 WALKERS_MPI

SIGMA=1.5

FILE=HILLS_abbb

... PBMETAD

PRINT STRIDE=500 FILE=BIAS

ARG=abbb

Figure S1. Convergence of metadynamics simulations. The plot shows the one-dimensional free energy profile associated with the collective variable alphabeta for the metadynamics simulations to enhance the sampling of loops H2-H3 for Nb VHH-72 and cAb CR3022. Convergence is shown by plottig the free energy surface at different simulation times. Alphabeta CV measures a distance including pbc between the instantaneous values of a set of torsional angles and set of reference values, here defined is π rad, as used by (LÖHR; SORMANNI; VENDRUSCOLO, 2022). Alphabeta calculates the following quantity:

$$s = \frac{1}{2}[1 + cos(\theta_i - \theta_i^{ref})] \tag{4.1}$$

Where θ_i values are the instantaneous values for the ψ angles, and θ_i^{ref} is the reference value for the torsional angle.

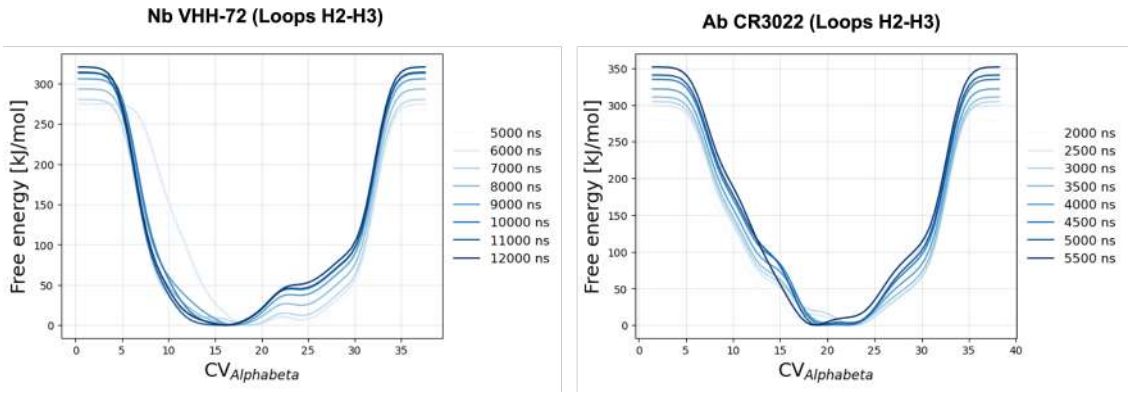


Figure S2. Evaluation metrics for the predicted structures using RoseTTaFold (left) Sequence coverage for the multiple sequence alignment for the designed Nb. The sequence coverage refers to the extent to which the amino acid sequence of a protein is covered or predicted by the model. It indicates the percentage of the protein sequence for which the model provides structural predictions. The low sequence coverage for the last part of the protein corresponds to the CDR3, which is a hypervariable loop in sequence, and therefore it is less covered. (right) Predicted Local Distance Difference Test (LDDT) for each position of the sequence. The LDDT score ranges from 0 to 1, where a higher score indicates a better agreement between the predicted distances and the ground truth distances. A score of 1 implies a perfect match between the predicted structure and the reference structure, while a score closer to 0 indicates a larger discrepancy. the LDDT score is typically computed by comparing the predicted distances within the protein structure to an internal statistical potential derived from the training data used to train the RoseTTaFold model. This statistical potential captures the typical distances observed between pairs of atoms in known protein structures.

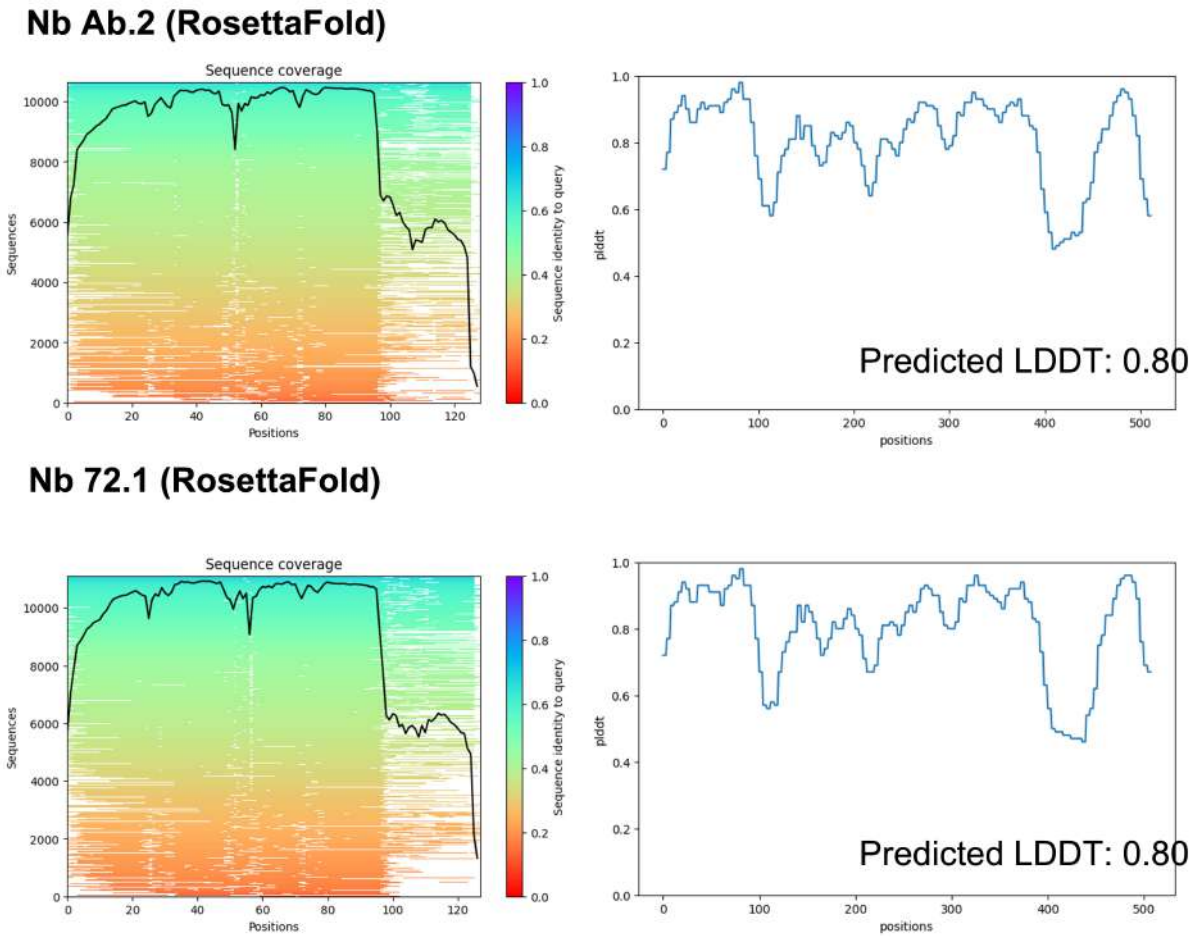
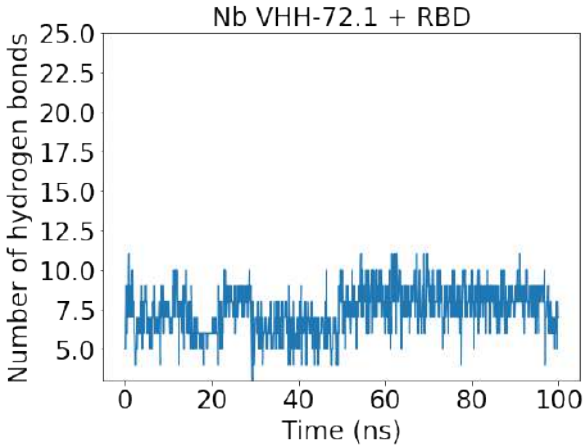
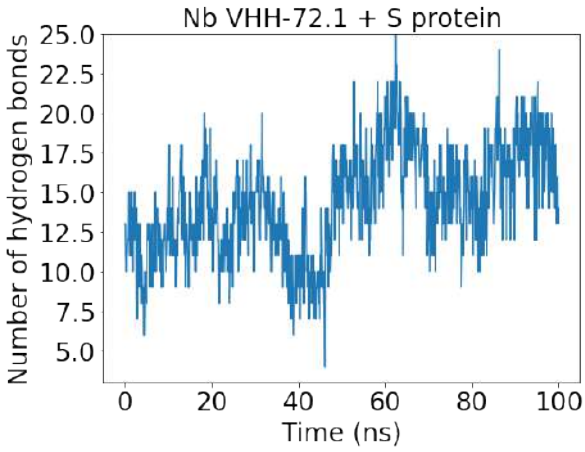


Figure S3. Time evolution series of the number of hydrogen bonds between the Nb VHH-72.1 and the RBD/S protein of SARS-CoV-2. The adopted criterion to be defined as a hydrogen bond was a distance cutoff of 3.5 Å between the donor and acceptor atoms, and an angle cutoff between hydrogen - donor - acceptor smaller than 30°.





5 COMPUTING UNBINDING KINETICS AND DISSOCIATION PATHWAYS

5.1 BACKGROUND

The drug-target residence time model has shifted the focus in pharmacology from the strength of the interaction (k_D) between a drug and its target to the duration of the interaction, as measured by the residence time (τ) (COPELAND, 2016; BERNETTI et al., 2019b). Slow dissociation rates, which result in longer residence times, are generally desirable (DURRANT; MCCAMMON, 2011), as τ has been demonstrated that it serves as a valuable surrogate indicator of *in vivo* target occupancy and exhibiting a strong correlation with clinical efficacy (GUO et al., 2012; LEE et al., 2019). In addition, Vauquelin et al. (2012) have shown that τ is also related to the toxicity of specific inhibitors (VAUQUELIN et al., 2012).

While numerous experimental methods exist to measure τ, they typically lack the ability to access the structural factors that contribute to the transition states involved in ligand unbinding. Obtaining such information would be essential for the development of ligands with longer residence times. In this context, MD simulations offer comprehensive means to investigate protein-ligand interactions and the atomic rearrangements responsible for ligand unbinding (BRUCE et al., 2018; BERGER et al., 2021), shedding light on the underlying mechanisms driving the unbinding process (SHAO; ZHU, 2019; BERNETTI et al., 2019a; KOKH et al., 2020; MARTÍNEZ et al., 2005; SONODA et al., 2008). However, calculating long-term kinetics using all-atom MD models has proven more challenging than computing the binding affinity as it demands the solution of three difficult tasks at the same time: i) the significant discrepancy in timescales between the binding process and the timescales accessible to conventional MD simulations (SOHRABY; NUNES-ALVES, 2022); ii) there is the need to repeatedly sample the slowest transitions, including the intermediates conformational arrangements during the unbinding pathway; iii) the computation of unbiased transition rates from simulation data is a complex task, relying on the applicability of macroscopic rate theories (BERNE; BORKOVEC; STRAUB, 1988).

In this context, numerous enhanced sampling methods have been developed to overcome these issues, including Weighted Ensemble Sampling, the Adaptive Multilevel Splitting Method, and metadynamics (BRUCE et al., 2018; NUNES-ALVES; KOKH; WADE, 2020). These methods have been successfully used to simulate ligand dissociation in several systems, such as trypsin, kinases, HIV-1 protease, and adenosine A2A receptor (SOHRABY; NUNES-ALVES, 2022). Recently, τ-RAMD technique was introduced to compute relative target residence

times of drug-like compounds using nanoseconds-timescale simulations (KOKH et al., 2018). Unlike the methods mentioned previously, τ-RAMD does not need any prior knowledge of the dissociation pathway, nor does it require extensive parameter fitting. Even though τ-RAMD has been successfully applied to a number of different type of systems, it has not been tested for protein-peptide or protein-protein systems, which have been less studied in the literature regarding the prediction of residence times. As discussed in the introduction of this thesis, these systems present an extra challenge for the calculation of unbinding kinetics parameters.

Up until now, various methods like Gaussian accelerated MD, weighted ensemble, and Markov state modeling have been utilized to calculate residence time for protein-peptide and protein-protein complexes (WANG et al., 2023). However, these techniques require extensive sampling, often exceeding tens of microseconds and have so far been an overlooked field.

During the Ph.D. period, in addition to the results presented in this chapter, τ-RAMD simulations were used to predict the relative τ for protein-ligand towards the design of novel inhibitors against 3-Hydroxykynurenine Transaminase from *Aedes aegypti* and *Anopheles gambiae* (MACIEL et al., 2023), an important target against arboviral infections such as Zika, chikungunya, dengue, and yellow fever. Also, we have also exploited the use of τ-RAMD simulations to calculate τ for protein-protein complexes. Notably, we have carried out τ-RAMD simulations for very big systems consisting of more than a million of atoms, and simulations were shown to be computationally tractable. Simulations were carried out to dissociate the hACE2 protein from the fully glycosylated SARS-CoV-2 S homotrimer in Giulia Paiardi (Heidelberg Institue for Theoretical Studies). Heparin has shown to act as an antiviral agent against SARS-CoV-2 (CLAUSEN et al., 2020), and MD simulations have elucidated the mechanism by which it exerts its antiviral effect (PAIARDI et al., 2022). However, the mechanism by which the trimeric spike/heparin/ACE2 complex forms remains unclear. Thus, RAMD simulations were used to simulate the detachment of the hACE2 unbinding to better characterize the dissociation mechanism (Figure 39) and shed light on important aspects of SARS-CoV-2 infection.

Figure 39 – Illustration of the application of RAMD simulations for studying protein-protein interactions by simulating the ternary complex SARS-CoV-2 S homotrimeric protein (homotrimers are shown in teal, purple, and light blue), hACE2 (in green), and heparin (in red). S and hACE2 are represented as surface, heparin as van der Waals, and glycans as licorice. The figure displays (a) relevant interactions within the system, and (b) the most representative metastable states observed along the unbinding pathway from the RAMD trajectories. The yellow arrows indicate the direction of the movement. Despite the systems contain over a million atoms, simulations were conducted for 5 ns within a 24-hour time frame, which is the allotted computation time on the in-house clusters at HITS. Remarkably, even with this limited duration, multiple 5 ns simulations effectively sampled the unbinding trajectories, demonstrating the technical utility of RAMD for big systems with significantly more degrees of freedom, compared to protein-small molecule interactions, which the method was originally developed for.

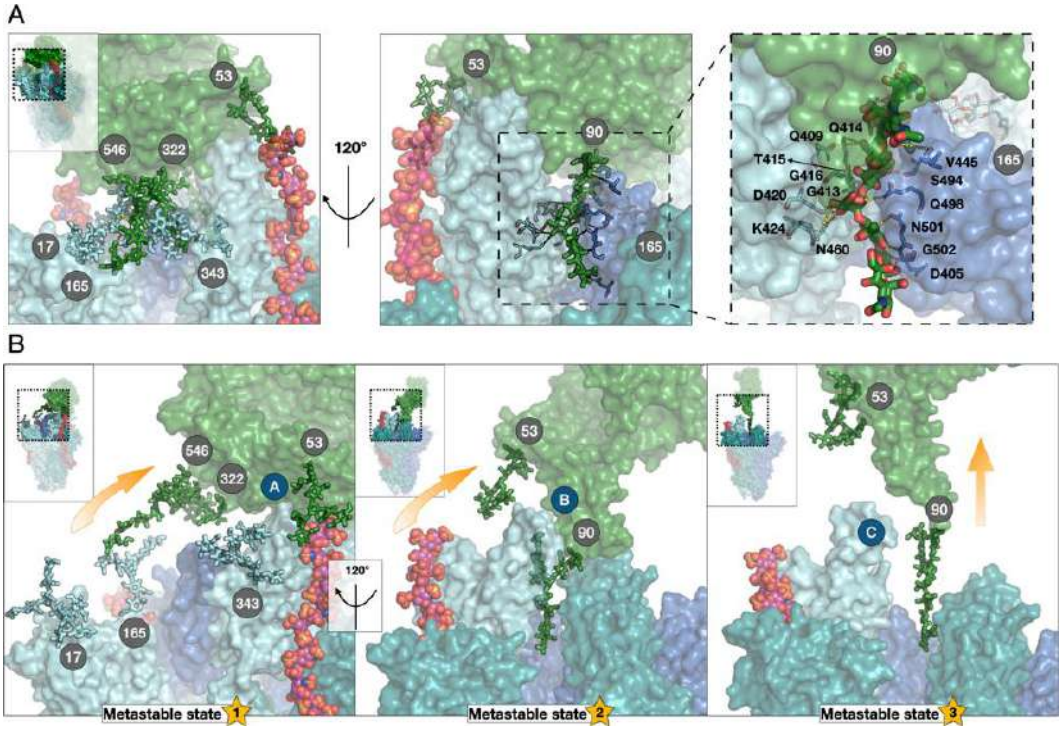


Figure by Dr. Giulia Paiardi

Based on the successful technical aspects of applying τ-RAMD simulations for protein-protein complexes, in this chapter, we propose using τ-RAMD simulations to compute relative residence times for protein-peptide and protein-protein interactions. Firstly, the validation of τ-RAMD for protein-peptide complexes is presented using peptides bound to MHC class I. Next, we apply τ-RAMD to understand the molecular determinants of the binding kinetics for a nanobody binding the SARS-CoV-1/2 receptor binding domain (protein-protein), and to shed light on its neutralization selectivity.

5.2 PEPTIDES BOUND TO MAJOR HISTOCOMPATIBILITY COMPLEX CLASS I: DISSOCIATION MECHANISMS AND OFF-RATES FROM τ-RAMD SIMULATIONS

5.2.1 Introduction

Major histocompatibility complex class I (MHC-I) molecules play a critical role in immune surveillance. They are located on the surface of antigen-presenting cells, where they present peptides that come from extracellular proteins to T helper cells (WIECZOREK et al., 2017). To detect anomalies such as infections, CD8+ T-cells scrutinize peptides presented on MHC-I molecules (ATTAF et al., 2015). While many peptide-MHC complexes are found on antigen-presenting cells, only those recognized by T-cell receptors can trigger an immune response, and these are referred to as T-cell epitopes (REGNER, 2001). Identifying T-cell epitopes is vital for a better understanding of cellular immunity and the development of peptide-based diagnostics, therapeutics, and vaccines. Determining which peptides are presented on MHC-I molecules has been the focus of many efforts in recent years (KNAPP et al., 2009b). The identification of peptides that potentially trigger immune responses against invading pathogens has mostly relied on the binding affinity between the peptide and the MHC-I receptor (KNAPP et al., 2009a). Thus, computational tools employing both sequence-based and structure-based methods have been developed to predict MHC-I and peptide interactions and have made it possible to screen and identify specific antigens from a very large sample space (OCHOA; LAIO; COSSIO, 2019; BLOODWORTH et al., 2022; MARZELLA et al., 2022; MIKHAYLOV; LEVINE, 2023).

Despite the importance of the binding affinity to predict immunogenic peptides, in which so far all the available methods have been aimed at this goal, there is growing evidence that kinetic stability of peptide-MHC-I is a key determinant of immunogenicity and is a better predictor than affinity (HARNDAHL et al., 2012; BLAHA et al., 2019). However, the prediction of immunogenic peptides using kinetic parameters, such as the residence time ($\tau = 1/k_{off}$) has been so far overlooked. Recently, a set of fluorescence polarization data for seventeen peptides binding to major histocompatibility (MHC) class I from mouse (H-2K^b) at different temperatures has become available (GARSTKA et al., 2015). This data includes crystal structures for some of the systems and information on k_D, k_{on}, and k_{off} at 26°C and 32°C. Analysis of the peptide on- and off-rates has revealed that although most peptides exhibit similar on-rates at both temperatures, there can be significant differences in off-rates between the two temperatures, suggesting the importance in taking the off-rates into consideration.

Given the importance of drug-target binding kinetics, the pursuit of understanding structure-kinetics relationship has driven the development of novel techniques for computing kinetic rate constants involved in receptor-ligand binding processes and gaining insights into the pathways of binding and unbinding, as well as the factors influencing structure-kinetic relationships (BRUCE et al., 2018). However, predicting biomolecular binding kinetic rates with high throughput still remains a significant challenge for both experimental and computational approaches (SOHRABY; NUNES-ALVES, 2022). Although peptides are involved in up to 40% of protein-protein interactions in higher eukaryotes, (PETSALAKI; RUSSELL, 2008) and there has been a recent increase in licensed peptide-based drugs (LEE et al., 2019), there are fewer studies on protein-peptide binding compared to protein-small molecule binding. This is mainly because the interactions between peptides and proteins are different from those between small molecules and proteins, and from those between protein-protein. Small molecules can bind to deep and buried sites in proteins, while peptides typically bind to the protein surface, particularly in larger pockets, usually shallow and highly solvent-exposed. Furthermore, protein-peptide interactions are generally weaker compared to protein-protein interactions, as they have a smaller binding interface and usually just a few residues from the peptide are considered hot spots, mainly the central and large residues (e.g., tyrosine and tryptophan). Additionally, given the arge number of rotatable bonds, most peptides lack stable structures before they form complexes (ROBUSTELLI; PIANA; SHAW, 2020; ZOU et al., 2020), making it challenging to incorporate their flexibility and conformational changes into computational modeling.

Up to know, just a few computational approaches have been developed specifically to predict peptide binding kinetic rates, such as infrequent metadynamics (ZOU et al., 2020), Weighted Ensemble (ZWIER et al., 2016), Markov state modeling (ZHOU et al., 2017), and Peptide Gaussian accelerated MD (WANG; MIAO, 2020). In these studies, simulations ranging from 27 μs to a total of 831 μs were employed. This highlights the fact that current enhanced sampling approaches are computationally expensive and time-consuming for characterizing peptide binding kinetics. Moreover, many of these methods rely on the use of collective variables, which can be challenging to select and they present a substantial effect on the outcomes and understanding of the underlying mechanisms. Therefore, there is a need for new methods and techniques that can improve the efficiency of simulations and accelerate the prediction of peptide binding kinetics.

Recently, the τ-random acceleration MD (RAMD) method (KOKH et al., 2018) has been developed to predict the relative residence times for protein-small molecule complexes. The τ-

RAMD method accelerates the dissociation of two molecular partners by applying a randomly oriented force to the center of mass of the ligand, expediting the unbinding event to the nanosecond timescale. By randomly selecting the force orientation, the method avoids any prior assumption on the direction and ensures that the unbinding pathways sampled are unbiased. τ-RAMD has been employed for different biological systems, including T4 lysozyme mutants (NUNES-ALVES; KOKH; WADE, 2021) and G Protein-Coupled Receptor (KOKH; WADE, 2021). However, it has not been validated for protein-peptide systems.

In this work, we evaluated the τ-RAMD method's ability to predict the relative residence time of flexible ligands, such as peptides, at different temperatures using the peptide-MHC data set. Our results show that the peptide binding kinetic rate constants obtained from τ-RAMD simulations matched the experimental data. Additionally, we combined τ-RAMD simulations with molecular dynamics interaction fingerprint (MD-IFP) to gain deeper insights into the peptide-protein binding mechanism at the atomic level. These findings suggest that the τ-RAMD method can be a valuable tool for predicting the binding kinetics of flexible ligands and providing mechanistic insights into their interactions with proteins.

5.2.2 Computational procedures

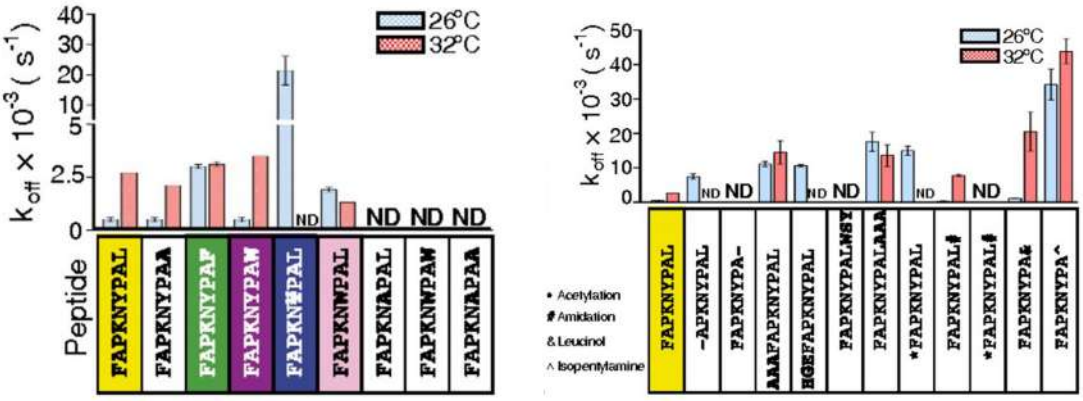
5.2.2.1 Data set

In this study we have used a data set of k_{off} values for 17 MHC Class I molecules (Figure 40) (GARSTKA et al., 2015), each bound to a modified Sendai virus nucleoprotein peptide. Among the data set, crystal structures for five peptides (FAPGNYPAL, FAPGNWPAL, FAPGN(3,5-diiodotyrosine)PAL, FAPGNYPAF, and FAPGNYPAW) bound to MHC-I were available in the Protein Data Bank (PDB) with resolutions ranging from 2.0 to 2.8 Å. The corresponding PDB IDs for these peptides were 4PG9, 4PGB, 4PGC, 4PGD, and 4PGE.

For the fluorescence polarization measurements, G4 was substitute by a lysine, resulting in the sequence FAPKNYPAL. A fluorescent probe - TAMRA - was added at K4. This means the NH3+ would not be present in the assays but TAMRA has two quaternary nitrogens. It is worth noting that the crystal structure was obtained for FAPGNYPAL. Upon replacement from G4 to K4 in our simulations, the TAMRA fluorophore was not included. It is worth noting that the K4 sidechain does not interact with the protein. In addition, the effect of TAMRA (and in our simulations the introduced K side chain) is assumed to be the same for all peptides

so would not affect relative residence times.

Figure 40 – Temperature-dependent kinetics of modified peptide binding to H-2K^b. The dissociation rate (koff) values were obtained via Fluorescence polarization measurements at 26 °C and 32 °C. The average koff values, along with their standard deviations ($\pm$ SD), are presented, derived from a minimum of two independent experiments. ND indicates no binding observed at the highest concentration of H-2Kb (4 μM).



Three peptides (**AAA**FAPGNYPAL, **HGE**FAPGNYPAL, and FAPGNYPAL**AAA**) were removed from our analysis due to large steric clashes resulting from the modeling of extra residues. This suggested significant conformational changes in the crystal structure of the MHC-I, rendering the data unreliable. As a result, the peptides listed in Table 7 comprised the final data set.

Table 7 – k_{off} values of peptides binding to MHC-I.

Peptide	Temperature [°C]	k_{off} x 10^{-3} $[s]$
FAPKNYPAL	26	0.5 ± 0.1
	32	2.7 ± 0.0
FAPKNYPAA	26	0.5 ± 0.1
	32	2.1 ± 0.0
FAPKNYPAF	26	3.0 ± 0.1
	32	3.1 ± 0.1
FAPKNYPAW	26	0.5 ± 0.1
	32	3.5 ± 0.0
FAPKN¥PAW	26	21.4 ± 4.8
FAPKNWPAL	26	1.9 ± 0.1
	32	1.3 ± 0.0
-APKNYPAL	26	7.4 ± 0.8
*FAPKNYPAL	26	15 ± 1.3
FAPKNYPAL#	26	0.3 ± 0.1
	32	7.8 ± 0.3
FAPKNYPA&	26	1.2 ± 0.1
	32	20.6 ± 5.6
FAPKNYPA^	26	34.2 ± 4.5
	32	43.9 ± 3.5

Where * refers to acetylation, # to amidation, & to leucinol, and ^to isopentylamine.

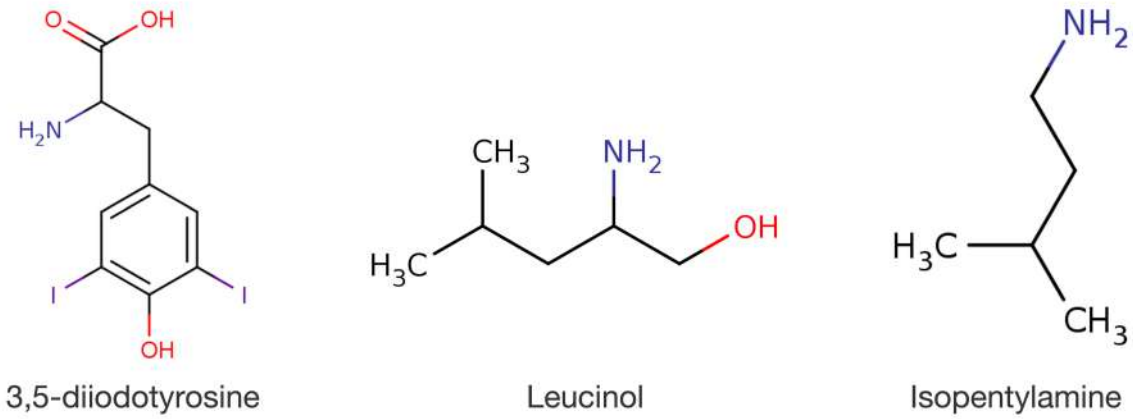
5.2.2.2 Modelling of mutants

The structures of the complexes of peptides bound to MHC-I were retrieved from the PDB under the accession code previously mentioned. For peptides without available crystal structures, we introduced point mutations using the PackRotamersMover tool of the Rosetta package v. 3.12. We repacked the residues within a 5 Å sphere surrounding the mutation site. Both the crystal and modeled structures underwent four rounds of the Rosetta relax protocol, which involves packing sidechains and minimizing torsional degrees of freedom (CONWAY et al., 2014). To ensure accuracy, we used harmonic constraints to fix the backbone atoms during relaxation, and assigned the protonation state of titratable residues at pH 7.5, the pH of the experiments, using PDB2PQR (DOLINSKY et al., 2007).

5.2.2.3 *Parametrization of the modified residues' charges*

The structure of the modified amino acids (3,5-diiodotyrosine, leucinol, and isopentylamine) (Figure 5.2.2.3) were drawn and capped as necessary with acetyl (ACE) and/or N-methyl (NME) residues (CORNELL et al., 1995), in the N and C terminus respectively, using the Avogadro software. The geometry-optimization was carried out in vacuum through GAMESS (SCHMIDT et al., 1993) using the density functional theory approach with the Pople-style 6-31g (d,p) basis set under the B3LYP correlation functional. As the 6-31g basis set is not available for iodine, the basis set used for the 3,5-diiodotyrosine was the Karlsruhe basis set valence triple-zeta polarization (def2-TZVP) (WEIGEND; AHLRICHS, 2005). The def2-TZVP basis set was exported to GAMESS using the basis set exchange web server (<https://www.basissetexchange.org/>) (PRITCHARD et al., 2019). Effective core potentials (ECPs) (STOLL; METZ; DOLG, 2002) were used to replace the core electrons in the calculations involving iodine. The partial atomic charges were computed from the electron density population at the HF/6-31G* level of theory followed by RESP fitting. Distances, angles, and dihedrals were derived from the quantum chemical data.

Figure 41 – Chemical structure for the modified amino acids found in the noncanonical peptides. The atoms and chemical groups are color-coded as follows: carbon (black), amino (blue), hydroperoxy (red), and iodine (purple).



5.2.2.4 *System's set up and equilibration*

The LEaP program (CASE et al., 2005) was used to prepare the systems for MD simulations, employing the AMBER14SB force field (MAIER et al., 2015). The complexes were positioned

at the center of a box with a 25 Å distance from the solute to the edges and solvated with explicit TIP3P water molecules (JORGENSEN et al., 1983). Neutralization of the total charge was achieved by the addition of sodium and chloride ions to create a buffer of saline solution at 150 nM.

The systems were energy-minimized, heated, and equilibrated using the AMBER18 engine. Energy minimization was carried out in a step-wise fashion using the steepest descent algorithm. The heavy atoms of the protein initially restrained with 500 kcal/mol/Å^2, and gradually decreasing to 100 kcal/mol/Å^2, 5 kcal/mol/Å^2, and ultimately being released. Each of these steps were carried out for 1,500 iterations. Then, the temperature was gradually increased to the desired temperature of 299 or 305 K using the Nose-Hoover thermostat (NOSÉ, 1984; MARTYNA; KLEIN; TUCKERMAN, 1992), with 50 kcal/(mol Å^2) harmonic restraints on all heavy atoms. After the initial energy minimization and heating, the systems underwent equilibration using the NPT ensemble. A positional restraint of 50 kcal/(mol Å^2) was applied for 1 ns while keeping the pressure and temperature constant using the Berendsen barostat (BERENDSEN et al., 1984) and Nose-Hoover thermostat (NOSÉ, 1984; MARTYNA; KLEIN; TUCKERMAN, 1992), respectively. The restraints were then released and the simulation was conducted for 1 more ns.

5.2.2.5 τ-random acceleration molecular dynamics

The final snapshot of the AMBER equilibration was used as the starting point for equilibration using GROMACS and sampling the bound state for replica generation. The AMBER-formatted files (topology and coordinates) were converted to GROMACS format using the ParmEd library (<https://parmed.github.io/ParmEd/html/index.html>). The first step of equilibration was performed for 10 ns in the NVT ensemble controlling the temperature using the Berendsen thermostat with relaxation time of 1 ps. Long-range electrostatic interactions were treated using the Particle Mesh Ewald (PME) method (DARDEN; YORK; PEDERSEN, 1993a) with a real-space cutoff or the Coulomb interactions and Fourier spacing of 12 Å. A cut-off scheme for the short-range van der Waals interactions was defined as 12 Å. Covalent bonds involving hydrogen atoms were restrained using the LINCS method (HESS et al., 1997). Then, 20 ns of production using GROMACS was carried out to sample the bound state. The production simulations were performed in the NPT ensemble using the Nose-Hoover thermostat and Parrinello-Rahman barostat.

From the production trajectories, snapshots were collected each 4 ns and was used as input to simulate the peptides dissociation using the RAMD method. Thus, five different structures were used to simulate the peptide's dissociation. For each of the five replicas, 15 RAMD dissociation trajectories were generated, resulting in a total of 75 trajectories for each system. An additional force of random direction with a magnitude of 16.7 kcal/(mol Å) was applied to the center of mass (COM) of the peptides. During the RAMD simulations, the force direction was changed randomly every 100 fs, but only if the ligand center of mass (COM) had not moved more than 0.025 Å. If the COM had moved more than 0.025 Å, the force direction was retained. The simulations were stopped, and the ligand was considered dissociated when the distance between the ligand and protein COMs was greater than 70 Å. The time required for ligand dissociation was recorded for each trajectory, and these times were used to calculate the relative residence times. Example of a GROMACS input file for τ-RAMD is found in the supplemental information.

5.2.2.6 Trajectory analysis

Within our analysis, we focused on identifying protein-peptide contact residues. These were defined as the potential interactions between pairs of protein-peptide interface residues that occurred when they were within a certain distance threshold of 15 Å. Our calculations considered the following interactions: aromatic, hydrophobic, H-Bond donor/acceptor, and cationic/anionic interactions, as defined in (KOKH et al., 2020). The interface residues were computed for each frame in both the equilibration replicas and the RAMD simulations. We also stored additional information, such as protein RMSD and COM position, for each snapshot in a matrix. Binding site contacts were derived from the interface residues matrix and referred to the close contacts that played a crucial role in determining protein interactions. Specifically, we considered pairs of contacts that were consistently found within a closer distance threshold of dr-r=5.5 Åfor more than half of the equilibration trajectories.

To investigate the regions that were frequently visited prior to dissociation and to explore possible dissociation pathways, we performed a clustering analysis in the protein interaction fingerprint space. In this analysis, we converted the contact matrices of PP-REs from all RAMD trajectories into a binary matrix. Contacts below a certain distance threshold (dr-r) were marked as 1, while others were assigned a value of 0. To determine similarities and dissimilarities between RAMD frames based on their PP-REs content, we utilized the Jaccard

Index. This index enabled us to quantify the degree of similarity between frames. Subsequently, we employed the k-means algorithm from the scikit-learn package (PEDREGOSA et al., 2011b) with default parameter values to cluster the RAMD frames based on their Jaccard Index scores. This clustering process allowed us to identify distinct groups representing similar interaction patterns and potential dissociation pathways.

5.2.2.7 Computation of τ

In addition to the standard τ-RAMD protocol, which stops a RAMD trajectory when the COM-COM distance surpasses 70 Å, different criteria for calculating protein dissociation times were also considered.

- τ-RAMD: Conventional criteria adopted for protein-small ligand, in which it measures dissociation by tracking the COM-COM distance until it exceeds 70 Å.
- Few contacts first: Dissociation occurs when the number of contacts drops below 50% of the equilibration simulations in the first frame.
- Many contacts last: The dissociation time is identified as the last frame where the number of contacts remains above 50% of the number of contacts in the equilibration simulations.
- By residue last: The dissociation time is determined by identifying the last snapshot where the average distance between half of the binding site residues is less than a given threshold.
- By residue first: The dissociation time is identified as the first snapshot where the average distance between half of the binding site residues exceeds a given threshold.

The residence time was defined as the duration for half of the trajectories to dissociate, representing 50% of the cumulative distribution function. To obtain the residence time for each replica, denoted as τ_{repl}, along with its corresponding standard deviation, $_{repl}$, we employed a bootstrapping procedure. This procedure involved 50,000 rounds, with 80% of samples randomly selected in each round. Subsequently, we calculated the average residence time, τ_{RAMD}, and its corresponding standard deviation, SD_{RAMD}, by considering all the replicas.

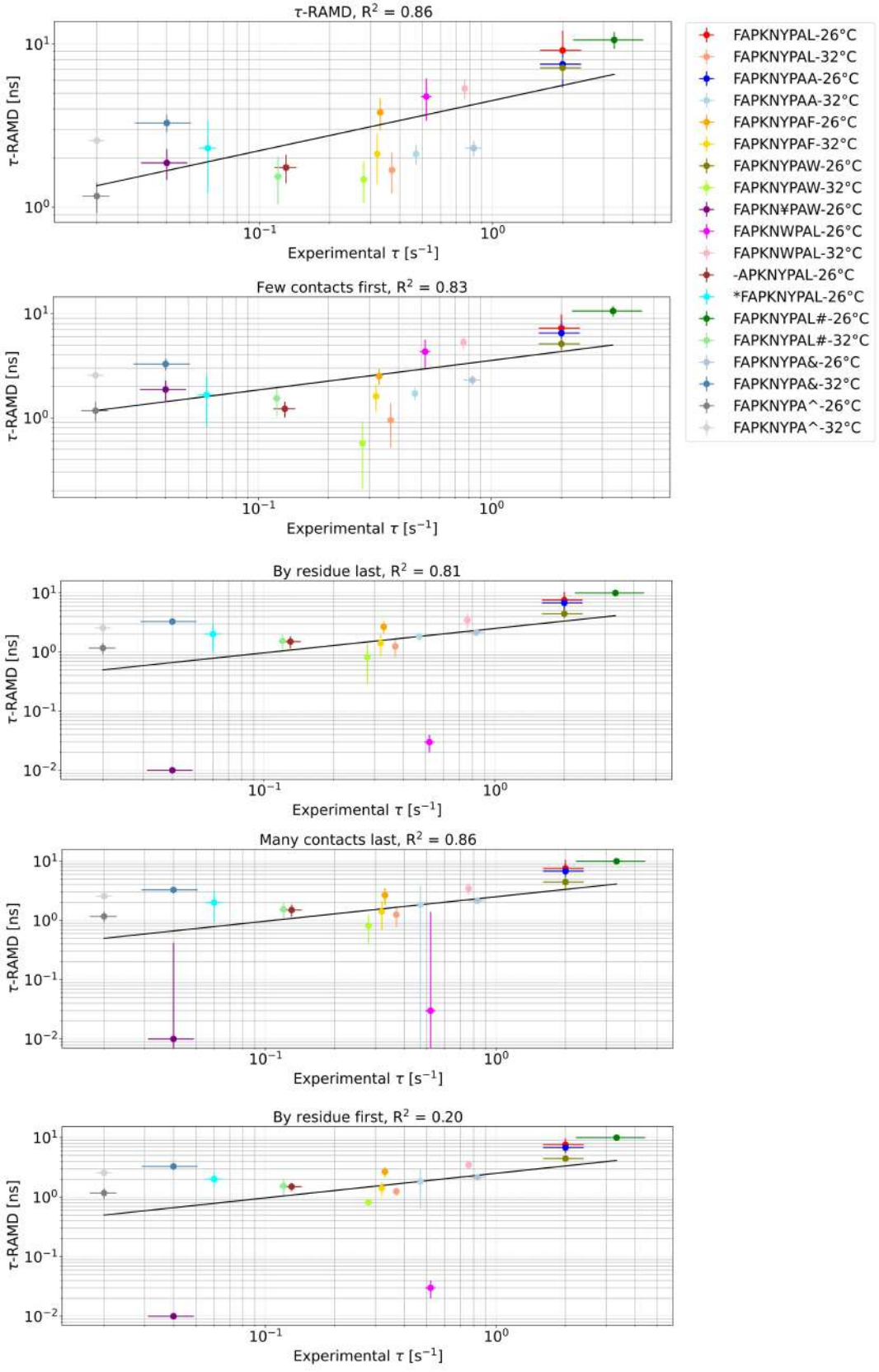
5.2.3 Results and discussion

5.2.3.1 Relative residence times from τ-RAMD simulations correlate with the experimentally measured residence times

Most evaluations of methodologies for computing structure-based protein-peptide residence time have been limited to small data sets, typically consisting of only one case of protein-peptide complex (WANG; MIAO, 2020; ZHOU et al., 2017). The lack of data sets that include both unbinding kinetics and corresponding crystal structures has hindered the assessment of these methods on a larger scale. As a result, methods that compute protein-peptide residence time have not been thoroughly benchmarked. To the best of our knowledge, this is the first study to comprehensively test a method to compute residence time for 19 cases, resulting in over 1400 short simulations. The τ-RAMD protocol was used to record egress times for different peptides from the MHC-I binding groove in a set of 75 trajectories for each peptide. The computed residence time (τ_{RAMD}) and its standard deviation were calculated using various residence time definitions as defined in the methodological section.

We observed a good correlation ($R^2 > 0.85$) between the computed residence time (τ_{RAMD}) and the experimental residence time (τ_{Exp}) from fluorescent polarization assays for all residence time definitions, except for residue first which presented an R^2 of 0.59 (Figure 42). Interestingly, the residue first definition has shown great suitability for protein-protein systems, as indicated by unpublished data from Dr. Giulia D'Arrigo. However, when it comes to protein-peptide systems, this definition proves to be less effective. The discrepancy arises from the fact that in this definition, simulations are halted when half of the binding site residues exceed a certain distance threshold, indicating dissociation. However, peptides exhibit high flexibility, meaning that in some frames, half of the residues may meet the distance criteria for dissociation, but the peptide itself is not fully dissociated. As a result, this definition underestimates the actual dissociation time for peptides.

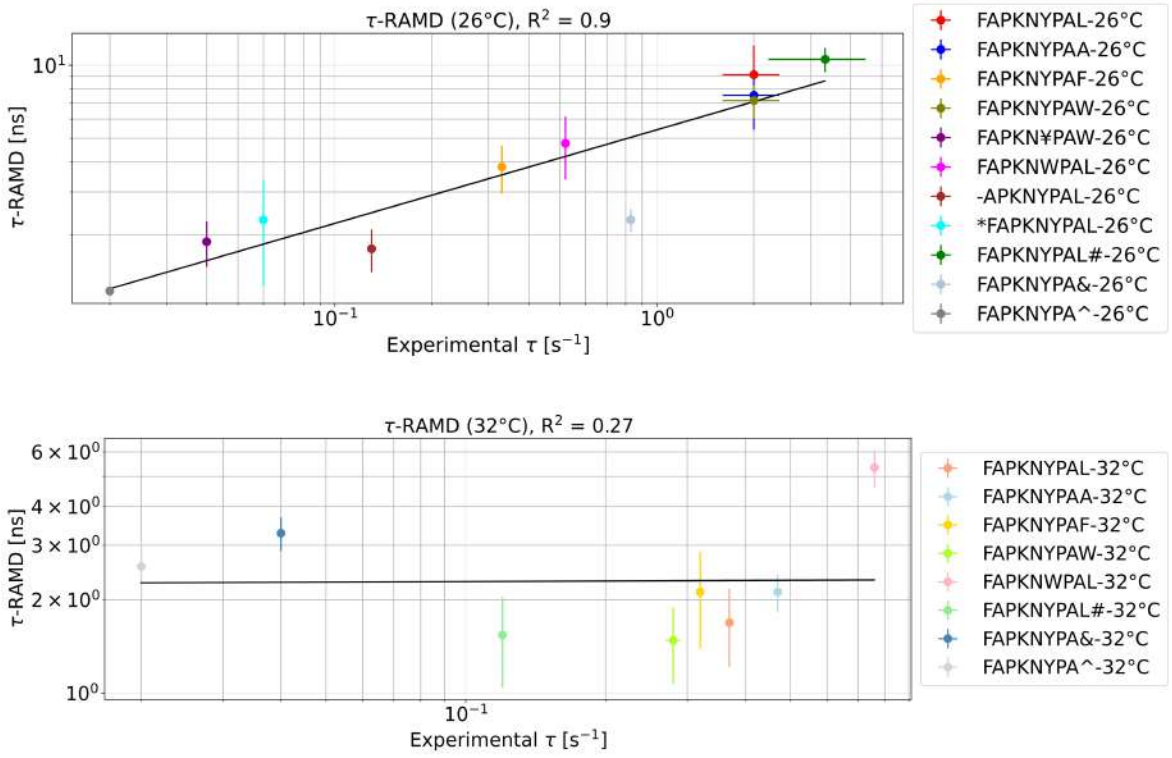
Figure 42 – Comparison of computed (τ-RAMD) and measured (experimental τ) residence times for the peptides-MHC systems at 26 and 32°C. The values are plotted on a logarithmic scale for the different definitions of τ. A linear fitting of the computed to experimental data is depicted by a black line. Standard deviation is presented as error bars.



5.2.3.2 *Lower correlation between the predicted and experimental residence time at 32°C is observed*

τ-RAMD simulations were able to discriminate between fast- and slow-dissociation peptides, even at different temperatures, despite a relatively small temperature difference of 6°C. This capability makes τ-RAMD a valuable tool in the screening of immunogenic peptides. However, as illustrated in Figure 43, the simulations performed at a higher temperature exhibit a weak correlation with the experimental residence times, while at a lower temperature, the R^2 value exceeds 0.9. Plots are shown for the τ-RAMD definition, as it is the classical definition for the τ-RAMD protocol and it yielded consistent results for the complete data set. This observation suggests that τ-RAMD simulations may not accurately capture the molecular determinants of unbinding kinetics at higher temperatures. One potential explanation for this discrepancy could be linked to the force used in the simulations, which was set at 700 kJ.(mol.nm)$^{-1}$ and may be a high force considering that at the higher temperature the dissociation takes place much faster.

Figure 43 – Comparison of computed (τ-RAMD) and measured (experimental τ) residence times for the peptides-MHC systems at 26 and 32°C separately using the classical definition for τ.



One of the possible explanations for this behavior at 32°C could be potentially related to melting temperature of the peptide-MHC complexes. Many of the studied peptides-MHC

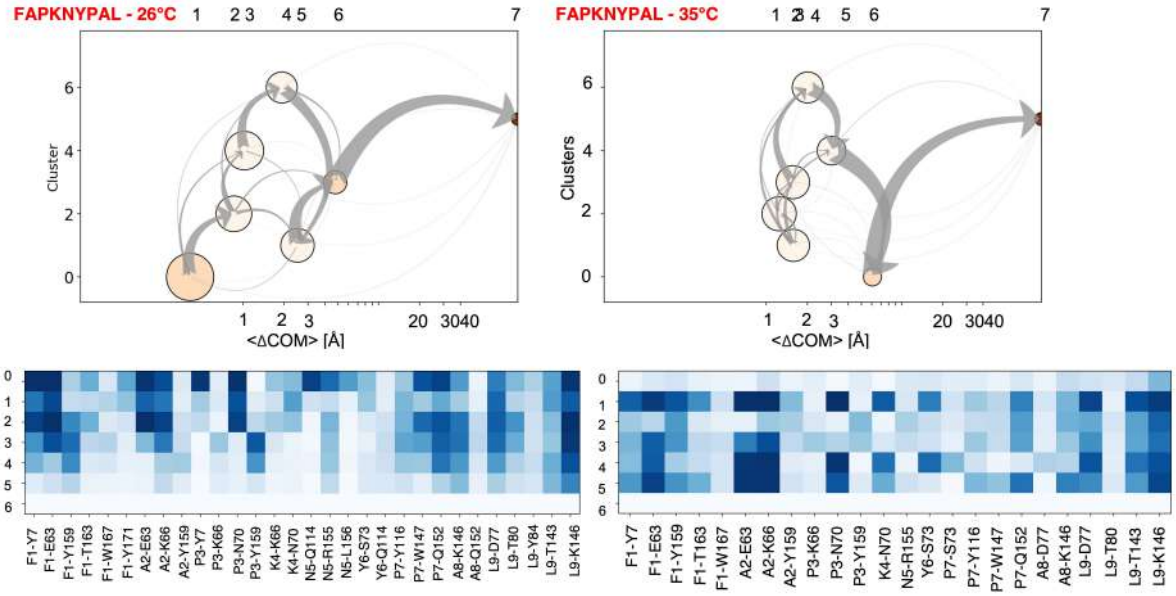
systems in this chapter present a melting temperature below 37°C, which may impact the accuracy of experimental measurements as the systems will be unstable and prone to dissociate faster.

Therefore, it is highly desirable to conduct a systematic benchmark at higher temperatures, such as 37°C, to accurately predict the binding stability of peptide-MHC complexes under conditions closer to the *in vivo* environment. However, the τ-RAMD simulations were able to discriminate between fast from slow binding peptides as well as capture the temperature effect for 26°C and 32°C for the same system.

5.2.3.3 Dissociation is driven by electrostatics and hydrophobic interactions

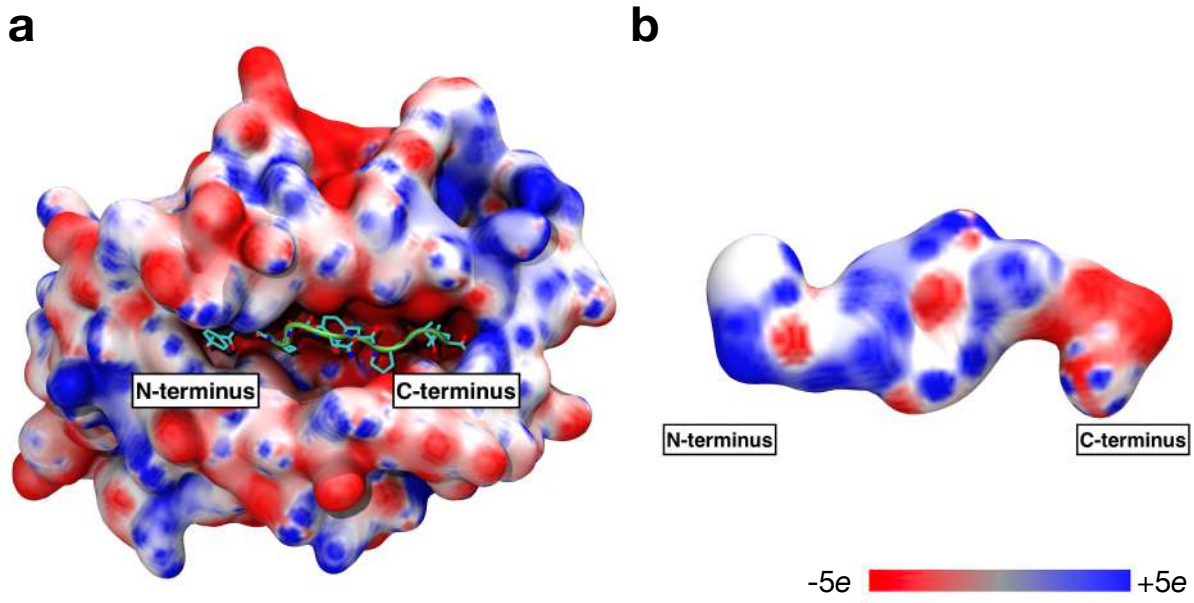
To explore the pathways by which peptides exit the binding site, we generated molecular dynamics interaction fingerprints (MD-IFPs) using the last 300 snapshots of each RAMD trajectory. Our analysis primarily focused on the parental peptide (FAPKNYPAL), as the same behavior was observed for the remaining systems, which can be found in the supplemental material. Thus, the insights gained for the parental peptide allowed us to extend our conclusions to the other systems. The primary dissociation mechanism observed for the peptide at both temperatures involves directly exiting the binding site in a perpendicular direction. Figure 44 depicts the dissociation mechanism using a graph-representation. Each node in the graph represents a cluster or metastable state. These nodes are colored and positioned based on the increasing mean RMSD of the ligand within the cluster relative to the starting complex. The clusters were identified by clustering frames from egress trajectories in IFP space. The size of each node corresponds to the population of the cluster, while the transitions between nodes are depicted by arrows representing simulations. Furthermore, the gray arrows indicate the net transition flux between nodes, with their thickness reflecting the magnitude of the flux. Each node is represented as a line in the heat maps, which indicate the contribution of interactions (with the color palette ranging from white to dark blue representing increasing contribution), show that the C-terminal residue is consistently the last contact to be released for both temperature conditions. This trend was also observed in other systems, suggesting the significance of this position in the unbinding kinetics.

Figure 44 – Analysis of the peptide unbinding from MHC-I at 26 and 32°C in RAMD trajectories. Above, the dissociation pathways are visualized using a graph representation, where each node corresponds to a cluster or metastable state. Nodes are colored and positioned based on the increasing mean RMSD of the ligand within the cluster compared to the starting complex. The size of each node represents the cluster population, and transitions between nodes are depicted by arrows. Below, the heat maps illustrate the composition of clusters in terms of ligand-protein contacts. The color palette, ranging from white to dark blue, represents an increasing contribution of the contacts.



To investigate the role of the C-terminus in binding, the plot of potential electrostatics on the protein surface reveals that the negatively charged deprotonated main chain of the leucine residue (COO^-) interacts persistently with a lysine residue of the MHC (Figure 45). This finding indicates that electrostatic interactions play a crucial role in the unbinding kinetics. This observation is in line with previous studies of peptide-MHC dissociation. It is worth noting that the immune response activation relies on the interface complementarity of peptide-MHC (LIU; GAO, 2011) and electrostatic potential distribution, has been deemed a key physicochemical factor. It has been previously shown that modifications on the electrostatic potential of MHC-binding peptides could favor the induction of different immune responses (AGUDELO et al., 2009; AGUDELO; GALINDO; PATARROYO, 2011). In addition, electrostatics has shown to be a key factor to predict the cross reactivity between peptides and MHC-I (MENDES et al., 2022).

Figure 45 – Electrostatic potential plotted onto the surface of (a) the MHC-I and (b) the peptide FAPKNYPAL.Negatively charged regions are shown in red, positively charged regions in blue, and neutral regions in white. Plotted potential ranged from -5 to +5 $kJ.mol^{-1}.e^{-1}$. In (a) the peptide is shown in in both new cartoon (cyan) and licorice (atom color-coded) representations.

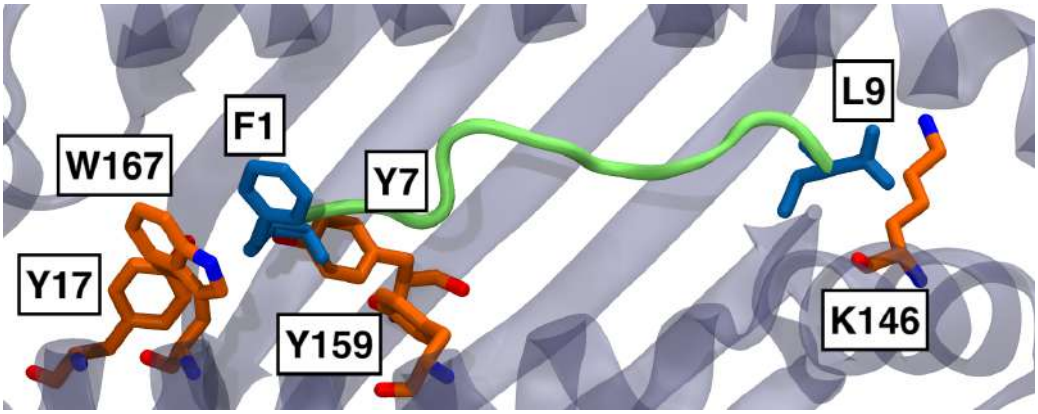


In contrast, we also observe that in the N-terminus, the phenylalanine residue, F1, is situated within a region enriched with tryptophan and tyrosine, creating a hydrophobic pocket that promotes dispersion interactions due to the aromatic ring (Figure 46). It is not unexpected that the N-terminus is detached firstly than the C-terminus, as typically electrostatics are long range, differently from hydrophobic interactions. However, the importance of the hydrophobic interacion in the N-terminus is evidenced by the change in τ due to the deletion of F1 in the first position. In the case of the peptide -APKNYPAL, a significant decrease of the residence time is observed, as well as in *FAPKNYPAL, in which the interactions between the F1 and the hydrophobic pocket are slightly shifted. On the other hand, mutations in the C-terminal leucine, L9, does not impact τ for the peptides, except for FAPKNYPAF. In the crystal structure for FAPGNYPAF, the deprotonated carboxyl group is shifted by ca 1 Å, as compared to the other structures. In addition, the N- and C terminus for FAPKNYPAF in the equilibrated part of the 26°C equilibration trajectory are slightly more flexible, as computed by the all-atom except hydrogen RMSF (Figure S1), than the parental peptide FAPKNYPAL, and FAPKNYPAW, a residue with a more voluminous C-terminus side chain.

In the case of the electrostatic interaction from the C-terminus, it arises from the deprotonated carboxyl from the main chain, in such a way all deprotonated residue will present the same interaction. Furthermore, when the charge of the C-terminus is removed, for example, by

the inclusion of the isopentylamine group, τ is severely reduced. This is not observed for the charge removal using leucinol for capping, which it is likely that the presence of the hydroxyl group can contribute to a negatively charged environment. Lastly, the peptide containing a modified tyrosine, FAPKN(3,5-diiodotyrosine)PAL, also presented a low τ even at low temperature, which could be explained by its high dynamism, as demonstrated by significantly structural disorder of the three N-terminal residues as shown by lack of clear electron density (GARSTKA et al., 2015). In addition, the presence of 3,5-diiodotyrosine promotes significant structural alterations, including the rotation of the proline ring at position P3 by approximately 90°, distinguishing it from the orientations observed in the other structures. Also, the width of the peptide-binding groove is altered, especially around the N terminus, which has shown to be relevant for τ.

Figure 46 – Several key interactions between the peptide are MHC-I. The peptide is in green and a portion of the MHC-I is in the background in purple. The N- and C terminus of the peptide are colored in blue, and the key residues in the MHC-I are color coded as follows: carbon (orange), nitrogen (blue), and oxygen (red).

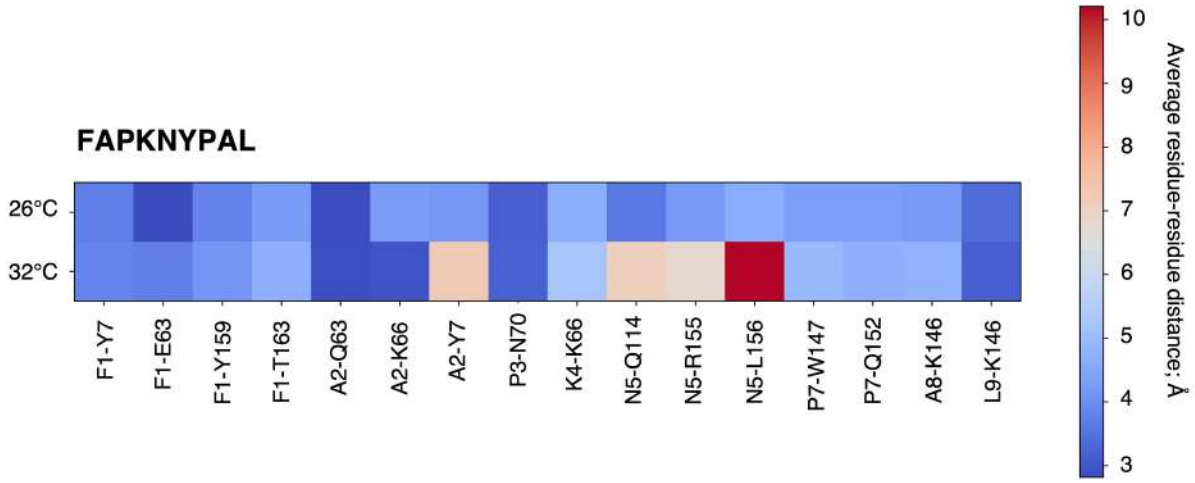


5.2.3.4 At higher temperatures, interactions are lost, facilitating dissociation

The graph-representation of the dissociation pathway based on clustering of trajectories frames (Figure 44) reveals that at higher temperatures, even in the bound state, most clusters exhibit a greater distance between the center-of-mass of the peptide and MHC-I, indicating a loosely bound state. When the temperature increases, molecules gain more thermal energy, causing increased vibrations and movement. Figure 47 demonstrates that in the equilibration trajectories of the bound state of the peptides-MHC, an increase in temperature leads to a greater distance between asparagine N5, situated at the geometric center of mass of the

peptide, and L156. This disrupts a hydrogen bond between these two residues. In addition, N5 also shows increased distance for Q114 and R115. Thus, it suggests that the loosely bound state can be a consequence of the loss of interactions in the center-of-mass of the peptide.It should be emphasized that, for the equilibration simulations at both temperatures, the initial coordinates for the MD remained identical.

Figure 47 – Average residue-residue distance for the contacts in the binding site between the peptide FAPKNYPAL and MHC-I during the equilibration trajectory. The color scheme ranges from blue (low distance) to red (high).



5.2.4 Conclusions

The binding of peptides to MHC-I molecules is a crucial step in the MHC class I antigen-processing and presentation pathway. Early studies focused on understanding the stability and affinity of these interactions, but the available data on peptide-MHC stability is limited due to the labor-intensive and low-throughput nature of current biochemical measurement methods. Up to date, there are only two available methods to predict the kinetics stability for protein-MHC. The BIMAS predictor, developed by Parker et al. (1994) (PARKER; BEDNAREK; COLIGAN, 1994), combines experimental peptide binding data to generate coefficients representing the contributions of each amino acid residue at specific positions within the peptide. However, this method has not been updated since 1997. Another method, NetMHCstab (JØRGENSEN et al., 2014), is a sequence-based tool that uses artificial neural network to predict the stability. Thus, NetMHCstab lacks structural information and therefore does not account for the chemical interactions in the binding site and cannot be used for peptides containing noncanonical amino acids or those with post-translational modifications like glycosylation or

phosphorylation. These modifications have a significant impact on peptide immunogenicity, and their exclusion hinders a comprehensive understanding of peptide-MHC binding dynamics.

In the context of protein-peptide structure-based kinetics, the computation of residence time (τ) has been largely overlooked. To address this gap, we introduce the τ-RAMD protocol for calculating residence time in protein-peptide interactions, specifically peptide-MHC I. Multiple short simulations were carried out to simulate the dissociation of the peptides from MHC-I, affording a faster approach when compared to other enhanced sampling techniques (e.g., Markov state modelling), despite the number of simulations required. Our results demonstrate the reliability of this method by successfully matching experimental and predicted data. These findings have significant implications for the screening of peptides in vaccine applications.

Furthermore, our analysis reveals that electrostatic and dispersion interactions, particularly at the C- and N terminus, respectively, are crucial factors influencing the kinetics of peptide-MHC unbinding. This aligns with previous research on the unbinding mechanisms of peptide-MHC complexes (PAPAKYRIAKOU et al., 2018; SONG; XU; DA, 2023). Additionally, we observe that higher temperatures increase the dynamics of the system, resulting in accelerated dissociation. This phenomenon could potentially be attributed to the higher temperature approaching the melting points, leading to increased instability.

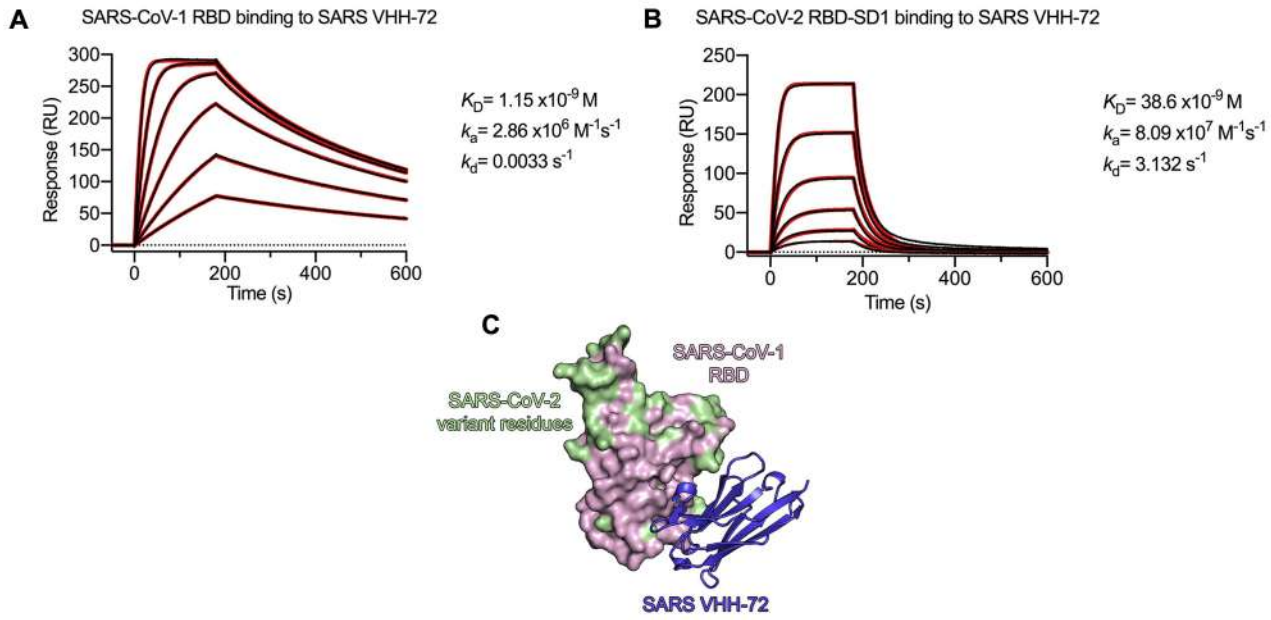
5.3 APPLICATION OF τ-RAMD SIMULATIONS FOR PROTEIN-PROTEIN: ELUCIDATING THE MOLECULAR SELECTIVITY OF VHH-72 AGAINST SARS-COV-1, BUT NOT SARS-COV-2 RECEPTOR BINDING DOMAIN

5.3.1 Introduction

As presented in chapter 4 of this thesis, the Nb refereed to as VHH-72 was isolated from a llama immunized with prefusion-stabilized coronavirus spikes. VHH-72 showed a strong affinity for SARS-CoV-1 and was found to have cross-reactivity with SARS-CoV-2 (WRAPP et al., 2020a). However, although VHH-72 could potently neutralize SARS-CoV-1, it was not as effective against SARS-CoV-2, except for in its dimeric form. Surface plasmon resonance (SPR) assays (Figure 48) showed a 1,000 times higher dissociation constant (k_{off}) of VHH-72 from SARS-CoV-2 RBD, and this could be the primary reason for its reduced neutralization efficacy. From the crystallographic structure, Wrapp et al (2020) suggest that this behavior

could be linked to the absence of a salt-bridge interaction present in the VHH-72-SARS-CoV-1 interface but missing in SARS-CoV-2, which is the only difference in the binding interface of both RBDs, impairing electrostatic complementary between the Nb and the SARS-CoV-2 RBD. On the other hand, electrostatics play a significant role in the diffusional encounters of biomolecules, by usually enhancing protein-protein association rates (ELCOCK; SEPT; MCCAMMON, 2001; CHANG et al., 2006; SPAAR et al., 2006). Therefore, it would be reasonable to expect that k_{on} for VHH-72 and SARS-CoV-1 RBD should be higher, but the SPR data does not support this hypothesis. Accordingly, the detailed clarification of the molecular-level binding selectivity is still absent. A thorough understanding of the chemical mechanisms responsible for the free energy and kinetics of binding between the Nb and the RBDs can provide a basis for the development of novel therapeutic Nbs with the capability of efficiently binding and potently neutralizing SARS-CoV-2 wild-type and variants of concern.

Figure 48 – Sensorgrams obtained from SPR experiments demonstrating the binding interactions between VHH-72 and the RBDs of SARS-CoV-1 (A) and SARS-CoV-2 (B). The binding curves are indicated in black, while the red curve represents the fitting of the data to a 1:1 binding model. (C) Crystal structure depiction of VHH-72 bound to the SARS-CoV-1 RBD, with VHH-72 shown as dark blue ribbons and the RBD represented as a pink molecular surface. Amino acids that differ between SARS-CoV-1 and SARS-CoV-2 are highlighted in green.



Source: Wrapp et al. (WRAPP et al., 2020a)

The binding between VHH-72 and the RBDs of both SARS-CoV-1 and SARS-CoV-2 is primarily driven by a conserved beta-sheet pairing at the interface of each binding partner, which is characterized by hydrophobic and polar interactions. However, due to the high

similarity in the binding interface, investigating the factors that contribute to the binding kinetics presents a challenge. Based on the crystal structure alone, it is difficult to discern the underlying mechanisms that give rise to the observed 1,000-fold higher residence time of VHH-72 with SARS-CoV-1 compared to SARS-CoV-2. Therefore, we postulate that structural dynamism could be a critical factor that contributes to the kinetic selectivity of VHH-72 towards SARS-CoV-1.

For this purpose, parallel computing was used to unveil, at a molecular-level scale, the factors responsible for the fast dissociation of the Nb, and consequently, the ineffective neutralization. To investigate the protein dynamism as a response to thermal fluctuations, MD simulations can be employed. However, simulating unbinding kinetics for protein-protein by brute-force MD is challenging is a rather complex process, as discussed in the introduction of this thesis. In part due to the inability to sample the timescale involved, typically orders of magnitude longer than what is trackable by conventional MD approaches. In addition, to simulate the dissociation mechanisms, the sampling of intermediate transition states between the bound and unbound states is also required. To overcome these issues, enhanced sampling simulations were employed. τ-RAMD and metadynamics were used to simulate the dissociation of the VHH-72 from the RBDs and compute the relative residence times ($1/k_{off}$ and to reconstruct the dissociation free energy landscape, respectively. So far, to the best of our knowledge, τ-RAMD simulations have not been used to compute the relative residence time values for protein-protein systems. The molecular dynamics interaction fingerprint (MD-IFP) approach was used to post-process the dissociation trajectories and identify molecular features that affect unbinding kinetics. In addition, to assess the odd behavior of lower k_{on} for the binding between VHH-72 and SARS-CoV-1 Brownian dynamics simulations were employed to simulate the diffusional association by numerically solving the diffusion equation.

This combination of techniques allowed us to relate the structural and dynamic differences between the two SARS-CoV RBDs to the kinetic selectivity of VHH-72. Jointly, these techniques were able to qualitatively reproduce and explain the SPR data from Wrapp et al. and reveal the detailed unbinding pathways, allowing for the identification of relevant interactions that could fine-tune the binding properties of the Nb towards SARS-CoV-2 RBD. Our results contribute insights for the engineering of novel therapeutic Nbs for efficient binding and potent neutralization of SARS-CoV-2.

5.3.2 Computational procedures

5.3.2.1 System set up

The starting point for the VHH-72 bound with SARS-CoV-1 RBD was the PDB ID 6WAQ, while the modelled complex of VHH-72 bound with SARS-CoV-2 RBD (Chapter 4) was utilized. To incorporate glycosylation, both systems were processed using CHARMM-GUI. In the case of SARS-CoV-1 RBD, only N330 was glycosylated. This decision was based on the fact that the modelled SARS-CoV-2 RBD contains a glycosylated site at the corresponding position. Since this region is conserved in both RBDs, the same glycan was employed.

The following sequences were used for SARS-CoV-1 and -2 RBDs, respetively: 320-502 of SARS-CoV-1 S (Tor2 strain) and residues 319-591.

5.3.2.2 Atomistic simulations

Classical atomistic simulations to equilibrate the structure of the complexes VHH-72 bound to SARS-CoV-1 and -2 RBDs were carried out for 500 ns as described in Chapter 4. For τ-RAMD simulations, the procedure was the same as that applied for peptide-MHC systems in chapter 5.2. A constant force of magnitude 21 kcal/(mol Å^2) was applied to the center-of-mass of the VHH-72.

5.3.2.3 Brownian dynamics

Based on the produced MD trajectories, the most representative snapshot, as from cluster analyses, was extracted from the simulations of the VHH-72 bound to the SARS-CoV-1 and -2 RBD. The protonation states for the proteins were assigned at the experimental pH of 8.0 (WRAPP et al., 2020a) using PROPKA3 (OLSSON et al., 2011). The association constant, k_{on} was computed using SDA v.7 (Simulation of Diffusional Association) (MARTINEZ et al., 2015). To this end, 50,000 runs of BD simulation for each system was carried out. The protein pairs were treated as rigid bodies, and the interaction energies and forces were described by electrostatic interactions with using the effective charge model (GABDOULLINE; WADE, 1996), along with the electrostatic desolvation term as proposed by Elcock et al (ELCOCK et al., 1999). The electrostatic potential of the interacting species was calculated through the numerical

solution of the linearized Poisson-Boltzmann equation under the single Debye-Hückel dielectric boundary condition using the APBS solver, with partial atomic charges retrieved from the AMBER force field (CASE et al., 2005). Electrostatics potential were obtained with a grid spacing of 1 Å, and ionic concentration of 150 mM, as well as a solvent dielectric constant of 78, a solute dielectric constant of 2 and an ionic radius of 1.5 Å. Electrostatic and hydrophobic desolvation grids were generated with a grid spacing of 1 Å. The ionic strength was set to 150 mM for computing the electrostatic grid, with electrostatic and hydrophobic desolvation factors of 1.67 and -0.0065 assigned, respectively. Hydropro (TORRE; HUERTAS; CARRASCO, 2000) was used to calculate the translational and rotational diffusion coefficients with an atomic element radius of 2.9 Å. The Northrup-Allison-McCammon (NORTHRUP; ALLISON; MCCAMMON, 1984) expression was used, defining the encounter complex when two independent native contacts were distant by 6 Å (GABDOULLINE; WADE, 2001). The BD trajectories were initiated at an intermolecular separation of b = 100 Å and truncated at a distance q = 500 Å (GABDOULLINE; WADE, 1997; GABDOULLINE; WADE, 1998). The reaction criteria was were calculated between polar non-hydrogen atoms (atoms within 6 Å on the other solute) (YU et al., 2015).

5.3.3 Results and discussion

5.3.3.1 Dissociation rates

The results of the τ-RAMD simulations were able to qualitatively reproduce the experimental trends in residence times for Nb-72 with the RBDs of SARS-CoV-1 and -2 (Table 8). The predicted residence time for Nb-72 and SARS-CoV-1 RBD was 8.02 $\pm$ 0.93 ns (compared to an experimental value of 303 s or 5.05 min), while for SARS-CoV-2 RBD it was 1.46 $\pm$ 0.5 ns (compared to an experimental value of 0.0003 s or 0.000005 min) Although the experimental difference in residence times is significant, spanning three orders of magnitude, it is important to note that τ-RAMD simulations were designed to calculate relative residence times on a ns timescale. Therefore, the simulations successfully captured the trends in residence times. Statistics for the τ calculations are found in the supplemental information. To improve the Kolmogorov-Smirnov (KS) test statistics, a few more simulations for each trajectory should be run.

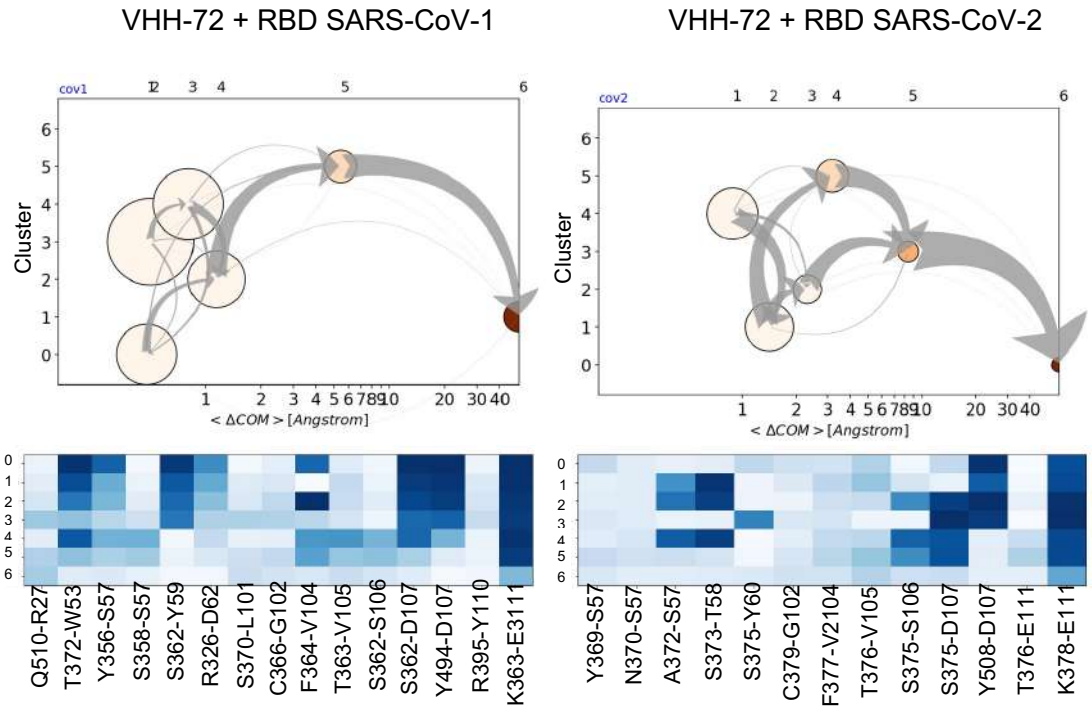
Table 8 – Comparison of experimental and predicted τ for VHH 72 dissociation from SARS-CoV-1/2 RBDs

System	**Experimental τ [s]**	**τ-RAMD [ns]**
VHH 72 + RBD CoV-1	303	8.02 $\pm$ 0.93
VHH 72 + RBD CoV-2	0.0003	1.46 $\pm$ 0.5

5.3.3.2 Unbinding mechanism

To explore the means by which VHH-72 exits the binding site of the RBDs for both SARS-CoV-1 and -2, we generated interaction fingerprints for the last 300 snapshots of each dissociation trajectory. As the dissociation took place, we hierarchically clustered the structures based on their structural similarity. Figure 49 illustrates the dissociation pathways via a graph representation, along with a heat map that represent the structures within each cluster. As it can be seen, the bound state for the VHH-72 is more favourable for SARS-CoV-1 RBD than for SARS-CoV-2, as the clusters in the bound state are bigger, suggesting more stronger binding. In contrast, for SARS-CoV-2, VHH-72 seems to be more loosely bound as dissociation happens in multiple states (intermediates). This observation is opposite to what has been observed for the τ-RAMD simulations for protein-small molecules (NUNES-ALVES; KOKH; WADE, 2021).

Figure 49 – Trajectory analysis of the VHH-72 dissociation from SARS-CoV-1 (left) and -2 (right) RBD in RAMD trajectories. Above, the dissociation pathways are visualized using a graph representation, where each node corresponds to a cluster or metastable state. Nodes are colored and positioned based on the increasing mean RMSD of the ligand within the cluster compared to the starting complex. The size of each node represents the cluster population, and transitions between nodes are depicted by arrows. Below, the heat maps illustrate the composition of clusters in terms of ligand-protein contacts. The color palette, ranging from white to dark blue, represents an increasing contribution of the contacts.



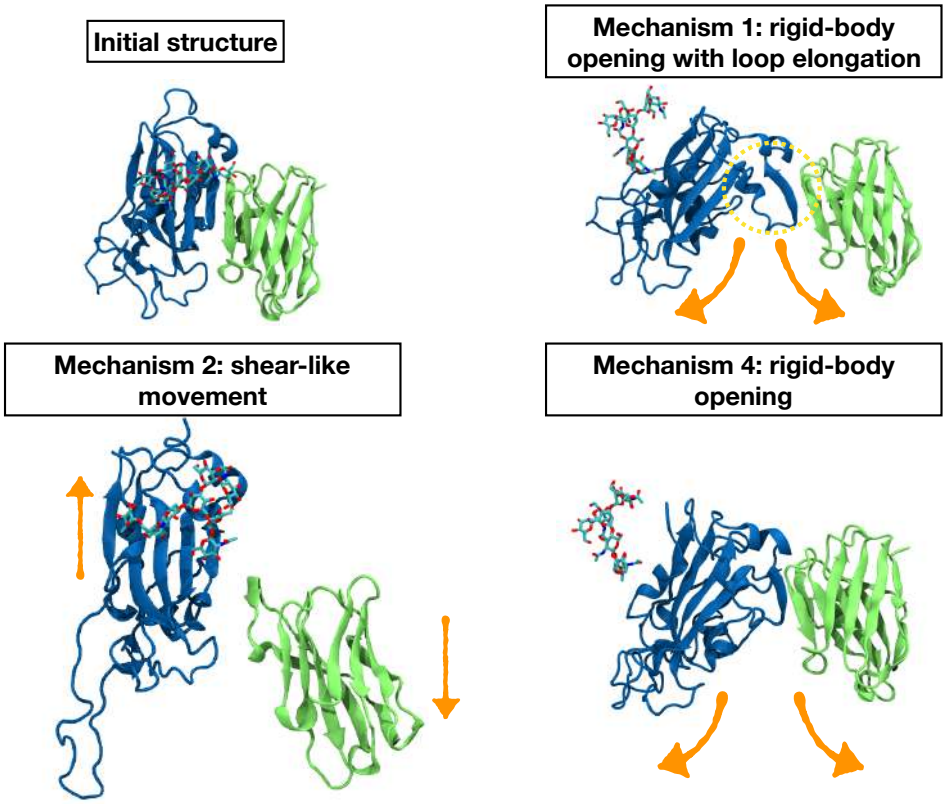
In both cases, the final interaction to be lost is an electrostatic interaction between K and E. This suggests that this particular interaction is the determining factor for the kinetics of unbinding. It is important to note that the interaction between R326 and D52 in SARS-CoV-2, which is associated with longer residence time, is one of the first interactions to be lost during dissociation. This indicates that it may not be the main driving factor for the difference in τ.

Another distinction is observed in the SARS-CoV-1 RBD, where there is a persistent hydrophobic interaction between F394 and V015. This suggests that hydrophobic interactions may contribute to prolonging τ. It is worth mentioning that although the RBD-bound glycans were not considered in the MD-IFP analysis, they are the first to detach from the VHH-72 prior to unbinding. This suggests that their role is more significant in stabilizing the structures rather than contributing directly to τ.

Unbinding trajectories reveal three main dissociation pathways (Figure 50): i. a rigid-body opening with loop elongation during dissociation; ii. a shear-like movement; and iii. rigid-

body opening without elongation. Short 100 ns metadynamics simulations were carried out to reconstruct the free energy landscape for VHH-72 unbinding (Figure S3), and the preferred dissociation route from the multiple binding/unbinding events was mechanism 3, in agreement with the τ-RAMD simulations, as mechanism 3 is the most frequently observed for SARS-CoV-1 RBD, while mechanism 1 is mostly associated to SARS-CoV-2 RBD in the τ-RAMD trajectories.

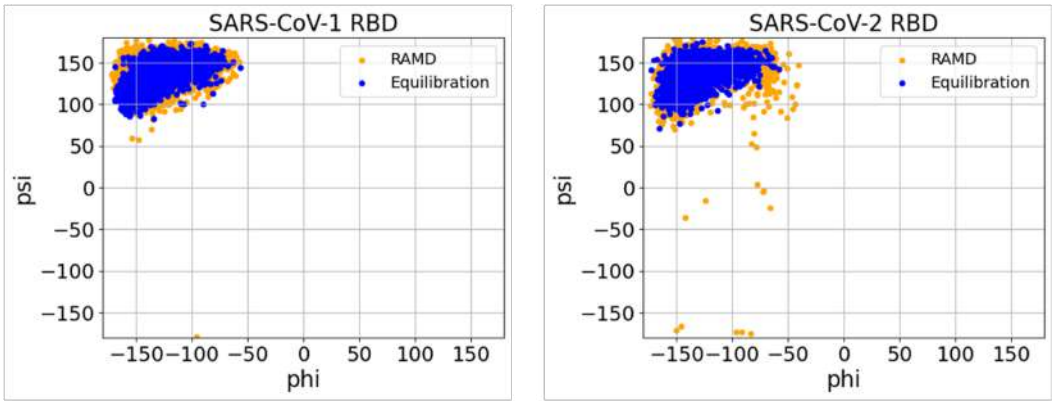
Figure 50 – Representative snapshots of the different dissociation mechanisms. VHH-72 is shown in green and the RBD is shown in blue. Both are represented as cartoon. Glycans are shown in licorice representation coloured as: carbon (cyan), oxygen (red), and nitrogen (blue). Orange arrows indicate the direction of the displacement of the proteins. A dashed yellow circle highlighths the loop alongment



It is crucial to emphasize an additional observation during the dissociation of VHH-72 from SARS-CoV-2 RBD. This behavior is consistently present in most trajectories, where the beta strand in the RBD, which is paired with the beta strand in VHH-72, becomes distorted. Figure 51 illustrates the distribution of ϕ and ψ angles for the residues that form the beta strand (residues 375-379). During dissociation, the angle distribution exhibits a pattern similar to that observed in the equilibration trajectories, occupying typical ϕ-ψ angles for a beta strand

according to the Ramachandran plot. In contrast, for SARS-CoV-2, these residues tend to occupy regions of the Ramachandran plot other than those typical for a beta strand. This instability in the binding to SARS-CoV-2 RBD may also contribute to the observed faster τ upon VHH-72 dissociation.

Figure 51 – Distribution of ϕ and ψ angles in the Ramachandran plot for the backbone of residues 375-379, which form a beta strand, in the RBD of SARS-CoV-1/2. The blue distribution represents the angles obtained from the equilibration trajectories, while the orange distribution corresponds to the angles obtained from the τ-RAMD trajectories.



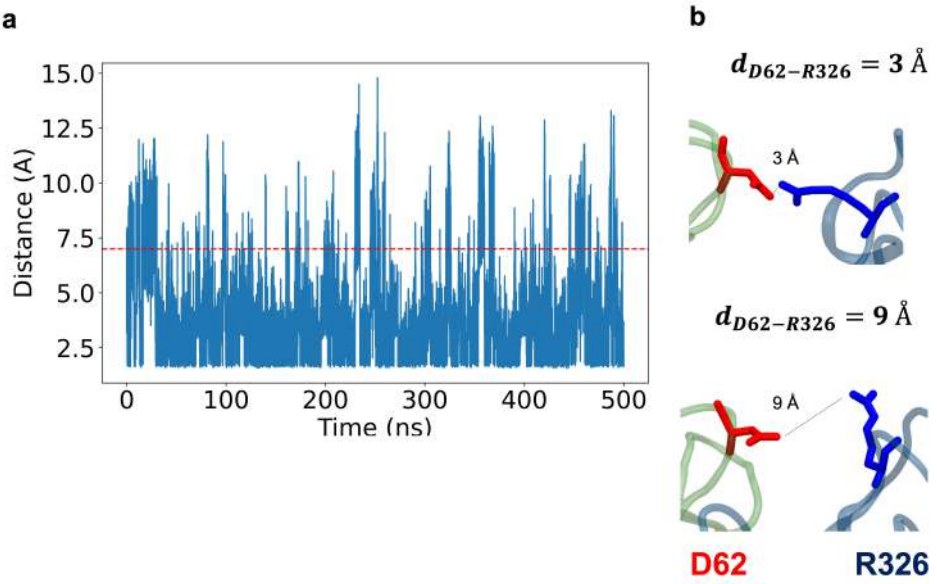
5.3.3.3 Association rates

The SPR assays data from Wrapp et al. (2020) (WRAPP et al., 2020a) revealed an intriguing observation: although an additional salt bridge interaction is present in the interface between VHH-72 and the RBD of SARS-CoV-1 as seen in the crystal structure, the association rate (k_{on}) of VHH-72 with the RBD of SARS-CoV-2 was found to be higher. It is well established that long-range interactions, such as electrostatics, have a significant impact on k_{on} (SCHREIBER; SHAUL; GOTTSCHALK, 2006; RADIC et al., 1997; ELCOCK; SEPT; MCCAMMON, 2001; GABDOULLINE; WADE, 2002; METZGER et al., 2014). Thus, we set out to investigate the lower k_{on} value for the RBD of SARS-CoV-1.

Next, we have simulated the structure for the complex VHH-72 bound to SARS-CoV-1 RBD and computed the distance along the time for the nitrogen atom of R326 sidechain (NH2), and the oxygen atom of the D62 sidechain (OD1). We have monitored these residues because they are the main difference in the binding interface of VHH-72 bound to SARS-CoV-1 and -2 RBD. This interaction is present with the RBD of SARS-CoV-1 and not -2. We observe that despite D62 and R326 are involved in a salt-link most of the simulated time,

in a significant number of snapshots, this interaction is lost, considering a cutoff of 7 Å for the range of electrostatics interaction. Thus, these transient interactions could be a factor to lower the k_{on} value.

Figure 52 – Simulation of the complex VHH-72 and the RBD of SARS-CoV-1 (a) Time-dependent distance between the nitrogen atom of the R326 sidechain (NH2) and the oxygen atom of the D62 sidechain (OD1). The red dashed line represents the cutoff for the range of electrostatic interactions. (b) Representation of the distance ($d_{D62-R_{326}}$) between D62 (shown as red licorice) and R326 (shown as blue licorice). The distance measurement for d = 3 Å was obtained from the most representative frame identified through cluster analysis (frame at 290 ns). For d = 9 Å, the frame was obtained directly from the trajectory without using cluster analysis (frame at 430 ns).



To investigate the impact of this interaction, we have carried out BD simulations to predict the k_{on} values. To this end, two different initial structures were used to simulate the diffusional association of VHH-72 and SARS-CoV-1 RBD: a representative structure obtained from cluster analysis, in which the salt-link D62-R326 is at 3 Å of distance, and one when this distance is 9 Å. For the association between VHH-72 and SARS-CoV-2 RBD, a representative structure from the atomistic MD simulation was used. As it can be seen from Table 9, BD simulations yielded the same order of magnitude as the experimental measurements for the association with SARS-CoV-2. However, for the association with SARS-CoV-1, when considering the salt-link, k_{on} value was overestimated. Upon the removal of the salt-link, by removing it from the list of reaction criteria and using a structure where the salt link is not present, BD simulations results were closer to the experimental value. These results supports our hypothesis that the transiency of this salt-link could be one of the factors that lower the k_{on} value. However, it

is not clear yet the main driving force that makes VHH-72 associated slower to the RBD of SARS-CoV-1 instead of SARS-CoV-2.

Table 9 – Comparison of experimental and predicted k_{on} for VHH 72 dissociation from SARS-CoV-1/2 RBDs

System	k_{on} **Exp** $[M^{-1}s^{-1}]$*	k_{on} **SDA** $[M^{-1}s^{-1}]$
VHH-72 + RBD CoV-2 (Cluster)	8.09 x 10^7	2.50 x 10^7 $\pm$ 3.3 x 10^3
VHH-72 + RBD CoV-1 (Cluster)	2.86 x 10^6	1.80 x 10^7 $\pm$ 7.5 x 10^3
VHH-72 + RBD CoV-1 ($d_{D_{62}-R_{326}} = 9$ Å)	2.86 x 10^6	7.18 x 10^6 $\pm$ 1.7 x 10^3

* (WRAPP et al., 2020a). The authors did not provide standard deviations for the measurements in the original publication.

5.3.4 Conclusions

In this study, we focused on understanding the molecular factors that determine the binding kinetics of VHH-72 to the RBDs of both SARS-CoV-1 and SARS-CoV-2. It is important to investigate these determinants because VHH-72 shows rapid dissociation from the SARS-CoV-2 RBD, making it unsuitable for therapeutic or diagnostic applications. In fact, the fast dissociation of VHH-72 is so pronounced that binding cannot even be detected in ELISA assays. Thus, in neutralization experiments, it did not succeeded (WRAPP et al., 2020a). Despite the challenges, the epitope targeted by VHH-72 on the S RBD is relatively conserved, making it an attractive target for potential broad-spectrum applications against multiple variants of concern, both existing and future ones. It is worth noting that VHH-72 binds to the SARS-CoV-1 S RBD through an intricate network of hydrogen bonds involving CDR loops 2 and 3. Although this network is expected to be conserved in the binding of VHH-72 to the SARS-CoV-2 S RBD, the dissociation kinetics are faster and the affinity is reduced in this case (WRAPP et al., 2020a).

Our analyses revealed that the extensive network of hydrogen bonds is indeed crucial for the dissociation of both VHH-72-RBD complexes. However, for the SARS-CoV-1 RBD, an additional hydrophobic interaction also plays a significant role in the dissociation process. On the other hand, when VHH-72 is bound to the SARS-CoV-2 RBD, it exhibits higher instability. Therefore, in order to improve the binding properties of VHH-72, it would be beneficial to introduce stabilizing mutations that specifically target this epitope, in addition to enhancing hydrophobic interactions.

Furthermore, our findings propose a hypothesis to elucidate the slower association of VHH-72 with the SARS-CoV-1 RBD, despite its superior electrostatic complementarity. This hypothesis emphasizes the importance of dynamic factors in the binding process. Therefore, for the development of novel VHH-72 variants, our approach, which utilizes metadynamics to enhance sampling (Chapter 4), could potentially facilitate the design of enhanced variants with improved binding kinetics.

However, it is important to note that further simulations are required to gain a more comprehensive understanding of the (un)binding kinetics. To accomplish this, an ensemble of conformations obtained from the MD simulation will be used to compute k_{on} via BD simulations to account for solute flexibility. Additionally, the application of long-scale metadynamics simulations can offer valuable insights into the intricate aspects of binding kinetics. The findings presented in this study have the potential to provide valuable insights for the development of enhanced biotherapeutics for COVID-19, focusing on improving stability and binding kinetics.

SUPPLEMENTAL INFORMATION - CHAPTER 5

SI: PEPTIDES BOUND TO MAJOR HISTOCOMPATIBILITY COMPLEX CLASS I: DISSOCIATION MECHANISMS AND OFF-RATES FROM τ-RAMD SIMULATIONS

Example of τ-RAMD input file (.mdp) for Gromacs

integrator = md

comm-mode = Linear

nstcomm = 100

comm_grps = System

tinit = 0.000

dt = 0.002

nsteps = 10000000

nstxout = 5000

nstvout = 5000

nstlog = 5000

nstenergy = 5000

nstxtcout = 5000

nstfout = 5000

compressed-x-precision = 1000

xtc_grps = SYSTEM

pbc = xyz

rlist = 1.10

coulombtype = PME

cutoff-scheme = Verlet

fourierspacing = 0.12

pme_order = 4

ewald_geometry = 3d

ewald-rtol = 1e-5

ewald-rtol-lj = 1e-5

optimize_fft = yes

vdw-type = Cut-off

vdw-modifier = Potential-shift

rvdw-switch = 0.0

rvdw = 1.00 tcoupl = Nose-Hoover

tc_grps = Protein Water_and_ions

tau_t = 1.0 1.0

ref_t = 299 299

Pcoupl = Parrinello-Rahman

pcoupltype = isotropic

tau_p = 2

compressibility = 4.5e-5

ref_p = 1

gen_vel = no continuation = no

constraints = h-bonds

constraint_algorithm = lincs

lincs_order = 4

lincs_iter = 1

lincs_warnangle = 60

DispCorr = EnerPres

ramd = yes; RAMD will be applied.

ramd-seed = 98XX; Seed for random direction generator.

ramd-ngroups = 1; The number of ramd groups defining the ligand-receptor pair

ramd-group1-receptor = r_1-375; Receptor for the first RAMD group.

ramd-group1-ligand = r_376-384; Ligand for the first RAMD group.

ramd-group1-force = 800 ; The force constant in kJ/mol/nm. Default value is 600 kJ/mol/nm

ramd-group1-r-min-dist = 0.0025; This parameter affect absolute dissociation time but have less effect on the relative dissociation times of different compounds. It is recommended to use default value.

ramd-group1-max-dist = 7.0; This value has to be adjusted for the system studied: no protein-ligand contacts should be observed in the last snapshot of a dissociation trajectory

ramd-eval-freq = 50; This parameter affect absolute dissociation time but have less effect on the relative dissociation times of different compounds. It is recommended to use default value.

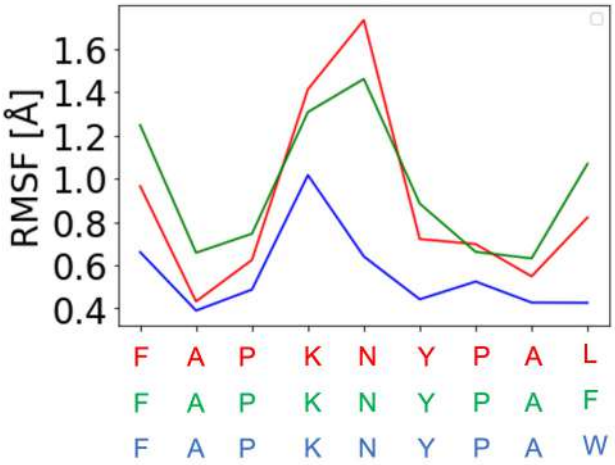
ramd-force-out-freq = 10; This ramd parameter resets pull-nstxout and pull-nstfout.

ramd-pbc-ref-prev-step-com = yes; The value will be forwarded to pull-pbc-ref-prev-step-com. Default value is 'yes'.

ramd-group1-ligand-pbcatom = 6075; The value will be forwarded to the ligand pull group of the receptor. Default takes the middle atom

ramd-group1-receptor-pbcatom = 3001; The value will be forwarded to the associated pull group of the receptor. Default takes the middle atom

Figure S1. Root mean square fluctuation (RMSF) per residue during the equilibration trajectories. RMSF computed for the last 30 ns of simulations. Data is shown for the peptides FAPKNYPAL (red), FAPKNYPAF (green), and FAPKWNYPAW (blue). The RMSF was calculated for all atoms except for the hydrogens.



SI: APPLICATION OF τ-RAMD SIMULATIONS FOR PROTEIN-PROTEIN: ELUCIDATING THE MOLECULAR SELECTIVITY OF VHH-72 AGAINST SARS-COV-1, BUT NOT SARS-COV-2 RECEPTOR BINDING DOMAIN

Figure S1-S2 present statistical plots illustrating the τ-RAMD protocol used to compute the residence time for the VHH-72 with the SARS-CoV-1/2 RBDs. The following information is depicted: In the first row, cumulative distribution functions demonstrate the RAMD dissociation times for the six replicas. The red line represents the effective residence time, indicating the simulation duration at which dissociation occurred in 50% of the runs. In the second row, the distribution function for the effective residence time is displayed after post-bootstrapping (in blue). The black line represents a Gaussian distribution, while the mean residence time is indicated by red lines. The third row presents a comparison between the Poisson cumulative distribution function (black line) and the empirical cumulative density function (blue points). Each replica's residence time is represented by a red line, and the Kolmogorov-Smirnov (KS) test is utilized to calculate the distance between the Poisson and empirical distribution functions. In the fourth row, bar plots showcase the relative residence times averaged over each replica. Whiskers represent the range of the data, while outliers are depicted as individual points. The median and mean of the residence time are denoted by orange and dashed red lines, respectively.

Figure S1. Statistics plots for the τ-RAMD protocol for computing the residence time for VHH-72 and the RBD of SARS-CoV-1

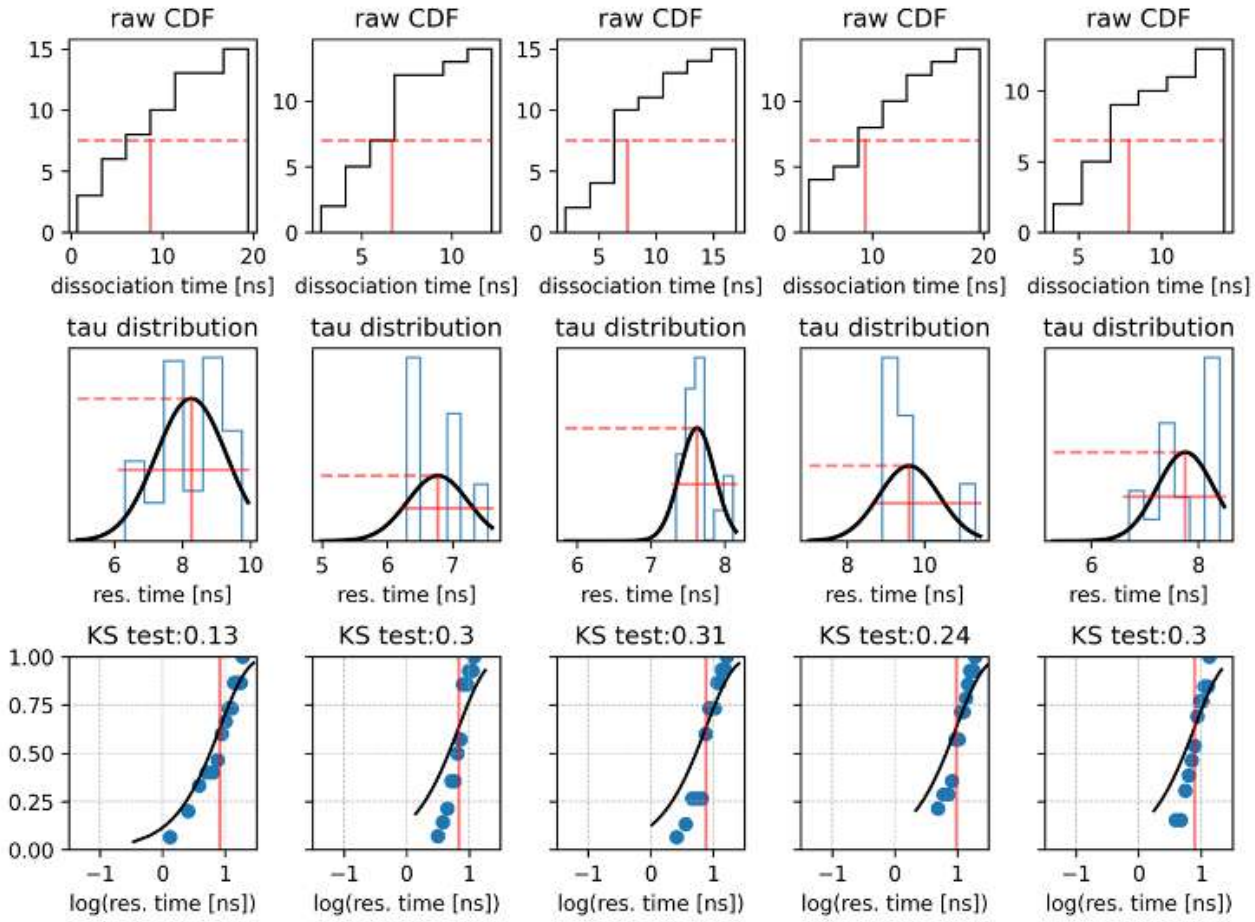


Figure S2. Statistics plots for the τ-RAMD protocol for computing the residence time for VHH-72 and the RBD of SARS-CoV-2

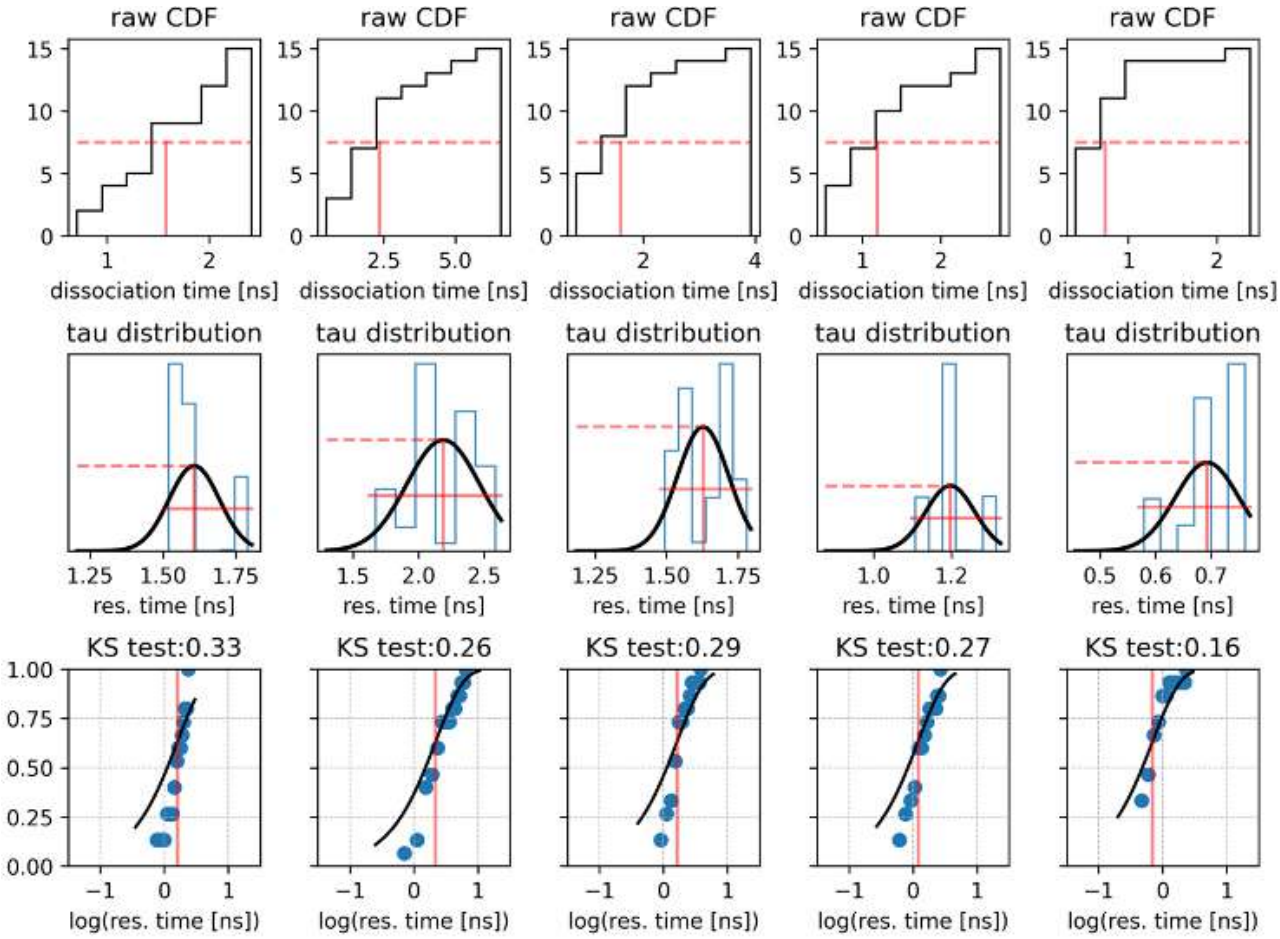
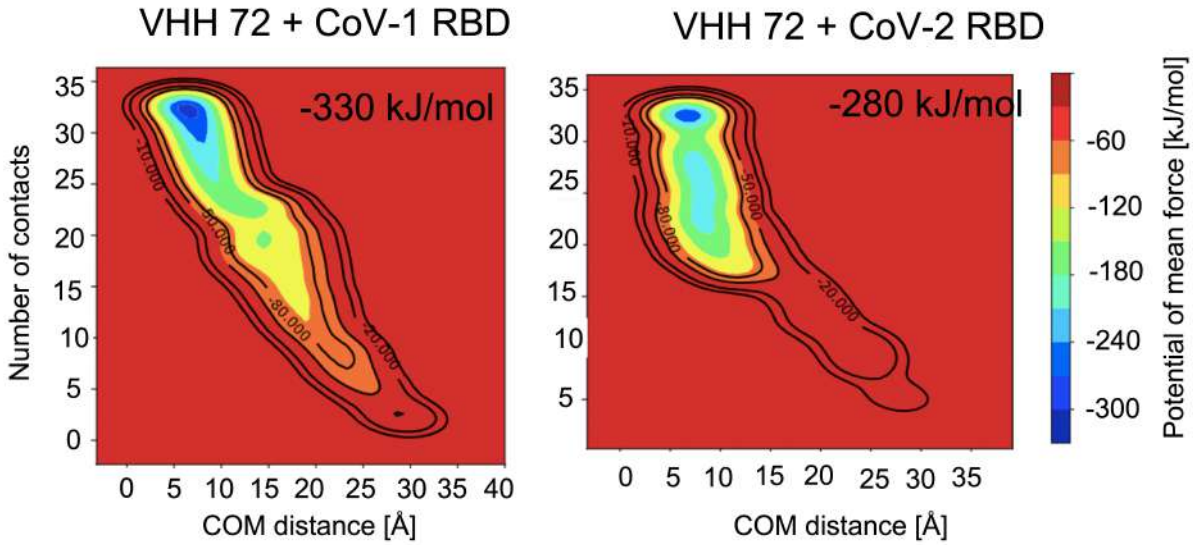


Figure S3. Well-tempered metadynamics simulations were carried out for 100 ns using GROMACS 2020 with PLUMED 2.5 plugin (TRIBELLO et al., 2014). To reconstruct the free energy surface, two collective variables (CVs) were used: CV1 was set as the distance between the center-of-mass of heavy atoms in in the interface of RBDs and VHH-72, as from the last snapshot of the equilbiration simulations, and CV2 was defined as the contact value between residues in RBDs and VHH-72. The contact value for a conformation X sampled in the simulation is calculated as follows (BEST; HUMMER; EATON, 2013):

$$S(X) = \sum_{(i,j)\in(i,j)_0} \frac{1}{1 + exp(\beta(r_{ij} - \lambda \cdot r_{ij}^0))} \tag{5.1}$$

Where β and λ are defined as 50 nm^{-1} and 1.8, respectively. r_{ij} represents the center-of-mass distance between residues i and j, and r_{ij}^0 was defined was 5.5 Å.

The Gaussian potential presented a width of 0.25 and 1.0 for CV1 and CV2, respectively. The Gaussian height was defined as 0.5 kJ/mol, and bias factor of 32 was used. CVs and Gaussian parameters were reproduced from (CHEN et al., 2021). Prior to metadynamics simulations, 500 ns of unrestrained MD simulations were used to equilibrate the systems. Longer metadynamics simulations are being carried out until reaching convergence.



As it can be seen, the free energy surface as a function of the 2 CVs, shows the binding of VHH-72 to SARS-CoV-1 RBD is more stable than with SARS-CoV-2. In addition, the bound state has a more deep and spread energy well, in agreement with the τ-RAMD simulations. However, longer simulations are still required.

6 FINAL CONSIDERATIONS

The field of computational biophysics has proven to be dynamics and rapidly evolving, laying the foundations to prepare for the next pandemic. Most of the research presented in this thesis was conducted during the COVID-19 pandemic, including the work carried out in Germany. Despite the challenges posed by the pandemic, researchers from academia and the pharmaceutical industry worldwide collaborated extensively, pushing the boundaries of computational biophysics. This global collaboration led to the development of several innovative approaches that played a crucial role in the rapid development and deployment of COVID-19 vaccines. Particularly, digital research infrastructure, ranging from massively parallel simulations of huge system to hosting, storing and analyzing data with cloud infrastructure solutions, was essential in the fast response to the coronavirus pandemic. Vaccination campaigns started worldwide at an unprecedented speed, showcasing the remarkable achievements that can be accomplished through collective efforts in computational biophysics and related fields.

Within the human body, the replication of viruses relies on intricate interactions between viral proteins and host proteins. These interactions are influenced by various factors, including the host's immune response, the administration of antiviral compounds, and mutations that can occur in viral proteins. To gain a comprehensive understanding of the mechanisms underlying these interactions and their susceptibility to mutations, it is essential to investigate the structures and dynamics of the proteins involved. By delving into the molecular details of these interactions, we can uncover the underlying mechanisms and potentially develop strategies to disrupt or modulate them for therapeutic purposes. This knowledge is of utmost importance in our ongoing efforts to combat and control viral infections. These strategies include the development of diagnostic tools, therapeutics, vaccines, and structured surveillance programs.

Despite significant progress in experimental characterization of these systems, molecular simulations have emerged as a vital and complementary approach. Therefore, in this thesis we have developed, applied, and validated computational methods based on both molecular dynamics simulations and ML to address the structure and dynamics of virus and their targets. To achieve this goal, we have used a combination of different levels of spatial and temporal resolution, ranging from implicit solvation Brownian dynamics and Monte Carlo simulations, up to an atomistic description of big systems in explicit solvent with enhanced sampling.

The main contributions of this thesis are summarized below:

1. Development of a fast and accurate tool for predicting the ΔG of binding for protein-protein systems. This tool has demonstrated efficiency in predicting the absolute ΔG for computer-engineered proteins targeting CHIKV (Chapter 3) and SARS-CoV-2 (Chapter 4);

2. Application of machine learning to unravel the molecular interactions within the E6/E6AP/p53 complex, enabling estimation of the oncogenic potential of unclassified HPV types in HPV-related cancer surveillance and prevention strategies;

3. Development of a general pipeline for designing nanobodies using enhanced sampling simulations, metadynamics, and machine learning. This protocol was applied to design anti-SARS-CoV-2 nanobodies, resulting in proteins with predicted binding affinities comparable to experimental measurements. Binding assays with SARS-CoV-2 virus-like particles indicated that one of the designed nanobodies exhibited affinity similar to neutralizing antibodies, suggesting its potential as a neutralizing biotherapeutic, and demonstrating the potential of biopharmaceuticals generation using this approach. Additionally, we propose considering the full SARS-CoV-2 spike protein rather than only the receptor binding domain in computational protein design;

4. Validation and application of τ-RAMD to predict τ for protein-peptide and protein-protein complexes, respectively, expanding the domain of applications of this technique. Up until now, structure-based methods to compute protein-protein and protein-peptide complexes unbinding kinetics are overlooked mainly to the extensive sampling required. Using τ-RAMD multiple short simulations in the ns timescale are employed to bypass the use of long timescale simulations. These results have direct implication for protein engineering of virus-targeting proteins, peptide immunogenecity prediction for vaccine development, and in shedding light to relevant biomolecular phenomena.

In summary, our contributions primarily focus on the methodological aspects of computational protein design for biopharmaceuticals. We have also utilized molecular simulations and machine learning tools to gain a deep understanding of protein-protein and protein-peptide systems at the molecular level. As we enter the era of powerful exascale computing, we can expect to perform longer and more detailed simulations, using all-atom descriptions of complete systems more frequently. Additionally, the large amount of data being generated and shared opens the door for novel machine learning architectures. In our future plans, we aim to

develop new machine learning methods using larger data sets. We also intend to incorporate direct three-dimensional structural information using graph neural networks. In addition, we plan to test the designed molecules in both *in vitro* and *in vivo*, allowing us the continuous optimization of the designs. Furthermore, we aim to explore further the binding mechanism between the VHH-72 and the RBD of SARS-CoV-1/2 to propose a computational protocol to improve binding based on structure-based binding kinetics. So far, most of the methods rely on affinity maturation.

In addition to the findings presented in this thesis, we have dedicated our efforts to study other problems as follow:

1. The comprehension of the multi-mechanistic infectivity of SARS-CoV-2. Specifically, we have studied three mechanisms:
   - Immune evasion, which involves the electrostatic remodeling of the SARS-CoV-2 receptor binding domain without compromising hACE2 affinity, and insertion or deletion of bases on the N-terminal domain;
   - Higher rate of processing viral proteins by an increased affinity to the furin site, which is correlated with enhanced spike cleavage;
   - The role of heparin in the infection mechanism.
2. Participation in drug discovery projects;
3. Engineering of proteins targeting different viruses. Although some of these engineered proteins did not fold to achieve the desired function as expected, they have provided crucial insights in the design of novel proteins.

$$V(r^N) = \sum_{i\ \epsilon\ bonds} k_{bi}(l_i - l_i^0)^2 + \sum_{i\ \epsilon\ angles} k_{ai}(\theta_i - \theta_i^0)^2 + \sum_{i\ \epsilon\ torsions} \sum_{i\ \epsilon\ n} \frac{1}{2} V_i^n [1 + cos(n\omega_i - \gamma_i)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \quad (6.1)$$

REFERENCES

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; GOODFELLOW, I.; HARP, A.; IRVING, G.; ISARD, M.; JIA, Y.; JOZEFOWICZ, R.; KAISER, L.; KUDLUR, M.; LEVENBERG, J.; MANé, D.; MONGA, R.; MOORE, S.; MURRAY, D.; OLAH, C.; SCHUSTER, M.; SHLENS, J.; STEINER, B.; SUTSKEVER, I.; TALWAR, K.; TUCKER, P.; VANHOUCKE, V.; VASUDEVAN, V.; VIéGAS, F.; VINYALS, O.; WARDEN, P.; WATTENBERG, M.; WICKE, M.; YU, Y.; ZHENG, X. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Available at: <https://www.tensorflow.org/>.

ABEL, R.; YOUNG, T.; FARID, R.; BERNE, B. J.; FRIESNER, R. A. Role of the active-site solvent in the thermodynamics of factor xa ligand binding. *Journal of the American Chemical Society*, ACS Publications, v. 130, n. 9, p. 2817–2831, 2008.

ADCOCK, S. A.; MCCAMMON, J. A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, v. 106, n. 5, p. 1589–1615, 2006.

ADESHINA, Y. O.; DEEDS, E. J.; KARANICOLAS, J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 117, n. 31, p. 18477–18488, 2020.

AGUDELO, W. A.; GALINDO, J. F.; ORTIZ, M.; VILLAVECES, J. L.; DAZA, E. E.; PATARROYO, M. E. Variations in the electrostatic landscape of class ii human leukocyte antigen molecule induced by modifications in the myelin basic protein peptide: a theoretical approach. *PLoS One*, Public Library of Science San Francisco, USA, v. 4, n. 1, p. e4164, 2009.

AGUDELO, W. A.; GALINDO, J. F.; PATARROYO, M. E. Electrostatic potential as a tool to understand interactions between malaria vaccine candidate peptides and mhc ii molecules. *Biochemical and Biophysical Research Communications*, Elsevier, v. 410, n. 3, p. 410–415, 2011.

AHMAD, K.; RIZZI, A.; CAPELLI, R.; MANDELLI, D.; LYU, W.; CARLONI, P. Enhanced-sampling simulations for the estimation of ligand binding kinetics: Current status and perspective. *Frontiers in molecular biosciences*, Frontiers Media SA, v. 9, 2022.

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. T cells and mhc proteins. In: *Molecular Biology of the Cell. 4th edition*. [S.l.]: Garland Science, 2002.

ALDER, B. J.; WAINWRIGHT, T. E. Studies in molecular dynamics. i. general method. *The Journal of Chemical Physics*, American Institute of Physics, v. 31, n. 2, p. 459–466, 1959.

ALEKSANDROV, A.; ROUX, B.; JR, A. D. M. p k a calculations with the polarizable drude force field and poisson–boltzmann solvation model. *Journal of chemical theory and computation*, ACS Publications, v. 16, n. 7, p. 4655–4668, 2020.

ALFORD, R. F.; LEAVER-FAY, A.; JELIAZKOV, J. R.; O'MEARA, M. J.; DIMAIO, F. P.; PARK, H.; SHAPOVALOV, M. V.; RENFREW, P. D.; MULLIGAN, V. K.; KAPPEL, K. et

al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, ACS Publications, v. 13, n. 6, p. 3031–3048, 2017.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. *Journal of molecular biology*, Elsevier, v. 215, n. 3, p. 403–410, 1990.

ALTSCHUL, S. F.; MADDEN, T. L.; SCHÄFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, Oxford University Press, v. 25, n. 17, p. 3389–3402, 1997.

ALVIZO, O.; MAYO, S. L. Evaluating and optimizing computational protein design force fields using fixed composition-based negative design. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 105, n. 34, p. 12242–12247, 2008.

ANAND, N.; EGUCHI, R.; MATHEWS, I. I.; PEREZ, C. P.; DERRY, A.; ALTMAN, R. B.; HUANG, P.-S. Protein sequence design with a learned potential. *Nature communications*, Nature Publishing Group, v. 13, n. 1, p. 1–11, 2022.

ANANDHARAJ, A.; EKSHYYAN, O.; MOORE-MEDLIN, T.; MEHTA, V.; NATHAN, C.-A. O. Human papillomavirus (hpv) biomarkers in head and neck: Squamous cell carcinoma (hnscc). In: *Biomarkers in Disease: Methods, Discoveries and Applications: Biomarkers in Cancer*. [S.l.]: Springer Netherlands, 2015. p. 709–728.

ANDERSEN, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of chemical physics*, American Institute of Physics, v. 72, n. 4, p. 2384–2393, 1980.

ANNUNZIATA, C.; STELLATO, G.; GREGGI, S.; SANNA, V.; CURCIO, M. P.; LOSITO, S.; BOTTI, G.; BUONAGURO, L.; BUONAGURO, F. M.; TORNESELLO, M. L. Prevalence of "unclassified" hpv genotypes among women with abnormal cytology. *Infectious Agents and Cancer*, Springer, v. 13, p. 1–8, 2018.

ANSARI, T.; BRIMER, N.; POL, S. B. V. Peptide interactions stabilize and restructure human papillomavirus type 16 e6 to interact with p53. *Journal of virology*, Am Soc Microbiol, v. 86, n. 20, p. 11386–11391, 2012.

ANTONIOU, A. N.; POWIS, S. J.; ELLIOTT, T. Assembly and export of mhc class i peptide ligands. *Current opinion in immunology*, Elsevier, v. 15, n. 1, p. 75–81, 2003.

APRILE, F. A.; SORMANNI, P.; PODPOLNY, M.; CHHANGUR, S.; NEEDHAM, L.-M.; RUGGERI, F. S.; PERNI, M.; LIMBOCKER, R.; HELLER, G. T.; SNEIDERIS, T. et al. Rational design of a conformation-specific antibody for the quantification of aβ oligomers. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 117, n. 24, p. 13509–13518, 2020.

ARANTES, P. R.; SAHA, A.; PALERMO, G. *Fighting COVID-19 using molecular dynamics simulations*. [S.l.]: ACS Publications, 2020.

ARDILA, D.; KIRALY, A. P.; BHARADWAJ, S.; CHOI, B.; REICHER, J. J.; PENG, L.; TSE, D.; ETEMADI, M.; YE, W.; CORRADO, G. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, Nature Publishing Group US New York, v. 25, n. 6, p. 954–961, 2019.

ARMACOST, K. A.; RINIKER, S.; COURNIA, Z. *Exploring novel directions in free energy calculations*. [S.l.]: ACS Publications, 2020. 5283–5286 p.

ASHTAWY, H. M.; MAHAPATRA, N. R. A comparative assessment of predictive accuracies of conventional and machine learning scoring functions for protein-ligand binding affinity prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, v. 12, n. 2, p. 335–347, 2014.

ATTAF, M.; LEGUT, M.; COLE, D. K.; SEWELL, A. K. The t cell antigen receptor: the swiss army knife of the immune system. *Clinical & Experimental Immunology*, Oxford University Press, v. 181, n. 1, p. 1–18, 2015.

AUSTIN, H. P.; ALLEN, M. D.; DONOHOE, B. S.; RORRER, N. A.; KEARNS, F. L.; SILVEIRA, R. L.; POLLARD, B. C.; DOMINICK, G.; DUMAN, R.; OMARI, K. E. et al. Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 115, n. 19, p. E4350–E4357, 2018.

BABIN, V.; ROLAND, C.; DARDEN, T. A.; SAGUI, C. The free energy landscape of small peptides as obtained from metadynamics with umbrella sampling corrections. *The Journal of chemical physics*, American Institute of Physics, v. 125, n. 20, p. 204909, 2006.

BAEK, M.; DIMAIO, F.; ANISHCHENKO, I.; DAUPARAS, J.; OVCHINNIKOV, S.; LEE, G. R.; WANG, J.; CONG, Q.; KINCH, L. N.; SCHAEFFER, R. D. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, American Association for the Advancement of Science, v. 373, n. 6557, p. 871–876, 2021.

BAKER, D. An exciting but challenging road ahead for computational enzyme design. *Protein science: a publication of the Protein Society*, Wiley-Blackwell, v. 19, n. 10, p. 1817, 2010.

BAKER, N. A.; SEPT, D.; JOSEPH, S.; HOLST, M. J.; MCCAMMON, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 98, n. 18, p. 10037–10041, 2001.

BARAN, D.; PSZOLLA, M. G.; LAPIDOTH, G. D.; NORN, C.; DYM, O.; UNGER, T.; ALBECK, S.; TYKA, M. D.; FLEISHMAN, S. J. Principles for computational design of binding antibodies. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 114, n. 41, p. 10900–10905, 2017.

BARDUCCI, A.; BUSSI, G.; PARRINELLO, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*, APS, v. 100, n. 2, p. 020603, 2008.

BARILLARI, C.; TAYLOR, J.; VINER, R.; ESSEX, J. W. Classification of water molecules in protein binding sites. *Journal of the American Chemical Society*, ACS Publications, v. 129, n. 9, p. 2577–2587, 2007.

BARLOW, K. A.; CONCHUIR, S. O.; THOMPSON, S.; SURESH, P.; LUCAS, J. E.; HEINONEN, M.; KORTEMME, T. Flex ddg: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *The Journal of Physical Chemistry B*, ACS Publications, v. 122, n. 21, p. 5389–5399, 2018.

BARROS, E. P.; RIES, B.; BÖSELT, L.; CHAMPION, C.; RINIKER, S. Recent developments in multiscale free energy simulations. *Current Opinion in Structural Biology*, Elsevier, v. 72, p. 55–62, 2022.

BARTH, E.; SCHLICK, T. Overcoming stability limitations in biomolecular dynamics. i. combining force splitting via extrapolation with langevin dynamics in ln. *The Journal of chemical physics*, American Institute of Physics, v. 109, n. 5, p. 1617–1632, 1998.

BARTHELEMY, P. A.; RAAB, H.; APPLETON, B. A.; BOND, C. J.; WU, P.; WIESMANN, C.; SIDHU, S. S. Comprehensive analysis of the factors contributing to the stability and solubility of autonomous human vh domains. *Journal of Biological Chemistry*, ASBMB, v. 283, n. 6, p. 3639–3654, 2008.

BAUER, M. S.; GRUBER, S.; HAUSCH, A.; GOMES, P. S.; MILLES, L. F.; NICOLAUS, T.; SCHENDEL, L. C.; NAVAJAS, P. L.; PROCKO, E.; LIETHA, D. et al. A tethered ligand assay to probe sars-cov-2: Ace2 interactions. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 119, n. 14, p. e2114397119, 2022.

BAYLY, C. I.; CIEPLAK, P.; CORNELL, W.; KOLLMAN, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry*, ACS Publications, v. 97, n. 40, p. 10269–10280, 1993.

BEACH, F. H. A gradient method for function minimization. *Journal of the Society for Industrial and Applied Mathematics*, SIAM, v. 1, n. 2, p. 149–158, 1953.

BEGHEIN, E.; GETTEMANS, J. Nanobody technology: a versatile toolkit for microscopic imaging, protein–protein interaction analysis, and protein function exploration. *Frontiers in immunology*, Frontiers Media SA, v. 8, p. 771, 2017.

BENDER, B. J.; III, A. C.; DURAN, A. M.; FINN, J. A.; FU, D.; LOKITS, A. D.; MUELLER, B. K.; SANGHA, A. K.; SAUER, M. F.; SEVY, A. M. et al. Protocols for molecular modeling with rosetta3 and rosettascripts. *Biochemistry*, ACS Publications, v. 55, n. 34, p. 4748–4763, 2016.

BENKOULOUCHE, M.; IMEDDOURENE, A. B.; BAREL, L.-A.; LEFEBVRE, D.; FANUEL, M.; ROGNIAUX, H.; ROPARTZ, D.; BARBE, S.; GUIEYSSE, D.; MULARD, L. A. et al. Computer-aided engineering of a branching sucrase for the glucodiversification of a tetrasaccharide precursor of s. flexneri antigenic oligosaccharides. *Scientific reports*, Nature Publishing Group, v. 11, n. 1, p. 1–14, 2021.

BENNETT, N. R.; COVENTRY, B.; GORESHNIK, I.; HUANG, B.; ALLEN, A.; VAFEADOS, D.; PENG, Y. P.; DAUPARAS, J.; BAEK, M.; STEWART, L. et al. Improving de novo protein binder design with deep learning. *Nature Communications*, Nature Publishing Group UK London, v. 14, n. 1, p. 2625, 2023.

BERENDSEN, H. J. C.; POSTMA, J. P. M.; GUNSTEREN, W. F. van; DINOLA, A.; HAAK, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, v. 81, n. 8, p. 3684–3690, 1984.

BERGER, B.-T.; AMARAL, M.; KOKH, D. B.; NUNES-ALVES, A.; MUSIL, D.; HEINRICH, T.; SCHRÖDER, M.; NEIL, R.; WANG, J.; NAVRATILOVA, I. et al. Structure-kinetic relationship reveals the mechanism of selectivity of fak inhibitors over pyk2. *Cell Chemical Biology*, Elsevier, v. 28, n. 5, p. 686–698, 2021.

BERNARDI, R. C.; DURNER, E.; SCHOELER, C.; MALINOWSKA, K. H.; CARVALHO, B. G.; BAYER, E. A.; LUTHEY-SCHULTEN, Z.; GAUB, H. E.; NASH, M. A. Mechanisms of nanonewton mechanostability in a protein complex revealed by molecular dynamics simulations and single-molecule force spectroscopy. *Journal of the American Chemical Society*, ACS Publications, v. 141, n. 37, p. 14752–14763, 2019.

BERNARDI, R. C.; MELO, M. C.; SCHULTEN, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, Elsevier, v. 1850, n. 5, p. 872–877, 2015.

BERNE, B. J.; BORKOVEC, M.; STRAUB, J. E. Classical and modern methods in reaction rate theory. *The Journal of Physical Chemistry*, ACS Publications, v. 92, n. 13, p. 3711–3725, 1988.

BERNETTI, M.; MASETTI, M.; RECANATINI, M.; AMARO, R. E.; CAVALLI, A. An integrated markov state model and path metadynamics approach to characterize drug binding processes. *Journal of chemical theory and computation*, ACS Publications, v. 15, n. 10, p. 5689–5702, 2019.

BERNETTI, M.; MASETTI, M.; ROCCHIA, W.; CAVALLI, A. Kinetics of drug binding and residence time. *Annual review of physical chemistry*, Annual Reviews, v. 70, p. 143–171, 2019.

BEST, R. B.; HUMMER, G.; EATON, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 110, n. 44, p. 17874–17879, 2013.

BINDER, A.; BOCKMAYR, M.; HÄGELE, M.; WIENERT, S.; HEIM, D.; HELLWEG, K.; ISHII, M.; STENZINGER, A.; HOCKE, A.; DENKERT, C. et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence*, Nature Publishing Group UK London, v. 3, n. 4, p. 355–366, 2021.

BLAHA, D. T.; ANDERSON, S. D.; YOAKUM, D. M.; HAGER, M. V.; ZHA, Y.; GAJEWSKI, T. F.; KRANZ, D. M. High-throughput stability screening of neoantigen/hla complexes improves immunogenicity predictionsneoantigen/hla complexes. *Cancer immunology research*, AACR, v. 7, n. 1, p. 50–61, 2019.

BLOODWORTH, N.; BARBARO, N. R.; MORETTI, R.; HARRISON, D. G.; MEILER, J. Rosetta flexpepdock to predict peptide-mhc binding: An approach for non-canonical amino acids. *Plos one*, Public Library of Science San Francisco, CA USA, v. 17, n. 12, p. e0275759, 2022.

BORDNER, A.; ABAGYAN, R. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 57, n. 2, p. 400–413, 2004.

BORN, M. Volumen und hydratationswärme der ionen. *Zeitschrift für physik*, Springer-Verlag Berlin/Heidelberg, v. 1, n. 1, p. 45–48, 1920.

BORN, M.; HEISENBERG, W. Zur quantentheorie der molekeln. In: *Original Scientific Papers Wissenschaftliche Originalarbeiten*. [S.l.]: Springer, 1985. p. 216–246.

BOYOGLU-BARNUM, S.; ELLIS, D.; GILLESPIE, R. A.; HUTCHINSON, G. B.; PARK, Y.-J.; MOIN, S. M.; ACTON, O. J.; RAVICHANDRAN, R.; MURPHY, M.; PETTIE, D. et al. Quadrivalent influenza nanoparticle vaccines induce broad protection. *Nature*, Nature Publishing Group, v. 592, n. 7855, p. 623–628, 2021.

BROOKS, B. R.; BRUCCOLERI, R. E.; OLAFSON, B. D.; STATES, D. J.; SWAMINATHAN, S. a.; KARPLUS, M. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, Wiley Online Library, v. 4, n. 2, p. 187–217, 1983.

BROOKS, B. R.; JANEŽIČ, D.; KARPLUS, M. Harmonic analysis of large systems. i. methodology. *Journal of computational chemistry*, Wiley Online Library, v. 16, n. 12, p. 1522–1542, 1995.

BRUCE, N. J.; GANOTRA, G. K.; KOKH, D. B.; SADIQ, S. K.; WADE, R. C. New approaches for computing ligand–receptor binding kinetics. *Current opinion in structural biology*, Elsevier, v. 49, p. 1–10, 2018.

BURD, E. M. Human papillomavirus and cervical cancer. *Clinical microbiology reviews*, Am Soc Microbiol, v. 16, n. 1, p. 1–17, 2003.

BUSSI, G.; LAIO, A. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, Nature Publishing Group, v. 2, n. 4, p. 200–212, 2020.

BUSSI, G.; LAIO, A.; PARRINELLO, M. Equilibrium free energies from nonequilibrium metadynamics. *Physical review letters*, APS, v. 96, n. 9, p. 090601, 2006.

BZHALAVA, D.; EKLUND, C.; DILLNER, J. International standardization and classification of human papillomavirus types. *Virology*, Elsevier, v. 476, p. 341–344, 2015.

CALLAWAY, E. 'it will change everything': Deepmind's ai makes gigantic leap in solving protein structures. *Nature*, Nature Publishing Group, v. 588, n. 7837, p. 203–205, 2020.

CALLAWAY, E. Revolutionary cryo-em is taking over structural biology. *Nature*, Nature Publishing Group, v. 578, n. 7794, p. 201–202, 2020.

CAO, L.; GORESHNIK, I.; COVENTRY, B.; CASE, J. B.; MILLER, L.; KOZODOY, L.; CHEN, R. E.; CARTER, L.; WALLS, A. C.; PARK, Y.-J. et al. De novo design of picomolar sars-cov-2 miniprotein inhibitors. *Science*, American Association for the Advancement of Science, v. 370, n. 6515, p. 426–431, 2020.

CASALINO, L.; GAIEB, Z.; GOLDSMITH, J. A.; HJORTH, C. K.; DOMMER, A. C.; HARBISON, A. M.; FOGARTY, C. A.; BARROS, E. P.; TAYLOR, B. C.; MCLELLAN, J. S. et al. Beyond shielding: the roles of glycans in the sars-cov-2 spike protein. *ACS central science*, ACS Publications, v. 6, n. 10, p. 1722–1734, 2020.

CASE, D. A.; III, T. E. C.; DARDEN, T.; GOHLKE, H.; LUO, R.; JR, K. M. M.; ONUFRIEV, A.; SIMMERLING, C.; WANG, B.; WOODS, R. J. The amber biomolecular simulation programs. *Journal of computational chemistry*, Wiley Online Library, v. 26, n. 16, p. 1668–1688, 2005.

CERIOTTI, M.; CLEMENTI, C.; LILIENFELD, O. Anatole von. *Introduction: Machine Learning at the Atomic Scale*. [S.l.]: ACS Publications, 2021. 9719–9721 p.

CHAN, J. F.-W.; YUAN, S.; KOK, K.-H.; TO, K. K.-W.; CHU, H.; YANG, J.; XING, F.; LIU, J.; YIP, C. C.-Y.; POON, R. W.-S. et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The lancet*, Elsevier, v. 395, n. 10223, p. 514–523, 2020.

CHAN, P. K.; ZHANG, C.; PARK, J.-S.; SMITH-MCCUNE, K. K.; PALEFSKY, J. M.; GIOVANNELLI, L.; COUTLÉE, F.; HIBBITTS, S.; KONNO, R.; SETTHEETHAM-ISHIDA, W. et al. Geographical distribution and oncogenic risk association of human papillomavirus type 58 e6 and e7 sequence variations. *International journal of cancer*, Wiley Online Library, v. 132, n. 11, p. 2528–2536, 2013.

CHANG, C.-E.; SHEN, T.; TRYLSKA, J.; TOZZINI, V.; MCCAMMON, J. A. Gated binding of ligands to hiv-1 protease: Brownian dynamics simulations in a coarse-grained model. *Biophysical journal*, Elsevier, v. 90, n. 11, p. 3880–3885, 2006.

CHAUDHURY, S.; BERRONDO, M.; WEITZNER, B. D.; MUTHU, P.; BERGMAN, H.; GRAY, J. J. Benchmarking and analysis of protein docking performance in rosetta v3. 2. *PloS one*, Public Library of Science San Francisco, USA, v. 6, n. 8, p. e22477, 2011.

CHEN, D.; OEZGUEN, N.; URVIL, P.; FERGUSON, C.; DANN, S. M.; SAVIDGE, T. C. Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Science advances*, American Association for the Advancement of Science, v. 2, n. 3, p. e1501240, 2016.

CHEN, H.; KANG, Y.; DUAN, M.; HOU, T. Regulation mechanism for the binding between the sars-cov-2 spike protein and host angiotensin-converting enzyme ii. *The Journal of Physical Chemistry Letters*, ACS Publications, v. 12, n. 27, p. 6252–6261, 2021.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794.

CHEVALIER, A.; SILVA, D.-A.; ROCKLIN, G. J.; HICKS, D. R.; VERGARA, R.; MURAPA, P.; BERNARD, S. M.; ZHANG, L.; LAM, K.-H.; YAO, G. et al. Massively parallel de novo protein design for targeted therapeutics. *Nature*, Nature Publishing Group, v. 550, n. 7674, p. 74–79, 2017.

CHILDERS, M. C.; DAGGETT, V. Insights from molecular dynamics simulations for computational protein design. *Molecular systems design & engineering*, Royal Society of Chemistry, v. 2, n. 1, p. 9–33, 2017.

CHRIST, C. D.; MARK, A. E.; GUNSTEREN, W. F. V. Basic ingredients of free energy calculations: a review. *Journal of computational chemistry*, Wiley Online Library, v. 31, n. 8, p. 1569–1582, 2010.

CISNEROS, G. A.; KARTTUNEN, M.; REN, P.; SAGUI, C. Classical electrostatics for biomolecular simulations. *Chemical reviews*, ACS Publications, v. 114, n. 1, p. 779–814, 2014.

CLAESEN, M.; MOOR, B. D. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015.

CLARK, A. J.; NEGRON, C.; HAUSER, K.; SUN, M.; WANG, L.; ABEL, R.; FRIESNER, R. A. Relative binding affinity prediction of charge-changing sequence mutations with fep in protein–protein interfaces. *Journal of molecular biology*, Elsevier, v. 431, n. 7, p. 1481–1493, 2019.

CLAUSEN, T. M.; SANDOVAL, D. R.; SPLIID, C. B.; PIHL, J.; PERRETT, H. R.; PAINTER, C. D.; NARAYANAN, A.; MAJOWICZ, S. A.; KWONG, E. M.; MCVICAR, R. N. et al. Sars-cov-2 infection depends on cellular heparan sulfate and ace2. *Cell*, Elsevier, v. 183, n. 4, p. 1043–1057, 2020.

CLIFFORD, G.; FRANCESCHI, S.; DIAZ, M.; MUÑOZ, N.; VILLA, L. L. Hpv type-distribution in women with and without cervical neoplastic diseases. *Vaccine*, Elsevier, v. 24, p. S26–S34, 2006.

CONRADY, M. C.; SUAREZ, I.; GOGL, G.; FRECOT, D. I.; BONHOURE, A.; KOSTMANN, C.; COUSIDO-SIAH, A.; MITSCHLER, A.; LIM, J.; MASSON, M. et al. Structure of high-risk papillomavirus 31 e6 oncogenic protein and characterization of e6/e6ap/p53 complex formation. *Journal of Virology*, Am Soc Microbiol, v. 95, n. 2, p. e00730–20, 2020.

CONWAY, P.; TYKA, M. D.; DIMAIO, F.; KONERDING, D. E.; BAKER, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein science*, Wiley Online Library, v. 23, n. 1, p. 47–55, 2014.

COPELAND, R. A. The drug–target residence time model: a 10-yexar retrospective. *Nature Reviews Drug Discovery*, Nature Publishing Group UK London, v. 15, n. 2, p. 87–95, 2016.

COPELAND, R. A.; POMPLIANO, D. L.; MEEK, T. D. Drug–target residence time and its implications for lead optimization. *Nature reviews Drug discovery*, Nature Publishing Group, v. 5, n. 9, p. 730–739, 2006.

CORNELL, W. D.; CIEPLAK, P.; BAYLY, C. I.; GOULD, I. R.; MERZ, K. M.; FERGUSON, D. M.; SPELLMEYER, D. C.; FOX, T.; CALDWELL, J. W.; KOLLMAN, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, ACS Publications, v. 117, n. 19, p. 5179–5197, 1995.

CORNELL, W. D.; CIEPLAK, P.; BAYLY, C. I.; KOLLMAN, P. A. Application of resp charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society*, ACS Publications, v. 115, n. 21, p. 9620–9631, 2002.

CORREIA, B. E.; BATES, J. T.; LOOMIS, R. J.; BANEYX, G.; CARRICO, C.; JARDINE, J. G.; RUPERT, P.; CORRENTI, C.; KALYUZHNIY, O.; VITTAL, V. et al. Proof of principle for epitope-focused vaccine design. *Nature*, Nature Publishing Group UK London, v. 507, n. 7491, p. 201–206, 2014.

COVA, T. F.; PAIS, A. A. Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Frontiers in chemistry*, Frontiers Media SA, v. 7, p. 809, 2019.

COX, S.; WILLIAMS, D. Representation of the molecular electrostatic potential by a net atomic charge model. *Journal of Computational chemistry*, Wiley Online Library, v. 2, n. 3, p. 304–323, 1981.

CUNHA, K. C.; RUSU, V. H.; VIANA, I. F.; MARQUES, E. T.; DHALIA, R.; LINS, R. D. Assessing protein conformational sampling and structural stability via de novo design and molecular dynamics simulations. *Biopolymers*, Wiley Online Library, v. 103, n. 6, p. 351–361, 2015.

CUTCLIFFE, J. W.; HELLMANN, E.; HEYL, A.; RASHOTTE, A. M. Crfs form protein–protein interactions with each other and with members of the cytokinin signalling pathway in arabidopsis via the crf domain. *Journal of Experimental Botany*, Oxford University Press, v. 62, n. 14, p. 4995–5002, 2011.

DARDEN, T.; YORK, D.; PEDERSEN, L. Particle mesh ewald: An n · log(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, v. 98, n. 12, p. 10089–10092, 1993. Available at: <https://doi.org/10.1063/1.464397>.

DARDEN, T.; YORK, D.; PEDERSEN, L. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, American Institute of Physics, v. 98, n. 12, p. 10089–10092, 1993.

DAS, R.; BAKER, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, v. 77, p. 363–382, 2008.

DAS, S.; CHAKRABARTI, S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Scientific reports*, Nature Publishing Group, v. 11, n. 1, p. 1–12, 2021.

DAUPARAS, J.; ANISHCHENKO, I.; BENNETT, N.; BAI, H.; RAGOTTE, R. J.; MILLES, L. F.; WICKY, B. I.; COURBET, A.; HAAS, R. J. de; BETHEL, N. et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, American Association for the Advancement of Science, v. 378, n. 6615, p. 49–56, 2022.

DAVID, A.; RAZALI, R.; WASS, M. N.; STERNBERG, M. J. Protein–protein interaction sites are hot spots for disease-associated nonsynonymous snps. *Human mutation*, Wiley Online Library, v. 33, n. 2, p. 359–363, 2012.

DAVIS, I. W.; III, W. B. A.; RICHARDSON, D. C.; RICHARDSON, J. S. The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure*, v. 14, n. 2, p. 265–274, 2006.

DECHERCHI, S.; CAVALLI, A. Thermodynamics and kinetics of drug-target binding by molecular simulation. *Chemical Reviews*, ACS Publications, v. 120, n. 23, p. 12788–12833, 2020.

DILL, K. A.; MACCALLUM, J. L. The protein-folding problem, 50 years on. *science*, American Association for the Advancement of Science, v. 338, n. 6110, p. 1042–1046, 2012.

DING, W.; NAKAI, K.; GONG, H. Protein design via deep learning. *Briefings in bioinformatics*, Oxford University Press, v. 23, n. 3, p. bbac102, 2022.

DOLINSKY, T. J.; CZODROWSKI, P.; LI, H.; NIELSEN, J. E.; JENSEN, J. H.; KLEBE, G. Pdb2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic acids research*, Oxford University Press, v. 35, n. suppl_2, p. W522–W525, 2007.

DOMMER, A.; CASALINO, L.; KEARNS, F.; ROSENFELD, M.; WAUER, N.; AHN, S.-H.; RUSSO, J.; OLIVEIRA, S.; MORRIS, C.; BOGETTI, A. et al. # covidisairborne: Ai-enabled multiscale computational microscopy of delta sars-cov-2 in a respiratory aerosol. *bioRxiv*, Cold Spring Harbor Laboratory Preprints, 2021.

DOORBAR, J. The papillomavirus life cycle. *Journal of clinical virology*, Elsevier, v. 32, p. 7–15, 2005.

DOORSLAER, K. V.; TAN, Q.; XIRASAGAR, S.; BANDARU, S.; GOPALAN, V.; MOHAMOUD, Y.; HUYEN, Y.; MCBRIDE, A. A. The papillomavirus episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic acids research*, Oxford University Press, v. 41, n. D1, p. D571–D578, 2012.

DORMESHKIN, D.; SHAPIRA, M.; DUBOVIK, S.; KAVALEUSKI, A.; KATSIN, M.; MIGAS, A.; MELESHKO, A.; SEMYONOV, S. Isolation of an escape-resistant sars-cov-2 neutralizing nanobody from a novel synthetic nanobody library. *Frontiers in Immunology*, Frontiers, p. 5421, 2022.

DREWS, C. M.; BRIMER, N.; POL, S. B. V. Multiple regions of e6ap (ube3a) contribute to interaction with papillomavirus e6 proteins and the activation of ubiquitin ligase activity. *PLoS Pathogens*, Public Library of Science San Francisco, CA USA, v. 16, n. 1, p. e1008295, 2020.

DU, Z.; SU, H.; WANG, W.; YE, L.; WEI, H.; PENG, Z.; ANISHCHENKO, I.; BAKER, D.; YANG, J. The trrosetta server for fast and accurate protein structure prediction. *Nature protocols*, Nature Publishing Group UK London, v. 16, n. 12, p. 5634–5651, 2021.

DURRANT, J. D.; FRIEDMAN, A. J.; ROGERS, K. E.; MCCAMMON, J. A. Comparing neural-network scoring functions and the state of the art: applications to common library screening. *Journal of chemical information and modeling*, ACS Publications, v. 53, n. 7, p. 1726–1735, 2013.

DURRANT, J. D.; MCCAMMON, J. A. Molecular dynamics simulations and drug discovery. *BMC biology*, BioMed Central, v. 9, n. 1, p. 1–9, 2011.

DYSON, N.; HOWLEY, P. M.; MÜNGER, K.; HARLOW, E. The human papilloma virus-16 e7 oncoprotein is able to bind to the retinoblastoma gene product. *Science*, American Association for the Advancement of Science, v. 243, n. 4893, p. 934–937, 1989.

EGAWA, N.; DOORBAR, J. The low-risk papillomaviruses. *Virus research*, Elsevier, v. 231, p. 119–127, 2017.

ELCOCK, A. H.; GABDOULLINE, R. R.; WADE, R. C.; MCCAMMON, J. A. Computer simulation of protein-protein association kinetics: acetylcholinesterase-fasciculin. *Journal of molecular biology*, Elsevier, v. 291, n. 1, p. 149–162, 1999.

ELCOCK, A. H.; SEPT, D.; MCCAMMON, J. A. *Computer simulation of protein- protein interactions*. [S.l.]: ACS Publications, 2001. 1504–1518 p.

ENSING, B.; KLEIN, M. L. Perspective on the reactions between f–and ch3ch2f: The free energy landscape of the e2 and sn2 reaction channels. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 102, n. 19, p. 6755–6759, 2005.

ERIJMAN, A.; ROSENTHAL, E.; SHIFMAN, J. M. How structure defines affinity in protein-protein interactions. *PLOS one*, Public Library of Science San Francisco, USA, v. 9, n. 10, p. e110085, 2014.

ERMAK, D. L.; MCCAMMON, J. A. Brownian dynamics with hydrodynamic interactions. *The Journal of chemical physics*, American Institute of Physics, v. 69, n. 4, p. 1352–1360, 1978.

FALK, K.; RÖTZSCHKE, O.; STEVANOVIÉ, S.; JUNG, G.; RAMMENSEE, H.-G. Allele-specific motifs revealed by sequencing of self-peptides eluted from mhc molecules. *Nature*, Nature Publishing Group UK London, v. 351, n. 6324, p. 290–296, 1991.

FAVRE, M. Structural polypeptides of rabbit, bovine, and human papillomaviruses. *Journal of virology*, Am Soc Microbiol, v. 15, n. 5, p. 1239–1247, 1975.

FENG, J.; JESTER, B. W.; TINBERG, C. E.; MANDELL, D. J.; ANTUNES, M. S.; CHARI, R.; MOREY, K. J.; RIOS, X.; MEDFORD, J. I.; CHURCH, G. M. et al. A general strategy to construct small molecule biosensors in eukaryotes. *Elife*, eLife Sciences Publications Limited, v. 4, p. e10606, 2015.

FERNÁNDEZ-QUINTERO, M. L.; KRAML, J.; GEORGES, G.; LIEDL, K. R. Cdr-h3 loop ensemble in solution–conformational selection upon antibody binding. In: TAYLOR & FRANCIS. *MAbs*. [S.l.], 2019. v. 11, n. 6, p. 1077–1088.

FERNÁNDEZ-QUINTERO, M. L.; KROELL, K. B.; HOFER, F.; RICCABONA, J. R.; LIEDL, K. R. Mutation of framework residue h71 results in different antibody paratope states in solution. *Frontiers in Immunology*, Frontiers, p. 243, 2021.

FERNÁNDEZ-QUINTERO, M. L.; LOEFFLER, J. R.; KRAML, J.; KAHLER, U.; KAMENIK, A. S.; LIEDL, K. R. Characterizing the diversity of the cdr-h3 loop conformational ensembles in relationship to antibody binding properties. *Frontiers in immunology*, Frontiers Media SA, v. 9, p. 3065, 2019.

FERRAZ, M. V.; MOREIRA, E. G.; COÊLHO, D. F.; WALLAU, G. L.; LINS, R. D. Immune evasion of sars-cov-2 variants of concern is driven by low affinity to neutralizing antibodies. *Chemical Communications*, Royal Society of Chemistry, v. 57, n. 49, p. 6094–6097, 2021.

FERRAZ, M. V.; NETO, J. C.; LINS, R. D.; TEIXEIRA, E. S. An artificial neural network model to predict structure-based protein–protein free energy of binding from rosetta-calculated properties. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry, v. 25, n. 10, p. 7257–7267, 2023.

FERRAZ, M. V. F.; ADAN, W. C. d. S.; LINS, R. D. Unraveling the role of nanobodies tetrad on their folding and stability assisted by machine and deep learning algorithms. In: SPRINGER. *Brazilian Symposium on Bioinformatics*. [S.l.], 2020. p. 93–104.

FERRAZ, M. V. F.; VIANA, I. F. T.; COÊLHO, D. F.; CRUZ, C. H. B. da; LIMA, M. de A.; ARAGÃO, M. A. de L.; LINS, R. D. Association strength of e6 to e6ap/p53 complex correlates with hpv-mediated oncogenesis risk. *Biopolymers*, Wiley Online Library, v. 113, n. 10, p. e23524, 2022.

FIORIN, G.; KLEIN, M. L.; HÉNIN, J. Using collective variables to drive molecular dynamics simulations. *Molecular Physics*, Taylor & Francis, v. 111, n. 22-23, p. 3345–3362, 2013.

FLEISHMAN, S. J.; LEAVER-FAY, A.; CORN, J. E.; STRAUCH, E.-M.; KHARE, S. D.; KOGA, N.; ASHWORTH, J.; MURPHY, P.; RICHTER, F.; LEMMON, G. et al. Rosettascripts: a scripting language interface to the rosetta macromolecular modeling suite. *PloS one*, Public Library of Science San Francisco, USA, v. 6, n. 6, p. e20161, 2011.

FLEISHMAN, S. J.; WHITEHEAD, T. A.; STRAUCH, E.-M.; CORN, J. E.; QIN, S.; ZHOU, H.-X.; MITCHELL, J. C.; DEMERDASH, O. N.; TAKEDA-SHITAKA, M.; TERASHI, G. et al. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *Journal of molecular biology*, Elsevier, v. 414, n. 2, p. 289–302, 2011.

FRENZ, B.; RÄMISCH, S.; BORST, A. J.; WALLS, A. C.; ADOLF-BRYFOGLE, J.; SCHIEF, W. R.; VEESLER, D.; DIMAIO, F. Automatically fixing errors in glycoprotein structures with rosetta. *Structure*, Elsevier, v. 27, n. 1, p. 134–139, 2019.

FRIEDLAND, G. D.; LINARES, A. J.; SMITH, C. A.; KORTEMME, T. A simple model of backbone flexibility improves modeling of side-chain conformational variability. *Journal of molecular biology*, Elsevier, v. 380, n. 4, p. 757–774, 2008.

FU, L.; DOORSLAER, K. V.; CHEN, Z.; RISTRIANI, T.; MASSON, M.; TRAVE, G.; BURK, R. D. Degradation of p53 by human alphapapillomavirus e6 proteins shows a stronger correlation with phylogeny than oncogenicity. *PloS one*, Public Library of Science San Francisco, USA, v. 5, n. 9, p. e12816, 2010.

GABDOULLINE, R. R.; WADE, R. C. Effective charges for macromolecules in solvent. *The Journal of Physical Chemistry*, ACS Publications, v. 100, n. 9, p. 3868–3878, 1996.

GABDOULLINE, R. R.; WADE, R. C. Simulation of the diffusional association of barnase and barstar. *Biophysical journal*, Elsevier, v. 72, n. 5, p. 1917–1929, 1997.

GABDOULLINE, R. R.; WADE, R. C. Brownian dynamics simulation of protein–protein diffusional encounter. *Methods*, Elsevier, v. 14, n. 3, p. 329–341, 1998.

GABDOULLINE, R. R.; WADE, R. C. Protein-protein association: investigation of factors influencing association rates by brownian dynamics simulations. *Journal of molecular biology*, Elsevier, v. 306, n. 5, p. 1139–1155, 2001.

GABDOULLINE, R. R.; WADE, R. C. Biomolecular diffusional association. *Current opinion in structural biology*, Elsevier, v. 12, n. 2, p. 204–213, 2002.

GAGE, J.; MEYERS, C.; WETTSTEIN, F. The e7 proteins of the nononcogenic human papillomavirus type 6b (hpv-6b) and of the oncogenic hpv-16 differ in retinoblastoma protein binding and other properties. *Journal of virology*, Am Soc Microbiol, v. 64, n. 2, p. 723–730, 1990.

GAINZA-CIRAUQUI, P.; CORREIA, B. E. Computational protein design—the next generation tool to expand synthetic biology applications. *Current opinion in biotechnology*, Elsevier, v. 52, p. 145–152, 2018.

GALLOWAY, D. A.; LAIMINS, L. A. Human papillomaviruses: shared and distinct pathways for pathogenesis. *Current opinion in virology*, Elsevier, v. 14, p. 87–92, 2015.

GARSTKA, M. A.; FISH, A.; CELIE, P. H.; JOOSTEN, R. P.; JANSSEN, G. M.; BERLIN, I.; HOPPES, R.; STADNIK, M.; JANSSEN, L.; OVAA, H. et al. The first step of peptide selection in antigen presentation by mhc class i molecules. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 112, n. 5, p. 1505–1510, 2015.

GEMAN, S.; BIENENSTOCK, E.; DOURSAT, R. Neural networks and the bias/variance dilemma. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 4, n. 1, p. 1–58, 1992.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: " O'Reilly Media, Inc.", 2022.

GERVASIO, F. L.; LAIO, A.; PARRINELLO, M. Flexible docking in solution using metadynamics. *Journal of the American Chemical Society*, ACS Publications, v. 127, n. 8, p. 2600–2607, 2005.

GHITTONI, R.; ACCARDI, R.; HASAN, U.; GHEIT, T.; SYLLA, B.; TOMMASINO, M. The biological properties of e6 and e7 oncoproteins from human papillomaviruses. *Virus genes*, Springer, v. 40, p. 1–13, 2010.

GIANNAKOULIAS, S.; SHRINGARI, S. R.; FERRIE, J. J.; PETERSSON, E. J. Biomolecular simulation based machine learning models accurately predict sites of tolerability to the unnatural amino acid acridonylalanine. *Scientific reports*, Nature Publishing Group, v. 11, n. 1, p. 1–12, 2021.

GILSON, M. K.; DAVIS, M. E.; LUTY, B. A.; MCCAMMON, J. A. Computation of electrostatic forces on solvated molecules using the poisson-boltzmann equation. *The Journal of Physical Chemistry*, ACS Publications, v. 97, n. 14, p. 3591–3600, 1993.

GILSON, M. K.; GIVEN, J. A.; BUSH, B. L.; MCCAMMON, J. A. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical journal*, Elsevier, v. 72, n. 3, p. 1047–1069, 1997.

GILSON, M. K.; SHARP, K. A.; HONIG, B. H. Calculating the electrostatic potential of molecules in solution: method and error assessment. *Journal of computational chemistry*, Wiley Online Library, v. 9, n. 4, p. 327–335, 1988.

GŌ, N.; SCHERAGA, H. A. Analysis of the contribution of internal vibrations to the statistical weights of equilibrium conformations of macromolecules. *The Journal of Chemical Physics*, American Institute of Physics, v. 51, n. 11, p. 4751–4767, 1969.

GODOY, M. O. d.; NOGUEIRA, V. H. R.; FREIRE, M. C. L. C.; SOUZA, G. E. d.; FASSIO, A. V.; FERRAZ, M.; OLIVA, G.; LINS, R. D.; GUIDO, R. V. C. Integration of virtual screening and experimental method to identify new mpro of sars-cov-2 inhibitors. *Anais eletrônicos*, 2022.

GOLDENZWEIG, A.; GOLDSMITH, M.; HILL, S. E.; GERTMAN, O.; LAURINO, P.; ASHANI, Y.; DYM, O.; UNGER, T.; ALBECK, S.; PRILUSKY, J. et al. Automated structure-and sequence-based design of proteins for high bacterial expression and stability. *Molecular cell*, Elsevier, v. 63, n. 2, p. 337–346, 2016.

GOSSEN, J.; ALBANI, S.; HANKE, A.; JOSEPH, B. P.; BERGH, C.; KUZIKOV, M.; COSTANZI, E.; MANELFI, C.; STORICI, P.; GRIBBON, P. et al. A blueprint for high affinity sars-cov-2 mpro inhibitors from activity-based compound library screening guided by analysis of protein dynamics. *ACS pharmacology & translational science*, ACS Publications, v. 4, n. 3, p. 1079–1095, 2021.

GRAHAM, S. V. Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies. *Future microbiology*, Future Medicine, v. 5, n. 10, p. 1493–1506, 2010.

GRAY, J. J.; MOUGHON, S.; WANG, C.; SCHUELER-FURMAN, O.; KUHLMAN, B.; ROHL, C. A.; BAKER, D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, Elsevier, v. 331, n. 1, p. 281–299, 2003.

GROMIHA, M. M.; YUGANDHAR, K.; ; JEMIMAH, S. Protein–protein interactions: scoring schemes and binding affinity. *Current opinion in structural biology*, Elsevier, v. 44, p. 31–38, 2017.

GULLI, A.; PAL, S. *Deep learning with Keras*. [S.l.]: Packt Publishing Ltd, 2017.

GUNSTEREN, W. F. V.; DAURA, X.; MARK, A. E. Computation of free energy. *Helvetica Chimica Acta*, Wiley Online Library, v. 85, n. 10, p. 3113–3129, 2002.

GUNSTEREN, W. F. V.; KARPLUS, M. Effect of constraints on the dynamics of macromolecules. *Macromolecules*, ACS Publications, v. 15, n. 6, p. 1528–1544, 1982.

GUO, D.; MULDER-KRIEGER, T.; IJZERMAN, A. P.; HEITMAN, L. H. Functional efficacy of adenosine a2a receptor agonists is positively correlated to their receptor residence time. *British journal of pharmacology*, Wiley Online Library, v. 166, n. 6, p. 1846–1859, 2012.

GÜTTLER, T.; AKSU, M.; DICKMANNS, A.; STEGMANN, K. M.; GREGOR, K.; REES, R.; TAXER, W.; RYMARENKO, O.; SCHÜNEMANN, J.; DIENEMANN, C. et al. Neutralization of sars-cov-2 by highly potent, hyperthermostable, and mutation-tolerant nanobodies. *The EMBO journal*, v. 40, n. 19, p. e107985, 2021.

HADDEN, J. A.; PERILLA, J. R. All-atom virus simulations. *Current opinion in virology*, Elsevier, v. 31, p. 82–91, 2018.

HARNDAHL, M.; RASMUSSEN, M.; RODER, G.; PEDERSEN, I. D.; SØRENSEN, M.; NIELSEN, M.; BUUS, S. Peptide-mhc class i stability is a better predictor than peptide affinity of ctl immunogenicity. *European journal of immunology*, Wiley Online Library, v. 42, n. 6, p. 1405–1416, 2012.

HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. V. D.; GOMMERS, R.; VIRTANEN, P.; COURNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J. et al. Array programming with numpy. *Nature*, Nature Publishing Group, v. 585, n. 7825, p. 357–362, 2020.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009.

HATEREN, A. V.; JAMES, E.; BAILEY, A.; PHILLIPS, A.; DALCHAU, N.; ELLIOTT, T. The cell biology of major histocompatibility complex class i assembly: towards a molecular understanding. *Tissue antigens*, Wiley Online Library, v. 76, n. 4, p. 259–275, 2010.

HAUSEN, H. Z. Papillomavirus infections—a major cause of human cancers. *Biochimica et biophysica acta (BBA)-reviews on cancer*, Elsevier, v. 1288, n. 2, p. F55–F78, 1996.

HAUSEN, H. Z. Papillomaviruses and cancer: from basic studies to clinical application. *Nature reviews cancer*, Nature Publishing Group UK London, v. 2, n. 5, p. 342–350, 2002.

HAUSEN, H. zur. Papillomaviruses in the causation of human cancers—a brief historical account. *Virology*, Elsevier, v. 384, n. 2, p. 260–265, 2009.

HEAD, J. D.; ZERNER, M. C. A broyden—fletcher—goldfarb—shanno optimization procedure for molecular geometries. *Chemical physics letters*, Elsevier, v. 122, n. 3, p. 264–270, 1985.

HENDERSON, R.; EDWARDS, R. J.; MANSOURI, K.; JANOWSKA, K.; STALLS, V.; GOBEIL, S. M.; KOPP, M.; LI, D.; PARKS, R.; HSU, A. L. et al. Controlling the sars-cov-2 spike glycoprotein conformation. *Nature structural & molecular biology*, Nature Publishing Group US New York, v. 27, n. 10, p. 925–933, 2020.

HESS, B.; BEKKER, H.; BERENDSEN, H. J.; FRAAIJE, J. G. Lincs: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, v. 18, n. 12, p. 1463–1472, 1997.

HEWITT, E. W. The mhc class i antigen presentation pathway: strategies for viral immune evasion. *Immunology*, Wiley Online Library, v. 110, n. 2, p. 163–169, 2003.

HILL, A. V. The combinations of haemoglobin with oxygen and with carbon monoxide. i. *Biochemical Journal*, Portland Press Ltd, v. 7, n. 5, p. 471, 1913.

HOLLINGSWORTH, S. A.; DROR, R. O. Molecular dynamics simulation for all. *Neuron*, Elsevier, v. 99, n. 6, p. 1129–1143, 2018.

HONIG, B.; NICHOLLS, A. Classical electrostatics in biology and chemistry. *Science*, American Association for the Advancement of Science, v. 268, n. 5214, p. 1144–1149, 1995.

HOPKINS, C. W.; GRAND, S. L.; WALKER, R. C.; ROITBERG, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of chemical theory and computation*, ACS Publications, v. 11, n. 4, p. 1864–1874, 2015.

HOPPE-SEYLER, K.; BOSSLER, F.; BRAUN, J. A.; HERRMANN, A. L.; HOPPE-SEYLER, F. The hpv e6/e7 oncogenes: key factors for viral carcinogenesis and therapeutic targets. *Trends in microbiology*, Elsevier, v. 26, n. 2, p. 158–168, 2018.

HUANG, B.; LILIENFELD, O. A. von. Ab initio machine learning in chemical compound space. *Chemical reviews*, ACS Publications, v. 121, n. 16, p. 10001–10036, 2021.

HUANG, J.; JR, A. D. M. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of computational chemistry*, Wiley Online Library, v. 34, n. 25, p. 2135–2145, 2013.

HUANG, P.-S.; BOYKEN, S. E.; BAKER, D. The coming of age of de novo protein design. *Nature*, Nature Publishing Group UK London, v. 537, n. 7620, p. 320–327, 2016.

HUBER, T.; TORDA, A. E.; GUNSTEREN, W. F. V. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *Journal of computer-aided molecular design*, Springer, v. 8, p. 695–708, 1994.

HUIBREGTSE, J. M.; SCHEFFNER, M.; HOWLEY, P. M. A cellular protein mediates association of p53 with the e6 oncoprotein of human papillomavirus types 16 or 18. *The EMBO journal*, v. 10, n. 13, p. 4129–4135, 1991.

HUIBREGTSE, J. M.; SCHEFFNER, M.; HOWLEY, P. M. Localization of the e6-ap regions that direct human papillomavirus e6 binding, association with p53, and ubiquitination of associated proteins. *Molecular and cellular biology*, Am Soc Microbiol, v. 13, n. 8, p. 4918–4927, 1993.

HÜNENBERGER, P. H. Lattice-sum methods for computing electrostatic interactions in molecular simulations. In: AMERICAN INSTITUTE OF PHYSICS. *AIP Conference Proceedings*. [S.l.], 1999. v. 492, n. 1, p. 17–83.

HÜNENBERGER, P. H. Calculation of the group-based pressure in molecular simulations. i. a general formulation including ewald and particle-particle–particle-mesh electrostatics. *The Journal of chemical physics*, AIP, v. 116, n. 16, p. 6880–6897, 2002.

HÜNENBERGER, P. H. Thermostat algorithms for molecular dynamics simulations. In: *Advanced computer simulation*. [S.l.]: Springer, 2005. p. 105–149.

HUNT, D. F.; HENDERSON, R. A.; SHABANOWITZ, J.; SAKAGUCHI, K.; MICHEL, H.; SEVILIR, N.; COX, A. L.; APPELLA, E.; ENGELHARD, V. H. Characterization of peptides bound to the class i mhc molecule hla-a2. 1 by mass spectrometry. *Science*, American Association for the Advancement of Science, v. 255, n. 5049, p. 1261–1263, 1992.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

HUO, J.; BAS, A. L.; RUZA, R. R.; DUYVESTEYN, H. M.; MIKOLAJEK, H.; MALINAUSKAS, T.; TAN, T. K.; RIJAL, P.; DUMOUX, M.; WARD, P. N. et al. Neutralizing nanobodies bind sars-cov-2 spike rbd and block interaction with ace2. *Nature structural & molecular biology*, Nature Publishing Group US New York, v. 27, n. 9, p. 846–854, 2020.

HUTCHINGS, C. J.; COLUSSI, P.; CLARK, T. G. Ion channels as therapeutic antibody targets. In: TAYLOR & FRANCIS. *MAbs*. [S.l.], 2019. v. 11, n. 2, p. 265–296.

HUTTER, F.; HOOS, H.; LEYTON-BROWN, K. An efficient approach for assessing hyperparameter importance. In: PMLR. *International conference on machine learning*. [S.l.], 2014. p. 754–762.

INC., P. T. *Collaborative data science*. Montreal, QC: Plotly Technologies Inc., 2015. Available at: <https://plot.ly>.

JARMOSKAITE, I.; ALSADHAN, I.; VAIDYANATHAN, P. P.; HERSCHLAG, D. How to measure and evaluate binding affinities. *Elife*, eLife Sciences Publications, Ltd, v. 9, p. e57264, 2020.

JARZYNSKI, C. Nonequilibrium equality for free energy differences. *Physical Review Letters*, APS, v. 78, n. 14, p. 2690, 1997.

JEAN-CHARLES, A.; NICHOLLS, A.; SHARP, K.; HONIG, B.; TEMPCZYK, A.; HENDRICKSON, T. F.; STILL, W. C. Electrostatic contributions to solvation energies: Comparison of free energy perturbation and continuum calculations. *Journal of the American Chemical Society*, ACS Publications, v. 113, n. 4, p. 1454–1455, 1991.

JEFFERYS, E. E.; SANSOM, M. S. Computational virology: molecular simulations of virus dynamics and interactions. In: *Physical Virology*. [S.l.]: Springer, 2019. p. 201–233.

JIANG, L.; ALTHOFF, E. A.; CLEMENTE, F. R.; DOYLE, L.; RÖTHLISBERGER, D.; ZANGHELLINI, A.; GALLAHER, J. L.; BETKER, J. L.; TANAKA, F.; BARBAS, C. F. et al. De novo computational design of retro-aldol enzymes. *science*, v. 319, n. 5868, p. 1387–1391, 2008.

JIANG, W.; TANG, L. Atomic cryo-em structures of viruses. *Current opinion in structural biology*, Elsevier, v. 46, p. 122–129, 2017.

JING, Z.; LIU, C.; CHENG, S. Y.; QI, R.; WALKER, B. D.; PIQUEMAL, J.-P.; REN, P. Polarizable force fields for biomolecular simulations: Recent advances and applications. *Annual Review of biophysics*, NIH Public Access, v. 48, p. 371, 2019.

JONES, E. Y. Mhc class i and class ii structures. *Current opinion in immunology*, Elsevier, p. 75–79, 1997.

JONES, S.; THORNTON, J. M. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 93, n. 1, p. 13–20, 1996.

JØRGENSEN, K. W.; RASMUSSEN, M.; BUUS, S.; NIELSEN, M. Net mhc stab–predicting stability of peptide–mhc-i complexes; impacts for cytotoxic t lymphocyte epitope discovery. *Immunology*, Wiley Online Library, v. 141, n. 1, p. 18–26, 2014.

JORGENSEN, W. L. Pulled from a protein's embrace. *Nature*, Nature Publishing Group, v. 466, n. 7302, p. 42–43, 2010.

JORGENSEN, W. L.; CHANDRASEKHAR, J.; MADURA, J. D.; IMPEY, R. W.; KLEIN, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, American Institute of Physics, v. 79, n. 2, p. 926–935, 1983.

JORGENSEN, W. L.; TIRADO-RIVES, J. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, ACS Publications, v. 110, n. 6, p. 1657–1666, 1988.

JOVČEVSKA, I.; MUYLDERMANS, S. The therapeutic potential of nanobodies. *BioDrugs*, Springer, v. 34, n. 1, p. 11–26, 2020.

JR, R. L. D.; KARPLUS, M. Backbone-dependent rotamer library for proteins application to side-chain prediction. *Journal of molecular biology*, v. 230, n. 2, p. 543–574, 1993.

JUBB, H. C.; PANDURANGAN, A. P.; TURNER, M. A.; OCHOA-MONTAÑO, B.; BLUNDELL, T. L.; ASCHER, D. B. Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in biophysics and molecular biology*, Elsevier, v. 128, p. 3–13, 2017.

JUMPER, J.; EVANS, R.; PRITZEL, A.; GREEN, T.; FIGURNOV, M.; RONNEBERGER, O.; TUNYASUVUNAKOOL, K.; BATES, R.; ŽÍDEK, A.; POTAPENKO, A. et al. Highly accurate protein structure prediction with alphafold. *Nature*, Nature Publishing Group, v. 596, n. 7873, p. 583–589, 2021.

KADURIN, A.; NIKOLENKO, S.; KHRABROV, K.; ALIPER, A.; ZHAVORONKOV, A. drugan: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular pharmaceutics*, ACS Publications, v. 14, n. 9, p. 3098–3104, 2017.

KALITA, P.; TRIPATHI, T.; PADHI, A. K. Computational protein design for covid-19 research and emerging therapeutics. *ACS Central Science*, ACS Publications, v. 9, n. 4, p. 602–613, 2023.

KAMENIK, A. S.; LINKER, S. M.; RINIKER, S. Enhanced sampling without borders: on global biasing functions and how to reweight them. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry, v. 24, n. 3, p. 1225–1236, 2022.

KARPLUS, M.; MCCAMMON, J. A. Molecular dynamics simulations of biomolecules. *Nature structural biology*, Nature Publishing Group, v. 9, n. 9, p. 646–652, 2002.

KASTRITIS, P. L.; BONVIN, A. M. Are scoring functions in protein- protein docking ready to predict interactomes? clues from a novel binding affinity benchmark. *Journal of proteome research*, ACS Publications, v. 9, n. 5, p. 2216–2225, 2010.

KASTRITIS, P. L.; MOAL, I. H.; HWANG, H.; WENG, Z.; BATES, P. A.; BONVIN, A. M.; JANIN, J. A structure-based benchmark for protein–protein binding affinity. *Protein Science*, Wiley Online Library, v. 20, n. 3, p. 482–491, 2011.

KAUFMANN, K. W.; LEMMON, G. H.; DELUCA, S. L.; SHEEHAN, J. H.; MEILER, J. Practically useful: what the rosetta protein modeling suite can do for you. *Biochemistry*, v. 49, n. 14, p. 2987–2998, 2010.

KEITH, J. A.; VASSILEV-GALINDO, V.; CHENG, B.; CHMIELA, S.; GASTEGGER, M.; MULLER, K.-R.; TKATCHENKO, A. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical reviews*, ACS Publications, v. 121, n. 16, p. 9816–9872, 2021.

KHOURY, D. S.; DOCKEN, S. S.; SUBBARAO, K.; KENT, S. J.; DAVENPORT, M. P.; CROMER, D. Predicting the efficacy of variant-modified covid-19 vaccine boosters. *Nature Medicine*, Nature Publishing Group US New York, p. 1–5, 2023.

KING, E.; AITCHISON, E.; LI, H.; LUO, R. Recent developments in free energy calculations for drug discovery. *Frontiers in Molecular Biosciences*, Frontiers Media SA, v. 8, p. 712085, 2021.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. *science*, American association for the advancement of science, v. 220, n. 4598, p. 671–680, 1983.

KIRSCHNER, K. N.; YONGYE, A. B.; TSCHAMPEL, S. M.; GONZÁLEZ-OUTEIRIÑO, J.; DANIELS, C. R.; FOLEY, B. L.; WOODS, R. J. Glycam06: a generalizable biomolecular force field. carbohydrates. *Journal of computational chemistry*, Wiley Online Library, v. 29, n. 4, p. 622–655, 2008.

KLEIN, S.; CORTESE, M.; WINTER, S. L.; WACHSMUTH-MELM, M.; NEUFELDT, C. J.; CERIKAN, B.; STANIFER, M. L.; BOULANT, S.; BARTENSCHLAGER, R.; CHLANDA, P. Sars-cov-2 structure and replication characterized by in situ cryo-electron tomography. *Nature communications*, Nature Publishing Group, v. 11, n. 1, p. 1–10, 2020.

KNAPP, B.; OMASITS, U.; BOHLE, B.; MAILLERE, B.; EBNER, C.; SCHREINER, W.; JAHN-SCHMID, B. 3-layer-based analysis of peptide–mhc interaction: In silico prediction, peptide binding affinity and t cell activation in a relevant allergen-specific model. *Molecular immunology*, Elsevier, v. 46, n. 8-9, p. 1839–1844, 2009.

KNAPP, B.; OMASITS, U.; FRANTAL, S.; SCHREINER, W. A critical cross-validation of high throughput structural binding prediction methods for pmhc. *Journal of Computer-Aided Molecular Design*, Springer, v. 23, p. 301–307, 2009.

KOENIG, P.-A.; DAS, H.; LIU, H.; KÜMMERER, B. M.; GOHR, F. N.; JENSTER, L.-M.; SCHIFFELERS, L. D.; TESFAMARIAM, Y. M.; UCHIMA, M.; WUERTH, J. D. et al. Structure-guided multivalent nanobodies block sars-cov-2 infection and suppress mutational escape. *Science*, American Association for the Advancement of Science, v. 371, n. 6530, p. eabe6230, 2021.

KOKH, D. B.; AMARAL, M.; BOMKE, J.; GRADLER, U.; MUSIL, D.; BUCHSTALLER, H.-P.; DREYER, M. K.; FRECH, M.; LOWINSKI, M.; VALLEE, F. et al. Estimation of drug-target residence times by τ-random acceleration molecular dynamics simulations. *Journal of chemical theory and computation*, ACS Publications, v. 14, n. 7, p. 3859–3869, 2018.

KOKH, D. B.; DOSER, B.; RICHTER, S.; ORMERSBACH, F.; CHENG, X.; WADE, R. C. A workflow for exploring ligand dissociation from a macromolecule: Efficient random acceleration molecular dynamics simulation and interaction fingerprint analysis of ligand trajectories. *The Journal of chemical physics*, AIP Publishing LLC, v. 153, n. 12, p. 125102, 2020.

KOKH, D. B.; WADE, R. C. G protein-coupled receptor–ligand dissociation rates and mechanisms from τramd simulations. *Journal of Chemical Theory and Computation*, ACS Publications, v. 17, n. 10, p. 6610–6623, 2021.

KOLLMAN, P. A.; MASSOVA, I.; REYES, C.; KUHN, B.; HUO, S.; CHONG, L.; LEE, M.; LEE, T.; DUAN, Y.; WANG, W. et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research*, ACS Publications, v. 33, n. 12, p. 889–897, 2000.

KONWARH, R. Nanobodies: prospects of expanding the gamut of neutralizing antibodies against the novel coronavirus, sars-cov-2. *Frontiers in Immunology*, Frontiers Media SA, v. 11, p. 1531, 2020.

KORTEMME, T.; BAKER, D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 99, n. 22, p. 14116–14121, 2002.

KORTEMME, T.; BAKER, D. Computational design of protein–protein interactions. *Current opinion in chemical biology*, v. 8, n. 1, p. 91–97, 2004.

KRANJEC, C.; DOORBAR, J. Human papillomavirus infection and induction of neoplasia: a matter of fitness. *Current Opinion in Virology*, Elsevier, v. 20, p. 129–136, 2016.

KRYSHTAFOVYCH, A.; SCHWEDE, T.; TOPF, M.; FIDELIS, K.; MOULT, J. Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 87, n. 12, p. 1011–1020, 2019.

KUHLMAN, B.; BAKER, D. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, v. 97, n. 19, p. 10383–10388, 2000.

KUHLMAN, B.; BRADLEY, P. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, Nature Publishing Group UK London, v. 20, n. 11, p. 681–697, 2019.

KUHLMAN, B.; DANTAS, G.; IRETON, G. C.; VARANI, G.; STODDARD, B. L.; BAKER, D. Design of a novel globular protein fold with atomic-level accuracy. *science*, v. 302, n. 5649, p. 1364–1368, 2003.

KULSKI, J. K.; SHIINA, T.; ANZAI, T.; KOHARA, S.; INOKO, H. Comparative genomic analysis of the mhc: the evolution of class i duplication blocks, diversity and complexity from shark to man. *Immunological reviews*, Wiley Online Library, v. 190, n. 1, p. 95–122, 2002.

KUMARI, R.; KUMAR, R.; CONSORTIUM, O. S. D. D.; LYNN, A. g_mmpbsa a gromacs tool for high-throughput mm-pbsa calculations. *Journal of chemical information and modeling*, ACS Publications, v. 54, n. 7, p. 1951–1962, 2014.

KUNZ, P.; FLOCK, T.; SOLER, N.; ZAISS, M.; VINCKE, C.; STERCKX, Y.; KASTELIC, D.; MUYLDERMANS, S.; HOHEISEL, J. D. Exploiting sequence and stability information for directing nanobody stability engineering. *Biochimica et Biophysica Acta (BBA)-General Subjects*, Elsevier, v. 1861, n. 9, p. 2196–2205, 2017.

KURSA, M. B.; JANKOWSKI, A.; RUDNICKI, W. R. Boruta–a system for feature selection. *Fundamenta Informaticae*, IOS Press, v. 101, n. 4, p. 271–285, 2010.

LAINE, R. F.; GOODFELLOW, G.; YOUNG, L. J.; TRAVERS, J.; CARROLL, D.; DIBBEN, O.; BRIGHT, H.; KAMINSKI, C. F. Structured illumination microscopy combined with machine learning enables the high throughput analysis and classification of virus structure. *Elife*, eLife Sciences Publications Limited, v. 7, p. e40183, 2018.

LAIO, A.; GERVASIO, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, IOP Publishing, v. 71, n. 12, p. 126601, 2008.

LAIO, A.; PARRINELLO, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, v. 99, n. 20, p. 12562–12566, 2002.

LAIO, A.; RODRIGUEZ-FORTEA, A.; GERVASIO, F. L.; CECCARELLI, M.; PARRINELLO, M. Assessing the accuracy of metadynamics. *The journal of physical chemistry B*, ACS Publications, v. 109, n. 14, p. 6714–6721, 2005.

LAN, J.; GE, J.; YU, J.; SHAN, S.; ZHOU, H.; FAN, S.; ZHANG, Q.; SHI, X.; WANG, Q.; ZHANG, L. et al. Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *nature*, Nature Publishing Group UK London, v. 581, n. 7807, p. 215–220, 2020.

LAZARIDIS, T. Inhomogeneous fluid approach to solvation thermodynamics. 1. theory. *The Journal of Physical Chemistry B*, ACS Publications, v. 102, n. 18, p. 3531–3541, 1998.

LAZARIDIS, T.; KARPLUS, M. Effective energy function for proteins in solution. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 35, n. 2, p. 133–152, 1999.

LEACH, A. Molecular modeling principles and applications. *2nd, editor.: Pearson Education Limited*, 2001.

LEAVER-FAY, A.; BUTTERFOSS, G. L.; SNOEYINK, J.; KUHLMAN, B. Maintaining solvent accessible surface area under rotamer substitution for protein design. *Journal of Computational Chemistry*, Wiley Online Library, v. 28, n. 8, p. 1336–1341, 2007.

LEAVER-FAY, A.; O'MEARA, M. J.; TYKA, M.; JACAK, R.; SONG, Y.; KELLOGG, E. H.; THOMPSON, J.; DAVIS, I. W.; PACHE, R. A.; LYSKOV, S. et al. Scientific benchmarks for guiding macromolecular energy function improvement. In: *Methods in enzymology*. [S.l.]: Elsevier, 2013. v. 523, p. 109–143.

LEAVER-FAY, A.; TYKA, M.; LEWIS, S. M.; LANGE, O. F.; THOMPSON, J.; JACAK, R.; KAUFMAN, K. W.; RENFREW, P. D.; SMITH, C. A.; SHEFFLER, W.; DAVIS, I. W.; COOPER, S.; TREUILLE, A.; MANDELL, D. J.; RICHTER, F.; BAN, Y.-E. A.; FLEISHMAN, S. J.; CORN, J. E.; KIM, D. E.; LYSKOV, S.; BERRONDO, M.; MENTZER, S.; POPOVIć, Z.; HAVRANEK, J. J.; KARANICOLAS, J.; DAS, R.; MEILER, J.; KORTEMME, T.; GRAY, J. J.; KUHLMAN, B.; BAKER, D.; BRADLEY, P. Chapter nineteen - rosetta3: An object-oriented software suite for the simulation and design of macromolecules. In: JOHNSON, M. L.; BRAND, L. (Ed.). *Computer Methods, Part C*. Academic Press, 2011, (Methods in Enzymology, v. 487). p. 545–574. Available at: <https://www.sciencedirect.com/science/article/pii/B9780123812704000196>.

LEAVER-FAY, A.; TYKA, M.; LEWIS, S. M.; LANGE, O. F.; THOMPSON, J.; JACAK, R.; KAUFMAN, K. W.; RENFREW, P. D.; SMITH, C. A.; SHEFFLER, W. et al. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. In: *Methods in enzymology*. [S.l.: s.n.], 2011. v. 487, p. 545–574.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998.

LEE, A. C.-L.; HARRIS, J. L.; KHANNA, K. K.; HONG, J.-H. A comprehensive review on current advances in peptide drug development and design. *International journal of molecular sciences*, MDPI, v. 20, n. 10, p. 2383, 2019.

LEE, J.; CHENG, X.; SWAILS, J. M.; YEOM, M. S.; EASTMAN, P. K.; LEMKUL, J. A.; WEI, S.; BUCKNER, J.; JEONG, J. C.; QI, Y. et al. Charmm-gui input generator for namd, gromacs, amber, openmm, and charmm/openmm simulations using the charmm36 additive force field. *Journal of chemical theory and computation*, ACS Publications, v. 12, n. 1, p. 405–413, 2016.

LEE, K. S. S.; YANG, J.; NIU, J.; NG, C. J.; WAGNER, K. M.; DONG, H.; KODANI, S. D.; WAN, D.; MORISSEAU, C.; HAMMOCK, B. D. Drug-target residence time affects in vivo target occupancy through multiple pathways. *ACS central science*, ACS Publications, v. 5, n. 9, p. 1614–1624, 2019.

LEE, T.-S.; CERUTTI, D. S.; MERMELSTEIN, D.; LIN, C.; LEGRAND, S.; GIESE, T. J.; ROITBERG, A.; CASE, D. A.; WALKER, R. C.; YORK, D. M. Gpu-accelerated molecular dynamics and free energy methods in amber18: performance enhancements and new features. *Journal of chemical information and modeling*, ACS Publications, v. 58, n. 10, p. 2043–2050, 2018.

LEHMANN, M.; PASAMONTES, L.; LASSEN, S. a.; WYSS, M. The consensus concept for thermostability engineering of proteins. *Biochimica et Biophysica Acta (BBA)-protein structure and molecular enzymology*, Elsevier, v. 1543, n. 2, p. 408–415, 2000.

LEMAN, J. K.; WEITZNER, B. D.; LEWIS, S. M.; ADOLF-BRYFOGLE, J.; ALAM, N.; ALFORD, R. F.; APRAHAMIAN, M.; BAKER, D.; BARLOW, K. A.; BARTH, P. et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, Nature Publishing Group, v. 17, n. 7, p. 665–680, 2020.

LEVITT, M.; LIFSON, S. Refinement of protein conformations using a macromolecular energy minimization procedure. *Journal of molecular biology*, Elsevier, v. 46, n. 2, p. 269–279, 1969.

LEVY, Y.; ONUCHIC, J. N. Water mediation in protein folding and molecular recognition. *Annual review of biophysics and biomolecular structure*, Palo Alto, Calif.: Annual Reviews Inc., c1992-, v. 35, n. 1, p. 389–415, 2006.

LIANG, G.; FOX, P. C.; BOWEN, J. P. Parameter analysis and refinement toolkit system and its application in mm3 parameterization for phosphine and its derivatives. *Journal of computational chemistry*, Wiley Online Library, v. 17, n. 8, p. 940–953, 1996.

LIPPOW, S. M.; TIDOR, B. Progress in computational protein design. *Current opinion in biotechnology*, Elsevier, v. 18, n. 4, p. 305–311, 2007.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: IEEE. *2008 eighth ieee international conference on data mining*. [S.l.], 2008. p. 413–422.

LIU, J.; GAO, G. F. Major histocompatibility complex: Interaction with peptides. *eLS*, John Wiley & Sons, Ltd, 2011.

LIU, J.; WANG, R. Classification of current scoring functions. *Journal of chemical information and modeling*, ACS Publications, v. 55, n. 3, p. 475–482, 2015.

LÖHR, T.; SORMANNI, P.; VENDRUSCOLO, M. Conformational entropy as a potential liability of computationally designed antibodies. *Biomolecules*, MDPI, v. 12, n. 5, p. 718, 2022.

LÜDEMANN, S. K.; CARUGO, O.; WADE, R. C. Substrate access to cytochrome p450cam: A comparison of a thermal motion pathway analysis with molecular dynamics simulation data. *Molecular modeling annual*, Springer, v. 3, p. 369–374, 1997.

LÜDEMANN, S. K.; LOUNNAS, V.; WADE, R. C. How do substrates enter and products exit the buried active site of cytochrome p450cam? 1. random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *Journal of molecular biology*, Elsevier, v. 303, n. 5, p. 797–811, 2000.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, v. 30, 2017.

LYNCH, D. L.; PAVLOVA, A.; FAN, Z.; GUMBART, J. C. Understanding virus structure and dynamics through molecular simulations. *Journal of Chemical Theory and Computation*, ACS Publications, 2023.

MACCALLUM, R. M.; MARTIN, A. C.; THORNTON, J. M. Antibody-antigen interactions: contact analysis and binding site topography. *Journal of molecular biology*, Elsevier, v. 262, n. 5, p. 732–745, 1996.

MACHADO, M. R.; PANTANO, S. Fighting viruses with computers, right now. *Current Opinion in Virology*, Elsevier, v. 48, p. 91–99, 2021.

MACIEL, L. G.; FERRAZ, M. V.; OLIVEIRA, A. A.; LINS, R. D.; ANJOS, J. V. dos; GUIDO, R. V.; SOARES, T. A. Inhibition of 3-hydroxykynurenine transaminase from aedes aegypti and anopheles gambiae: A mosquito-specific target to combat the transmission of arboviruses. *ACS Bio & Med Chem Au*, ACS Publications, 2023.

MAGLIERY, T. J. Protein stability: computation, sequence statistics, and new experimental methods. *Current opinion in structural biology*, Elsevier, v. 33, p. 161–168, 2015.

MAGUIRE, J. B.; BOYKEN, S. E.; BAKER, D.; KUHLMAN, B. Rapid sampling of hydrogen bond networks for computational protein design. *Journal of chemical theory and computation*, ACS Publications, v. 14, n. 5, p. 2751–2760, 2018.

MAIER, J. A.; MARTINEZ, C.; KASAVAJHALA, K.; WICKSTROM, L.; HAUSER, K. E.; SIMMERLING, C. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, American Chemical Society, v. 11, n. 8, p. 3696–3713, 2015.

MARCANDALLI, J.; FIALA, B.; OLS, S.; PEROTTI, M.; SCHUEREN, W. de van der; SNIJDER, J.; HODGE, E.; BENHAIM, M.; RAVICHANDRAN, R.; CARTER, L. et al. Induction of potent neutralizing antibody responses by a designed protein nanoparticle vaccine for respiratory syncytial virus. *Cell*, Elsevier, v. 176, n. 6, p. 1420–1431, 2019.

MARCHAND, A.; HALL-BEAUVAIS, A. K. V.; CORREIA, B. E. Computational design of novel protein–protein interactions–an overview on methodological approaches and applications. *Current Opinion in Structural Biology*, Elsevier, v. 74, p. 102370, 2022.

MARILLET, S.; LEFRANC, M.-P.; BOUDINOT, P.; CAZALS, F. Novel structural parameters of ig–ag complexes yield a quantitative description of interaction specificity and binding affinity. *Frontiers in immunology*, Frontiers Media SA, v. 8, p. 34, 2017.

MARSHALL, S. A.; VIZCARRA, C. L.; MAYO, S. L. One-and two-body decomposable poisson-boltzmann methods for protein design calculations. *Protein Science*, Wiley Online Library, v. 14, n. 5, p. 1293–1304, 2005.

MARSILI, S.; BARDUCCI, A.; CHELLI, R.; PROCACCI, P.; SCHETTINO, V. Self-healing umbrella sampling: a non-equilibrium approach for quantitative free energy calculations. *The Journal of Physical Chemistry B*, ACS Publications, v. 110, n. 29, p. 14011–14013, 2006.

MARTEL, C. D.; PLUMMER, M.; VIGNAT, J.; FRANCESCHI, S. Worldwide burden of cancer attributable to hpv by site, country and hpv type. *International journal of cancer*, Wiley Online Library, v. 141, n. 4, p. 664–670, 2017.

MARTÍNEZ, L.; SONODA, M. T.; WEBB, P.; BAXTER, J. D.; SKAF, M. S.; POLIKARPOV, I. Molecular dynamics simulations reveal multiple pathways of ligand dissociation from thyroid hormone receptors. *Biophysical journal*, Elsevier, v. 89, n. 3, p. 2011–2023, 2005.

MARTINEZ, M.; BRUCE, N. J.; ROMANOWSKA, J.; KOKH, D. B.; OZBOYACI, M.; YU, X.; ÖZTÜRK, M. A.; RICHTER, S.; WADE, R. C. Sda 7: A modular and parallel implementation of the simulation of diffusional association software. *Journal of computational chemistry*, Wiley Online Library, v. 36, n. 21, p. 1631–1645, 2015.

MARTINEZ-ZAPIEN, D.; RUIZ, F. X.; POIRSON, J.; MITSCHLER, A.; RAMIREZ, J.; FORSTER, A.; COUSIDO-SIAH, A.; MASSON, M.; POL, S. V.; PODJARNY, A. et al. Structure of the e6/e6ap/p53 complex required for hpv-mediated degradation of p53. *Nature*, Nature Publishing Group UK London, v. 529, n. 7587, p. 541–545, 2016.

MARTYNA, G. J.; KLEIN, M. L.; TUCKERMAN, M. Nosé–hoover chains: The canonical ensemble via continuous dynamics. *The Journal of chemical physics*, American Institute of Physics, v. 97, n. 4, p. 2635–2643, 1992.

MARZELLA, D. F.; PARIZI, F. M.; TILBORG, D. V.; RENAUD, N.; SYBRANDI, D.; BUZATU, R.; RADEMAKER, D. T.; HOEN, P. A.; XUE, L. C. Pandora: a fast, anchor-restrained modelling protocol for peptide: Mhc complexes. *Frontiers in Immunology*, Frontiers Media SA, v. 13, 2022.

MARZINEK, J. K.; HUBER, R. G.; BOND, P. J. Multiscale modelling and simulation of viruses. *Current opinion in structural biology*, Elsevier, v. 61, p. 146–152, 2020.

MATOS, I. d. A.; PINTO, A. C. G.; FERRAZ, M. V. F.; ADAN, W. C. S.; RODRIGUES, R. P.; SANTOS, J. X. D.; KITAGAWA, R. R.; LINS, R. D.; OLIVEIRA, T. B.; JUNIOR, N. B. d. C. Identification of potential staphylococcus aureus dihydrofolate reductase inhibitors using qsar, molecular docking, dynamics simulations and free energy calculation. *Journal of Biomolecular Structure and Dynamics*, Taylor & Francis, p. 1–12, 2022.

MATSUMURA, M.; FREMONT, D. H.; PETERSON, P. A.; WILSON, I. A. Emerging principles for the recognition of peptide antigens by mhc class i molecules. *Science*, American Association for the Advancement of Science, v. 257, n. 5072, p. 927–934, 1992.

MATSUOKA, D.; NAKASAKO, M. Probability distributions of hydration water molecules around polar protein atoms obtained by a database analysis. *The Journal of Physical Chemistry B*, ACS Publications, v. 113, n. 32, p. 11274–11292, 2009.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.

MCINNES, L.; HEALY, J.; SAUL, N.; GROSSBERGER, L. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, v. 3, n. 29, p. 861, 2018.

MCKINNEY, W. et al. Data structures for statistical computing in python. In: AUSTIN, TX. *Proceedings of the 9th Python in Science Conference*. [S.l.], 2010. v. 445, p. 51–56.

MCKINNEY, W. et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, Seattle, v. 14, n. 9, p. 1–9, 2011.

MCQUARRIE, D. A. *Statistical mechanics*. [S.l.]: Sterling Publishing Company, 2000.

MCQUARRIE, D. A.; SIMON, J. D. *Physical chemistry: a molecular approach*. [S.l.]: University science books Sausalito, CA, 1997.

MEISAL, R.; ROUNGE, T. B.; CHRISTIANSEN, I. K.; EIELAND, A. K.; WORREN, M. M.; MOLDEN, T. F.; KOMMEDAL, Ø.; HOVIG, E.; LEEGAARD, T. M.; AMBUR, O. H. Hpv genotyping of modified general primer-amplicons is more analytically sensitive and specific by sequencing than by hybridization. *PloS one*, Public Library of Science San Francisco, CA USA, v. 12, n. 1, p. e0169074, 2017.

MELO, M. C.; BERNARDI, R. C. Fostering discoveries in the era of exascale computing: How the next generation of supercomputers empowers computational and experimental biophysics alike. *Biophysical Journal*, Elsevier, 2023.

MENDES, M. F. D. A.; BRAGATTE, M. A. d. S.; VIANNA, P.; FREITAS, M. V. D.; PÖHNER, I.; RICHTER, S.; WADE, R.; SALZANO, F. M.; VIEIRA, G. F. Matchtope: A tool to predict the cross reactivity of peptides complexed with major histocompatibility complex i. *Frontiers in Immunology*, Frontiers, p. 6262, 2022.

MENDOZA, M. N.; JIAN, M.; KING, M. T.; BROOKS, C. L. Role of a noncanonical disulfide bond in the stability, affinity, and flexibility of a vhh specific for the listeria virulence factor inlb. *Protein Science*, Wiley Online Library, v. 29, n. 4, p. 990–1003, 2020.

MESPLÈDE, T.; GAGNON, D.; BERGERON-LABRECQUE, F.; AZAR, I.; SÉNÉCHAL, H.; COUTLÉE, F.; ARCHAMBAULT, J. p53 degradation activity, expression, and subcellular localization of e6 proteins from 29 human papillomavirus genotypes. *Journal of virology*, Am Soc Microbiol, v. 86, n. 1, p. 94–107, 2012.

METZGER, V. T.; EUN, C.; KEKENES-HUSKEY, P. M.; HUBER, G.; MCCAMMON, J. A. Electrostatic channeling in p. falciparum dhfr-ts: Brownian dynamics and smoluchowski modeling. *Biophysical journal*, Elsevier, v. 107, n. 10, p. 2394–2402, 2014.

MICHELETTI, C.; LAIO, A.; PARRINELLO, M. Reconstructing the density of states by history-dependent metadynamics. *Physical Review Letters*, APS, v. 92, n. 17, p. 170601, 2004.

MIKHAYLOV, V.; LEVINE, A. J. Accurate modeling of peptide-mhc structures with alphafold. *bioRxiv*, Cold Spring Harbor Laboratory, p. 2023–03, 2023.

MIR, M. A.; MEHRAJ, U.; SHEIKH, B. A.; HAMDANI, S. S. Nanobodies: The "magic bullets" in therapeutics, drug delivery and diagnostics. *Human antibodies*, IOS Press, v. 28, n. 1, p. 29–51, 2020.

MITCHELL, L. S.; COLWELL, L. J. Comparative analysis of nanobody sequence and structure data. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 86, n. 7, p. 697–706, 2018.

MITCHELL, T. *Machine learning*. [S.l.]: McGraw-hill New York, 1997.

MITTAL, S.; BANKS, L. Molecular mechanisms underlying human papillomavirus e6 and e7 oncoprotein-induced cell transformation. *Mutation Research/Reviews in Mutation Research*, Elsevier, v. 772, p. 23–35, 2017.

MOAL, I. H.; AGIUS, R.; BATES, P. A. Protein–protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, Oxford University Press, v. 27, n. 21, p. 3002–3009, 2011.

MOLLICA, L.; DECHERCHI, S.; ZIA, S. R.; GASPARI, R.; CAVALLI, A.; ROCCHIA, W. Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Scientific reports*, Nature Publishing Group UK London, v. 5, n. 1, p. 11539, 2015.

MOODY, C. A.; LAIMINS, L. A. Human papillomavirus oncoproteins: pathways to transformation. *Nature Reviews Cancer*, Nature Publishing Group, v. 10, n. 8, p. 550–560, 2010.

MORRISON, C. Nanobody approval gives domain antibodies a boost. *Nat Rev Drug Discov*, v. 18, n. 7, p. 485–487, 2019.

MOULT, J.; FIDELIS, K.; KRYSHTAFOVYCH, A.; SCHWEDE, T.; TRAMONTANO, A. Critical assessment of methods of protein structure prediction (casp)—round xii. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 86, p. 7–15, 2018.

MÜNGER, K.; WERNESS, B.; DYSON, N.; PHELPS, W.; HARLOW, E.; HOWLEY, P. Complex formation of human papillomavirus e7 proteins with the retinoblastoma tumor suppressor gene product. *The EMBO journal*, v. 8, n. 13, p. 4099–4105, 1989.

MUÑOZ, N. International agency for research on cancer multicenter cervical cancer study group. epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med*, v. 348, p. 518–527, 2003.

MURIN, C. D.; WILSON, I. A.; WARD, A. B. Antibody responses to viral infections: a structural perspective across three different enveloped viruses. *Nature microbiology*, Nature Publishing Group UK London, v. 4, n. 5, p. 734–747, 2019.

MUYLDERMANS, S. Single domain camel antibodies: current status. *Reviews in molecular Biotechnology*, Elsevier, v. 74, n. 4, p. 277–302, 2001.

MUYLDERMANS, S. Nanobodies: Natural single-domain antibodies. *Annual Review of Biochemistry*, v. 82, n. 1, p. 775–797, 2013.

MUYLDERMANS, S. et al. Nanobodies: natural single-domain antibodies. *Annu Rev Biochem*, v. 82, n. 1, p. 775–797, 2013.

NAVECA, F. G.; NASCIMENTO, V.; SOUZA, V.; CORADO, A. d. L.; NASCIMENTO, F.; SILVA, G.; MEJÍA, M. C.; BRANDÃO, M. J.; COSTA, Á.; DUARTE, D. et al. Spread of gamma (p. 1) sub-lineages carrying spike mutations close to the furin cleavage site and deletions in the n-terminal domain drives ongoing transmission of sars-cov-2 in amazonas, brazil. *Microbiology spectrum*, Am Soc Microbiol, v. 10, n. 1, p. e02366–21, 2022.

NAYAK, J.; MISHRA, M.; NAIK, B.; SWAPNAREKHA, H.; CENGIZ, K.; SHANMUGANATHAN, V. An impact study of covid-19 on six different industries: Automobile, energy and power, agriculture, education, travel and tourism and consumer electronics. *Expert systems*, Wiley Online Library, v. 39, n. 3, p. e12677, 2022.

NEAL, R. M. Bayesian sampling. *Bayesian Learning for Neural Networks*, Springer, p. 179–207, 1996.

NGUYEN, D. D.; WANG, B.; WEI, G.-W. Accurate, robust, and reliable calculations of poisson–boltzmann binding energies. *Journal of computational chemistry*, Wiley Online Library, v. 38, n. 13, p. 941–948, 2017.

NOMINÉ, Y.; MASSON, M.; CHARBONNIER, S.; ZANIER, K.; RISTRIANI, T.; DERYCKÈRE, F.; SIBLER, A.-P.; DESPLANCQ, D.; ATKINSON, R. A.; WEISS, E. et al. Structural and functional analysis of e6 oncoprotein: insights in the molecular pathways of human papillomavirus-mediated pathogenesis. *Molecular cell*, Elsevier, v. 21, n. 5, p. 665–678, 2006.

NORRBY, E. Nobel prizes and the emerging virus concept. *Archives of virology*, Springer, v. 153, n. 6, p. 1109–1123, 2008.

NORTHRUP, S. H.; ALLISON, S. A.; MCCAMMON, J. A. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *The Journal of Chemical Physics*, American Institute of Physics, v. 80, n. 4, p. 1517–1524, 1984.

NOSÉ, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, American Institute of Physics, v. 81, n. 1, p. 511–519, 1984.

NUNES-ALVES, A.; KOKH, D. B.; WADE, R. C. Recent progress in molecular simulation methods for drug binding kinetics. *Current Opinion in Structural Biology*, Elsevier, v. 64, p. 126–133, 2020.

NUNES-ALVES, A.; KOKH, D. B.; WADE, R. C. Ligand unbinding mechanisms and kinetics for t4 lysozyme mutants from τramd simulations. *Current Research in Structural Biology*, Elsevier, v. 3, p. 106–111, 2021.

OCHOA, R.; LAIO, A.; COSSIO, P. Predicting the affinity of peptides to major histocompatibility complex class ii by scoring molecular dynamics simulations. *Journal of Chemical Information and Modeling*, ACS Publications, v. 59, n. 8, p. 3464–3473, 2019.

OLGA, P. et al. Development of neutralizing nanobodies to the hemagglutinin stem domain of influenza a viruses. *Acta Naturae ( )*, -, v. 13, n. 4, p. 33–41, 2021.

OLGA, P.; YU, L. D. et al. Nanobodies are potential therapeutic agents for the ebola virus infection. *Acta Naturae ( )*, -, v. 13, n. 4, p. 53–63, 2021.

OLSSON, M. H.; SØNDERGAARD, C. R.; ROSTKOWSKI, M.; JENSEN, J. H. Propka3: consistent treatment of internal and surface residues in empirical p k a predictions. *Journal of chemical theory and computation*, ACS Publications, v. 7, n. 2, p. 525–537, 2011.

ONSAGER, L. Electric moments of molecules in liquids. *Journal of the American Chemical Society*, ACS Publications, v. 58, n. 8, p. 1486–1493, 1936.

OOSTENBRINK, C.; VILLA, A.; MARK, A. E.; GUNSTEREN, W. F. V. A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6. *Journal of computational chemistry*, Wiley Online Library, v. 25, n. 13, p. 1656–1676, 2004.

O'MEARA, M. J.; LEAVER-FAY, A.; TYKA, M. D.; STEIN, A.; HOULIHAN, K.; DIMAIO, F.; BRADLEY, P.; KORTEMME, T.; BAKER, D.; SNOEYINK, J. et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. *Journal of chemical theory and computation*, ACS Publications, v. 11, n. 2, p. 609–622, 2015.

PAIARDI, G.; RICHTER, S.; ORESTE, P.; URBINATI, C.; RUSNATI, M.; WADE, R. C. The binding of heparin to spike glycoprotein inhibits sars-cov-2 infection by three mechanisms. *Journal of Biological Chemistry*, ASBMB, v. 298, n. 2, 2022.

PÁLL, S.; ZHMUROV, A.; BAUER, P.; ABRAHAM, M.; LUNDBORG, M.; GRAY, A.; HESS, B.; LINDAHL, E. Heterogeneous parallelization and acceleration of molecular dynamics simulations in gromacs. *The Journal of Chemical Physics*, AIP Publishing LLC, v. 153, n. 13, p. 134110, 2020.

PAN, A. C.; JACOBSON, D.; YATSENKO, K.; SRITHARAN, D.; WEINREICH, T. M.; SHAW, D. E. Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 116, n. 10, p. 4244–4249, 2019.

PAN, X.; KORTEMME, T. Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry*, ASBMB, v. 296, 2021.

PANDAY, S. K.; ALEXOV, E. Protein–protein binding free energy predictions with the mm/pbsa approach complemented with the gaussian-based method for entropy estimation. *ACS omega*, ACS Publications, v. 7, n. 13, p. 11057–11067, 2022.

PAPAKYRIAKOU, A.; REEVES, E.; BETON, M.; MIKOLAJEK, H.; DOUGLAS, L.; COOPER, G.; ELLIOTT, T.; WERNER, J. M.; JAMES, E. The partial dissociation of mhc class i–bound peptides exposes their n terminus to trimming by endoplasmic reticulum aminopeptidase 1. *Journal of Biological Chemistry*, ASBMB, v. 293, n. 20, p. 7538–7548, 2018.

PARK, H.; BRADLEY, P.; JR, P. G.; LIU, Y.; MULLIGAN, V. K.; KIM, D. E.; BAKER, D.; DIMAIO, F. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of chemical theory and computation*, ACS Publications, v. 12, n. 12, p. 6201–6212, 2016.

PARK, S.-J.; LEE, J.; QI, Y.; KERN, N. R.; LEE, H. S.; JO, S.; JOUNG, I.; JOO, K.; LEE, J.; IM, W. Charmm-gui glycan modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology*, Oxford University Press, v. 29, n. 4, p. 320–331, 2019.

PARKER, K. C.; BEDNAREK, M. A.; COLIGAN, J. E. Scheme for ranking potential hla-a2 binding peptides based on independent binding of individual peptide side-chains. *Journal of immunology (Baltimore, Md.: 1950)*, v. 152, n. 1, p. 163–175, 1994.

PARRINELLO, M.; RAHMAN, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, v. 52, n. 12, p. 7182–7190, 1981.

PAUL, F.; WEHMEYER, C.; ABUALROUS, E. T.; WU, H.; CRABTREE, M. D.; SCHÖNEBERG, J.; CLARKE, J.; FREUND, C.; WEIKL, T. R.; NOÉ, F. Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations. *Nature communications*, Nature Publishing Group UK London, v. 8, n. 1, p. 1095, 2017.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011.

PETSALAKI, E.; RUSSELL, R. B. Peptide-mediated interactions in biological systems: new discoveries and applications. *Current opinion in biotechnology*, Elsevier, v. 19, n. 4, p. 344–350, 2008.

PHELAN, A. L.; KATZ, R.; GOSTIN, L. O. The novel coronavirus originating in wuhan, china: challenges for global health governance. *Jama*, American Medical Association, v. 323, n. 8, p. 709–710, 2020.

PHILLIPS, J. C.; HARDY, D. J.; MAIA, J. D.; STONE, J. E.; RIBEIRO, J. V.; BERNARDI, R. C.; BUCH, R.; FIORIN, G.; HÉNIN, J.; JIANG, W. et al. Scalable molecular dynamics on cpu and gpu architectures with namd. *The Journal of chemical physics*, AIP Publishing LLC, v. 153, n. 4, p. 044130, 2020.

PIETRUCCI, F.; LAIO, A. A collective variable for the efficient exploration of protein beta-sheet structures: application to sh3 and gb1. *Journal of Chemical Theory and Computation*, ACS Publications, v. 5, n. 9, p. 2197–2201, 2009.

PIOT, P.; BARTOS, M.; GHYS, P. D.; WALKER, N.; SCHWARTLÄNDER, B. The global impact of hiv/aids. *Nature*, Nature Publishing Group, v. 410, n. 6831, p. 968–973, 2001.

PLATT, J. C. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, p. 185–208, 1998.

POHORILLE, A.; JARZYNSKI, C.; CHIPOT, C. Good practices in free-energy calculations. *The Journal of Physical Chemistry B*, ACS Publications, v. 114, n. 32, p. 10235–10253, 2010.

POIRSON, J.; BIQUAND, E.; STRAUB, M.-L.; CASSONNET, P.; NOMINÉ, Y.; JONES, L.; WERF, S. van D.; TRAVÉ, G.; ZANIER, K.; JACOB, Y. et al. *Mapping the interactome of HPV E6 and E7 oncoproteins with the ubiquitin-proteasome system*. [S.l.]: Wiley Online Library, 2017. 3171–3201 p.

POL, S. B. V.; KLINGELHUTZ, A. J. Papillomavirus e6 oncoproteins. *Virology*, Elsevier, v. 445, n. 1-2, p. 115–137, 2013.

POLLOCK, E.; GLOSLI, J. Comments on p3m, fmm, and the ewald method for large periodic coulombic systems. *Computer Physics Communications*, Elsevier, v. 95, n. 2-3, p. 93–110, 1996.

PRITCHARD, B. P.; ALTARAWY, D.; DIDIER, B.; GIBSON, T. D.; WINDUS, T. L. New basis set exchange: An open, up-to-date resource for the molecular sciences community. *Journal of chemical information and modeling*, ACS Publications, v. 59, n. 11, p. 4814–4820, 2019.

PROCKO, E.; BERGUIG, G. Y.; SHEN, B. W.; SONG, Y.; FRAYO, S.; CONVERTINE, A. J.; MARGINEANTU, D.; BOOTH, G.; CORREIA, B. E.; CHENG, Y. et al. A computationally designed inhibitor of an epstein-barr viral bcl-2 protein induces apoptosis in infected cells. *Cell*, Elsevier, v. 157, n. 7, p. 1644–1656, 2014.

QUIJANO-RUBIO, A.; YEH, H.-W.; PARK, J.; LEE, H.; LANGAN, R. A.; BOYKEN, S. E.; LAJOIE, M. J.; CAO, L.; CHOW, C. M.; MIRANDA, M. C. et al. De novo design of modular and tunable protein biosensors. *Nature*, Nature Publishing Group UK London, v. 591, n. 7850, p. 482–487, 2021.

RADIC, Z.; KIRCHHOFF, P. D.; QUINN, D. M.; MCCAMMON, J. A.; TAYLOR, P. Electrostatic influence on the kinetics of ligand binding to acetylcholinesterase: Distinctions between active center ligands and fasciculin. *Journal of Biological Chemistry*, ASBMB, v. 272, n. 37, p. 23265–23277, 1997.

RAY, D.; STONE, S. E.; ANDRICIOAEI, I. Markovian weighted ensemble milestoning (m-wem): Long-time kinetics from short trajectories. *Journal of Chemical Theory and Computation*, ACS Publications, v. 18, n. 1, p. 79–95, 2021.

REGNER, M. Cross-reactivity in t-cell antigen recognition. *Immunology and cell biology*, Wiley Online Library, v. 79, n. 2, p. 91–100, 2001.

REICHMANN, D.; RAHAT, O.; ALBECK, S.; MEGED, R.; DYM, O.; SCHREIBER, G. The modular architecture of protein–protein binding interfaces. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 102, n. 1, p. 57–62, 2005.

RENFREW, P. D.; CHOI, E. J.; BONNEAU, R.; KUHLMAN, B. Incorporation of noncanonical amino acids into rosetta and use in computational protein-peptide interface design. *PLoS One*, Public Library of Science San Francisco, USA, v. 7, n. 3, p. e32637, 2012.

REVETS, H.; BAETSELIER, P. D.; MUYLDERMANS, S. Nanobodies as novel agents for cancer therapy. *Expert opinion on biological therapy*, Taylor & Francis, v. 5, n. 1, p. 111–124, 2005.

REYNOLDS, C. A.; ESSEX, J. W.; RICHARDS, W. G. Atomic charges for variable molecular conformations. *Journal of the American Chemical Society*, ACS Publications, v. 114, n. 23, p. 9075–9079, 1992.

RINIKER, S. Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: an overview. *Journal of chemical information and modeling*, ACS Publications, v. 58, n. 3, p. 565–578, 2018.

ROBBIANI, D. F.; GAEBLER, C.; MUECKSCH, F.; LORENZI, J. C.; WANG, Z.; CHO, A.; AGUDELO, M.; BARNES, C. O.; GAZUMYAN, A.; FINKIN, S. et al. Convergent antibody responses to sars-cov-2 in convalescent individuals. *Nature*, Nature Publishing Group UK London, v. 584, n. 7821, p. 437–442, 2020.

ROBUSTELLI, P.; PIANA, S.; SHAW, D. E. Mechanism of coupled folding-upon-binding of an intrinsically disordered protein. *Journal of the American Chemical Society*, ACS Publications, v. 142, n. 25, p. 11092–11101, 2020.

ROHL, C. A.; STRAUSS, C. E.; MISURA, K. M.; BAKER, D. Protein structure prediction using rosetta. In: *Methods in enzymology*. [S.l.: s.n.], 2004. v. 383, p. 66–93.

ROMERO, P. A.; ARNOLD, F. H. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology*, Nature Publishing Group UK London, v. 10, n. 12, p. 866–876, 2009.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.

ROUET, R.; DUDGEON, K.; CHRISTIE, M.; LANGLEY, D.; CHRIST, D. Fully human vh single domains that rival the stability and cleft recognition of camelid antibodies. *Journal of Biological Chemistry*, ASBMB, v. 290, n. 19, p. 11905–11917, 2015.

RUDORFF, G. F. von; LILIENFELD, O. A. von. Simplifying inverse materials design problems for fixed lattices with alchemical chirality. *Science Advances*, American Association for the Advancement of Science, v. 7, n. 21, p. eabf1173, 2021.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. *Learning internal representations by error propagation*. [S.l.], 1985.

SALONEN, L. M.; ELLERMANN, M.; DIEDERICH, F. Aromatic rings in chemical and biological recognition: energetics and structures. *Angewandte Chemie International Edition*, Wiley Online Library, v. 50, n. 21, p. 4808–4842, 2011.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, v. 3, n. 3, p. 210–229, 1959.

SANCHEZ-LENGELING, B.; ASPURU-GUZIK, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, American Association for the Advancement of Science, v. 361, n. 6400, p. 360–365, 2018.

SCHAFER, T. M.; SETTANNI, G. Data reweighting in metadynamics simulations. *Journal of chemical theory and computation*, ACS Publications, v. 16, n. 4, p. 2042–2052, 2020.

SCHEFFNER, M.; HUIBREGTSE, J. M.; VIERSTRA, R. D.; HOWLEY, P. M. The hpv-16 e6 and e6-ap complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell*, Elsevier, v. 75, n. 3, p. 495–505, 1993.

SCHIFFMAN, M.; CASTLE, P. E.; JERONIMO, J.; RODRIGUEZ, A. C.; WACHOLDER, S. Human papillomavirus and cervical cancer. *The lancet*, Elsevier, v. 370, n. 9590, p. 890–907, 2007.

SCHIFFMAN, M.; CLIFFORD, G.; BUONAGURO, F. M. Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline. *Infectious agents and cancer*, BioMed Central, v. 4, n. 1, p. 1–8, 2009.

SCHLICK, T. *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*. [S.l.]: Springer Science & Business Media, 2010.

SCHMID, N.; EICHENBERGER, A. P.; CHOUTKO, A.; RINIKER, S.; WINGER, M.; MARK, A. E.; GUNSTEREN, W. F. van. Definition and testing of the gromos force-field versions 54a7 and 54b7. *European biophysics journal*, Springer, v. 40, n. 7, p. 843–856, 2011.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural networks*, Elsevier, v. 61, p. 85–117, 2015.

SCHMIDT, M. W.; BALDRIDGE, K. K.; BOATZ, J. A.; ELBERT, S. T.; GORDON, M. S.; JENSEN, J. H.; KOSEKI, S.; MATSUNAGA, N.; NGUYEN, K. A.; SU, S. et al. General atomic and molecular electronic structure system. *Journal of computational chemistry*, Wiley Online Library, v. 14, n. 11, p. 1347–1363, 1993.

SCHOOF, M.; FAUST, B.; SAUNDERS, R. A.; SANGWAN, S.; REZELJ, V.; HOPPE, N.; BOONE, M.; BILLESBØLLE, C. B.; PUCHADES, C.; AZUMAYA, C. M. et al. An ultrapotent synthetic nanobody neutralizes sars-cov-2 by stabilizing inactive spike. *Science*, American Association for the Advancement of Science, v. 370, n. 6523, p. 1473–1479, 2020.

SCHREIBER, G.; SHAUL, Y.; GOTTSCHALK, K. E. Electrostatic design of protein–protein association rates. *Protein Design: Methods and Applications*, Springer, p. 235–249, 2006.

SEIDEL, S. A.; DIJKMAN, P. M.; LEA, W. A.; BOGAART, G. van den; JERABEK-WILLEMSEN, M.; LAZIC, A.; JOSEPH, J. S.; SRINIVASAN, P.; BAASKE, P.; SIMEONOV, A. et al. Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. *Methods*, Elsevier, v. 59, n. 3, p. 301–315, 2013.

SEIDLER, C. A.; KOKOT, J.; FERNÁNDEZ-QUINTERO, M. L.; LIEDL, K. R. Structural characterization of nanobodies during germline maturation. *Biomolecules*, MDPI, v. 13, n. 2, p. 380, 2023.

SENIOR, A. W.; EVANS, R.; JUMPER, J.; KIRKPATRICK, J.; SIFRE, L.; GREEN, T.; QIN, C.; ŽÍDEK, A.; NELSON, A. W.; BRIDGLAND, A. et al. Improved protein structure prediction using potentials from deep learning. *Nature*, Nature Publishing Group UK London, v. 577, n. 7792, p. 706–710, 2020.

SHAJAHAN, A.; SUPEKAR, N. T.; GLEINICH, A. S.; AZADI, P. Deducing the n-and o-glycosylation profile of the spike protein of novel coronavirus sars-cov-2. *Glycobiology*, Oxford University Press, v. 30, n. 12, p. 981–988, 2020.

SHAN, Y.; KIM, E. T.; EASTWOOD, M. P.; DROR, R. O.; SEELIGER, M. A.; SHAW, D. E. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, ACS Publications, v. 133, n. 24, p. 9181–9183, 2011.

SHAO, Q.; ZHU, W. Exploring the ligand binding/unbinding pathway by selectively enhanced sampling of ligand in a protein–ligand complex. *The Journal of Physical Chemistry B*, ACS Publications, v. 123, n. 38, p. 7974–7983, 2019.

SHRINGARI, S. R.; GIANNAKOULIAS, S.; FERRIE, J. J.; PETERSSON, E. J. Rosetta custom score functions accurately predict $\delta\delta$ g of mutations at protein–protein interfaces using machine learning. *Chemical Communications*, Royal Society of Chemistry, v. 56, n. 50, p. 6774–6777, 2020.

SIGFRIDSSON, E.; RYDE, U. Comparison of methods for deriving atomic charges from the electrostatic potential and moments. *Journal of Computational Chemistry*, Wiley Online Library, v. 19, n. 4, p. 377–395, 1998.

SILVA, D.-A.; CORREIA, B. E.; PROCKO, E. Motif-driven design of protein–protein interfaces. In: *Computational Design of Ligand Binding Proteins*. [S.l.]: Springer, 2016. p. 285–304.

SINGH, R. K.; TIWARI, M. K.; SINGH, R.; LEE, J.-K. From protein engineering to immobilization: promising strategies for the upgrade of industrial enzymes. *International journal of molecular sciences*, Molecular Diversity Preservation International (MDPI), v. 14, n. 1, p. 1232–1277, 2013.

SITTEL, F.; STOCK, G. Perspective: Identification of collective variables and metastable states of protein dynamics. *The Journal of chemical physics*, AIP Publishing LLC, v. 149, n. 15, p. 150901, 2018.

SMITH, C. A.; KORTEMME, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of molecular biology*, v. 380, n. 4, p. 742–756, 2008.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and computing*, Springer, v. 14, n. 3, p. 199–222, 2004.

SO, K. A.; KIM, M. J.; LEE, K.-H.; LEE, I.-H.; KIM, M. K.; LEE, Y. K.; HWANG, C.-S.; JEONG, M. S.; KEE, M.-K.; KANG, C. et al. The impact of high-risk hpv genotypes other than hpv 16/18 on the natural course of abnormal cervical cytology: a korean hpv cohort study. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, Korean Cancer Association, v. 48, n. 4, p. 1313–1320, 2016.

SOARES, T. A.; NUNES-ALVES, A.; MAZZOLARI, A.; RUGGIU, F.; WEI, G.-W.; MERZ, K. *The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods*. [S.l.]: ACS Publications, 2022. 5317–5320 p.

SOHRABY, F.; NUNES-ALVES, A. Advances in computational methods for ligand binding kinetics. *Trends in Biochemical Sciences*, Elsevier, 2022.

SOLER, M. A.; FORTUNA, S.; MARCO, A. D.; LAIO, A. Binding affinity prediction of nanobody–protein complexes by scoring of molecular dynamics trajectories. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry, v. 20, n. 5, p. 3438–3444, 2018.

SONG, K.; XU, H.; DA, L.-T. Loading dynamics of one sars-cov-2-derived peptide into mhc-ii revealed by kinetic models. *Biophysical Journal*, Elsevier, 2023.

SONODA, M. T.; MARTÍNEZ, L.; WEBB, P.; SKAF, M. S.; POLIKARPOV, I. Ligand dissociation from estrogen receptor is mediated by receptor dimerization: evidence from molecular dynamics simulations. *Molecular endocrinology*, Oxford University Press, v. 22, n. 7, p. 1565–1578, 2008.

SOÑORA, M.; MARTINEZ, L.; PANTANO, S.; MACHADO, M. R. Wrapping up viruses at multiscale resolution: optimizing packmol and sirah execution for simulating the zika virus. *Journal of Chemical Information and Modeling*, ACS Publications, v. 61, n. 1, p. 408–422, 2021.

SPAAR, A.; DAMMER, C.; GABDOULLINE, R. R.; WADE, R. C.; HELMS, V. Diffusional encounter of barnase and barstar. *Biophysical journal*, Elsevier, v. 90, n. 6, p. 1913–1924, 2006.

SRINIVASAN, J.; CHEATHAM, T. E.; CIEPLAK, P.; KOLLMAN, P. A.; CASE, D. A. Continuum solvent studies of the stability of dna, rna, and phosphoramidate- dna helices. *Journal of the American Chemical Society*, ACS Publications, v. 120, n. 37, p. 9401–9409, 1998.

SRINIVASAN, J.; MILLER, J.; KOLLMAN, P. A.; CASE, D. A. Continuum solvent studies of the stability of rna hairpin loops and helices. *Journal of Biomolecular Structure and Dynamics*, Taylor & Francis, v. 16, n. 3, p. 671–682, 1998.

STEIPE, B.; SCHILLER, B.; PLÜCKTHUN, A.; STEINBACHER, S. *Sequence statistics reliably predict stabilizing mutations in a protein domain*. [S.l.]: Elsevier, 1994. 188–192 p.

STOLL, H.; METZ, B.; DOLG, M. Relativistic energy-consistent pseudopotentials—recent developments. *Journal of computational chemistry*, Wiley Online Library, v. 23, n. 8, p. 767–778, 2002.

STRANGES, P. B.; KUHLMAN, B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Science*, Wiley Online Library, v. 22, n. 1, p. 74–82, 2013.

STROKAPPE, N. M.; HOCK, M.; RUTTEN, L.; MCCOY, L. E.; BACK, J. W.; CAILLAT, C.; HAFFKE, M.; WEISS, R. A.; WEISSENHORN, W.; VERRIPS, T. Super potent bispecific llama vhh antibodies neutralize hiv via a combination of gp41 and gp120 epitopes. *Antibodies*, MDPI, v. 8, n. 2, p. 38, 2019.

SUGITA, Y.; OKAMOTO, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, Elsevier, v. 314, n. 1-2, p. 141–151, 1999.

SWANSON, J. M.; HENCHMAN, R. H.; MCCAMMON, J. A. Revisiting free energy calculations: a theoretical connection to mm/pbsa and direct calculation of the association free energy. *Biophysical journal*, Elsevier, v. 86, n. 1, p. 67–74, 2004.

SZABO, A.; OSTLUND, N. S. *Modern quantum chemistry: introduction to advanced electronic structure theory*. [S.l.]: Courier Corporation, 2012.

SZTAIN, T.; AHN, S.-H.; BOGETTI, A. T.; CASALINO, L.; GOLDSMITH, J. A.; SEITZ, E.; MCCOOL, R. S.; KEARNS, F. L.; ACOSTA-REYES, F.; MAJI, S. et al. A glycan gate controls opening of the sars-cov-2 spike protein. *Nature chemistry*, Nature Publishing Group, v. 13, n. 10, p. 963–968, 2021.

TAKA, E.; YILMAZ, S. Z.; GOLCUK, M.; KILINC, C.; AKTAS, U.; YILDIZ, A.; GUR, M. Critical interactions between the sars-cov-2 spike glycoprotein and the human ace2 receptor. *The Journal of Physical Chemistry B*, ACS Publications, v. 125, n. 21, p. 5537–5548, 2021.

TANFORD, C.; KIRKWOOD, J. G. Theory of protein titration curves. i. general equations for impenetrable spheres. *Journal of the American Chemical Society*, ACS Publications, v. 79, n. 20, p. 5333–5339, 1957.

TANG, B.; PAN, Z.; YIN, K.; KHATEEB, A. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in genetics*, Frontiers Media SA, v. 10, p. 214, 2019.

THOMSEN, M. C. F.; NIELSEN, M. Seq2logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic acids research*, Oxford University Press, v. 40, n. W1, p. W281–W287, 2012.

TIANA, G. Estimation of microscopic averages from metadynamics. *The European Physical Journal B*, Springer, v. 63, n. 2, p. 235–238, 2008.

TILLER, K. E.; CHOWDHURY, R.; LI, T.; LUDWIG, S. D.; SEN, S.; MARANAS, C. D.; TESSIER, P. M. Facile affinity maturation of antibody variable domains using natural diversity mutagenesis. *Frontiers in immunology*, Frontiers Media SA, v. 8, p. 986, 2017.

TIRONI, I. G.; SPERB, R.; SMITH, P. E.; GUNSTEREN, W. F. van. A generalized reaction field method for molecular dynamics simulations. *The Journal of chemical physics*, AIP, v. 102, n. 13, p. 5451–5459, 1995.

TOMAIĆ, V.; PIM, D.; BANKS, L. The stability of the human papillomavirus e6 oncoprotein is e6ap dependent. *Virology*, Elsevier, v. 393, n. 1, p. 7–10, 2009.

TORRE, J. G. de la; HUERTAS, M. L.; CARRASCO, B. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophysical journal*, Elsevier, v. 78, n. 2, p. 719–730, 2000.

TORRIE, G. M.; VALLEAU, J. P. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, Elsevier, v. 23, n. 2, p. 187–199, 1977.

TRIBELLO, G. A.; BONOMI, M.; BRANDUARDI, D.; CAMILLONI, C.; BUSSI, G. Plumed 2: New feathers for an old bird. *Computer physics communications*, Elsevier, v. 185, n. 2, p. 604–613, 2014.

TSO, S.-C.; CHEN, Q.; VISHNIVETSKIY, S. A.; GUREVICH, V. V.; IVERSON, T.; BRAUTIGAM, C. A. Using two-site binding models to analyze microscale thermophoresis data. *Analytical biochemistry*, Elsevier, v. 540, p. 64–75, 2018.

TUCKERMAN, M. *Statistical mechanics: theory and molecular simulation*. [S.l.]: Oxford university press, 2010.

TUROŇOVÁ, B.; SIKORA, M.; SCHÜRMANN, C.; HAGEN, W. J.; WELSCH, S.; BLANC, F. E.; BÜLOW, S. von; GECHT, M.; BAGOLA, K.; HÖRNER, C. et al. In situ structural analysis of sars-cov-2 spike reveals flexibility mediated by three hinges. *Science*, American Association for the Advancement of Science, v. 370, n. 6513, p. 203–208, 2020.

TZOU, P. L.; TAO, K.; NOUHIN, J.; RHEE, S.-Y.; HU, B. D.; PAI, S.; PARKIN, N.; SHAFER, R. W. Coronavirus antiviral research database (cov-rdb): an online database designed to facilitate comparisons between candidate anti-coronavirus compounds. *Viruses*, MDPI, v. 12, n. 9, p. 1006, 2020.

VALENZUELA-NIETO, G.; MIRANDA-CHACON, Z.; SALINAS-REBOLLEDO, C.; JARA, R.; CUEVAS, A.; BERKING, A.; ROJAS-FERNANDEZ, A. Nanobodies: Covid-19 and future perspectives. *Frontiers in Drug Discovery*, Frontiers, p. 14, 2022.

VALIEV, M.; BYLASKA, E. J.; GOVIND, N.; KOWALSKI, K.; STRAATSMA, T. P.; DAM, H. J. J. V.; WANG, D.; NIEPLOCHA, J.; APRÀ, E.; WINDUS, T. L. et al. Nwchem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Computer Physics Communications*, Elsevier, v. 181, n. 9, p. 1477–1489, 2010.

VANGONE, A.; BONVIN, A. M. Contacts-based prediction of binding affinity in protein–protein complexes. *eLife*, eLife Sciences Publications, Ltd, v. 4, p. e07454, jul 2015. ISSN 2050-084X. Available at: <https://doi.org/10.7554/eLife.07454>.

VAPNIK, V.; VAPNIK, V. Statistical learning theory. *New York*, v. 1, n. 624, p. 2, 1998.

VARNAI, A. D.; BOLLMANN, M.; BANKFALVI, A.; GRIEFINGHOLT, H.; PFENING, N.; SCHMITT, C.; PAJOR, L.; BOLLMANN, R. The spectrum of cervical diseases induced by low-risk and undefined-risk hpvs: implications for patient management. *Anticancer research*, International Institute of Anticancer Research, v. 27, n. 1B, p. 563–570, 2007.

VAUQUELIN, G.; BOSTOEN, S.; VANDERHEYDEN, P.; SEEMAN, P. Clozapine, atypical antipsychotics, and the benefits of fast-off d 2 dopamine receptor antagonism. *Naunyn-Schmiedeberg's archives of pharmacology*, Springer, v. 385, p. 337–372, 2012.

VERLI, H. *Bioinformática: da biologia à flexibilidade molecular*. [S.l.]: Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

VIANA, I. F.; CRUZ, C. H.; ATHAYDE, D.; ADAN, W. C. S.; XAVIER, L. S.; ARCHER, M.; LINS, R. D. In vitro neutralisation of zika virus by an engineered protein targeting the viral envelope fusion loop. *Molecular Systems Design & Engineering*, Royal Society of Chemistry, 2023.

VIANA, I. F.; SOARES, T. A.; LIMA, L. F.; MARQUES, E. T.; KRIEGER, M. A.; DHALIA, R.; LINS, R. D. De novo design of immunoreactive conformation-specific hiv-1 epitopes based on top7 scaffold. *Rsc Advances*, Royal Society of Chemistry, v. 3, n. 29, p. 11790–11800, 2013.

VILLIERS, E.-M. D.; FAUQUET, C.; BROKER, T. R.; BERNARD, H.-U.; HAUSEN, H. Z. Classification of papillomaviruses. *Virology*, Elsevier, v. 324, n. 1, p. 17–27, 2004.

VINCKE, C.; LORIS, R.; SAERENS, D.; MARTINEZ-RODRIGUEZ, S.; MUYLDERMANS, S.; CONRATH, K. General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold. *Journal of Biological Chemistry*, ASBMB, v. 284, n. 5, p. 3273–3284, 2009.

VINCKE, C.; MUYLDERMANS, S. Introduction to heavy chain antibodies and derived nanobodies. *Single Domain Antibodies: Methods and Protocols*, Springer, p. 15–26, 2012.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; van der Walt, S. J.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; POLAT, İ.; FENG, Y.; MOORE, E. W.; VanderPlas, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020.

VIVO, M. D.; MASETTI, M.; BOTTEGONI, G.; CAVALLI, A. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry*, ACS Publications, v. 59, n. 9, p. 4035–4061, 2016.

WALLS, A. C.; PARK, Y.-J.; TORTORICI, M. A.; WALL, A.; MCGUIRE, A. T.; VEESLER, D. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, Elsevier, v. 181, n. 2, p. 281–292, 2020.

WALLS, A. C.; XIONG, X.; PARK, Y.-J.; TORTORICI, M. A.; SNIJDER, J.; QUISPE, J.; CAMERONI, E.; GOPAL, R.; DAI, M.; LANZAVECCHIA, A. et al. Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. *Cell*, Elsevier, v. 176, n. 5, p. 1026–1039, 2019.

WALMSLEY, T.; ROSE, A.; WEI, D. The impacts of the coronavirus on the economy of the united states. *Economics of disasters and climate change*, Springer, v. 5, n. 1, p. 1–52, 2021.

WANG, C.; BRADLEY, P.; BAKER, D. Protein–protein docking with backbone flexibility. *Journal of molecular biology*, Elsevier, v. 373, n. 2, p. 503–519, 2007.

WANG, J.; BHATTARAI, A.; DO, H. N.; MIAO, Y. Challenges and frontiers of computational modelling of biomolecular recognition. *QRB Discovery*, Cambridge University Press, v. 3, p. e13, 2022.

WANG, J.; CAI, Q.; LI, Z.-L.; ZHAO, H.-K.; LUO, R. Achieving energy conservation in poisson–boltzmann molecular dynamics: Accuracy and precision with finite-difference algorithms. *Chemical physics letters*, Elsevier, v. 468, n. 4-6, p. 112–118, 2009.

WANG, J.; CIEPLAK, P.; KOLLMAN, P. A. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *Journal of computational chemistry*, Wiley Online Library, v. 21, n. 12, p. 1049–1074, 2000.

WANG, J.; DO, H. N.; KOIRALA, K.; MIAO, Y. Predicting biomolecular binding kinetics: A review. *Journal of Chemical Theory and Computation*, v. 0, n. 0, p. null, 2023. PMID: 36989090. Available at: <https://doi.org/10.1021/acs.jctc.2c01085>.

WANG, J.; ISHCHENKO, A.; ZHANG, W.; RAZAVI, A.; LANGLEY, D. A highly accurate metadynamics-based dissociation free energy method to calculate protein–protein and protein–ligand binding potencies. *Scientific reports*, Nature Publishing Group, v. 12, n. 1, p. 1–11, 2022.

WANG, J.; LISANZA, S.; JUERGENS, D.; TISCHER, D.; WATSON, J. L.; CASTRO, K. M.; RAGOTTE, R.; SARAGOVI, A.; MILLES, L. F.; BAEK, M. et al. Scaffolding protein functional sites using deep learning. *Science*, American Association for the Advancement of Science, v. 377, n. 6604, p. 387–394, 2022.

WANG, J.; MIAO, Y. Peptide gaussian accelerated molecular dynamics (pep-gamd): Enhanced sampling and free energy and kinetics calculations of peptide binding. *The Journal of Chemical Physics*, AIP Publishing LLC, v. 153, n. 15, p. 154109, 2020.

WANG, L.; WU, Y.; DENG, Y.; KIM, B.; PIERCE, L.; KRILOV, G.; LUPYAN, D.; ROBINSON, S.; DAHLGREN, M. K.; GREENWOOD, J. et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern

free-energy calculation protocol and force field. *Journal of the American Chemical Society*, ACS Publications, v. 137, n. 7, p. 2695–2703, 2015.

WANG, R.; FANG, X.; LU, Y.; WANG, S. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, ACS Publications, v. 47, n. 12, p. 2977–2980, 2004.

WASKOM, M. L. seaborn: statistical data visualization. *Journal of Open Source Software*, The Open Journal, v. 6, n. 60, p. 3021, 2021. Available at: <https://doi.org/10.21105/joss.03021>.

WATANABE, Y.; ALLEN, J. D.; WRAPP, D.; MCLELLAN, J. S.; CRISPIN, M. Site-specific glycan analysis of the sars-cov-2 spike. *Science*, American Association for the Advancement of Science, v. 369, n. 6501, p. 330–333, 2020.

WEDDERBURN, R. W. Maximum likelihood estimation by iterative refinement. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 22, n. 1, p. 69–82, 1960.

WEIGEND, F.; AHLRICHS, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry, v. 7, n. 18, p. 3297–3305, 2005.

WEINER, S. J.; KOLLMAN, P. A.; CASE, D. A.; SINGH, U. C.; GHIO, C.; ALAGONA, G.; PROFETA, S.; WEINER, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, ACS Publications, v. 106, n. 3, p. 765–784, 1984.

WIECZOREK, M.; ABUALROUS, E. T.; STICHT, J.; ÁLVARO-BENITO, M.; STOLZENBERG, S.; NOÉ, F.; FREUND, C. Major histocompatibility complex (mhc) class i and mhc class ii proteins: conformational plasticity in antigen presentation. *Frontiers in immunology*, Frontiers Media SA, v. 8, p. 292, 2017.

WOO, H.-J.; ROUX, B. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 102, n. 19, p. 6825–6830, 2005.

WORD, J. M.; LOVELL, S. C.; RICHARDSON, J. S.; RICHARDSON, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology*, Elsevier, v. 285, n. 4, p. 1735–1747, 1999.

WRAPP, D.; VLIEGER, D. D.; CORBETT, K. S.; TORRES, G. M.; WANG, N.; BREEDAM, W. V.; ROOSE, K.; SCHIE, L. van; COVID, V.-C.; TEAM, R. et al. Structural basis for potent neutralization of betacoronaviruses by single-domain camelid antibodies. *Cell*, Elsevier, v. 181, n. 5, p. 1004–1015, 2020.

WRAPP, D.; WANG, N.; CORBETT, K. S.; GOLDSMITH, J. A.; HSIEH, C.-L.; ABIONA, O.; GRAHAM, B. S.; MCLELLAN, J. S. Cryo-em structure of the 2019-ncov spike in the prefusion conformation. *Science*, American Association for the Advancement of Science, v. 367, n. 6483, p. 1260–1263, 2020.

WRENBECK, E. E.; FABER, M. S.; WHITEHEAD, T. A. Deep sequencing methods for protein engineering and design. *Current opinion in structural biology*, Elsevier, v. 45, p. 36–44, 2017.

WU, F.; ZHAO, S.; YU, B.; CHEN, Y.-M.; WANG, W.; SONG, Z.-G.; HU, Y.; TAO, Z.-W.; TIAN, J.-H.; PEI, Y.-Y. et al. A new coronavirus associated with human respiratory disease in china. *Nature*, Nature Publishing Group, v. 579, n. 7798, p. 265–269, 2020.

WU, N. C.; YUAN, M.; BANGARU, S.; HUANG, D.; ZHU, X.; LEE, C.-C. D.; TURNER, H. L.; PENG, L.; YANG, L.; BURTON, D. R. et al. A natural mutation between sars-cov-2 and sars-cov determines neutralization by a cross-reactive antibody. *PLoS pathogens*, Public Library of Science San Francisco, CA USA, v. 16, n. 12, p. e1009089, 2020.

WU, Y.; SCHMITT, J. D.; CAR, R. Mapping potential energy surfaces. *The Journal of chemical physics*, American Institute of Physics, v. 121, n. 3, p. 1193–1200, 2004.

XUE, L. C.; RODRIGUES, J. P.; KASTRITIS, P. L.; BONVIN, A. M.; VANGONE, A. Prodigy: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, Oxford University Press, v. 32, n. 23, p. 3676–3678, 2016.

XUE, W.; FU, T.; DENG, S.; YANG, F.; YANG, J.; ZHU, F. Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram. *ACS chemical neuroscience*, ACS Publications, v. 13, n. 3, p. 340–351, 2022.

XUE, W.; YANG, F.; WANG, P.; ZHENG, G.; CHEN, Y.; YAO, X.; ZHU, F. What contributes to serotonin–norepinephrine reuptake inhibitors' dual-targeting mechanism? the key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS chemical neuroscience*, ACS Publications, v. 9, n. 5, p. 1128–1140, 2018.

YANG, J.; ANISHCHENKO, I.; PARK, H.; PENG, Z.; OVCHINNIKOV, S.; BAKER, D. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 117, n. 3, p. 1496–1503, 2020.

YANG, L.; SHAMI, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, Elsevier, v. 415, p. 295–316, 2020.

YANG, Y. I.; SHAO, Q.; ZHANG, J.; YANG, L.; GAO, Y. Q. Enhanced sampling in molecular dynamics. *The Journal of chemical physics*, AIP Publishing LLC, v. 151, n. 7, p. 070902, 2019.

YU, L.; MAJERCIAK, V.; ZHENG, Z.-M. Hpv16 and hpv18 genome structure, expression, and post-transcriptional regulation. *International journal of molecular sciences*, MDPI, v. 23, n. 9, p. 4943, 2022.

YU, X.; MARTINEZ, M.; GABLE, A. L.; FULLER, J. C.; BRUCE, N. J.; RICHTER, S.; WADE, R. C. websda: a web server to simulate macromolecular diffusional association. *Nucleic Acids Research*, Oxford University Press, v. 43, n. W1, p. W220–W224, 2015.

YUAN, M.; WU, N. C.; ZHU, X.; LEE, C.-C. D.; SO, R. T.; LV, H.; MOK, C. K.; WILSON, I. A. A highly conserved cryptic epitope in the receptor binding domains of sars-cov-2 and sars-cov. *Science*, American Association for the Advancement of Science, v. 368, n. 6491, p. 630–633, 2020.

YUE, K.; DILL, K. A. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 89, n. 9, p. 4163–4167, 1992.

ZAJC, C. U.; DOBERSBERGER, M.; SCHAFFNER, I.; MLYNEK, G.; PÜHRINGER, D.; SALZER, B.; DJINOVIĆ-CARUGO, K.; STEINBERGER, P.; LINHARES, A. D. S.; YANG, N. J. et al. A conformation-specific on-switch for controlling car t cells with an orally available drug. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 117, n. 26, p. 14926–14935, 2020.

ZANIER, K.; CHARBONNIER, S.; SIDI, A. O. M. O.; MCEWEN, A. G.; FERRARIO, M. G.; POUSSIN-COURMONTAGNE, P.; CURA, V.; BRIMER, N.; BABAH, K. O.; ANSARI, T. et al. Structural basis for hijacking of cellular lxxll motifs by papillomavirus e6 oncoproteins. *Science*, American Association for the Advancement of Science, v. 339, n. 6120, p. 694–698, 2013.

ZAVRTANIK, U.; HADŽI, S. A non-redundant data set of nanobody-antigen crystal structures. *Data in brief*, Elsevier, v. 24, p. 103754, 2019.

ZAVRTANIK, U.; LUKAN, J.; LORIS, R.; LAH, J.; HADŽI, S. Structural basis of epitope recognition by heavy-chain camelid antibodies. *Journal of molecular biology*, Elsevier, v. 430, n. 21, p. 4369–4386, 2018.

ZHANG, G.; ANDERSEN, J.; GERONA-NAVARRO, G. Peptidomimetics targeting protein-protein interactions for therapeutic development. *Protein and Peptide Letters*, Bentham Science Publishers, v. 25, n. 12, p. 1076–1089, 2018.

ZHAO, F.; LI, S.; STERNER, B. W.; XU, J. Discriminative learning for protein conformation sampling. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 73, n. 1, p. 228–240, 2008.

ZHENG, Z.-M.; BAKER, C. C. et al. Papillomavirus genome structure, expression, and post-transcriptional regulation. *Front Biosci*, v. 11, n. 1, p. 2286–2302, 2006.

ZHOU, G.; PANTELOPULOS, G. A.; MUKHERJEE, S.; VOELZ, V. A. Bridging microscopic and macroscopic mechanisms of p53-mdm2 binding with kinetic network models. *Biophysical journal*, Elsevier, v. 113, n. 4, p. 785–793, 2017.

ZHOU, H.-X. Rate theories for biologists. *Quarterly reviews of biophysics*, Cambridge University Press, v. 43, n. 2, p. 219–293, 2010.

ZHU, N.; ZHANG, D.; WANG, W.; LI, X.; YANG, B.; SONG, J.; ZHAO, X.; HUANG, B.; SHI, W.; LU, R. et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*, Mass Medical Soc, 2020.

ZIMMERMANN, I.; EGLOFF, P.; HUTTER, C. A.; ARNOLD, F. M.; STOHLER, P.; BOCQUET, N.; HUG, M. N.; HUBER, S.; SIEGRIST, M.; HETEMANN, L. et al. Synthetic single domain antibodies for the conformational trapping of membrane proteins. *elife*, eLife Sciences Publications, Ltd, v. 7, p. e34317, 2018.

ZIMMERMANN, I.; EGLOFF, P.; HUTTER, C. A.; KUHN, B. T.; BRÄUER, P.; NEWSTEAD, S.; DAWSON, R. J.; GEERTSMA, E. R.; SEEGER, M. A. Generation of synthetic nanobodies against delicate proteins. *Nature protocols*, Nature Publishing Group, v. 15, n. 5, p. 1707–1741, 2020.

ZOU, R.; ZHOU, Y.; WANG, Y.; KUANG, G.; ÅGREN, H.; WU, J.; TU, Y. Free energy profile and kinetics of coupled folding and binding of the intrinsically disordered protein p53 with mdm2. *Journal of chemical information and modeling*, ACS Publications, v. 60, n. 3, p. 1551–1558, 2020.

ZWANZIG, R. W. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, American Institute of Physics, v. 22, n. 8, p. 1420–1426, 1954.

ZWIER, M. C.; PRATT, A. J.; ADELMAN, J. L.; KAUS, J. W.; ZUCKERMAN, D. M.; CHONG, L. T. Efficient atomistic simulation of pathways and calculation of rate constants for a protein–peptide binding process: Application to the mdm2 protein and an intrinsically disordered p53 peptide. *The journal of physical chemistry letters*, ACS Publications, v. 7, n. 17, p. 3440–3445, 2016.

ANNEX A – FIRST PAGE OF EACH PUBLICATION FROM THE PH.D. PERIOD

ACS BIO&MED CHEM Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/biomedchemau Article

Inhibition of 3-Hydroxykynurenine Transaminase from *Aedes aegypti* and *Anopheles gambiae*: A Mosquito-Specific Target to Combat the Transmission of Arboviruses

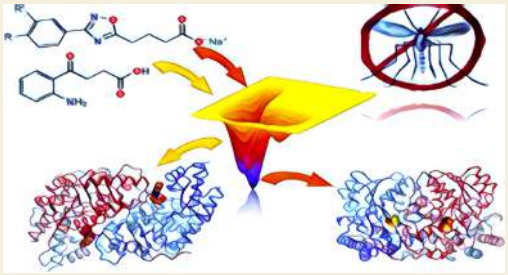
Larissa G. Maciel,[#] Matheus V. F. Ferraz,[#] Andrew A. Oliveira, Roberto D. Lins, Janaína V. dos Anjos, Rafael V. C. Guido,* and Thereza A. Soares*

[#]L.G.M. and M.V.F.F. contributed equally to this work.
[#]L.G.M. and M.V.F.F. contributed equally to this work.

Cite This: *ACS Bio Med Chem Au* 2023, 3, 211–222 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Arboviral infections such as Zika, chikungunya, dengue, and yellow fever pose significant health problems globally. The population at risk is expanding with the geographical distribution of the main transmission vector of these viruses, the *Aedes aegypti* mosquito. The global spreading of this mosquito is driven by human migration, urbanization, climate change, and the ecological plasticity of the species. Currently, there are no specific treatments for *Aedes*-borne infections. One strategy to combat different mosquito-borne arboviruses is to design molecules that can specifically inhibit a critical host protein. We obtained the crystal structure of 3-hydroxykynurenine transaminase (AeHKT) from *A. aegypti*, an essential detoxification enzyme of the tryptophan metabolism pathway. Since AeHKT is found exclusively in mosquitoes, it provides the ideal molecular target for the development of inhibitors. Therefore, we determined and compared the free binding energy of the inhibitors 4-(2-aminophenyl)-4-oxobutyric acid (4OB) and sodium 4-(3-phenyl-1,2,4-oxadiazol-5-yl)butanoate (OXA) to AeHKT and AgHKT from *Anopheles gambiae*, the only crystal structure of this enzyme previously known. The cocrystallized inhibitor 4OB binds to AgHKT with K_i of 300 μM. We showed that OXA binds to both AeHKT and AgHKT enzymes with binding energies 2-fold more favorable than the crystallographic inhibitor 4OB and displayed a 2-fold greater residence time τ upon binding to AeHKT than 4OB. These findings indicate that the 1,2,4-oxadiazole derivatives are inhibitors of the HKT enzyme not only from *A. aegypti* but also from *A. gambiae*.



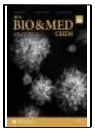
KEYWORDS: *noncompetitive inhibitor, metadynamics simulations, τRAMD, unbinding kinetics,, binding affinity, binding efficiency*

INTRODUCTION

The *Aedes aegypti* mosquito is the main transmission vector for several viruses, including the urban yellow fever (YFV), dengue (DENV), chikungunya (CHIKV), and Zika (ZIKV) viruses responsible for high rates of morbidity and mortality in tropical regions around the globe.[1] Just for the dengue fever disease, it is estimated a total of 390 million virus infections per year worldwide (95% credible interval 284−528 million), of which 96 million (67−136 million) manifest clinically.[2] *A. aegypti* is closely associated with human habitation, thriving in densely populated regions. The female mosquito not only feeds on humans but also prefers to lay eggs in artificial water containers (e.g., water tanks, flower vases, pot plant bases, discarded tires) typically found around or inside houses.[3,4] Furthermore, eggs can withstand desiccation for up to one year.[3] Although *A. aegypti* is intolerant to temperate winters and its eggs are sensitive to frost, the current global climate changes may expand the geographical distribution of this disease vector.[5,6] For instance, the European Centre for Disease Prevention and Disease Control considers that, with the increase on average temperatures, the coastal regions of the Mediterranean, Black Sea, Caspian Sea, and areas along large lowland rivers (Ebro, Garonne, Rhone, and Po) can become suitable habitats for *A. Aegypti*.[7]

As the effective immunization against arboviruses and all their many serotypes is not currently available,[1,8,9] populational control of vector species becomes the most effective and affordable way to prevent disease transmission. This is particularly important because some arboviruses (e.g.,

Received: November 30, 2022
Revised: February 7, 2023
Accepted: February 7, 2023
Published: February 16, 2023



https://doi.org/10.1021/acsbiomedchemau.2c00080
ACS Bio Med Chem Au 2023, 3, 211–222

PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d2cp05644e

An artificial neural network model to predict structure-based protein–protein free energy of binding from Rosetta-calculated properties†

Matheus V. F. Ferraz, [abc] José C. S. Neto, [d] Roberto D. Lins [ab] and Erico S. Teixeira *[d]

Received 3rd December 2022,
Accepted 29th January 2023

DOI: 10.1039/d2cp05644e

rsc.li/pccp

The prediction of the free energy (ΔG) of binding for protein–protein complexes is of general scientific interest as it has a variety of applications in the fields of molecular and chemical biology, materials science, and biotechnology. Despite its centrality in understanding protein association phenomena and protein engineering, the ΔG of binding is a daunting quantity to obtain theoretically. In this work, we devise a novel Artificial Neural Network (ANN) model to predict the ΔG of binding for a given three-dimensional structure of a protein–protein complex with Rosetta-calculated properties. Our model was tested using two data sets, and it presented a root-mean-square error ranging from 1.67 kcal mol^{-1} to 2.45 kcal mol^{-1}, showing a better performance compared to the available state-of-the-art tools. Validation of the model for a variety of protein–protein complexes is showcased.

1 Introduction

Protein–protein association is a pervasive phenomenon in physiological and biotechnological processes, ranging from mechanisms with cell interactions and disease modulation to industrial applications of metabolic control.[1–5] To characterize this association, Gibbs free energy (ΔG) of binding is one of the most fundamental thermodynamic quantities. In addition to the comprehension of biomolecular processes, the development of novel biomaterials (*e.g.*, towards biomedical,[6] vaccinal,[7,8] catalytic,[9] development of biosensors,[10] and industrial applications[11]) relies substantially on the binding affinity (k_D), which is directly related to the ΔG between the involved binding partners ($\Delta G = -RT\ln k_D$, where T is the absolute temperature and R is the universal gas constant).

Although the ΔG of binding is a central concern for protein modeling and design, and the calculations by computational means have been an active area of development since the beginning of the 1980s,[12,13] obtaining it accurately at a low computational cost remains a challenge for the molecular simulation community.[14] This difficulty is even more pronounced when trying to obtain the ΔG for liquids and flexible macromolecules, mainly due to insufficient sampling and inaccuracies in the description of the potential energy surface.[15]

However, despite the challenge, methods like molecular dynamics and Monte Carlo simulations afford a range of rigorous approaches such as thermodynamic integration or free energy perturbation,[16] which provide accurate results. In these techniques, nonphysical pathways *via* alchemical methods (*e.g.*, thermodynamic integration and free energy perturbation)[17] are simulated by connecting two end states by a coupling parameter. Even though these methods offer accurate values, for example, achieving an error of *ca.* 1.0 kcal mol^{-1},[18,19] a major hurdle is the amount of sampling needed to simulate the regions of the phase space that make important contributions to the ΔG. On the other hand, end-point state methods, such as molecular mechanics generalized Born surface area (MM-GBSA) and molecular mechanics Poisson Boltzmann surface area (MM-PBSA),[20–22] have been used extensively to compute the ΔG of binding,[23,24] as they exempt the need for simulating intermediate states as in alchemical methods and make use of implicit solvation, reducing the computational cost. Despite that, to account for entropic changes in MM-PBSA and MM-GBSA, typically, the standard normal mode is used, which is time-consuming and approximate.[25,26] Thus, in many studies, the entropic contribution is often neglected, leading to inconsistent results. Aiming at the improvement of computational costs and maintaining accuracy, enhanced sampling methods (*e.g.*, metadynamics[27] and umbrella sampling[28]) simulate a

[a] Department of Virology, Aggeu Magalhães Institute, Oswaldo Cruz Foundation, FIOCRUZ, Recife, PE, Brazil
[b] Department of Fundamental Chemistry, Federal University of Pernambuco, UFPE, Recife, PE, Brazil
[c] Heidelberg Institute for Theoretical Studies, HITS, Heidelberg, Germany
[d] Recife Center for Advanced Studies and Systems, CESAR, Recife, PE, Brazil. E-mail: est@cesar.school; Tel: +55-81-2123-7848

† Electronic supplementary information (ESI) available: Parsed command lines, Rosetta scripts, polar atom definitions, supplementary figures, and tables (Tables S1–S3 and Fig. S1–S5). See DOI: https://doi.org/10.1039/d2cp05644e

Received: 17 May 2022 | Revised: 26 July 2022 | Accepted: 1 August 2022
DOI: 10.1002/bip.23524

ARTICLE

Biopolymers WILEY

Association strength of E6 to E6AP/p53 complex correlates with HPV-mediated oncogenesis risk

Matheus Vitor Ferreira Ferraz[1,2] | Isabelle Freire Tabosa Viana[1] | Danilo Fernandes Coêlho[1,2] | Carlos Henrique Bezerra da Cruz[3] | Maíra de Arruda Lima[1] | Madson Allan de Luna Aragão[1] | Roberto Dias Lins[1,2]

[1]Aggeu Magalhães Institute, Oswaldo Cruz Foundation, Recife, Brazil

[2]Department of Fundamental Chemistry, Federal University of Pernambuco, Recife, Brazil

[3]Institute of Chemical and Biological Technology António Xavier, NOVA University Lisbon, Lisbon, Portugal

Correspondence
Roberto Dias Lins, Aggeu Magalhães Institute, Oswaldo Cruz Foundation, Recife, PE, 50.740-465, Brazil.
Email: roberto.neto@fiocruz.br

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Numbers: 303001/2018-6, 425997/2018-9; Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco, Grant/Award Number: APQ-0346-2.09/19; Fundação Oswaldo Cruz, Grant/Award Number: VPPCB-007-FIO-18-2-134; INCT-FCx, Grant/Award Number: 465259/2014-6

Abstract

Human papillomavirus (HPV) is recognized as the causative agent of cervical cancer in women, and it is associated with other anogenital and head/neck cancers. More than 120 types of HPV have been identified and many classified as high- or low-risk according to their oncogenic potential. One of its proteins, E6, has evolved to overcome the oncosuppressor functions of p53 by targeting this protein for degradation via interaction with the human ubiquitin-ligase E6AP. This study evaluates the correlation between the association strength of 40 HPV E6 types to the E6AP/p53 complex and the HPV oncogenesis risk using molecular simulations and machine and deep learning (ML/DL). In addition, a ML/DL-driven prediction is proposed for the HPV unclassified oncogenic risk type. The results indicate that thermodynamics play a pivotal role in the establishment of HPV-associated cancer and highlight the need to include some viral types in the HPV-related cancer surveillance and prevention strategies.

KEYWORDS
binding free energy, cervical cancer, human papillomavirus, machine learning, molecular dynamics

1 | INTRODUCTION

Papillomaviruses belong to the Papillomaviridae family and are classified into five genera (alpha-, beta- gamma, mu-, and nu-papillomavirus) encompassing 49 species and more than 200 viral types.[1] Human papillomaviruses (HPVs) are human parasites with tropism for squamous epithelium.[2,3] Over 120 types of HPVs have been identified and approximately one-third of those infect the squamous epithelia of the genital tract.[4] Among those, 15 are categorized as high risk and are considered the major cause of cervical cancer among women all over the world, with over 99% of cervical lesions containing viral sequences.[5] Within the high-risk group, the HPV16 and 18 types are the most prevalent ones, followed by the 31, 33, 35, 42, 52, and 58 types.[6] High-risk HPVs are also associated with many penile, vulvar, and anal carcinomas and contribute to 40% of oral cancers.[5] The remaining viral types are usually associated with benign lesions, such as warts and condylomas, and therefore are categorized as low-risk types.[7]

Human papillomavirus are non-enveloped, double-stranded DNA viruses with approximately 8 kb in size. Transcription is initiated from more than one promoter region and is polycistronic, yielding multiple mRNAs with several open reading frames (ORFs) divided into long control region (LCR), late (L), and early (E) ORFs.[8,9] The LCR region encompasses the gene regulation elements, such as promoters and transcription regulatory elements. The HPV genomes do not encode polymerases nor other enzymes required for proliferation and therefore depend on the host cell machinery to mediate viral DNA synthesis.[5] Late ORFs (L1 and L2) encode the viral capsid structural proteins, which are expressed in the final stages of cellular differentiation, allowing for virus particle assembling and release in the extracellular medium.[10–14] The early ORFs are named E1, E2, E4, E5, E6 and E7 and encode the viral cycle regulatory proteins at the early stages

Identification of potential *Staphylococcus aureus* dihydrofolate reductase inhibitors using QSAR, molecular docking, dynamics simulations and free energy calculation

Isaac de Araujo Matos[a], Ana Carolina Goes Pinto[a], Wenny Camilla Santos Adan[a], Ricardo Pereira Rodrigues[b], Juliane Xavier dos Santos[a], Rodrigo Rezende Kitagawa[b], Tiago Branquinho Oliveira[c] and Nivan Bezerra da Costa Junior[a].*

[a] Department of Chemistry, Graduate Program in Chemistry, Federal University of Sergipe-UFS, São Cristóvão-SE, Brazil; [b] Department of Pharmaceutical Sciences, Postgraduate Program in Pharmaceutical Sciences, Federal University of Espírito Santo-UFES, Vitória-ES, Brazil; [c] Department of Pharmacy, Graduate Program in Chemistry, Federal University of Sergipe-UFS, São Cristóvão-SE, Brazil.

*To whom correspondence should be addressed:
Nívan Bezerra da Costa Júnior or Isaac de Araújo Matos

Department of Chemistry, Graduate Program in Chemistry, Federal University of Sergipe – UFS, Sergipe, Brazil.

Av. Marechal Rondon, s/n, Jd. Rosa Elze.

Postal code 49100-000, Tel.: +55 79 3194-6600

Email: nivanjr@gmail.com, isaacbioquim@usp.br

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

Spread of Gamma (P.1) Sub-Lineages Carrying Spike Mutations Close to the Furin Cleavage Site and Deletions in the N-Terminal Domain Drives Ongoing Transmission of SARS-CoV-2 in Amazonas, Brazil

Felipe Gomes Naveca,[a,b] Valdinete Nascimento,[a] Victor Souza,[a] André de Lima Corado,[a] Fernanda Nascimento,[a] George Silva,[a,c] Matilde Contreras Mejía,[a] Maria Júlia Brandão,[a] Ágatha Costa,[a] Débora Duarte,[a] Karina Pessoa,[a] Michele Jesus,[d] Luciana Gonçalves,[a,e] Cristiano Fernandes,[e] Tirza Mattos,[f] Ligia Abdalla,[g] João Hugo Santos,[h] Alex Martins,[g] Fabiola Mendonça Chui,[g] Fernando Fonseca Val,[i] Gisely Cardoso de Melo,[g,i] Mariana Simão Xavier,[i,j] Vanderson de Souza Sampaio,[e,i] Maria Paula Mourão,[g,i] Marcus Vinícius Lacerda,[i,k] Érika Lopes Rocha Batista,[l] Alessandro Leonardo Álvares Magalhães,[l] Nathânia Dábilla,[m] Lucas Carlos Gomes Pereira,[n] Fernando Vinhal,[n] Fabio Miyajima,[o] Fernando Braga Stehling Dias,[o] Eduardo Ruback dos Santos,[p] Danilo Coêlho,[q] Matheus Ferraz,[q] Roberto Lins,[q] Gabriel Luz Wallau,[r] Edson Delatorre,[s] Tiago Gräf,[t] Marilda Mendonça Siqueira,[u] Paola Cristina Resende,[u] Gonzalo Bello[v] on behalf of Fiocruz COVID-19 Genomic Surveillance Network

[a]Laboratório de Ecologia de Doenças Transmissíveis na Amazônia, Instituto Leônidas e Maria Deane, Fiocruz, Manaus, Amazonas, Brazil
[b]Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Rio de Janeiro, Brazil
[c]Fundação Centro de Controle de Oncologia do Estado do Amazonas, Manaus, Amazonas, Brazil
[d]Laboratório de Diversidade Microbiana da Amazônia com Importância para a Saúde, Instituto Leônidas e Maria Deane, Fiocruz, Manaus, Amazonas, Brazil
[e]Fundação de Vigilância em Saúde do Amazonas - Dra. Rosemary Costa Pinto, Manaus, Amazonas, Brazil
[f]Laboratório Central de Saúde Pública do Amazonas, Manaus, Amazonas, Brazil
[g]Universidade do Estado do Amazonas, Manaus, Amazonas, Brazil
[h]Hospital Adventista de Manaus, Manaus, Amazonas, Brazil
[i]Fundação de Medicina Tropical Doutor Heitor Vieira Dourado, Manaus, Amazonas, Brazil
[j]Instituto Nacional de Infectologia Evandro Chagas, Fiocruz, Rio de Janeiro, Rio de Janeiro, Brazil
[k]Laboratório de Diagnóstico e Controle e Doenças Infecciosas da Amazônia, Instituto Leônidas e Maria Deane, Fiocruz, Manaus, Amazonas, Brazil
[l]Secretaria de Saúde de Aparecida de Goiânia, Goiás, Brazil
[m]Laboratório de Virologia e Cultivo Celular, Instituto de Patologia Tropical e Saúde Pública, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[n]HLAGYN-Laboratório de Imunologia de Transplantes de Goiás, Aparecida de Goiânia, Goiás, Brazil
[o]Laboratório Analitico de Competências Moleculares e Epidemiológicas, Fundação Oswaldo Cruz Ceará, Fiocruz, Eusébio, Ceará, Brazil
[p]Unidade de Apoio Diagnóstico à COVID-19, Fundação Oswaldo Cruz Ceará, Fiocruz, Eusébio, Ceará, Brazil
[q]Departamento de Virologia, Instituto Aggeu Magalhães, Fiocruz, Recife, Pernambuco, Brazil
[r]Departamento de Entomologia e Núcleo de Bioinformática, Instituto Aggeu Magalhães, Fiocruz, Recife, Pernambuco, Brazil
[s]Departamento de Biologia, Centro de Ciências Exatas, Naturais e da Saúde, Universidade Federal do Espírito Santo, Alegre, Espírito Santo, Brazil
[t]Instituto Gonçalo Moniz, Fiocruz, Salvador, Bahia, Brazil
[u]Laboratório de Vírus Respiratórios e do Sarampo (LVRS), Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Rio de Janeiro, Brazil
[v]Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Rio de Janeiro, Brazil

ABSTRACT The Amazonas was one of the most heavily affected Brazilian states by the COVID-19 epidemic. Despite a large number of infected people, particularly during the second wave associated with the spread of the Variant of Concern (VOC) Gamma (lineage P.1), SARS-CoV-2 continues to circulate in the Amazonas. To understand how SARS-CoV-2 persisted in a human population with a high immunity barrier, we generated 1,188 SARS-CoV-2 whole-genome sequences from individuals diagnosed in the Amazonas state from 1st January to 6th July 2021, of which 38 were vaccine breakthrough infections. Our study reveals a sharp increase in the relative prevalence of Gamma plus (P.1+) variants, designated Pango Lineages P.1.3 to P.1.6, harboring two types of additional Spike changes: deletions in the N-terminal (NTD) domain (particularly Δ144 or Δ141-144) associated with resistance to anti-NTD

Editor Bo Zhang, Wuhan Institute of Virology

Address correspondence to Felipe Gomes Naveca, felipe.naveca@fiocruz.br, or Gonzalo Bello, gbellobr@gmail.com.
The authors declare no conflict of interest.

Received 22 November 2021
Accepted 24 January 2022
Published 23 February 2022

VIRUS EVOLUTION

Virus Evolution, 2021, 7(2), 1–11
DOI: https://doi.org/10.1093/ve/veab069
Advance access publication date: 14 August 2021
Rapid Communication

The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the spike protein

Paola Cristina Resende,[1,†,‡] Felipe G. Naveca,[2,†] Roberto D. Lins,[3] Filipe Zimmer Dezordi,[4,5] Matheus V. F. Ferraz,[3,6] Emerson G. Moreira,[3,6] Danilo F. Coêlho,[3,6,§] Fernando Couto Motta,[1] Anna Carolina Dias Paixão,[1] Luciana Appolinario,[1] Renata Serrano Lopes,[1] Ana Carolina da Fonseca Mendonça,[1] Alice Sampaio Barreto da Rocha,[1] Valdinete Nascimento,[2] Victor Souza,[2] George Silva,[2] Fernanda Nascimento,[2] Lidio Gonçalves Lima Neto,[7] Fabiano Vieira da Silva,[7] Irina Riediger,[8] Maria do Carmo Debur,[8] Anderson Brandao Leite,[9] Tirza Mattos,[10] Cristiano Fernandes da Costa,[11] Felicidade Mota Pereira,[12] Cliomar Alves dos Santos,[13] Darcita Buerger Rovaris,[14] Sandra Bianchini Fernandes,[14] Adriano Abbud,[15,**] Claudio Sacchi,[15] Ricardo Khouri,[16,††] André Felipe Leal Bernardes,[17] Edson Delatorre,[18] Tiago Gräf,[19,‡‡] Marilda Mendonça Siqueira,[1] Gonzalo Bello,[20,†] and Gabriel L. Wallau[4,5,*,†,§§] on behalf of Fiocruz COVID-19 Genomic Surveillance Network

[1]Laboratory of Respiratory Viruses and Measles (LVRS), Instituto Oswaldo Cruz, FIOCRUZ-Rio de Janeiro, Av. Brasil, 4365 - Manguinhos, Rio de Janeiro 21040-900, Brazil, [2]Laboratório de Ecologia de Doenças Transmissíveis na Amazônia (EDTA), Instituto Leônidas e Maria Deane, FIOCRUZ-Amazonas, Rua Teresina, 476. Adrianópolis, Manaus 69.057-070, Brazil, [3]Department of Virology, Instituto Aggeu Magalhães, FIOCRUZ-Pernambuco, Av. Professor Moraes Rego, s/n – Cidade Universitária, Recife 50.740-465, Brazil, [4]Departamento de Entomologia, Instituto Aggeu Magalhães, FIOCRUZ-Pernambuco, Av. Professor Moraes Rego, s/n – Cidade Universitária, Recife 50.740-465, Brazil, [5]Núcleo de Bioinformática (NBI), Instituto Aggeu Magalhães FIOCRUZ-Pernambuco, Av. Professor Moraes Rego, s/n – Cidade Universitária, Recife 50.740-465, Brazil, [6]Department of Fundamental Chemistry, Federal University of Pernambuco, Av. Professor Moraes Rego, s/n – Cidade Universitária, Recife 50.740-560, Brazil, [7]Laboratório Central de Saúde Pública do Estado do Maranhão (LACEN-MA), Rua João Luís, Bairro Diamente, Sao Luis 65020-320, Brazil, [8]Laboratório Central de Saúde Pública do Estado do Paraná (LACEN-PR), Rua Ubaldino do Amaral 545 - Alto da XV, Curitiba 80060-190, Brazil, [9]Laboratório Central de Saúde Pública do Estado do Alagoas (LACEN-AL), Av. Marechal Castelo Branco, 1773 Jatiúca, Alagoas, 57030340 Brazil, [10]Laboratório Central de Saúde Pública do Amazonas (LACEN-AM), Rua Emílio Moreira, 528 - Centro, Manaus 69020-040, Brazil, [11]Fundação de Vigilância em Saúde do Amazonas, Av. Torquato Tapajós, 4.010 Colônia Santo Antônio, Manaus 69.093-018, Brazil, [12]Laboratório Central de Saúde Pública do Estado da Bahia (LACEN-BA), Rua Waldemar Falcão, 123 - Bairro Brotas, Salvador 40295-001, Brazil, [13]Laboratório Central de Saúde Pública do Estado de Sergipe (LACEN-SE), Rua Campo do Brito, 551 - Bairro São José, Aracajú, Sergipe 49020-380, Brazil, [14]Laboratório Central de Saúde Pública do Estado de Santa Catarina (LACEN-SC), Avenida Rio Branco, 152 – Fundos, Florianópolis, Santa Catarina 88015-201, Brazil, [15]Instituto Adolfo Lutz, Av. Dr. Arnaldo, 351, São Paulo 01246-000, Brazil, [16]Laboratório de Enfermidades Infecciosas Transmitidas por Vetores, Instituto Gonçalo Moniz, FIOCRUZ-Bahia, Rua Waldemar Falcão, 121, Candeal, Salvador, Bahia 40296-710 , Brazil, [17]Laboratório Central de Saúde Pública do Estado de Minas Gerais (LACEN-MG), Rua Conde Pereira Carneiro, 80 - Gameleira, Belo Horizonte 30510-010, Brazil, [18]Departamento de Biologia, Centro de Ciências Exatas, Naturais e da Saúde, Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 514 - Goiabeira, Alegre 29075-910, Brazil, [19]Plataforma de Vigilância Molecular, Instituto Gonçalo Moniz, FIOCRUZ-Bahia, Rua Waldemar Falcão, 121, Candeal, Salvador 40296-710, Brazil and [20]Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz, FIOCRUZ-Rio de Janeiro, Av. Brasil, 4365 - Manguinhos, Rio de Janeiro 21040-900, Brazil

[†]These authors contributed equally to this work.
[‡]http://orcid.org/0000-0002-2884-3662
[§]http://orcid.org/0000-0002-1111-0825
[**]http://orcid.org/0000-0003-1685-0619
[††]http://orcid.org/0000-0001-5664-4436
[‡‡]http://orcid.org/0000-0003-4921-7975
[§§]http://orcid.org/0000-0002-1419-5713
*Corresponding author: E-mail: gabriel.wallau@fiocruz.br

Abstract

Mutations at both the receptor-binding domain (RBD) and the amino (N)-terminal domain (NTD) of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Spike (S) glycoprotein can alter its antigenicity and promote immune escape. We identified that SARS-CoV-2 lineages circulating in Brazil with mutations of concern in the RBD independently acquired convergent deletions and insertions in the NTD of the S protein, which altered the NTD antigenic-supersite and other predicted epitopes at this region. Importantly, we detected the community transmission of different P.1 lineages bearing NTD indels Δ69-70 (which can impact several SARS-CoV-2 diagnostic protocols), Δ144 and ins214ANRN, and a new VOI N.10 derived from the B.1.1.33 lineage carrying three NTD deletions (Δ141–144, Δ211, and Δ256–258). These findings support that the ongoing widespread transmission of SARS-CoV-2 in Brazil generates new viral lineages that might be more resistant to antibody neutralization than parental variants of concern.

Key words: COVID-19; pandemics; antibody escape; SARS-CoV-2; community transmission

ChemComm

COMMUNICATION

View Article Online
View Journal

Cite this: DOI: 10.1039/d1cc01747k

Received 1st April 2021,
Accepted 14th May 2021

DOI: 10.1039/d1cc01747k

rsc.li/chemcomm

Immune evasion of SARS-CoV-2 variants of concern is driven by low affinity to neutralizing antibodies†

Matheus V. F. Ferraz, ‡[ab] Emerson G. Moreira, ‡[ab] Danilo F. Coêlho, ‡[ab] Gabriel L. Wallau ‡[a] and Roberto D. Lins ‡*[a]

SARS-CoV-2 VOC immune evasion is mainly due to lower cross-reactivity from previously elicited class I/II neutralizing antibodies, while increased affinity to hACE2 plays a minor role. The affinity between antibodies and VOCs is impacted by remodeling of the electrostatic surface potential of the Spike RBDs. The P.3 variant is a putative VOC.

The COVID-19 pandemic has dramatically impacted the world population since 2019 and currently accounts for more than 2 million deaths.[1] The genome evolution of its etiological agent, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has been closely monitored since the rapid sharing of the first genomic sequences in December 2019.[2] SARS-CoV-2 shows a relatively low mutation rate compared to other RNA viruses, and hence few genomic sites accumulated mutations and were fixed until the second quarter of 2020. However, a substantially different scenario emerged between September–December 2020 with the detection of independent variants of concern (VOC) from different lineages, including B.1.1.7,[3] B.1.351,[4] and P.1,[5] bearing multiple amino acid replacements (K417T, E484K, and N501Y) and indels in the Spike protein,[4] which some researchers hypothesized to have occurred due to a "global shift in the SARS-CoV-2 selective landscape".[6] Although large-scale immunological studies are not available so far, the main hypothesis to explain such a global shift takes into account the rising population immunity, which would naturally select escape mutants with a higher fitness compared to previous circulating lineages. To support this hypothesis, some evidence could be mentioned, such as the increasing number of reinfection cases with VOCs and variants of interest (VOI) carrying some of the same amino acid mutation (E484K),[7] the continuous emergence of new VOIs carrying E484K and N501Y during the first months of 2021[8] and the recurrent emergence of some of those Spike amino acid changes in SARS-CoV-2 experimental evolution settings challenged with monoclonal and polyclonal antibodies.

To enter the host cell, SARS-CoV-2 makes use of the glycoprotein Spike (S). Protein S is a homotrimer and each monomer has two subunits, S1 and S2. The S1 subunit contains the receptor-binding domain (RBD), which binds to the human receptor angiotensin-converting enzyme 2 (hACE2), thus allowing the fusion of membranes and entry into the cell. Among the 29 SARS-CoV-2 encoded proteins, the S protein has been investigated more thoroughly due to its key role in hACE2 binding, and because the RBD region is one of the main targets of neutralizing antibodies (nAbs) produced from the human immunological response against SARS-CoV-2. By deep mutational scanning of the RBD region, it has been identified that most amino acid changes are deleterious for hACE2 binding, whereas a few marginally enhance the affinity to hACE2,[9] including some that have been detected in VOCs, such as N501Y in the more transmissible and mortal B.1.1.7 lineage. On the other hand, amino acid changes such as K417N, E484K, and N501Y found in VOCs P.1 and B.1.351[4] have been shown to increase viral fitness by lowering the effectiveness of neutralizing monoclonal and/or polyclonal antibodies.[8] Therefore, the emergence and spread of more fit VOC lineages may be driven by a mechanism other than the often-proposed affinity increase between hACE2 and SARS-CoV-2 Spike protein.

Despite the extensive description of mutations occurring in the SARS-CoV-2 RBD, little is known about their impact on receptor recognition, namely hACE2. In this regard, Starr *et al.*[9] have systematically measured the impact of every amino acid in the RBD, by replacing the 20 amino acids in each position, towards hACE2 binding affinity, expressed as the $\Delta\log(K_D)$, in which K_D represents the dissociation constant. Changes in K_D upon single-point mutations were obtained from a deep

[a] *Aggeu Magalhães Institute, Oswaldo Cruz Foundation, Recife, PE, Brazil. E-mail: roberto.lins@cpqam.fiocruz.br*
[b] *Department of Fundamental Chemistry, Federal University of Pernambuco, Recife, PE, Brazil*

† Electronic supplementary information (ESI) available: Associated content includes computational details and sequence data. See DOI: 10.1039/d1cc01747k
‡ These authors contributed equally to this work.

Unraveling the Role of Nanobodies Tetrad on Their Folding and Stability Assisted by Machine and Deep Learning Algorithms

Matheus Vitor Ferreira Ferraz[1,2], Wenny Camilla dos Santos Adan[2], and Roberto Dias Lins[1,2( )]

[1] Department of Fundamental Chemistry, Federal University of Pernambuco, Recife, PE 50670-560, Brazil
[2] Department of Virology, Aggeu Magalhães Institute, Oswaldo Cruz Foundation, Recife, PE 50670-420, Brazil
roberto.lins@cpqam.fiocruz.br

Abstract. Nanobodies (Nbs) achieve high solubility and stability due to four conserved residues referred to as the Nb tetrad. While several studies have highlighted the importance of the Nbs tetrad to their stability, a detailed molecular picture of their role has not been provided. In this work, we have used the Rosetta package to engineer synthetic Nbs lacking the Nb tetrad and used the Rosetta Energy Function to assess the structural features of the native and designed Nbs concerning the presence of the Nb tetrad. To develop a classification model, we have benchmarked three different machine learning (ML) and deep learning (DL) algorithms and concluded that more complex models led to better binary classification for our dataset. Our results show that these two classes of Nbs differ significantly in features related to solvation energy and native-like structural properties. Notably, the loss of stability due to the tetrad's absence is chiefly driven by the entropic contribution.

Keywords: Camelid antibodies · Rosetta Energy Function · Machine learning

1 Introduction

Ever since their discovery, single-domain binding fragment of heavy-chain camelid antibodies [1], referred to as nanobodies (Nbs), have gained considerable attention in translational research as therapeutic and diagnostic tools against human diseases and pathogens [2]. Along with its small size (15 kDa) and favorable physical-chemical properties (e.g., thermal and environmental stabilities), Nbs display binding affinities equivalent to conventional antibodies (cAbs) [1, 3]. Moreover, its heterologous expression in bacteria allows overcoming cAbs production pitfalls, such as high production cost and need of animal facility [4, 5]. Hence, Nbs are considered as a promising tool against numerous diseases. A variety of Nbs is currently being investigated under pre-clinical and clinical stages against a wide range of viral infections [6, 7].

J. C. Setubal and W. M. Silva (Eds.): BSB 2020, LNBI 12558, pp. 93–104, 2020.
https://doi.org/10.1007/978-3-030-65775-8_9

Journal of Molecular Graphics and Modelling 93 (2019) 107442

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

The influence of biotinylation on the ability of a computer designed protein to detect B-cells producing anti-HIV-1 2F5 antibodies

Danilo F. Coêlho [a, b], Matheus V.F. Ferraz [a, b], Ernesto T.A. Marques [a, c], Roberto D. Lins [a, b, *], Isabelle F.T. Viana [a, **]

[a] Aggeu Magalhães Institute, Oswaldo Cruz Foundation, Recife, PE, 50670-465, Brazil
[b] Department of Fundamental Chemistry, Federal University of Pernambuco, Recife, PE, 50740-540, Brazil
[c] Department of Infectious Diseases and Microbiology, University of Pittsburgh, Pittsburgh, PA, 15261, USA

ARTICLE INFO

Article history:
Received 25 June 2019
Received in revised form
22 August 2019
Accepted 26 August 2019
Available online 27 August 2019

Keywords:
Top7
Antigen
Biotin
Vaccine
Molecular dynamics

ABSTRACT

Antibodies against the HIV-1 2F5 epitope are known as one of the most powerful and broadly protective anti-HIV antibodies. Therefore, vaccine strategies that include the 2F5 epitope in their formulation require a robust method to detect specific anti-2F5 antibody production by B cells. Towards this goal, we have biotinylated a previously reported computer-designed protein carrying the HIV-1 2F5 epitope aiming the further development of a platform to detect human B-cells expressing anti-2F5 antibodies through flow cytometry. Biophysical and immunological properties of our devised protein were characterized by computer simulation and experimental methods. Biotinylation did not affect folding and improved protein stability and solubility. The biotinylated protein exhibited similar binding affinity trends compared to its unbiotinylated counterpart and was recognized by anti-HIV-1 2F5 antibodies expressed on the surface of patient-derived peripheral blood mononuclear cells. Moreover, we present a high affinity marker for the identification of epitope-specific B cells that can be used to measure the efficacy of vaccine strategies based on the HIV-1 envelope protein.

1. Introduction

Antibodies are the most common markers used to detect and quantify binding and recognition of specific antigens. In natural infections, their production is typically elicited by pathogen proteins and/or carbohydrates. Prophylactic strategies often rely on the use of antigens (e.g., viral particles, attenuated viruses, fusion protein domains) to mimic the host immune response upon pathogen exposure. Nevertheless, conventional approaches have faltered for several viruses, such as the Human Immunodeficiency virus – HIV. This virus has evolved an arsenal of molecular tricks to avoid or mitigate immune responses [1], including the transient exposure of key epitopes capable of triggering the production of broadly neutralizing antibodies. One of these cryptic epitopes is termed 2F5 epitope, whose core is a highly conserved continuous 9-mer amino acid sequence in the ectodomain of the gp41 envelope protein from HIV, also known as membrane-proximal external region (MPER) [2]. While the antibodies raised by this epitope (anti-2F5 antibodies) are well known to be produced by only a minority of infected individuals after 2–3 years of exposure [3], these antibodies are one of the most powerful and broadly protective anti-HIV antibodies described so far. They can neutralize over than 50% of viral isolated panels and protect primate models upon challenge [4,5]. Accordingly, the protective potential of a vaccine strategy can be assessed by measuring its capability of eliciting anti-2F5 antibodies upon immunization.

The identification of anti-2F5 antibodies *in vitro* has been a difficult task since their detection relies on the recognition of the MPER peptide. When free in solution, this peptide is poorly recognized by human antibodies, which hampers the assessment of the protective value of the vaccines under development. To overcome this limitation, we have previously engineered a chimerical protein where a 9-mer (ELDKWASLW), based on the 2F5 epitope, was grafted onto Top7, a computationally devised protein [6]. The protein named Top7-2F5 was shown to be specifically recognized by the respective monoclonal antibody (mAb 2F5) by means of

* Corresponding author. Aggeu Magalhães Institute, Oswaldo Cruz Foundation, Recife, PE, 50670-465, Brazil.
** Corresponding author.
E-mail addresses: roberto.lins@cpqam.fiocruz.br (R.D. Lins), isabelle.viana@cpqam.fiocruz.br (I.F.T. Viana).

https://doi.org/10.1016/j.jmgm.2019.107442
1093-3263/

ANNEX B – *CURRICULUM VITAE*

Matheus Ferraz
Structural Bioinformatician @ NEC OncoImmunity

NEC OncoImmunity
E-mail: matheus@oncoimmunity.com
Forskningsparken, Gaustadalléen 21, 0349,
Oslo Norway

Google scholar: Matheus Ferraz
LinkedIn: www.linkedin.com/in/matheus-ferraz-15bbaab4/
Researchgate: www.researchgate.net/profile/Matheus-Ferraz
Orcid: https://orcid.org/0000-0002-6958-3115

RESEARCH PROFILE

Structural Bioinformatician at NEC OncoImmunity using artificial intelligence and computational simulations to develop personalized therapeutic vaccines for infectious diseases. Expertise in molecular simulations of biological systems and mostly interested in machine learning engineering with expertise in full stack development.

EMPLOYMENT HISTORY

- Structural bioinformatician, NEC OncoImmunity, Norway (2023 – present)

EDUCATION

- Ph.D. Chemistry, Federal University of Pernambuco, Brazil (2019-Present).
  Doctoral stay at the Heidelberg Institute for Theoretical Studies (2021 – 2023).
- MSc. Chemistry, Federal University of Pernambuco, Brazil (2018-2019)
- B.Sc. Chemical Engineering, Federal University of Pernambuco, Brazil (2013 – 2018).

RESEARCH TRAINING

- Visiting Ph.D. student, **Heidelberg Institute for Theoretical Studies**, Germany (2021 – 2023)
  Advisor: Rebecca C. Wade; Molecular and Cellular Modelling group
  Activities:
  - Calculation of binding kinetics and thermodynamics properties (k_{on}, k_{off}, and k_D) for protein-ligand, protein-protein (nanobodies targeting SARS-CoV-2 RBD; SARS-CoV-2 spike protein targeting hACE2 in the presence of heparin) and protein-peptide complexes (MHC Class I – peptides)

- Ph.D. student collaborator, **Oswaldo Cruz Foundation**, Brazil (2019 – Present)
  Advisor: Roberto D. Lins; Biomaterial modelling group, Department of Virology
  Activities:
  Drug discovery and design; computational design of proteins against viral infections (Zika virus, Coronavirus); free energy calculations via enhanced sampling simulations; machine learning development to compute protein – protein binding free energies; member of the Brazilian COVID-19 genomic surveillance team at FIOCRUZ.

- Undergraduate Scientific Initiation, **Oswaldo Cruz Foundation**, Brazil (2015 – 2018)
  Advisor: Roberto D. Lins; Biomaterial modelling group, Department of Virology
  Activities:
  - Modelling of peptide – mineral surfaces interactions using molecular dynamics, metadynamics, and steered MD; modelling of camelid nanobodies.

- Undergraduate Diploma work, **Federal University of Pernambuco**, Brazil (2017-2017)
  Advisor: Sérgio Lucena; Laboratory for Advanced Control and Processes Optmization, Department of Chemical Engineering
  Activities:
  - Modelling of biochemical processes via SuperPro Designer®.

TEACHING EXPERIENCE

- **Volunteer High school Teacher**, Gradação Project, Federal University of Pernambuco (2019): Taught physics and chemistry classes at high-school level for needy students to prepare for university entrance exams in Brazil. In the occasion, students with disabilities joined the project.
- **Teacher Assistant**, Post-graduate Chemistry Program, Federal University of Pernambuco (2021): Taught some of the practical and theoretical lessons of the course "Molecular Biophysical Chemistry" under the supervision of Danilo F. Coêlho.
- **Teacher Assistant**, Undergraduate Chemistry Program, Federal University of Pernambuco (2019): Taught lectures on "Introductory chemistry" to the undergraduate-level in chemistry major under the supervision of Prof. Ricardo Oliveira.
- **Minicourse,** Molecular modeling of antigen-antibodies (In Portuguese), National Online Congress of Immunology, 2022.
- **Minicourse,** Molecular modeling in biotechnology (In Portuguese), Oswaldo Cruz Foundation 2022
- **Minicourse**, Molecular dynamics: Theory and applications in biotechnology and health (in Portuguese), Oswaldo Cruz Foundation, 2021.
- **Lecture**, Post-graduate Chemistry Program, Federal University of Pernambuco (2021): Taught some of the theoretical lessons of the course "Molecular Biophysical Chemistry" under the supervision of Danilo F. Coêlho.
- **Lecture,** Introduction to Biophysics (in Portuguese), Postgraduate program in Nanobiosystems, Federal University of Rio de Janeiro, 2023.

PUBLICATIONS

1. **Ferraz, M.V.F.**; Cláudio, J.; Lins, R.D.; Teixeira, E.S. Artificial neural networks to predict structure-based protein-protein free energy of binding from Rosetta-calculated properties. **Physical Chemistry Chemical Physics**, 2023
2. Maciel, L.G.; **Ferraz, M.V.F**.; Oliveira, A.A.; Guido, R.V.C.; Lins, R.D.; Anjos, J.A.; Soares, T.A. The X-ray structure of the 3-hydroxykynurenine transaminase from *Aedes aegypti*: a novel target to combat the transmission of arboviruses. **RSC Medicinal chemistry**, 2023.
3. **Ferraz, M.V.F.,** Viana, I.F.T., Coêlho, D.F., Cruz, C.H.B., Lima, M.A., Aragão, M.A.L., Lins, R.D. Association strength of E6 to E6AP/p53 correlates with HPV-mediated oncogenesis risk. **Biopolymers**, 2022
4. Matos, I.A.; Pinto, A.C.G.; **Ferraz, M.V.F.**, Adan, W.C.S.; Rodrigues, R.P.; Santos, J.X.; Kitagawa,R.R.; Lins, R.D.; Oliveira, T.B. and Costa, N.B. Identification of potential *Staphylococcus aureus* dihydrofolate reductase inhibitors using QSAR, molecular docking, dynamics simulations and free energy calculation. **Journal of Biomolecular Structure and Dynamics**, 2022.
5. Naveca, Felipe Gomes; Nascimento, Valdinete; Souza, Victor; Corado, André De Lima; Nascimento, Fernanda Silva, George Mejía, Matilde Contreras Brandão, Maria Júlia Costa, Ágatha Duarte, Débora Pessoa, Karina Jesus, Michele Gonçalves, Luciana Fernandes, Cristiano Mattos, Tirza Abdalla, Ligia Santos, João Hugo Martins, Alex Chui, Fabiola Mendonça Val, Fernando Fonseca De Melo, Gisely Cardoso Xavier, Mariana Simão Sampaio, Vanderson De Souza Mourão, Maria Paula Lacerda, Marcus Vinícius , Batista, Érika Lopes Rocha Magalhães, Alessandro Leonardo Álvares Dábilla, Nathânia Pereira, Lucas Carlos Gomes Vinhal, Fernando Miyajima, Fabio Dias, Fernando Braga Stehling Dos Santos, Eduardo Ruback; Coêlho, Danilo; **Ferraz, M.V.F.;** Lins, Roberto; Wallau, Gabriel; Luz Delatorre, Edson; Gräf, Tiago; Siqueira, Marilda; Mendonça Resende, Paola Cristina Bello, Gonzalo ; Spread of Gamma (P.1) Sub-Lineages Carrying Spike Mutations Close to the Furin Cleavage Site and Deletions in the N-Terminal Domain Drives Ongoing Transmission of SARS-CoV-2 in Amazonas, Brazil. **Microbiology Spectrum**, 2022. (Article)
6. **Ferraz, M.V.F.**; Gonçalves, E.M; Coêlho, D.F; Wallau, G.; Lins R.D.; Immune evasion of SARS-CoV-2 variants of concern is driven by low affinity to neutralizing antibodies. **Chemical Communication**, 2021 (Article)
7. Resende, P.C., Naveca, F.G., Lins R.D., Dezordi F.Z., **Ferraz, M.V.F.**, Moreira, E.G., et al. The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revelaed by convergent indels in the amino (N)-terminal domain of the Spike protein, **Virus evolution**, 2021 (Article)
8. **Ferraz, M.V.F**, Adan, W.C.S.; Lins, R.D. Unraveling the Role of Nanobodies Tetrad on Their Folding and Stability Assisted by Machine and Deep Learning Algorithms. **Lecture notes in Computer Sciences**, 2020. (Book chapter)

9. Freire, M.C.L.C.; Silva, Y.A.M.; **Ferraz, M.V.F**.; Cruz, C.H.B.; Ferreira, L.S.; Pedrosa, M.F.F.; Barbosa, E.G. Molecular Basis of Tityus Stigmurus Alpha Toxin and Potassium Channel Kv1.2 Interactions. **Journal Of Molecular Graphics & Modelling**, V. 87, P. 197-203, 2019; (Article)
10. Coêlho, D.F.; **Ferraz, M.V.F**.; Marques, E.T.A.; Lins, R.D.; Viana, I.F.T. The Influence Of Biotinylation On The Ability Of A Computer Designed Protein To Detect B-Cells Producing Anti-HIV-1 2F5 Antibodies. **Journal Of Molecular Graphics & Modellin**g, V. 93, P. 107442, 2019. (Article)

MANUSCRIPTS IN PROGRESS

1. **Ferraz, M.V.F**; Wade, R.C. Recent advances in enhanced sampling and machine learning in computing protein-target binding kinetics. (Invited review from Journal of Chemical Information and Modeling, American chemical society)
2. **Ferraz, M.V.F**; Lins, R.D.; Wade, R.C. Peptides Bound to Major Histocompatibility Complex Class I: Dissociation Mechanisms and off-Rates from tau-RAMD Simulations (Article)
3. **Ferraz, M.V.F**; Adan, W.C.S.; Viana, I.F.T; Wade, R.C.; Lins, R.D. A data-centric and enhanced sampling-based approach to the design of artificial nanobodies. (Article)
4. Lima, T.E; **Ferraz, M.V.F**.; Brito, C.A.A; Ximenes, P.A.; Mariz C.A.; Lins, R.D.; Viana, I.F.T. Assessment of the IgG antibody profile of individuals infected with SARS-CoV-2 and determination of serological prognostic markers for disease severity. (Article)
5. Adan, W.C.S; **Ferraz, M.V.F**.; Lins, R.D. Assessing the dynamical variability and affinity of VHH antigen-binding loops by sampling their conformational ensemble. (Article)

TECHNICAL SKILLS

- Programming Languages: Python, JavaScript, SQL
- Computational skills: Cloud computing (AWS, Azure, GCP), PyTorch and TensorFlow, autoML, KubeFlow, docker, Django, HTML5, css, React, node.JS, Apache Airflow, MLFlow;
- Computational biophysics: Drug discovery and design; Machine/Deep Learning; Computational Protein Design; Molecular Dynamics; enhanced sampling in MD simulations (Metadynamics, Umbrella Sampling); Out-of-equilibrium MD simulations (Steered MD, Targeted MD); Brownian Dynamics; τ-random accelerated MD; Quantum-chemical calculations.
- Solid mathematical background.

FELLOWSHIPS

- Research Grants-Bi-nationally Supervise DoctoralDegrees/Cotutelle, Deutscher Akademischer Austauschdienst, DAAD, Germany, 2021
- Doctoral Scholarship, 142297/2019-4, National Council for Scientific and Technological Development CNPq, Brazil, 2019
- Master Scholarship, National Council for Scientific and Technological Development, CNPq, Brazil, 2018
- Technical Cooperation scholarship, Foundation for Scientific Affairs from Pernambuco, FACEPE, Brazil, 2016
- Scientific Initiation, Coordination for the Improvement of higher education, CAPES, Brazil, 2015

AWARDS

- Honorable mention at the LatinXchem, American Chemical Society
- Honorable Mention due to Innovative Idea at the Mimesis Hackathon, University of Pernambuco, Brazil, 2020;
- Travel Grant Award, Institute Pasteur, Uruguay, 2017;
- Travel Grant Award, University of Buenos Aires, 2017;
- Honorable Mention due to the presented work entitled: “Assessing Peptide Adhesion to Hematite Surface: Towards Biofuel Cell Developments” at the Scientific Computation National Laboratory, Brazil, 2016;

REFEREE ASSISTANT FOR SCIENTIFIC JOURNAL

- Journal of molecular recognition - Wiley online library

INVITED TALKS

- Protein design to fight COVID-19: opportunities in health within engineering (In Portuguese), Department of Chemical Engineering, Federal University of Pernambuco, 2021
- Not all chemists wear a white coat: COVID-19 pandemics in the computation in petaescale era (In portuguese), National Online Congress of Chemistry, Brazil, 2021.
- Protein engineering as a strategy against COVID-19 (In Portuguese), VI Chemistry school of the Federal University of Sergipe, 2020.
- Protein Engineering in the Interface of Thermodynamics and Machine Learning (In portuguese), CESAR School, Brazil, 2020.
- Impact of STEM within the universities: Zika Virus diagnostic (In portuguese), Federal University of Pernambuco Radio, Brazil, 2019.

CONFERENCES – ORAL PRESENTATIONS (selected)

- **Ferraz, M.V.F.;** Lins, R.D.; Insights into the binding mechanism reveal the molecular selectivity of VHH-72 against SARS-CoV-1 but not for SARS-CoV-2 RBD. In: Biological Diffusion and Brownian Dynamics Brainstorm Meeting 5 (BDBDB), 2021, Heidelberg (Online).
- **Ferraz, M.V.F**.; Adan, W. C. S.; Coêlho, D. F.; Lins, R. D. Machine Learning Associated with Enhanced Sampling Simulations to Engineering Immunoreactive Proteins. eSSENCE-EMCC Meeting on Multiscale Modelling of materials and molecules in complex systems, Uppsala University, Sweden (Online), 2020.
- **Ferraz, M. V. F.;** Coêlho, D. F. ; Adan, W. C. S. ; Carvalho, R. D. ; Lins, R. D. . Rational design of a high affinity nanobody binding ZIKV NS1 protein aiming at differential serological diagnostic. IV Advanced School on Biomolecular Simulation: protein engineering with Rosetta, from fundamental principles to tutorial, Brazil, 2019.
- **Ferraz, M. V. F.**; Lins, R. D. . Adhesion between peptides and mineral surfaces aiming at green energy production. 2nd Protein Biohpysics at the end of world, Argentina, 2017.
- **Ferraz, M. V. F.;** Viana, I. F. T. ; Lins, R. D. . Characterizing the temperature-dependent conformational transition of dengue virus envelope protein. III Advanced School on Biomolecular Simulation: Multiscale Methods from Fundamental Principles to Tutorials, Brazil, 2017.
- **Ferraz, M. V. F.;** Lins, R. D. . Assessing the Molecular Basis of Flavivirus Breathing and its Consequence to Antibody Sensitivity. Performing Molecular Simulations with SIRAH force field, Uruguay, 2017.
- **Ferraz, M. V. F.;** Lins, R. D. . Understanding Hematite Surface Adhering Peptides. II Advanced School on Biomolecular Simulation, Brazil, 2016.
- **Ferraz, M. V. F.;** Cunha, K. C. ; Coêlho, D. F. ; Lins, R. D. . Assessing the Structural Features of Engineered VHH Antibodies targeting a Recombinant Nucleoprotein of Araucaria Hantavirus. I Advanced School on Biomolecular Simulation, Brazil, 2015.
- **Ferraz, M. V. F.;** Lins, R. D. . Assessing structural features of VHH-based antibodies used on novel Hantavirus Pulmonary Syndrome diagnosis via molecular dynamics. V STINT Workshop on Understand Biocompatibility of Polymerics Surface, Brazil, 2015.

ABSTRACTS PUBLISHED IN CONGRESSES (selected)

- Godoy, M.O; Nogueira, V.H.R; Freire, M.; Souza, G; Fassio, A.V.; **Ferraz, M.V.F**; Oliva, G. Lins, R.D.; Guido, R.V. Integration of virtual screening and experimental method to identify new Mpro of SARS-CoV-2 inhibitors. Brazilian MedChem, 2022.
- Xavier, L.S.; Coelho, D.F.; **Ferraz, M.V.F**; Viana, I.F.T; Lins, R.D. Development of synthetic proteins for Zika vírus neutralization. SBBQ-SBBF, 2022.
- **Ferraz, M.V.F**; Adan, W.C.S; Viana, I.F.T.; Wade, R.C.; Lins, R.D. Structure-based computational design of nanomolar-binding VHHs targeting the SARS-CoV-22 spike protein. In: Flagship Initiative Engineering Molecular Systems Young Scientist Retreat, 2022, Dannenfels. Abstract Booklet, 2022. p. 20-20.
- **Ferraz, M.V.F;** Lins, R.D. Characterizing binding kientics and thermodynamics of computer-designed nanobodies targeting SARS-CoV-2 RBD. Biophysical Society, Biophysical Journal, 2021
- **Ferraz, M.V.F**; Lins, R.D.; Insights into the binding mechanism reveal the molecular selectivity of VHH-72 against SARS-CoV-1 but not for SARS-CoV-2 RBD. In: Biological Diffusion and Brownian Dynamics Brainstorm Meeting 5 (BDBDB), 2021, Heidelberg. V Biological Diffusion and Brownian Dynamics Brainstorm Meeting, 2021. p. 11-11

- **Ferraz, M.V.F**; Seitz, C.; Lins, R.D. Probing the Conformational Dynamics of a Conserved Epitope in the SARSCoV-1 and -2 Spike Protein Using Gaussian Fluctuations and Collective Motions. In: Workshop on Computer Simulation and Theory of Macromolecules, 2021, Hünfeld. Abstracts of poster contributions, 2021. p. 34-34.

COMITTEE PARTICIPATION (All the Works in Portuguese)

- Diploma Work Defense Committee of Júlia Silvestre de Sena. (Supervision and control of a process in an industry of powder detergent) Bachelor of industrial chemistry, Federal University of Pernambuco, 2022.
- Diploma Work Defense Committee of Vinícius Firmino dos Santos (Computational simulation of nucleotides adsorption of polymeric brushes of poly(dimethylaminoethyl acrylate) Bachelor of Chemistry, Federal University of Pernambuco, 2021.
- Diploma Work Defense Committee of Isabelle Cristine de Lima (Application of the Kano and PDCA method on the assessment of quality service of the Stone Pagamentos S.A.) Bachelor of Chemical Engineering, Federal University of Pernambuco, 2021.
- Diploma Work Defense Committee of Julia Gabriela da Silva (Evaluation and reduction of the overweight indicator in the production of PVC tubes by extrusion) Bachelor of Chemical Engineering, Federal University of Pernambuco, 2021.
- Diploma Work Defense Committee of Yamê Cavalcatin Bezerra (Implementing APPC plan in an oil refinery). Bachelor of Chemical Engineering, Federal University of Pernambuco, 2020.
- Diploma Work Defense Committee of Larissa Dias da Silva Santos (Design of novel antagonists of the 3-hydroxykynurenine transaminase enzyme: a prototype for a new larvicide). Bachelor of Chemistry, Federal University of Pernambuco, 2019.
- Poster Jury at the XXIX Scientific Initiation Congress from the Rural Federal University of Pernambuco, 2019.

OUTREACH AND NEWS

- https://www.cpqam.fiocruz.br/institucional/noticias/estudo-investiga-mecanismo-da-reinfeccao-por-covid-19 (Institutional News in Portuguese, 2021)
- https://www.youtube.com/watch?v=pf29TS2bdX8 (BSB Technical section, 2020)
- https://www.youtube.com/watch?v=XP644bvCgPw&t=1781s (VI escola de química, 2020)
- https://www.youtube.com/watch?v=IIZMuCG2qf0 (Discovery channel documentary – Brazil science, episode 3: supercomputer, In Portuguese, 2016)

LANGUAGES

- Native Portuguese
- Fluent English
- Advanced German

REFERENCES

Prof. Dr. Roberto Lins (roberto.lins@cpqam.fiocruz.br)
Aggeu Magalhães Institute, Brazil

Prof. Dr. Rebecca Wade (rebecca.wade@h-its.org)
Heidelberg Institute for Theoretical Studies, Germany

Prof. Dr. Thereza Amélia Soares (thereza.soares@usp.br)
Universidade de São Paulo, Brazil