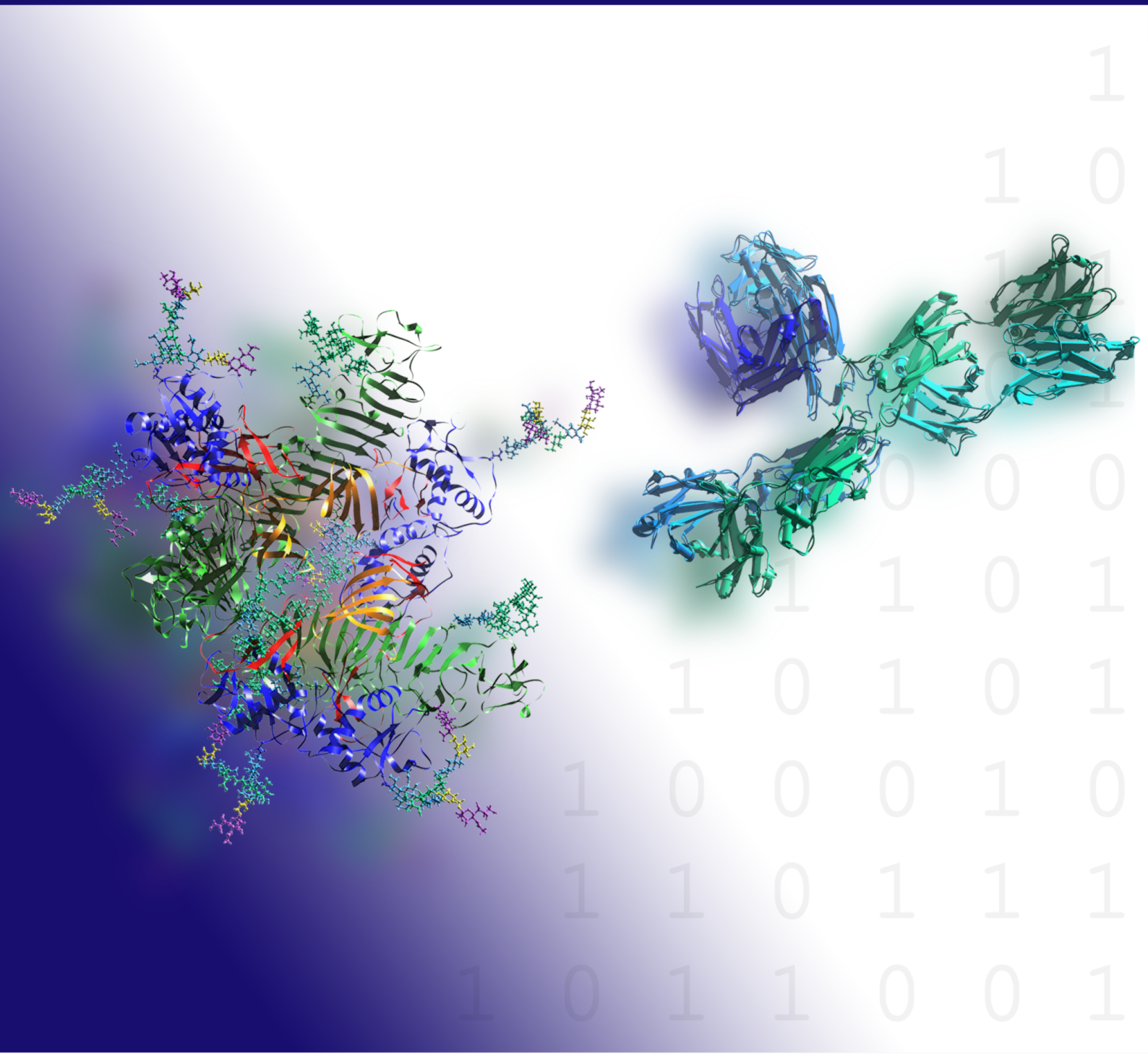
# Structural Bioinformatic Studies of the Zika Virus Non-structural Protein I and its Molecular Interactions



Sergio Alejandro Poveda Cuevas

# Structural Bioinformatics Studies of the Zika Virus Non-structural Protein 1 and its Molecular Interactions

Sergio Alejandro Poveda Cuevas

University of São Paulo

Ribeirão Preto

São Paulo - Brazil

2022

# Structural Bioinformatics Studies of the Zika Virus Non-structural Protein 1 and its Molecular Interactions

Sergio Alejandro Poveda Cuevas

Thesis Presented to the Interinstitutional Graduate Program in Bioinformatics

University of São Paulo

For Obtaining the Title in Doctor of Science

Advisor

Fernando Luís Barroso da Silva

Co-Advisor

Catherine Etchebest

Ribeirão Preto

São Paulo - Brazil

2022

Distribution by (name and address): Sergio Alejandro Poveda Cuevas*.

*e-mail: alejandropc@alumni.usp.br & seapovedac@gmail.com.

*website: https://www.researchgate.net/profile/Sergio-Poveda-Cuevas.

*ORCID: 0000-0002-3087-6674.

- Laboratory of Computational Biophysical Chemistry (https://barrosolab.fcfrp.usp.br/), Department of Biomolecular Sciences, School of Pharmaceutical Sciences at Ribeirão Preto, University of São Paulo, Ribeirão Preto - SP, Brazil (2022).

- Interinstitutional Graduate Program in Bioinformatics (https://www.ime.usp.br/pos-bioinformatica/), University of São Paulo, São Paulo - SP, Brazil (2022).

*I authorize the reproduction and total or partial disclosure of this work by any conventional or electronic media for study and research purposes provided the source is mentioned.*

© Sergio Alejandro Poveda Cuevas (2022)

Janus

Universidade de São Paulo

## ATA DE DEFESA



Ata de defesa de Tese do(a) Senhor(a) Sergio Alejandro Poveda Cuevas no Programa: Bioinformática, do(a) Interunidades em Bioinformática da Universidade de São Paulo.

Aos 09 dias do mês de março de 2022, no(a) realizou-se a Defesa da Tese do(a) Senhor(a) Sergio Alejandro Poveda Cuevas, apresentada para a obtenção do título de Doutor intitulada:

"ESTUDOS DE BIOINFORMÁTICA ESTRUTURAL DA PROTEÍNA NÃO ESTRUTURAL 1 DO VÍRUS ZIKA E SUAS INTERAÇÕES MOLECULARES"

Após declarada aberta a sessão, o(a) Sr(a) Presidente passa a palavra ao candidato para exposição e a seguir aos examinadores para as devidas arguições que se desenvolvem nos termos regimentais. Em seguida, a Comissão Julgadora proclama o resultado:

| Nome dos Participantes da Banca | Função | Sigla da CPG | Resultado |
|---|---|---|---|
| Fernando Luis Barroso da Silva | Presidente | FCFRP - USP | Não Votante |
| Leandro Ramos Souza Barbosa | Titular | IF - USP | Aprovado |
| Marie Flamand | Titular | IP - Externo | Aprovado |
| Eduardo Rocha de Almeida Lima | Suplente | UERJ - Externo | Aprovado |

**Resultado Final:** Aprovado

**Parecer da Comissão Julgadora** *

Eu, Fernando Luís Barroso da Silva, lavrei a presente ata, que assino juntamente com os(as) Senhores(as). São Paulo, aos 09 dias do mês de março de 2022.

Leandro Ramos Souza Barbosa

M. Flamand

Marie Flamand

Eduardo R. de A. Lima

Eduardo Rocha de Almeida Lima

Fernando Luis B. da Silva

Fernando Luis Barroso da Silva
Presidente da Comissão Julgadora

* Obs: Se o candidato for reprovado por algum dos membros, o preenchimento do parecer é obrigatório.

A defesa foi homologada pela Comissão de Pós-Graduação em 16.03.22 e, portanto, o(a) aluno(a) acima jus ao título de Doutor em Ciências obtido no Programa Bioinformática.

André Fujita

Presidente da Comissão de Pós-Graduação

*To my beloved family*

*... In tribute to all victims of Covid-19*

# Acknowledgements

I want to start by thanking the country that has embraced me during the last seven years: Brazil. Thank you for allowing me to fulfill one of my biggest dreams!

I am very grateful to my advisor, Fernando Luís Barroso da Silva, and my co-advisor, Catherine Etchebest for their valuable advice, guidance, and inspiration during my project. All the knowledge on computational stuff and biomolecular systems that I learned during this period of my life were because of your guidance as rigorous mentors. I also want to thank them for all the teachings in parallel that are not possible to translate into a text. Thanks to my academic parents!

I want to give my deepest thanks to Seti, my favorite human being, and my love too. Thank you for all these years of new experiences, for the support in my falls, and for always giving me the confidence to push myself forward where my path is full of love and peace. Your rationality and strength have been an inexhaustible source of inspiration for me, so thank you for being who you are and give me that indirect support as well.

My thanks and appreciation to my siblings: Lala, Leo, and Jackson, for always giving me encouragement and the right actions/words in critical situations. All the shared moments and teachings with each one I would like to repeat over and over again. I am what I am because of all these experiences. I am lucky because our existences have coincided in the immensity and randomness of the universe. We are the fantastic four ☺!

My sincere gratitude to the architects of my life, my parents Rosa and Fredy. The foundations of who I am, come from all the teachings from you. Reaching this point in my academic career was possible because your hands shaped my character, attitude, and curiosity for the life. Thank you!

Special thanks to my aunt Yolanda and my uncle Olmer, for all your caring, good vibes, and love.

Thanks to my friends Sudi, Javier, Airton, and Juliana for always giving me support in these years of intense academic experiences. Life with laughter and joy is more beautiful and they always gave me many beautiful moments without any reproach.

I am very grateful to CAPES for its financial support during this academic and transformative period. (Science needs funding, do not destroy one of the nuclei that society cries out for).

# Contents

**List of Figures** iii

**List of Tables** iv

**Symbols and Abbreviations** v

**List of Papers** ix

**1 Introduction** 1

**2 Biological Background** 8

2.1 Overview of flaviviruses . . . 9

2.2 Differential virulence . . . 11

2.3 Molecular constitution of flaviviruses . . . 12

2.3.1 Selection of study protein . . . 13

2.3.2 Non-structural protein 1 . . . 15

**3 Workflow** 18

**4 Computational Aspects** 21

4.1 Multiscale modeling . . . 22

4.1.1 Components of the molecular model . . . 22

4.1.2 Atomistic representation . . . 22

4.1.3 Coarse-grained representation at the four-to-one level . . . 26

4.1.4 Coarse-grained representation at the amino acid level . . . 27

4.1.5 Periodic boundary conditions . . . 30

4.2 Computational simulations . . . 30

4.2.1 Molecular Dynamics . . . 32

4.2.2 Metropolis Monte Carlo . . . 34

4.3 What is quantified in the computer simulations? . . . 35

4.3.1 Physical quantities measured in the MD runs . . . 35

4.3.1.1 Canonical measures . . . 36

4.3.1.2 Statistical measures . . . 37

4.3.2 Physical quantities measured in the MC runs . . . . . 39

**5 Preliminary insights about the NS1 electrostatic behavior 42**

**6 Electrostatic epitopes prediction through PROCEEDp$K_a$ 61**

**7 NS1 self-association 82**

**8 Preliminary insights about the NS1 dynamical behavior 119**

**9 NS1 interacting with a lipid bilayer 135**

**10 General Conclusions 168**

**A Comparison between experimental and simulation NS1 data 170**

**B Comparison between glycosylated and non-glycosylated proteins 172**

**C CPRISMA (Coloring PRoteins by Inputs and Multiple Alignments) 174**

**Bibliography 259**

# List of Figures


2.1 Flavivirus genome organization. . . . . 13
2.2 Three-dimensional representation of the nonstructural protein 1. . . . . 14
2.3 The NS1 trafficking within an infected cell. . . . . 16

3.1 Summary of the present study. . . . . 19

4.1 Schematically representation of the covalent and non-covalent bonds as used in atomistic models. . . . . 23
4.2 Example of a simulation box for a biomolecular system represented at the all-atoms modeling level. . . . . 24
4.3 Schematic representation of the model cell used in this study. . . . . 29
4.4 Example of a theoretical titration plot. . . . . 41

A.1 $B - factor$ as a function of the residue position for two $NS1_{ZIKV}$ from experimental and simulated data. . . . . 170

B.1 Schematic representation of two oligosaccharides used to construct NS1 glycosylated forms. . . . . 173
B.2 Principal component analysis for two $NS1_{ZIKV}$ at the presence and absence of sugar molecules. . . . . 173

# List of Tables

4.1 Properties of the coarse-grained amino acids. . . . . 28

5.1 Molecular modeling and measurements carried out in Chapter 5. . . . . 42

6.1 Molecular modeling and measurements carried out in Chapter 6. . . . . 61

7.1 Molecular modeling and measurements carried out in Chapter 7. . . . . 82

8.1 Molecular modeling and measurements carried out in Chapter 8. . . . . 119

9.1 Molecular modeling and measurements carried out in Chapter 9. . . . . 135

# Symbols and Abbreviations

| | **Symbols** |
|---|---|
| $A$ | Free energy of interaction (given in $k_BT$) |
| $\mathbf{A}$ | Eigenvectors |
| $\mathcal{A}$ | Thermodynamical ensemble |
| $b$ | Atomic bond length |
| $C, C_P$ | Charge regulation parameter (or protein capacitance) |
| $C_{ij}$ | Correlation coefficient |
| $\gamma$ | Phase shift |
| $\mathbf{C}$ | Covariance matrix |
| $c_k$ | Density of the mobile electrolyte specie $k$ |
| $\epsilon_s$ | Dielectric constant of the medium |
| $\epsilon_0$ | Dielectric constant of the vacuum (given in $C^2/N\ m^2$) |
| $\varepsilon_{LJ}$ | Depth-well of the Lennard-Jones potential (given in $k_BT$) |
| $e$ | Elementar charge (equivalent to $1.602 \times 10^{-19}$ C) |
| $\theta$ | Atomic angle |
| $\vec{F}$ | Force (given in kJ $mol^{-1}nm^{-1}$ ) |
| $g$ | Radial distribution function |
| $\kappa$ | Inverse Debye screening length (given in Å) |
| $k$ | Force constant (given in kcal/mol/$Å^2$) |
| $k_B$ | Boltzmann' constant (equivalent to $1.3807 \times 10^{-23}$ J $mol^{-1}$ $K^{-1}$) |
| $\lambda$ | Eigenvalues |
| $\lambda_{\mathrm{pH}}$ | Residue protonation |
| $l_B$ | Bjerrum length (given in Å) |
| $l_{cyl}$ | Cylinder height (given in Å) |
| $M$ | Total mass (given in g) |
| $Mw$ | Amino acid molecular weight (given in g/mol) |
| $\mu_P$ | Dipole moment number (given in Å) |
| $m$ | Particle mass (given in g) |
| $m_{acc}$ | Number of dots on a specific atom not occluded by neighboring beads |
| $N$ | Number of particles |
| $N_A$ | Avogadro's number (equivalent to $6.02 \times 10^{23}$ $mol^{-1}$) |

| | |
|---|---|
| $n$ | Multiplicity |
| $\mathrm{p}K_a$ | Dissociation constant |
| $\phi$ | Dihedral angle |
| $\psi$ | Electric potential of the macroion (given in V) |
| $Q^{GN}$ | Maximum modularity value |
| $q$ | Charge for a given chemical species (given in C) |
| $R$ | Distance between titratable sites (given in Å) |
| $R_a$ | Radius of Lennard-Jones sphere (given in Å) |
| $R_g$ | Radius of gyration (given in nm) |
| $\mathbf{R}$ | Protein center of mass |
| $\mathbf{r}, r_k$ | Particle position |
| $r$ | Separation distance (given in Å) |
| $r_{cyl}$ | Cylinder radius (given in Å) |
| $T$ | Temperature (given in K) |
| $t$ | Time (given in ps) |
| $U$ | Total interaction potential energy (given in J) |
| $u$ | Interaction potential energy (given in J) |
| $W$ | Edge weight between two nodes |
| $w$ | Potential of mean force (given in $k_BT$) |
| $x$ | Atomic coordinates |
| $Z, Z_P$ | Protein charge number |
| $\zeta$ | Zeta potential (given in V) |
| $z$ | Valence for a given chemical species |
| $\omega$ | Improper dihedral angle |

**Abbreviations**

| | |
|---|---|
| AA | All-atoms representation |
| Ab | Antibody |
| ADE | Antibody-dependent enhancement mechanism |
| C | Capsid protein |
| CG | Coarse-grained representation |
| Cryo-EM | Cryo-electron microscopy |
| DCCM | Dynamic cross-correlation map |
| DENV | Dengue virus |
| DH | Debye-Hückel |
| DHF | Dengue hemorrhagic fever |
| DLVO | Derjaguin, Landau, Verwey and Overbeek theory |
| E | Envelope protein |
| ED | Essential dynamics |
| EE | Electrostatic epitope |

| | |
|---|---|
| ELISA | Enzyme-linked immunosorbent assay |
| ElNeDyn | Elastic Network in Dynamics |
| EPR | Electron paramagnetic resonance |
| ER | Rough endoplasmic reticulum |
| FF | Force field |
| FPTS | Fast proton titration scheme |
| FRET | Förster resonance energy transfer |
| GAG | Glycosaminoglycans |
| GF | Greasy fingers |
| GPI | Glycosylphosphatidylinositol |
| hNSCs | Human fetal brain neural stem cells |
| IFT | Immuno-fluorescence tests |
| IL | Intertwined loop |
| JEV | Japanese encephalitis virus |
| KUV | Kunjin encephalitis virus |
| LD | Lipid droplet |
| LJ | Lennard-Jones |
| MC | Monte Carlo |
| MD | Molecular dynamics |
| MVEV | Murray Valley encephalitis virus |
| NMR | Nuclear magnetic resonance |
| NS | Non-structural |
| NTR | Non-translating regions |
| ORF | Open reading frames |
| PME | Particle Mesh Ewald |
| PBC | Periodic boundary conditions |
| PC | Principal component |
| PCA | Principal component analysis |
| prM | Premembrane protein |
| PRNT | Plaque reduction neutralization tests |
| PSN | Protein structure networks |
| RMSD | Root mean square deviation |
| RMSF | Root mean square fluctuation |
| RT-PCR | Reverse transcription polymerase chain reaction |
| RT-qPCR | Reverse transcription quantitative polymerase chain reaction |
| SASA | Solvent accessible surface area |
| SB | Structural bioinformatics |
| SLEV | St. Louis encephalitis virus |
| TBEV | Tick-borne encephalitis virus |
| TK | Tanford-Kirkwood model |

| | |
|---|---|
| WNV | West Nile virus |
| YFV | Yellow fever virus |
| ZIKV | Zika virus |

# List of Papers

Paper I Poveda-Cuevas, S.A.; Etchebest, C.; Barroso da Silva, F. L. Insights into the ZIKV NS1 Virology from Different Strains through a Fine Analysis of Physicochemical Properties. *ACS Omega*, 3(11), 16212–16229, 2018. doi: 10.1021/acsomega.8b02081.

Paper II Poveda-Cuevas, S.A.; Etchebest, C.; Barroso da Silva, F. L. Identification of Electrostatic Epitopes in Flavivirus by Computer Simulations: The PROCEEDp$K_a$ Method. *Journal of Chemical Information and Modeling*, 60(2), 944–963, 2020. doi: 10.1021/acs.jcim.9b00895.

Paper III Poveda-Cuevas, S.A.; Etchebest, C.; Barroso da Silva, F. L. Self-association features of NS1 proteins from different flaviviruses. *BioRxiv*, 1–35, 2022. doi: 10.1101/2022.03.15.484429v1.

Paper IV Poveda-Cuevas, S.A.; Barroso da Silva, F. L.; Etchebest, C. How the Strain Origin of Zika Virus NS1 Protein Impacts Its Dynamics and Implications to Their Differential Virulence. *Journal of Chemical Information and Modeling*, 61(3), 1516–1530, 2021. doi: 10.1021/acs.jcim.0c01377.

Paper V Poveda-Cuevas, S.A.; Barroso da Silva, F. L.; Etchebest, C. NS1 from Two Zika Virus Strains Differently Interacts with Membrane: A Plausible Explanation for their Differential Virulence. *BioRxiv*, 1–32, 2022. doi: 10.1101/2022.04.08.487413v1.

*"La creación intelectual es el más misterioso y solitario de los oficios humanos"*

Gabriel García Marquez

**Abstract**

The Flavivirus genus is a group with more than 70 virus species generally distributed in tropical zones. Over several years, flaviviruses have been representing a great public health problem, producing thousands of deaths and infections per year. The challenges to improve the diagnostic capacity and other interventions for viruses such as Zika require the comprehension of the physical interactions that occur at the molecular level. As a therapeutic target, the nonstructural protein 1 (NS1) has been a centerpiece since it is shown as an important element in the virulence and survival of the Zika virus. Using Structural Bioinformatics approaches, we aim to understand the molecular properties and mechanisms that occur at different biological interfaces of this viral protein and their relation with virulence. Employing a constant-pH Monte Carlo computational approach and classical molecular dynamics simulations, we described several physicochemical and dynamical-structural properties that have a close relationship with molecular mechanisms where NS1 is a key player. Our results contribute to a general physicochemical understanding of the macromolecular complexation and their vast role in cell biology providing evidence for possible pharmacological interventions.

**Keywords:** *NS1 protein*, *biomolecular complexes*, *electrostatic interactions*, *Monte Carlo*, *molecular dynamics*, *structural bioinformatics.*

**Resumo**

O gênero Flavivirus é um grupo com mais de 70 espécies virais, geralmente, distribuídas em zonas tropicais. Durante vários anos, os flavivirus vêm representando um grande problema de saúde pública, produzindo milhares de mortes e infecções anualmente. Os desafios para melhorar a capacidade de diagnóstico e outras intervenções para vírus, tais como o Zika, requerem a compreensão das interações físicas que ocorrem no nível molecular. Como um alvo terapêutico, a proteína não estrutural 1 (NS1) tem sido uma peça central, por ser um elemento importante na virulência e sobrevivência do vírus Zika. Usando abordagens de Bioinformática Estrutural, objetivamos entender as propriedades moleculares e os mecanismos que ocorrem em diferentes interfaces biológicas desta proteína viral e sua relação com a virulência. Empregando uma abordagem computacional através de simulações Monte Carlo a pH-constante e dinâmica molecular, descrevemos várias propriedades físico-químicas e dinâmico-estruturais que têm uma estreita relação com os processos de associação onde NS1 é um participante chave. Nossos resultados contribuem para o entendimento físico-químico geral da complexação macromolecular e seu vasto papel na biologia celular, fornecendo evidências para possíveis aplicações farmacológicas.

**Palavras-chave:** *proteína NS1*, *complexos biomoleculares*, *interações eletrostáticas*, *Monte Carlo*, *dinâmica molecular*, *bioinformática estrutural*.

## Chapter 1

# Introduction

*"My own suspicion is that the universe*
*is not only queerer than we suppose*
*but queerer than we can suppose"*
J. B. S. Haldane

Over the last decades, many efforts in biomedical areas have been dedicated to identify, treat, and eliminate pathogens. Although evolutionarily our bodies have "weapons" to fight against infectious agents (*i.e.* immune system), technological development has also accelerated and improved the fight against harmful pathogens such as viruses. From the pharmacological point of view, antiviral drugs and vaccines are the most suitable alternatives to treat viral infections. However, their eradication or treatment turn out to be challenging because the viruses adapt to therapeutic interventions. The changes in viruses are mostly associated with mutations that give them new biological features to survive, evade the host immune response, overcome the action of a drug, improve the viral particle preservation, or increase the chance of replication (Fields, 2014). All this is possible since viral molecular mechanisms are selected and propagated to future generations of viruses. Owing to that, in recent years, many studies have been devoted to uncover the biological processes that occur at the molecular level between viruses and the host cells.

To understand viruses, we need to study their biomolecular constituents. The basic structure of a virus is made up of a genome and proteins (Fields, 2014). Regarding proteins, these biomolecules is made up of amino acids, where each of these isolated molecules presents chiral carbon symmetry (except glycine). Depending on their functional group in the side chain R, it is possible to define 20 essential amino acids. From the physicochemical point of view, they can be classified as *a)* polar or apolar; *b)* positively, negatively, or neutrally charged; and *c)* hydrophobic or hydrophilic (Cox and Nelson, 2000). Both covalent and noncovalent interactions among residues[1] trigger emergent properties that result in a particular three-dimensional (3D) structure. Amino acids can

[1]Biochemical nomenclature that applies to amino acids covalently linked by peptide bonds.

also interact with an aqueous medium[2]. This makes globular proteins, for example, to have a particular residue distribution. Those amino acids with more hydrophilic properties tend to localize on the 3D surface, while the hydrophobic ones tend to remain buried at the interior (Tsuchida and Abe, 1982; Gregory, 1995). These types of patterns are due to electrostatic affinities produced between the water molecules and ionizable groups (hydrogen bonding) (Israelachvili, 2011). Moreover, the "repulsion" produced between the hydrophobic protein regions and the solvent also contributes to a particular structural pattern (Gregory, 1995; Israelachvili, 2011). However, the arrangement of residues in an aqueous medium can be a much more complex phenomenon and other agents contribute to the formation of a particular 3D structure (*e.g.* concentration of protons, salt ions, organic molecules, other biomolecules, etc) (Zhang et al., 2011; Deller et al., 2016; Nassar et al., 2021). Thereby, the protein itself and the medium manage the specificity, structure, dynamics, and stability of these macromolecules. Defining a specific 3D structure, proteins can perform a certain function such as transport, support, signaling, catalysis, and so on (Branden and Tooze, 1999; Cox and Nelson, 2000). Often, biological functionality is possible because of the formation of biomolecular complexes (*i.e.* the quaternary structure). Given that the complexation phenomena happen in a dynamical medium (*e.g.* charge variations can occur down a pH gradient), also the complex stability is affected (Barroso da Silva, 2000). The macromolecules "take advantage" of that to associate or dissociate according to the physicochemical condition of the medium (*e.g.* various enzymes have catalytic cycles that depend on metal ion concentration in a specific cellular location) (Cox and Nelson, 2000; Blanco and Blanco, 2017). At last, we can divide the complexes as *i)* homocomplexes (*e.g.* when two proteins with the same amino acid sequence associate), or *ii)* heterocomplexes (*e.g.* when a protein binds to a membrane) (Jones and Thornton, 1995).

To study protein structures, we can invoke an experimental approach. To date, 184,700 structures have been deposited in the RCSB Protein Data Bank (PDB) (Berman et al., 2000; Burley et al., 2021). Several of these 3D structures are native-like or complexes and they were resolved by X-ray crystallography (Shi, 2014), nuclear magnetic resonance (NMR) (Sugiki et al., 2017), or cryo-electron microscopy (Cryo-EM) (Murata and Wolf, 2018). These techniques constitute the basis of the research field called *structural biology* (Curry, 2015). This experimental area has helped to reveal important characteristics of the "structure-function" relationship of diverse biomolecules. Yet, some limitations are also highlighted since the solved structures represent a single "biomolecular snapshot" (especially structures determined by X-ray and Cryo-EM). Nowadays, it is well understood that protein functionality not only depends on a single native structure but also on an arrangement of macromolecules with different configurations (Clark, 2008). On the other hand, spectroscopy techniques such as NMR (Sugiki et al., 2017), electron paramagnetic

[2]The proteins are usually inside the cell, which can often have a certain viscosity.

resonance (EPR) (Nesmelov and Thomas, 2010), and Förster resonance energy transfer (FRET) (Bajar et al., 2016) were developed to understand the dynamics and molecular mechanisms on biomolecules as well. Advanced details in this regard were obtained through these experimental techniques. Nonetheless, some obstacles have also arisen due to expensive laboratory processes, limitations due to the size of the macromolecule, contamination of samples, etc (Howard, 1998; Le Breton et al., 2015; Teunissen et al., 2018; Sahu and Lorigan, 2020).

In a complementary way, theoretical approaches like computer simulations have aided to study proteins in situations where variations are important (*e.g.* protein motions or charge shifts in ionizable residues) (van der Kamp et al., 2008). When protein simulations are done in *large-scale* contexts (*e.g.* when different physicochemical conditions are explored) and employing their structures, we refer to it as an area of knowledge called *Structural Bioinformatics* (SB) (Chandra et al., 2010; Samish et al., 2015). The basis of SB is the molecular modeling of biomolecules present in nature. In order to have a reliable relationship between the theoretical approach and the observed physical phenomenon, the following points should be considered: *i)* the model to be used, *ii)* the degree of details of the model, *iii)* the desired statistical precision, *iv)* the computational costs, and *v)* the available computational resources (Luscombe et al., 2001). The first two items refer to the number of parameters needed to well-describe the system (*e.g.* in this thesis, the typical system is a protein in an electrolyte solution). The representation of the studied system can be classically approached from different perspectives, among them the best known being: *a)* an *all-atom* (AA) description of the system, or *b)* a *coarse-grained* (CG) modeling (Kmiecik et al., 2016; Barroso da Silva et al., 2020). The main difference relies on what is defined as the smaller subunit of the system. At the first level, the system is treated employing atomic coordinates and the interactions effectively average the degrees of freedom from the quantum level (Tozzini, 2009). The second modeling option refers to specific groupings of atoms (Voth, 2008). Their selection is a balance of the last three items (*i.e. iii*, *iv*, and *v*). After defining the appropriate molecular model, deterministic methods such as molecular dynamics (MD) calculations (Karplus and McCammon, 2002), or stochastic approaches including Monte Carlo (MC) simulations (Cramer, 2013) can be performed. The accuracy of a specific measured property will depend on the number of configurations that can be generated within the available cpu resources (more details about the molecular modeling applied in this research are mentioned in Chapter 4).

The development of SB has gone hand in hand with great advances in the areas of biotechnology, materials science, food, clinical, and pharmacy (Chou, 2004; Samish et al., 2015; Barroso da Silva et al., 2020). In the context of viruses, SB has provided key insights on the MD of viral proteins (Oliveira et al., 2016; Gonçalves et al., 2019), drug-design (Alhossary et al., 2018), elucidation of complexation processes (Machado et al., 2017), prediction of epitope binding sites (Coêlho et al., 2019), and evaluation of electrostatic

properties (Fuzo and Degrève, 2014; Giron et al., 2020, 2021). Currently, the SB and structural biology are one of the most important research areas in the fight against SARS-CoV-2 (Waman et al., 2021; Yang and Rao, 2021). SB has also been proposed to study a fairly diverse virus genus called *Flavivirus* (Mottin et al., 2017; Freire et al., 2017; Chuang et al., 2018; da Penha Valente et al., 2020). The attention that flaviviruses have received in this field is partly due to frequent outbreaks in recent years and dangerous diseases that they produce in the human body [*e.g.* Dengue (DENV), West Nile (WNV), Yellow Fever (YFV), or Zika (ZIKV) viruses] (Neufeldt et al., 2018).

Several SB studies have had a special emphasis on the MD aspects of various structural and non-structural (NS) flavivirus proteins. Dynamical properties have been studied in flavivirus in order to have accurate and effective vaccines. When vaccines are being developed it is necessary to determine regions for the recognition of antibodies (Ab) or *epitopes*. The work by Kim et al. (2021) have revealed that flexible protein parts are more likely to be epitopes, showing that dynamical features should be considered to have a suitable characterization of the antigenic regions. For instance, biomolecules such as the WNV envelope (E) protein, may have "cryptic epitopes" that become exposed after significant conformational changes occur (da Penha Valente et al., 2020). Once the candidate epitope regions are defined, viral peptides can be computationally designed and docked with host proteins to test the stability of intermolecular interactions (*e.g.* van der Waals or H-bonds) during MD simulations. These approaches were done by Mirza et al. (2016) where peptides of NS3 (helicase protein) and NS5 (polymerase system) of ZIKV in complex with T-cell receptors were tested. Dynamic characteristics of flavivirus proteins are also important for drug design. The interaction between NS5 of ZIKV and a specific drug (RNA analogue or SAH) was studied through sophisticated MD approaches (Chuang et al., 2018). These studies revealed that considerable conformational changes in particular regions of NS5 must occur in order to selectively bind one of the therapeutic drug targets. Ultimately, the question of the dynamics of the whole viral particles is also of main interest. This question has been addressed to examine DENV (Huber et al., 2021) and ZIKV (Machado et al., 2017) in CG representation and MD simulations. The work of Huber et al. (2021), have proposed important insights into the structural stability of E proteins and associated lipids that made up the viral particle. Besides, these investigations also show that flexible regions are essential for the assembly of E proteins.

The study of pH effects on protein molecules, implicitly deals with the electrostatic properties of the system (Jönsson et al., 2007). Through SB approaches, these topics can also be examined. Since several molecular mechanisms related to cell entry and release of flavivirus genomic material are associated with the E protein and dependent on pH variations, to date, most SB studies have focused on this biomolecule (specifically E protein from DENV) and in comprehending the mechanisms induced by pH. Computational simulations have shown that transitions from neutral to acidic pH regimes

are necessary to induce changes of oligomerization states from trimeric E protein to the dimeric E ones, respectively (Chaudhury et al., 2015). A detailed analysis demonstrated that these changes are related to variations of p$K_a$ in ionizable groups such as histidines (Prakash et al., 2010; Fuzo and Degrève, 2014; Chaudhury et al., 2015). The most recent investigation by Huber et al. (2021), also concludes the same but at the level of the whole viral particle of DENV. Electrostatic properties are also important for the affinity of a specific viral biomolecule with an Ab (Freire et al., 2017). Chapter 6 describes how the p$K_a$ shift ($\Delta$p$K_a$) can be used to define regions called "electrostatic epitopes" (EE) in NS1 proteins of two ZIKV variants. For the estimates of EE's the PROCEEDp$K_a$ (**PR**ediction **O**f ele**C**trostatic **E**pitopes bas**ED** on **p**$K_a$ shifts) method was implemented (see Chapter 6). Moreover, the use of PROCEEDp$K_a$ is extrapolatable to other viral systems such as spike protein of SARS-CoV-2 (Giron et al., 2020, 2021), showing that EE's are general features in this kind of biomolecules. The detection of EE's can participate with the development of a therapeutic alternative for the diagnosis and generation of vaccines (Neamtu et al., 2022).

Another important aspect that is highlighted in this study is related to comprehending the differential virulence expressed by species or strains/variants of flaviviruses through a SB approach. Indeed, cases of differential virulence for ZIKV are documented in Ugandan (ZIKV-UG, with African origin) and Brazilian (ZIKV-BR, with Asian origin) strains (see Section 2.2). The reasons for these discrepancies are not fully understood, but it is evident that they have a molecular basis origin. The differential tropism of a virus are related to mutations that are accumulated in some of the viral proteins expressed (see Section 2.3). Evolutionary studies show that NS1 is the biomolecule along the whole flavivirus genome that positively selects more point mutations (Sironi et al., 2016). The consequences of these mutations were exemplified by crystallography studies that revealed discrepant electrostatic profiles among 3D structures for different flavivirus NS1 proteins (from ZIKV, DENV1, DENV2, and WNV with identity and similarity sequences between 20%−40% and 60−80%, respectively) which were used to reveal discrepant electrostatic profiles (Song et al., 2016; Brown et al., 2016; Xu et al., 2016). This consequently should have an impact on how this viral protein interacts with the virus itself and the molecules of the host (the choice of this viral target is justified in Subsection 2.3.1). Indeed, when NS1 is translated in the rough endoplasmic reticulum (ER) it is secreted by conventional and unconventional routes to the extracellular space (see Subsection 2.3.2) (Watterson et al., 2016). During this process, NS1 depends on the formation of different oligomers (dimers, tetramers, or hexamers) and binding to different lipid membranes. In the extracellular space or the bloodstream, NS1 usually interacts with different proteins of the immune system. We assume that these three types of partnership [*i.e. a)* NS1 itself, *b)* lipid bilayer membrane, and *c)* antibodies] are crucial since the first two [*i.e. a)* and *b)*] define the next steps in the NS1 maturation process, and the remaining partner [*i.e. c)*] has direct

consequences in cell signaling cascades related to virulence. Therefore, the molecular mechanisms that govern the interactions between NS1 and the three partners must be better understood.

Several important SB works with a focus on NS1 and the three main partners were reported in the literature. For example, the dimer stability for $NS1_{ZIKV\text{-}UG}$ (Gonçalves et al., 2019), $NS1_{DENV2}$ (Oliveira et al., 2016; Gonçalves et al., 2019), and $NS1_{YFV}$ (Menezes et al., 2020) proteins was examined through AA MD simulations. These investigations showed important insights on the impact produced by mutations in the monomer-monomer interaction. Moreover, Gonçalves et al. (2019), found that $NS1_{ZIKV\text{-}UG}$ and $NS1_{DENV2}$ present differences in "druggable pockets" localized in the protein-protein interface as a consequence of discrepant MD profiles. From the point of view of NS1-immunoglobulin heterocomplexes, most of the *in silico* works have focused on the prediction of epitope regions using computational tools (Freire et al., 2017; Jones et al., 2017; Songprakhon et al., 2020). These studies were limited to structures but did not consider the impact neither of the dynamics (*i.e.* they are based on static NS1 3D structures) nor of the physicochemical features (*e.g.* the viral protein cannot titrate). However, they have been valuable since the predicted antigenic parts reproduced the epitope regions determined by laboratory experiments. For example, the investigation by Jones et al. (2017) showed that the antigenic predicted parts for the four serotypes of $NS1_{DENV}$ can selectively bind to the Abs 1H7.4, 1G5.3 and GUS2. The confirmations of these assays were done with the enzyme-linked immunosorbent assay (ELISA). Concerning NS1-membrane interactions, there are no computational studies reported in the literature so far. This topic has been examined assuming that the intertwined loop (IL) regions (localized in the wing domain) of different NS1 (from ZIKV-BR, DENV2, and WNV) should interact in different ways due to their residue composition, specific to each flavivirus (Xu et al., 2016).

In the current study, we aim to examine through a bioinformatical approach several physicochemical, energetics, structural and dynamical properties of the flavivirus NS1 of Zika from Uganda and Brazil. The biomolecular interactions of this viral toxin with the three main biological partners aforementioned were examined. This thesis is organized as follows. In Chapter 2, a quick review of the most general aspects of the biochemistry and molecular biology in flaviviruses is given. The relevance of NS1 in the viral life cycle and their association with their main biological partners are highlighted. In Chapter 3, the workflow of the present study is described. Chapter 4 describes the main methodological aspects and pinpoints some theoretical/analytical tools used here. In Chapters 5 we investigate the physicochemical properties of NS1 in an electrolyte solution (*i.e.* without any partner). Subsequently, in Chapter 6, the heterocomplex NS1-Ab was studied and the PROCEEDp$K_a$ method for the prediction of antigenic regions is detailed. The NS1 self-association in a dimerization/tetramerization process is examined in Chapter

7. Dynamical aspects for the NS1 in an electrolyte solution are covered in Chapter 8 and the interaction of the NS1 with a lipid bilayer is evaluated in Chapter 9. Finally, in Chapter 10 the main conclusions are mentioned. The results obtained in this research give new clues about the relationships between two NS1 proteins and the differential virulence expressed in Zika virus strains.

## Chapter 2

# Biological Background

Viruses are biological entities that necessarily require a host to survive (Fields, 2014). In short, a virus is a nucleic acid molecule encapsulated in a protein shell called capsid. Some of them may be enveloped by an organized set of glycoproteins and lipids that work as a vehicle for transporting both nucleic acids and the capsid (Govindjee et al., 2010; Fields, 2014). Since viruses can have different successive generations over time, they need a genetic material to carry out it. In most biological organisms known to date, the biomolecule responsible for the transmission of information is the deoxyribonucleic acid (DNA) and viruses are not the exception (Schmid et al., 2013). However, ribonucleic acid (RNA) can be present in some virus families as well (White et al., 2011). The organization of these biomolecules can occur as a single or double strand with a special polarity (Fields, 2014). Thereby, the replication process of the genetic viral material is dependent on its chemical and structural nature. Based on the molecular mechanisms to process the DNA or RNA to a messenger RNA (mRNA) molecule, the viruses are classified in the so-called "Baltimore classification system" (Baltimore, 1971). The International Committee on Taxonomy of Viruses to date groups 86 virus families within almost 9 orders (ICTV, 2021). Nevertheless, it is estimated that there are about $10^{31}$ viral species on earth, where many of them are still unclassified (Breitbart and Rohwer, 2005; Mushegian, 2020). This makes viruses an intriguing subject for researchers due to their high diversity. Moreover, special attention has been given to viral species infectives of humans. Their ability to spread is a consequence of multiple evolutionary factors that act on the selection of features and physicochemical properties suitable to survive (Fields, 2014). From the inoculation until the virulence phase in a host, it is clear that these biological entities must "overcome" several barriers. Thus, viruses should develop a specific tropism through their specialized molecular machinery generating cellular reprogramming[1] (Nomaguchi et al., 2012).

In this chapter, topics about flaviviruses, their ability to develop different virulence, and the NS1 protein as an important biomolecular target at a clinical and biological level are reported.

[1]For a more detailed view of general aspects of virology, we suggested the book by Fields (2014).

## 2.1 Overview of flaviviruses

One of the virus families with a higher versatility concerning cell reprogramming is the Flaviviridae family (Barrett, 2001; Ye et al., 2013). Within this family of viruses there are four different genera (Simmonds et al., 2017): *i) Flavivirus* (with 53 species), *ii) Hepacivirus* (with 14 species), *iii) Pegivirus* (with 11 species), and *iv) Pestivirus* (with 4 species). Several of them are pathogens of humans and important clinical topics of study. Flaviviruses can produce acute infections in humans that are either asymptomatic or indistinguishable among arboviruses (Musso and Desprès, 2020). They can also promote various damages in the circulatory (*e.g.* hemorrhage or vascular loss) or nervous (*e.g.* encephalitis or microcephaly) systems (Neufeldt et al., 2018). Examples of common flaviviruses are Zika (ZIKV), West Nile (WNV), Dengue (DENV), Japanese Encephalitis (JEV), Yellow Fever (YFV), Tick-borne Encephalitis (TBEV), Murray Valley Encephalitis (MVEV), St. Louis Encephalitis (SLEV), Kunjin Encephalitis (KUV) viruses, among others (Pierson and Diamond, 2020).

DENV leads to dengue fever and dengue hemorrhagic fever (DHF). The first reported case of DHF emerged as a public health problem in 1954 in Manila, Philippines (Gubler, 1998). Gradually, during the 80s and 90s, DENV had spread out to several countries around the world diversifying into 4 different serotypes (Malavige et al., 2004). Recent studies have shown the presence of a fifth variant, probably as a product of the coexistence of transmitting mosquitoes and human populations in sylvatic or emergency areas (Mustafa et al., 2015). This flavivirus is responsible for producing roughly 20 thousand deaths per year mainly in tropical and neotropical zones (WHO, 2021a). It is estimated that in 128 countries 4 billion people are at risk to suffer diseases related to DENV (Brady et al., 2012). Recently, some vaccines have been licensed and developed (Izmirly et al., 2020; Alves et al., 2021). However, their use is yet not available for young children or older adults (Hadinegoro et al., 2015).

Flaviviruses such as JEV, TBEV, WNV, and ZIKV can also be neurotropic (Neufeldt et al., 2018). The first isolation of ZIKV occurred in the Ziika forest of Uganda in 1947 from the Rhesus monkey (Dick et al., 1952). Between the 60s and 70s, ZIKV infections were confirmed in several countries in Africa and Asia (Beckham et al., 2016; Talero-Gutiérrez et al., 2018). Phylogenetic studies showed the presence of three lineages that are related to outbreaks in East Africa, West Africa, and Asia (Faye et al., 2014; Gong et al., 2016). Nonetheless, the first case of infection outside these two continents was reported in 2007 in Yap Island in the Federated States of Micronesia, Oceania (Lanciotti et al., 2008). Afterwards, ZIKV crossed the Pacific generating two new outbreaks in 2013 and 2014 in French Polynesia and the Americas, respectively (Cao-Lormeau et al., 2014; Faria et al., 2016). One of the most critical ZIKV outbreaks was reported in 2015 in the northeast of Brazil (BR), which involve a rapid spread of microcephaly cases in

newborns as a consequence of this flavivirus (Petersen et al., 2016; Mlakar et al., 2016). In addition to this neurological disorder, ZIKV is also known due to its capacity to produce immunopathies such as the Guillain-Barré syndrome (Cao-Lormeau et al., 2016). Although ZIKV infections have gradually decreased over the last few years, there is still a concern since new variants are being generated (Kasprzykowski et al., 2020). To date, and according to the World Health Organization, 86 countries are reported with risk or vulnerability to the flavivirus ZIKV (WHO, 2021b). Furthermore, there are no approved antiviral drugs for the treatment of ZIKV infections (Collins and Metz, 2017).

The vast number of geographic areas where these flaviviruses have reached out is primarily due to "exogenous" and "endogenous" agents. Regarding the first, it is well known that flaviviruses are transported in mosquitoes of the genus *Aedes*, where humans are infected by a bite and the virus is transferred to the bloodstream[2]. Since the distribution of *Aedes* mosquitoes covers a wide range of tropical areas (in part by uncontrolled population growth, urbanization, and inadequate treatment of wastewater), flaviviruses are transported to different places where human populations are most exposed to them (Guzmán and Kouri, 2002; Mansfield et al., 2009; Vasilakis and Weaver, 2017). On the other hand, the second reason that could explain such successful spreading could be related to the molecular plasticity presented in these biological entities. In general, viruses have a high mutation rate that promotes adaptations to novel tropisms in order to increase their chances to survive (Drake et al., 1998; Sniegowski et al., 2000). Evolutionary mechanisms like natural selection, genetic drift, or molecular recombination lead to an increase in the diversity within the taxonomic viral hierarchy (Manrubia and Lázaro, 2006). In this way, genetic variation initially affects a given virus generating new variants/strains, and later, by a successive accumulation of mutations, new species can appear with a convergent[3] functionality when compared to the ancestor (Geoghegan and Holmes, 2018).

To treat diseases and/or to prevent virus infections, it is necessary to precisely identify viruses (*e.g.* through viral genome sequencing) found in infected patients (*i.e.*, diagnosis) to formulate an appropriate treatment. In flaviviruses, the diagnosis is made through laboratory tests based on molecular detection of the viral genome in the first days of infection [*e.g.* by conventional reverse transcription-polymerase chain reaction (RT-PCR) or quantitative RT-PCR (RT-qPCR)], followed by identification of specific antibodies (Ab) employing serological tests after the first week of the virus' inoculation (Musso and Desprès, 2020). The high number of different flaviviruses co-circulating particularly in economically vulnerable regions make access to sophisticated PCR laboratories prohibited (Mackenzie et al., 2004; Paixão et al., 2018; Musso and Desprès, 2020). The diagnosis

[2]The vectors of some flaviviruses can also be ticks.

[3]When we use the word "convergent", it must interpreted from the context of evolutionary biology. Following the definition of Kasahara (2010), the evolutionary convergence is "*the process whereby distantly related organisms independently evolve similar traits to adapt to similar necessities*."

of the specialist has been considered as the principal strategy to determine a flavivirus infections in that areas. However, more rigorous approaches such as serological tests are often necessary in order to define the type of flavivirus present in the patient. Enzyme-linked immunosorbent assay (ELISA), immunofluorescence tests (IFT), or lateral flow are the most commonly used serological laboratory methods (Musso and Desprès, 2020). Nevertheless, serological methods tend to show cross-reactivity into IgM or IgG immunoglobulins (Beck et al., 2013; Priyamvada et al., 2016; Stettler et al., 2016; Dejnirattisai et al., 2016). Thus, other clinical trials like plaque reduction neutralization tests (PRNT) are necessary to complement and confirm the specific flavivirus target (Paixão et al., 2018). The cross-reactivity is a biological phenomenon that occur when a person previously infected with a flavivirus (*e.g.* DENV), in a subsequent infection with another member of the genus (*e.g.* ZIKV or even another DENV strain) generates a more aggressive immune response [autoimmune diseases, (John and Rathore, 2019)] or facilitates the mechanism called antibody-dependent enhancement [ADE, (Dejnirattisai et al., 2016)]. In some cases, diseases produced by cross-reactivity can be treated. In other situations, they can leave sequelae or be fatal [*e.g.* a recurrent dengue-infected patient can develop DHF (John and Rathore, 2019; Bonheur et al., 2021)].

## 2.2 Differential virulence

Briefly, the pathologies caused by flaviviruses can be divided into encephalitis (produced by ZIKV, WNV, or JEV, for example) and hemorrhagic (produced by the well-known DENV or YFV) diseases (Neufeldt et al., 2018). Although flavivirus pathology is not fully understood, several viral molecular mechanisms that affect the host are related to damages to the immune system or vascular endothelial dysfunction (Ye et al., 2013; Puerta-Guardo et al., 2019). As well described in the review study by Ye et al. (2013), the biochemical interactions of the different viral proteins can occur with several host biomolecules, such as, interferons, membrane receptors, gene factors, complement system proteins, or biomolecules involved in the apoptosis modulation. The interactions between the pathogen and the host macromolecules are the perfect "cocktail" whereby each flavivirus triggers a specific signaling cascade converging on a determined differential virulence.

So far, it is clear that at the species level, viruses behave differently. However, among strains of a specific virus, they may also express discrepant pathologies (despite high genome similarity and identity of the variants). Regarding ZIKV, it has been reported that the strains from African or Asian lineages induce distinguishable antiviral immune responses in human infected astrocytes (Hamel et al., 2017). This has also been proved in other cells like the human fetal brain neural stem cells (hNSCs) or trophoblasts (Simonin et al., 2016; Sheridan et al., 2017; McGrath et al., 2017). Similarly, model organisms such

as mice have shown different virulent phenotypes for African and Asian ZIKV strains (Shao et al., 2017; Aubry et al., 2021). The most remarkable hypothesis according to this experimental evidence is that the African ZIKV resulted in early termination of pregnancy, while Asian strains are less destructive allowing pregnancy to continue, resulting in a subsequent fetal microcephaly (Aubry et al., 2021). This fact has a high relationship with the abrupt peak of microcephaly cases reported in 2015 in northeast of Brazil in women infected with ZIKV[4] (Faria et al., 2016).

## 2.3 Molecular constitution of flaviviruses

Flaviviruses are enveloped viruses of approximately 50 nm with icosahedral symmetry (Kuhn et al., 2015). When viral particles enter into an organism, they bind to susceptible cells[5] via receptors or binding factors [*e.g.* concerning ZIKV the entry mechanisms can be given via interaction of the envelope protein (E) of the virus with the cell receptors such as C-type lectin, TIM (T cell immunoglobulin mucin domain protein 1), or TAM (tyrosine protein kinase receptor 3 (TYRO3)–AXL–MER)] (Cruz-Oliveira et al., 2015; Neufeldt et al., 2018; Agrelli et al., 2019). Subsequently, viruses are endocytosed via a clathrin-dependent entry pathway (Cruz-Oliveira et al., 2015), where the exposure to a more acidic pH medium [around pH 5.5 (Stiasny et al., 2002)] generates a rearrangement of the viral surface particle releasing the nucleocapsid and the genetic material into the cytoplasm (Modis et al., 2004). The flavivirus genome is a single positive-sense RNA or (+)ssRNA with roughly 10 kb, divided into different functional regions (Figure 2.1) (Lindenbach and Rice, 2003; Neufeldt et al., 2018). At extremes 5' and 3', there are non-translating regions (NTR). These NTR sequences can form RNA secondary structures that are involved in the recognition of the ribosomal machinery of the host (Lindenbach and Rice, 2003). Also, the 5' end has a cap-like structure that works for protection against ribonucleases (Liu et al., 2010). The single-stranded RNA presents one open reading frame (ORF) that translates to a polyprotein of three structural and seven non-structural (NS) proteins. The synthesis of the polyprotein is carried out in the lumen of the rough endoplasmic reticulum (ER).

The three structural proteins are the capsid protein (C) that binds viral RNA, the pre-membrane (prM) protein that helps to block the premature viral fusion and assists the folding of the E protein, and the E protein that mediates the membrane fusion, viral attachment, and viral particle assembly (Bressanelli et al., 2004). Concerning the non-structural proteins, the NS1 is a multifunctional viral factor (see Subsection 2.3.2); NS2A is involved in viral replication; NS2B is a cofactor of NS3 that helps its coupling with

[4]The Brazilian strain of ZIKV has Asian origin (Faye et al., 2014; Kochakarn et al., 2016).

[5]Examples of cells can be monocytes, skin dendritic (*e.g.* DENV), neurons, or trophoblasts (*e.g.* WNV and ZIKV) (Neufeldt et al., 2018).

the ER membrane; NS3 is a biomolecule with protease and helicase activities; NS4A is an integral membrane protein responsible for defining the membrane curvature; 2K is a signal peptide for co-translating NS4B and allows its insertion into the ER membrane; and NS5 is a protein with two structural domains with different functions, the first one (N-terminal) presents guanylyltransferase, guanine-N7-methyltransferase, and nucleoside-2'-O-methyltransferase activities, and the second one (C-terminal) is responsible for the viral RNA synthesis (Lindenbach and Rice, 2003; Fields, 2014; Cox et al., 2015; Neufeldt et al., 2018).

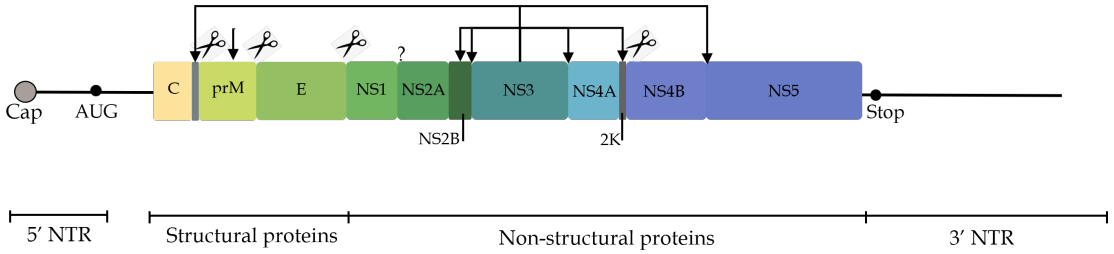
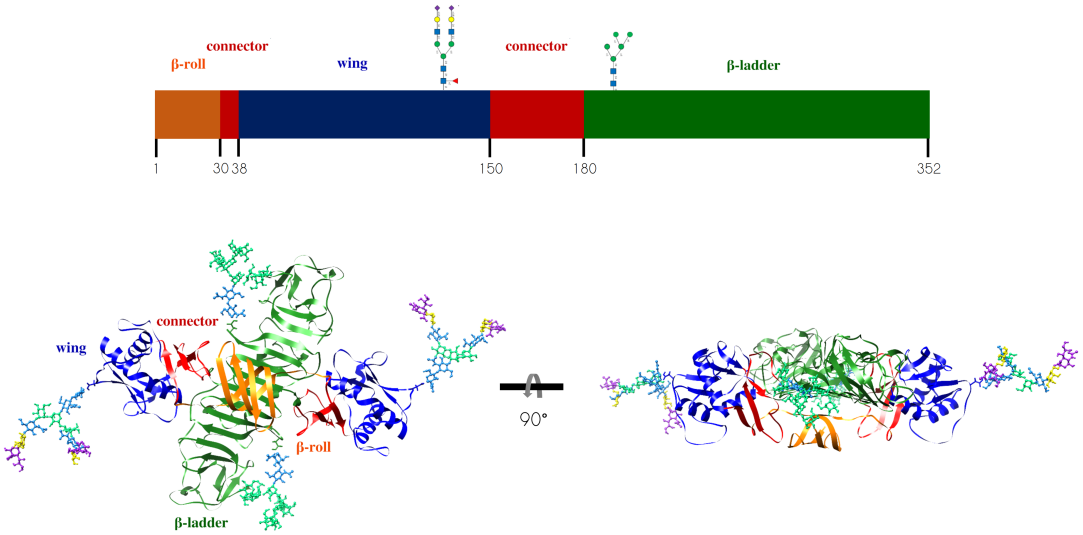



Figure 2.1: Flavivirus genome organization. Scissors and arrows are spot cleavages by, respectively, cellular signal peptidases and viral proteases, whereas the vertical small arrow close to prM and the question mark are cleavages that happen via Golgi apparatus-resident furin and an unknown protease, respectively.

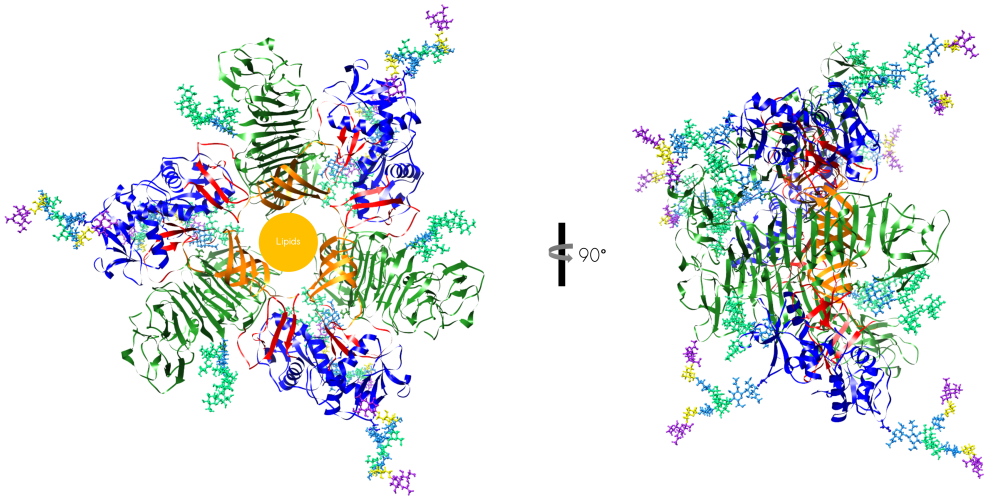
### 2.3.1 Selection of study protein

All flavivirus proteins fulfill some specific functions. Computational studies have shown that various interactions happen between the flavivirus macromolecules and the host proteins (Dey et al., 2017; Scaturro et al., 2018). Structural proteins such as prM or Env have been extensively studied to comprehend the processes of virus entry into the host cell (Cruz-Oliveira et al., 2015; Hu et al., 2021). Furthermore, structural proteins have been the main focus for the vaccine development (Anasir et al., 2020). On the other hand, non-structural proteins such as NS3 and NS5 have been important targets in studies because of their role in viral replication processes. This has motivated research for the design of antiviral drugs based on NS3 and NS5 (Bollati et al., 2010). Nevertheless, among all the networks of interactions and flavivirus players, the NS1 protein (Figure 2.2) catches attention since it is involved in the interaction with critical proteins for the innate and adaptive immune responses such as interferon type I (Ye et al., 2013), toll-like receptor 4 (TL4) (Modhiran et al., 2015), and neutralizing/non-neutralizing Abs (Muller and Young, 2013; Watterson et al., 2016). This viral toxin is responsible for triggering different reactions that lead to hyperpermeability on the human endothelial cells as well (Puerta-Guardo et al., 2016, 2019). Interestingly, depending on the type of NS1 flavivirus, endothelial damage can occur in a specific way [*e.g.* $NS1_{YFV}$ and $NS1_{WNV}$ have special affinity by liver and brain endothelial cells, respectively (Puerta-Guardo et al., 2019)]. In

addition, it is hypothesized that this viral protein creates the appropriate extracellular medium to facilitate the entry of virions to vital organs (Puerta-Guardo et al., 2016, 2019). For the specific case of ZIKV, it has been found that NS1 has a greater affinity for umbilical vein and brain endothelial cells (Puerta-Guardo et al., 2019), which may give clues about how this virus overcomes placental barriers (Jurado et al., 2016). From the point of view of replication, this protein is a key co-factor for the assembly of the replication complex (RC). It has also been shown that NS1 assists the formation of the replication organelle by facilitating membrane curvature (Ci et al., 2020, 2021).




(a)




(b)

Figure 2.2: Three-dimensional representation of the nonstructural protein 1. (a) X-ray crystal structure of $NS1_{ZIKV\text{-}BR}$ dimer (PDB id 5GS6). The NS1 domains are distinguished at the primary and tertiary structures. Glycosylated residues are shown. The 90° rotation about the $x$-axis for the dimeric NS1 represents the estimated position of interaction with the lipid bilayer. (b) The modeled putative $NS1_{ZIKV\text{-}BR}$ hexamer obtained by the PDB id 4O6D.

NS1 is a key biomarker used for flavivirus detection (Figueiredo et al., 2015; Cecchetto et al., 2017). Nevertheless, experimental assays have been shown that this antigen is susceptible of cross-reactivity. For instance, it has been detected that specific Abs that recognize $NS1_{DENV}$ can also associate with neurotropic flavivirus proteins such as $NS1_{ZIKV}$, $NS1_{WNV}$, or $NS1_{JEV}$ (Biering et al., 2021). Moreover, this phenomenon can also happen at the level of flavivirus variants (Stettler et al., 2016). Some studies using peptide microarrays seem promising strategies because they have high specificity distinguishing NS1 fragments of DENV variant 2 (DENV2) and ZIKV (Freire et al., 2017). Meanwhile, other techniques based on electrical detection open new paths in the development of new serological tests (Figueiredo et al., 2015). These virus identification techniques are shown as an alternative to the problem of cross-reactivity. Finally, in terms of vaccine design, NS1 is a good candidate because of the low risk of developing ADE (Sridhar et al., 2018; Li et al., 2018; Wessel et al., 2021).

The versatility of interactions that NS1 has with different biomolecules and its use as an important clinical target, make it an ideal candidate for the study of molecular mechanisms related to differential virulence.

### 2.3.2 Non-structural protein 1

The NS1 is a glycoprotein that consists of 352 amino acids (monomeric state) with a weight between 46 to 50 kDa[6] (see Figure 2.2) (Muller and Young, 2013; Watterson et al., 2016). When the NS1 is translated into the ER lumen its linear structure is spontaneously folded resulting in a 3D structure that is stabilized by disulfide bridges (Figure 2.3, steps 1−3) (Blitvich et al., 2001; Wallis et al., 2004). As shown in Figure 2.2a, the NS1 chain has three structural domains called β-roll (1−30), wing (38−150), β-ladder (181−352), and a connector subdomain (31−37 and 151−180) (Akey et al., 2014; Xu et al., 2016). On the wing and β-ladder there are two N-glycosylation sites in residues N130 and N207, respectively[7] (Blitvich et al., 2001).

In the cytosol, the monomeric NS1 has a hydrophilic character and undergoes translational modifications where the oligosaccharyltransferase complex links mannose glycans (Watterson et al., 2016). Later, an increase in the monomeric NS1 concentration leads to dimerization (see Figure 2.3, step 4) (Winkler et al., 1988). The NS1 in the dimeric state is more liposoluble, allowing it to have a higher affinity for the ER membrane (see Figure 2.3, step 5) (Winkler et al., 1989; Akey et al., 2014; Brown et al., 2016; Xu et al., 2016). The NS1 homodimer is resistant to ionic and non-ionic detergents (Winkler et al., 1989), and to be dissociated requires high temperatures or acidic pH media (Falconar and Young, 1990). The NS1 cell-membrane-associated (mNS1) presented in vesicular compartments is divided into three different populations (Watterson et al., 2016). A

[6]The weight is depending on the glycosylation degree (Somnuke et al., 2011).
[7]Depending on the flavivirus, the N-glycosylation sites can be two or three.

significant fraction assists the RC, where the mNS1 also produces a remodeling of the ER membrane to make vesicle packets (VPs) (see Figure 2.3, step 8) (Mackenzie et al., 1996; Ci et al., 2020). Moreover, it has been shown that the prM-E-NS1 complex is also responsible for the packaging of viral particles or virions (see Figure 2.3, step 9) (Scaturro et al., 2015).

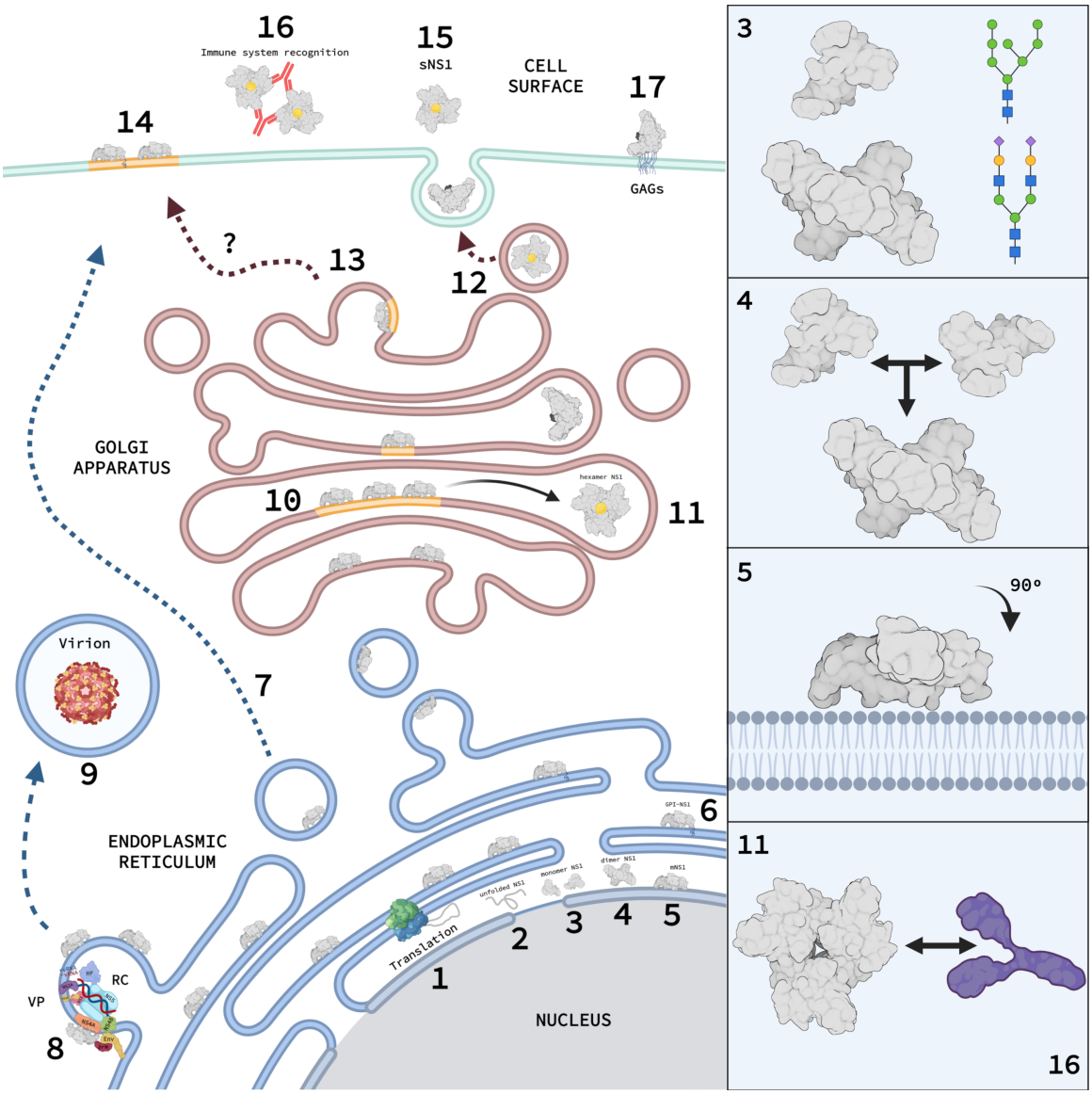



Figure 2.3: The NS1 protein trafficking within an infected cell. The steps highlighted on the right-hand side refer to the main aspects examined in this study. More details are given in the text. Question marks are used to indicate unknown routes of secretion. (Created with BioRender.com). (Created with BioRender.com).

A second minor population is modified with a glycosylphosphatidylinositol (GPI) anchor (see Figure 2.3, step 6) (Jacobs et al., 2000), to later be trafficked to the plasma membrane surface (see Figure 2.3, steps 7 and 14) (Winkler et al., 1989; Watterson et al., 2016). Note that the mNS1 without GPI can also be secreted as shown in step 7 (Figure 2.3).The third fraction of mNS1 transits from the ER to the Golgi apparatus and can

associate to microdomains or lipid rafts (see Figure 2.3, step 10) (Noisakran et al., 2008). Since NS1 has been shown to have a high affinity to triglycerides, it is speculated that mNS1 is preferentially attached to lipid droplet (LD) nascent (Gutsche et al., 2011). Later, this condition facilitates the assembly of dimers to become hexamers (see Figure 2.3, step 11). Simultaneously, Golgi enzymes edit glycans converting the high mannose sugars to complex carbohydrates (especially in N130 residue) (Watterson et al., 2016). In addition, the hexameric NS1 has a hydrophobic core that helps to give greater structural stability (Gutsche et al., 2011). This core bind several lipids that are interacting in the β-roll regions of the hexameric NS1 (see Figure 2.2b) (Akey et al., 2014). Next, the NS1 is secreted to the extracellular media in higher proportion as a hexamer (sNS1) and minor quantity as a dimer (see Figure 2.3, steps 12−13) (Flamand et al., 1999). Finally, the sNS1 can interact with several essential proteins of the host immune system [*e.g.* TL4 or Abs (see Figure 2.3, steps 15−16) (Muller and Young, 2013; Modhiran et al., 2015; Watterson et al., 2016)]. Moreover, sNS1 can also be bound to the cell surface utilizing membrane receptors [*e.g.* lectins, Fc, TL4 (Modhiran et al., 2015; Thiemmeca et al., 2016; Watterson et al., 2016)] or polysaccharides [*e.g.* glycosaminoglycans (GAGs) (see 2.3, steps 17) (Avirutnan et al., 2007; Puerta-Guardo et al., 2016)].

The trafficking and secretion processes of NS1 mainly depends on the association with three key biological partners: *i)* NS1 itself (self-association), *ii)* antibody, and *iii)* host cellular membrane. For this reason, these three target-macromolecules were investigated in this thesis. These most important biological systems are highlighted in the lateral part in 2.3.

## Chapter 3

# Workflow

In this chapter, the workflow followed during this study is described. Thus, we seek to give a quick overview on how the different biological systems were studied (*i.e.* the NS1 in an electrolyte solution, NS1 in a self-association process, NS1 interacting with an Ab, and NS1 binding to a lipid bilayer). Emphasis was given to the different computational models implemented for each case and the physical variables measured. The details of the computational models used for each system and the physical variables measured are more detailed in the following chapters. An outline of the workflow is shown in Figure 3.1 and is described below.

The starting point are crystal structures with PDB ids 5K6K (1.89 Å, pH 8.5) and 5GS6 (2.85 Å, pH 8.5) of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$, respectively (Figure 3.1, item 1). For other NS1 strains other 3D models were also built. The NS1 structures from Senegal, Central African Republic, Malaysia, Thailand, and Yap Island were modeled through an homology protocol (see Chapter 5). Note that the first two and the remaining variants have African and Asian origins, respectively. Moreover, the heterocomplex of $NS1_{WNV_176\text{-}352\text{-}Ab}$ [PDB id 4OII (2.85 Å, 5.4)] was also employed. $NS1_{WNV_176\text{-}352}$ is an incomplete structure where only the C-terminal region is present. Thus, for comparison purposes, the N-terminus of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ were removed. This was also the case for the prediction of antigenic regions (see Chapter 6). Regarding the NS1 self-association study, additional models were built for $NS1_{DENV2}$ and $NS1_{WNV}$ 3D structures (see Chapter 7).

Two complementary approaches were followed to computationally analyze the viral proteins: *i)* "*protein-protein complexation phenomena*", where Constant-pH (CpH) Monte Carlo (MC) simulations were employed (Figure 3.1, item 11) [approach *i*], this approach *i* being more focused on the electrostatic interactions also integrates semiquantitative DLVO [Derjaguin, Landau, Verwey, and Overbeek (Dahirel and Jardat, 2010)] analysis (Figure 3.1, item 12); and *ii)* "*dynamical features*" of the NS1 protein in different situations (*e.g.* as a monomer, as a dimer or interacting with a membrane) [approach *ii*]. We looked at motions and contacts at the local level (*e.g.* residue), considering essential dynamics, or cross-correlation dynamics, using classical Molecular Dynamics (MD) simulations (Figure 3.1, item 10).

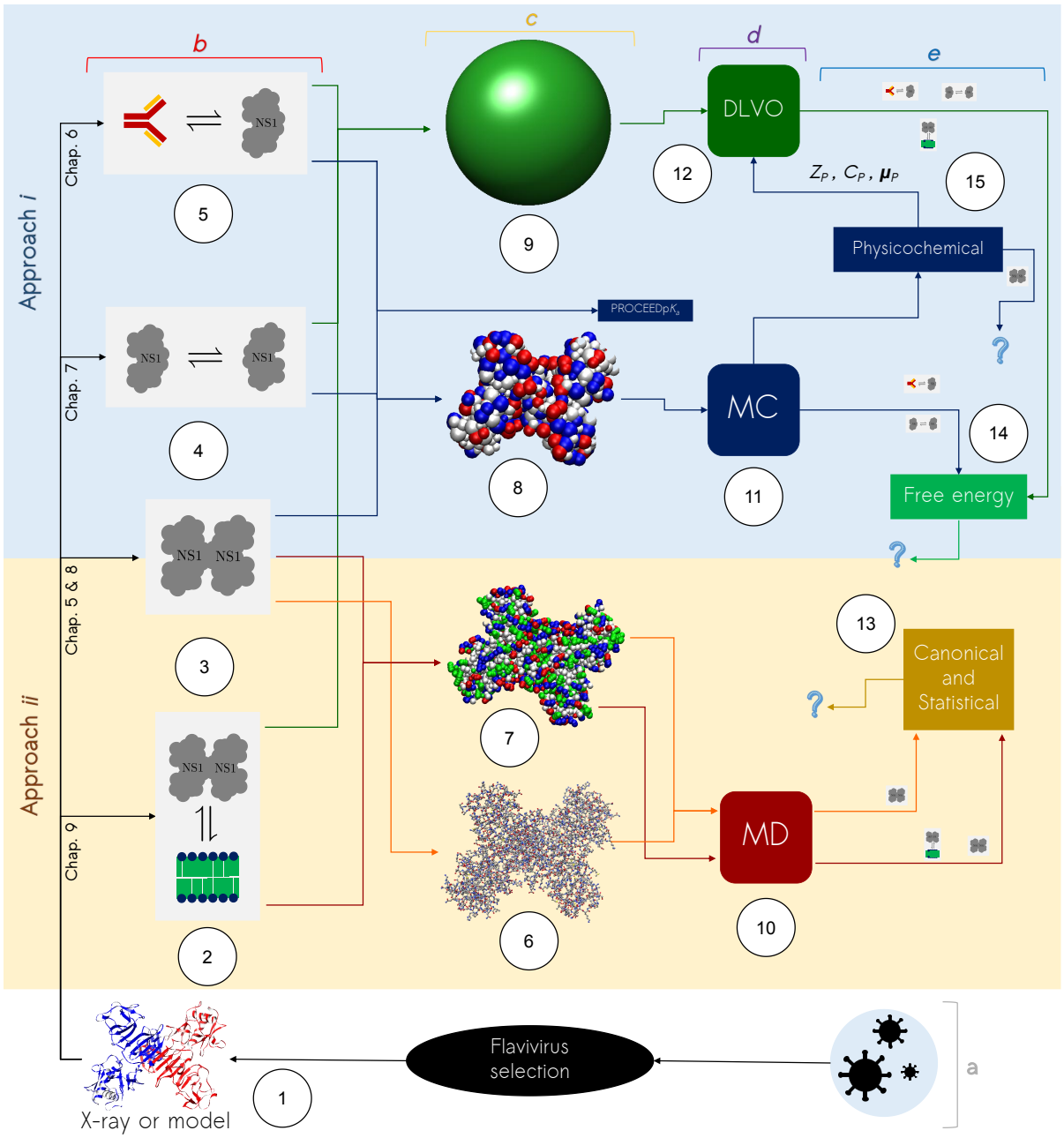



Figure 3.1: Summary of the present study. The "protein-protein complexation phenomena" (top) and "dynamical features" (down) approaches (*i* and *ii*, respectively) are illustrated in the diagram. The question marks are related to the queries formulated in Chapters 5-9. More details are discussed in the text.

Employing the first approach, we investigated the heterocomplex NS1-Ab (Figure 3.1, item 5), for the incomplete $NS1_{176\text{-}352}$ from ZIKV-UG, ZIKV-BR, and WNV (see Chapter 6). Moreover, the self-association of the NS1 protein in dimerization and tetramerization processes was also examined (Figure 3.1, item 4) for different strains of NS1, that is, NS1 from ZIKV-UG, ZIKV-BR, DENV2, and WNV (see Chapter 7). In addition, the monomeric and dimeric forms in the absence of a biological macromolecular partner were also explored (Figure 3.1, item 3) for several several $NS1_{ZIKV}$ with African and Asian

origins (see Chapter 5). For all these cases, the atomistic crystallographic coordinates given by the PDB for these protein structures were converted to a coarse-grained (CG) at the amino acid level (Figure 3.1, item 8). From these systems CpH MC simulations were performed and physicochemical properties such as the total protein charge number, charge regulation capacity, dipole moment number, and p$K_a$s[1] for all ionizable groups are obtained (Figure 3.1, item 15; and Chapter 4). Free energies of interactions for the protein-protein systems were also calculated (Figure 3.1, item 14). It is important to mention that from the MC calculations, the electrostatic properties obtained were used as an input for free energy estimates using the DLVO theory (see the transition from items 15 to 12 in Figure 3.1). Thanks to its lower cpu costs, the DLVO approach was useful to investigate the $NS1_{ZIKV}$ strains and gain initial insights about them (see Chapter 5). This simplified calculation identified important features and guided the next steps of the work. For instance, they allow us to define the strains from Uganda (African origin) and Brazil (Asian origin) as the most relevant for the next deeper investigations.

The dynamical properties of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ were evaluated by both AA (Figure 3.1, item 6) and CG (Figure 3.1, item 7) representations (some aspects about the models are discussed in Chapter 4). Here, instead of a CpH approach, constant-charge MD simulations were carried out. They provided not only time-dependent properties but were more generally used as a tool for sampling the conformational space. This part of the study aims to capture molecular motions or dynamical properties with biological relevance. As a first stage, MD simulations were performed to evaluate a single $NS1_{ZIKV}$ protein in an electrolyte solution in the absence of lipids and any other macromolecule (Figure 3.1, item 3). To investigate this system, canonical and statistical analyses were used (Figure 3.1, item 13). Chapter 8 discusses this part of the study. This work was further extended to include the presence of a membrane (see Chapter 9).

In summary, the stages of this study (Figure 3.1) follow the steps detailed below:

*a)* The coordinates deposited in the PDB for the viral protein of interest are downloaded.

*b)* A case of study (*i.e.* the NS1 in the presence or absence of a partner) is chosen.

*c)* The molecular model for the system of study is assigned (see Section 4.1).

*d)* Computational simulations based on deterministic or stochastic methods are performed (see Section 4.2).

*e)* Dynamical or physicochemical properties are calculated in order to answer a specific question (see Section 4.3).

[1]This electrostatic feature is so relevant in the systems of study. To evaluate its impact we used the software program developed by us called CPRISMA (**C**oloring **PR**oteins by **I**nputs and **S**ets of **M**ultiple **A**lignments). See Appendix C for more details.

## Chapter 4

# Computational Aspects

Biomolecular systems such as viral proteins can be represented by molecular models. A model is an abstraction or idealization of the reality that allows a specific process to be described. Models are often mathematical constructions that intend to reproduce an observable phenomenon (Gräbner, 2018). The so-called *molecular modeling* is the process by which molecules can be described through mathematical functions that represent the physical interactions of the sites[1] used to define the molecule (Andrew, 2001). The solution of these models allows the extraction of relevant physical information (*e.g.* the free energy of interactions for two proteins in an electrolyte solution, titration plots, and so on) (Barroso da Silva, 2000; Jönsson et al., 2007). In principle, one could solve these models by hand. Molecular modeling incorporates many parameters (*e.g.* atoms weight, equilibration positions in an interaction potential, particle sizes) and variables (*e.g.* radius of gyration, p$K_a$ in a titratable system, residue fluctuation). Computer simulations are performed to generate configurations for a given model. A vast number of different models are available in the literature to represent biomolecular systems depending on specific purposes (Elliott, 2011; Kmiecik et al., 2016; Barroso da Silva et al., 2020).

In Structural Bioinformatics (SB), models utilize three dimensional information of the molecules to describe many processes as conformational changes, folding, flexibility, stability, ligand binding, assembly pathways, among others (Andrew, 2001; Karplus and McCammon, 2002; Hollingsworth and Dror, 2018). Then, to describe this type of phenomenon, the so-called force fields (FFs) are employed. These are a set of equations and parameters that allow representing potentials of intramolecular and intermolecular potential energy of interactions. By defining a FF, the total energy associated with the system studied can be estimated, which in turn allows the calculation of physical properties of interest.

In the next sections, a description of the FF employed to examine the four different systems (*i.e.* NS1 in an electrolyte solution, NS1 in a self-association process, NS1 interacting with an Ab, and NS1 binding to a lipid bilayer) in this study and the corresponding approximations are exposed. In addition, some details of the two

[1] A site can be an atom or a group of atoms depending on the resolution to describe the system in a classical context.

computational simulation methods used here (*i.e.* MD and MC calculations) are given. Finally, the most frequently calculated and analyzed dynamic and physicochemical properties in our work are described.

# 4.1 Multiscale modeling

## 4.1.1 Components of the molecular model

The systems studied here generally consist of at least three components: *i)* a solvent (usually water), *ii)* salt ions (counterions and added salt), and *iii)* biomolecules (*e.g.* protein, membrane, or associated sugars). The atomic coordinates of the NS1 crystal structures are our starting point. The finite-sized protein is usually solvated with a few layers of water molecules ["explicit solvent model" (Lemkul, 2018)]. For the solute-solvent interactions to be correctly calculated through this approach, special methods are needed (see Subsection 4.1.5). However, the water solvation can also be represented implicitly or by a "continuum model", where the main properties of the molecular water are averaged out and effectively described by the relative permittivity or dielectric constant of the water (de Lima et al., 2009). Similarly, the ionic strength can be explicitly incorporated in the simulation box (*i.e.* the position and charge of each ion particles is included in the model) or implicitly [*i.e.* when a Debye-Hückel (DH) approximation is assumed] (Jönsson et al., 2007). The macromolecule is often represented in all-atoms (AA) in most classical FFs (Harrison et al., 2018). Yet, it can also be reduced to a coarse-grained (CG) modeling the atoms of a macromolecule are grouped into effective CG pseudoatom sites (Lund and Jönsson, 2003; Kmiecik et al., 2016), or even at a simple sphere [the typical approximation made by physicists (Phillies, 1974; Jönsson et al., 2007)]. These simplifications are generally applied to limit the number of calculations and to allow larger solute systems to be examined for the same amount of cpu time (Andrew, 2001). Next, important considerations about the FFs used in this thesis are highlighted.

## 4.1.2 Atomistic representation

Atomistic models assume the functional form for the interaction between the sites as described by the following formula,

$$U\left\{\mathbf{r}^N\right\} = u^{bond} + u^{angle} + u^{dihedral} + u^{improper} + u^{vdw} + u^{el}, \tag{4.1}$$

where $U\left\{\mathbf{r}^N\right\}$ is the total potential energy for the $N$ atoms at positions $\mathbf{r}$. The first four terms in Equation 4.1 describe the bond stretching $\left(u^{bond}\right)$, the angle bending $\left(u^{angle}\right)$, the torsion $\left(u^{dihedral}\right)$, and the improper torsion $(u^{improper})$ of the particles, respectively. Figures 4.1a, 4.1b, 4.1c, and 4.1d are representing, respectively, the variables that these

potentials depend on. The last two terms in Equation 4.1 are representing the non-covalent interactions that describe the van der Waals interactions and repulsion effects in close contact [that are generally modeled by Lennard-Jones (LJ) function, $\left(u^{vdw}\right)$], and the electrostatic interactions [generally approximated by Coulombic terms, $\left(u^{el}\right)$]. The terms in Equation 4.1 are better described below.

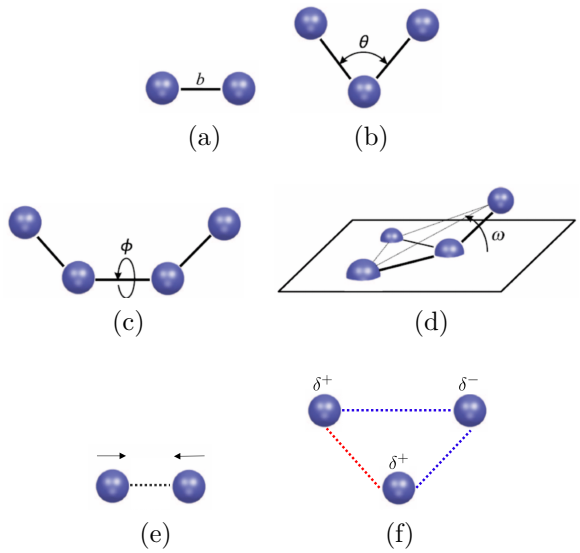



Figure 4.1: Schematically representation of the covalent and non-covalent bonds as used in atomistic models. (a) Bond stretching $\left(u^{bond}\right)$. (b) Angle bending $\left(u^{angle}\right)$. (c) Dihedral rotation $\left(u^{dihedral}\right)$. (d) Improper dihedral torsion $(u^{improper})$. (e) van der Waals Interactions $\left(u^{vdw}\right)$. (f) Electrostatic interactions $\left(u^{el}\right)$. See the text for more details.

These "bonded" terms are classically represented by harmonic potentials and trigonometric functions:

$$u^{bond}(b) = \sum_{bonds} k_b \left(b - b_{eq}\right)^2, \tag{4.2}$$

$$u^{angle}(\theta) = \sum_{angles} k_\theta \left(\theta - \theta_{eq}\right)^2, \tag{4.3}$$

$$u^{dihedral}(\theta) = \sum_{torsions} \frac{V_n}{2} \left[1 + \cos\left(n\phi - \gamma\right)\right]^2, \tag{4.4}$$

$$u^{improper}(\omega) = \sum k_\omega \left(\omega - \omega_{eq}\right)^2, \tag{4.5}$$

where $b$ is the bond length; $\theta$ is the atomic angle; $\phi$ is the dihedral angle; $\omega$ is the improper dihedral angles; $k_b$, $k_\theta$, and $k_\omega$ are force constants; $V_n$ is the $n$-fold rotation barrier, and $b_{eq}$, $\theta_{eq}$, $\omega_{eq}$ are equilibrium positions. The source of the force constants can be obtained from infrared or NMR spectra; or through quantum calculations such as potential energy

surface. Values of $k_b$, $k_\theta$, and $k_\omega$ are typically in the ranges of 300-600 kcal mol$^{-1}$Å$^2$, 100-150 kcal mol$^{-1}$rad$^2$, and 1-3 kcal mol$^{-1}$, respectively. Equilibrium positions can be established by data from crystallography, NMR, or electronic density maps. The three first terms correspond to short-range interactions involving atoms separated by at most 3 covalent bonds. The bonded stretching term $\left(u^{bond}\right)$ is based on Hooke's law and represent the interaction between pairs of bonded atoms. The $\left(u^{angle}\right)$ is also built from Hooke's law and represents the interaction and angle formed between three atoms $i$, $j$, and $k$ in which $i$ and $k$ are both bonded to $j$. The dihedral term $\left(u^{dihedral}\right)$ is a periodic function characterized by the multiplicity ($n$) and the phase shift ($\gamma$). The variable $n$ is a value related to the minimum points that the function can have when the bond is rotated through 360°, whereas $\gamma$ establishes when the torsion passes through the minima in the function (Andrew, 2001). The last term ($u^{improper}$) is introduced to correct chirality errors of asymmetric carbon or out-of-plane deformation of aromatic rings.

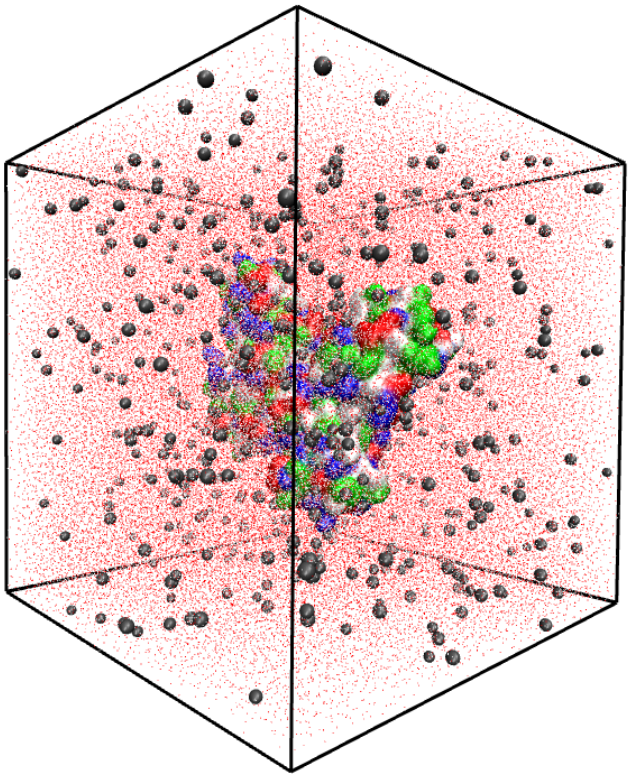


Figure 4.2: Example of a simulation box for a biomolecular system represented at the all-atoms modeling level. The $NS1_{ZIKV\text{-}BR}$ protein (PDB id 5GS6) is displayed in a surface representation. Salt ions and water molecules are appearing in colored spheres and red points representation, respectively.

The pair LJ and electrostatic potential energies of interaction are defined by Equations 4.6 and 4.7, respectively:

$$u^{vdw}(r) = \sum_{i<j}^{atoms} 4\varepsilon_{LJ}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]. \tag{4.6}$$

$$u^{el}(r) = \sum_{i<j}^{atoms} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}, \tag{4.7}$$

The LJ potential energy function requires two parameters: $\varepsilon_{LJ}$ that describes the potential-well, and $\sigma$ that is the minimum contact distance between the particles $i$ and $j$. The LJ potential energy depends also on the variable $r$ that is the separation distance between the $i$-th and $j$-th atoms. Note that the first term with the power 12 is repulsive and is supposed to mimic the Pauli-exclusion principle between the atoms at short separation distance. Meanwhile, the term with power 6 instantaneous dipole-induced dipoe interaction (London dispersion) (Figure 4.1e). Regarding the Coulombic equation, the $q$'s are the charges for the $i$-th and $j$-th atoms (given in C) and $\epsilon_0$ is the dielectric constant of the vacuum $\left(= 8.854 \times 10^{-12}\ \mathrm{C^2/N\ m^2}\right)$. Figure 4.1f shows a representation of the electrostatic interactions for a set of three particles. Note that while the partial charges ($\delta$) of two atoms have opposite signs an attraction is produced (see the blue dashed line in Figure 4.1f), while a repulsive behavior persists when the atoms have the same partial charge (see the red dashed line in Figure 4.1f).

Most AA FFs incorporate terms such as those described in Equation 4.1. The most popular atomistic classical FFs are CHARMM (Brooks et al., 1983), AMBER (Weiner et al., 1984), OPLS (Jorgensen et al., 1996), or GROMOS (Hermans et al., 1984). For the present study, we used CHARMM36m (see Chapter 8). This FF has been refined based on QM potential energy surfaces and NMR data in order to reproduce structural features such as secondary structures and intrinsically disordered regions. (Huang et al., 2017). This FF also considers water and salt molecules explicitly. CHARMM36m is specifically parameterized to match the water model TIP3P (Huang et al., 2017). A drawback is that the pH effect is not incorporated since atomic charges are kept constant[2]. Based on all these characteristics, the CHARMM36m FF is ideal for studying the dynamic properties of NS1 protein in an aqueous solution. For example, some intrinsically disordered regions of NS1 that are difficult to detect by X-ray assays (Appendix A) can be characterized by this FF. Furthermore, the results shown in Chapter 8 match with other computational approaches where FFs like AMBERff12SB (Gonçalves et al., 2019) or GROMOS 54A7 (Freire et al., 2017) were utilized as well. CHARMM36m has also been tested in several types of systems showing high transferability and reliability as mentioned in the original study by Huang and MacKerell Jr (2013). Moreover, proteins from Ebola, SARS-CoV-2, and MERS-CoV viruses have been studied using CHARMM36m giving important contributions in their understanding (Olson, 2018; Yu et al., 2021; Akbayrak et al.,

[2]It is important to clarify that the CHARMM atomistic FF can be used to run CpH simulations as well (Arthur and Brooks III, 2016).

2021). More information on the CHARMM36m is given in detail in the literature (Huang and MacKerell Jr, 2013; Lee et al., 2016; Huang et al., 2017). A typical sketch of the $NS1_{ZIKV\text{-}BR}$ dimer system with explicit solvent and salt in atomistic representation is shown in Figure 4.2.

### 4.1.3 Coarse-grained representation at the four-to-one level

In systems such as proteins, membranes, or large biomolecular complexes, the so-called "functional motions" may require long sampling times (Grudinin et al., 2020). CG models may be suitable to examine these phenomena. Different CG FFs are available in the literature (Kmiecik et al., 2016; Singh and Li, 2019). An extensively CG FF employed in biomolecular systems is Martini (also used by us here) (Periole and Marrink, 2013). Martini FF was originally developed for lipid membranes, then it was extended to proteins. This FF has usually been implemented for protein-membrane interactions showing correlation/reproducibility at the level of both AA models and experimental data (Marrink and Tieleman, 2013).

The grouping in the Martini FF is given by calculating the average of four atoms to have a single interaction center (four-to-one mapping). For example, Martini represents four water molecules as one, making a "super water" model (Marrink et al., 2004). But, this mapping is valid for various molecules, except those ring-like structures (*e.g.* benzene molecules) which require finer detail (two-to-one mapping), or a single salt ion that is represented by a unique bead. The clusterization of atoms allows the classification of four types of sites: polar (P), non-polar (N), apolar (C), and charged (Q). This results in 18 "building blocks" of CG particles (Periole and Marrink, 2013; Marrink and Tieleman, 2013). The full potential energy function of Martini is quite similar to the function shown in Equation 4.1, except that atoms are replaced by the Martini CG particles. The parameters of the FF were adjusted in order to balance the enthalpy and entropy terms (Monticelli et al., 2008). Alternative versions of the Martini FF include elastic network dynamics (Martini ElNeDyn) that helps to optimize the overall structure of the CG protein model through a spring constant (Periole et al., 2009). We have used this FF (see Chapter 9) because *i)* its parameterization is appropriate to study heterogeneous systems involving water molecules, salt ions, proteins, and lipid bilayers, *ii)* this FF allows the inclusion of a composition of lipid moieties similar to an ER membrane, and *iii)* large systems could be sampled ($>$ 55000 CG particles, for the NS1-membrane heterocomplex) for long periods of time (*e.g.* 20 $\mu$s) in order to detect functional movements. The recent version of this FF [Martini 3.0, (Souza et al., 2021)], has been refined to obtain accurate predictions of molecular packing and interactions in general. However, in our studies, we employed Martini 2.0 and some parameters were calibrated (specifically, the lower and upper elastic bond cutoffs) in order to fit some dynamical properties computed from AA

simulations (see Chapter 8). A version of Martini FF that may be interesting for us is one that incorporates titration (Bennett et al., 2013; Grünewald et al., 2020). This version has been more tested in membrane systems and some organic molecules. However, this FF for CpH simulations has not yet been well defined for proteins. Besides, some problems on the protonation/deprotonation of water molecules can be presented since the size of the beads (Grünewald et al., 2020). For these reasons, Martini FF for CpH simulations was not used in this study. Finally, more details on the parameterization of the interaction potential, topology, and calibration of the Martini FF are given in the interesting review article by Marrink and Tieleman (2013).

### 4.1.4 Coarse-grained representation at the amino acid level

CG models can also be further simplified removing bonded interactions, which leads to a rigid macromolecule description. This strategy is appealing to estimate free energy of interactions for protein-protein systems (Barroso da Silva et al., 2016, 2018, 2020). This kind of FF was used to describe the complexation process of two proteins A and B in an electrolyte solution as a function of solution pH. This approach allows an efficient incorporation of pH effects. So, constant-pH simulations can be done with these CG models in order to calculate several electrostatic properties (see Section 4.3.2). Note that the previously described models (see Sections 4.1.2 and 4.1.3) do not incorporate charge variations in the titratable residues as a function of solution pH. Conversely, this simplified model misses the description of conformational changes that can be very important especially at the short-range regime. The use of a simplified CG model at the amino acid level has shown to be adequate for simulations where the long-ranged electrostatic interactions are the main driving forces (Jönsson et al., 2007; Barroso da Silva and Jönsson, 2009; Teixeira et al., 2010; Delboni and Barroso da Silva, 2016; Barroso da Silva et al., 2016; Poveda-Cuevas, 2017; Barroso da Silva et al., 2017, 2018; Mendonça et al., 2019). This FF was also employed in viral proteins showing important insights on the molecular mechanisms and the electrostatic role in the protein-protein complexation (Giron et al., 2020, 2021).

For this CG model, the full configurational energy of the system $[U(\{r_k\})]$ considers only two terms as follows,

$$U(\{r_k\}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[u_{cg}^{el}(r_{ij}) + u_{cg}^{vdw}(r_{ij})\right], \tag{4.8}$$

where $\{r_k\}$ are amino acid positions, $N = N_A + N_B$ is the total numbers of residues ($N_A$ and $N_B$ are the number of residues for the proteins $A$ and $B$, respectively). The terms $u_{cg}^{el}$ and $u_{cg}^{vdw}$ are the Coulombic and LJ energies of interactions between amino acids $i$ and $j$ at a specific separation distance $r_{ij}$, respectively.

Table 4.1: Properties of the coarse-grained amino acids.

| Residue | $R_{ai}$[(a)] (Å) | $Mw$ (g/mol) | $pK_0$[(b)] | $z_{ai}$[(c)] |
|---|---|---|---|---|
| Ala | 3.1 | 66 | -.- | -.- |
| Arg | 4.0 | 144 | 12.0 | 0 and +1 |
| Asn | 3.6 | 108 | -.- | -.- |
| Asp | 3.6 | 110 | 4.0 | −1 and 0 |
| Cys[(d)] | 3.5 | 98 | 10.8 | −1 and 0 |
| Gln | 3.8 | 120 | -.- | -.- |
| Glu | 3.8 | 122 | 4.4 | −1 and 0 |
| Gly | 2.9 | 54 | -.- | -.- |
| His | 3.9 | 130 | 6.3 | 0 and +1 |
| Ile | 3.6 | 102 | -.- | -.- |
| Leu | 3.6 | 102 | -.- | -.- |
| Lys | 3.7 | 116 | 10.4 | 0 and +1 |
| Met | 3.8 | 122 | -.- | -.- |
| Phe | 3.9 | 138 | -.- | -.- |
| Pro | 3.4 | 90 | -.- | -.- |
| Ser | 3.3 | 82 | -.- | -.- |
| Thr | 3.5 | 94 | -.- | -.- |
| Trp | 4.3 | 176 | -.- | -.- |
| Tyr | 4.1 | 154 | 9.6 | −1and 0 |
| Val | 3.4 | 90 | -.- | -.- |
| Ctr | 2.0 | 45 | 2.6 | −1 and 0 |
| Ntr | 2.0 | 16 | 7.6 | 0 and +1 |

[(a)] $R_{ai}$ is the radius of the amino acid as given by Persson et al. (2010).
[(b)] p$K_0$ is the experimental intrinsic acid dissociation constant taken from Nozaki and Tanford (1967).
[(c)] $z_{ai}$ is the valence of the amino acid in a particular protonation.
[(d)] Cysteine which is not involved in disulfide bridge formation.

The long-ranged contribution $\left[u_{cg}^{el}\left(r_{ij}\right)\right]$ between a pair of amino acids is given by the screened Coulombic formula:

$$u_{cg}^{el}\left(r_{ij}\right) = \frac{l_B z_i z_j}{r_{ij}} \exp\left(-\kappa r_{i,j}\right), \tag{4.9}$$

where $z_i$ and $z_j$ are the valencies for, respectively, the $i$-th and $j$-th ionizable group obtained through the titration scheme named FPTS (see Subsection 4.2.2) (Teixeira et al., 2010; Barroso da Silva and Dias, 2017; Barroso da Silva and MacKernan, 2017), $l_B$ is the Bjerrum length (= 7.2 Å for water at room temperature), and $\kappa$ is the modified inverse Debye screening length that incorporates the counter ion concentration (Beresford-Smith and Chan, 1983; Schmitz, 1994). For this case, we are dealing with $z$ and not $q$ (compare Equations 4.9 and 4.7, respectively). Since [$z = q/e$, where $e$ is the elementary charge ($1.602 \times 10^{-19}$ C)], this is a dimensionless quantity (the valence). The charge state of each ionizable group is determined as a function of pH (Table 4.1). Thereby, we have assumed the dissociation constant (p$K_0$)

for each group. p$K_0$s are the intrinsic p$K_a$s values for "isolated" titratable groups and they were obtained from the experimental work by Nozaki and Tanford (1967).

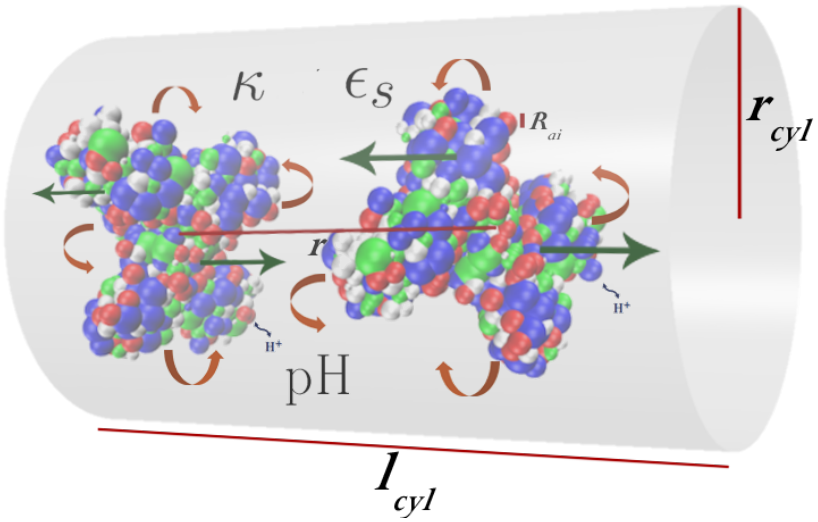



Figure 4.3: Schematic representation of the cell model used in this study. Each protein builds up by a collection of charged LJ spheres of radii ($R_{ai}$) and valence ($z_a$) is placed in an electroneutral open cylindrical cell of radius $r_{cyl}$ and height $l_{cyl}$. Inside the cylinder, the proteins can titrate in a specific pH solution. The green straight arrows indicate that the proteins can translate from back and forward (or vice versa) along one axis, and the orange curved arrows show that they can rotate. The solvent and salt ions are represented by its static dielectric constant $\epsilon_s$ and the Debye's term $\kappa$. Positive and negatively charged sites are colored in blue and red, respectively. Hydrophobic and neutral amino acids are colored in green and white, respectively.

The Bjerrum length parameter is given by,

$$l_B = \frac{e^2}{4\pi\epsilon_0\epsilon_s k_B T}, \tag{4.10}$$

where $\epsilon_s$ is the solvent static dielectric constant (= 78.7 at 298 K), $k_B$ is the Boltzmann's constant ($1.3807 \times 10^{-23}$ J mol$^{-1}$ K$^{-1}$), and $T$ is the temperature (in Kelvin).

The LJ term between two beads of this CG model $\left[u_{cg}^{vdw}\left(r_{ij}\right)\right]$ is quite similar to that one shown in Equation 4.6, except that the sites $i$ and $j$ are here the amino acids instead of atoms. Thus, $\sigma_{ij}$ is the arithmetical average of the radius $R_a$ of each amino acid (also known as Lorentz-Berthelot rules). The $R_a$ is calculated based on the study of Persson et al. (2010), where they assumed a density equal to 0.9 g ml$^{-1}$ and a molecular weights ($Mw$) for each amino acid (Table 4.1). The simulation box for this case is an electroneutral open cylindrical cell of radius $r_{cyl}$ and height $l_{cyl}$. Inside this cylinder the proteins A and B are placed, and the values of the solution pH, $\epsilon_s$, and $\kappa$ are assigned. An illustration of the simulation box for this CG model is shown in Figure 4.3.

### 4.1.5 Periodic boundary conditions

*In vitro* or macroscopic systems can be made up of a number of particles greater than Avogadro's number. This number of particles is prohibitive at the level of computational simulations due to a high cpu cost. Thereby, a fraction of particles is usually employed to represent the system. The studied systems are constituted by a number of particles between $10^1$ to $10^5$. As shown in Figures 4.2 or 4.3, the particles of the system are confined in a simulation box with a specific geometry. However, when it is introduced this approximation some artifacts may appear especially for the particles at the edge of the simulation box. This problem can be solved using *periodic boundary conditions* (PBC). PBC is a technique where the simulation box containing the particles is replicated three-dimensionally with infinite periodicity (Andrew, 2001; Cramer, 2013; Lemkul, 2018).

During the computational simulations the calculation of the intermolecular potentials must be considered for each replicated "image" of the box. For interactions that decay in proportion to $r^{-6}$ (as shown in the second term in the parentheses of the LJ potential energy, Equation 4.6) the *minimum-image convention* is used. Under this approach, the simulation box should be at least twice as big as the largest cut-off radius of the short-range interactions considered during the simulation (Andrew, 2001). Concerning long-range interactions (*i.e.* those that decay in proportion to $r^{-1}$, as seen in Equation 4.7 for the electrostatic potential energy), the treatment of the electrostatic interactions is done using more sophisticated methods such as Particle Mesh Ewald (PME), Smooth Particle Mesh Ewald, Reaction Field, etc (Essmann et al., 1995; Andrew, 2001). For example, through PME what is sought is to count all the electrostatic interactions of a system with PBC utilizing summations in the "real space" (where short-range electrostatic interactions are considered) and the "reciprocal space" (where long-range electrostatic interactions are taking into account) (Andrew, 2001; Lemkul, 2018).

## 4.2 Computational simulations

Once the molecular model is defined, it can be used to calculate physical properties that will be helpful to understand the studied system. The total energy of the study system must be calculated, to estimate any observable physical property. Generally, we start from a "microscopic" view of the system from which it is possible to compute thermodynamic or "macroscopic" information [*e.g.* temperature ($T$), volume ($V$), pressure ($P$)]. The link between the microscopic and macroscopic features of the system is possible thanks to the tools of the Statistical Mechanics. Thus, a "macroscopic" system with total energy ($E$) made up of $N$ particles will present some possible energy states ($E_1$, $E_2$, $E_3$,..., $E_i$) derived from the individual molecular energies ($\varepsilon_N$). The collection of all these energies for the system at thermal equilibrium is called as "ensemble". Different thermodynamics

ensembles may be explored (*e.g.* microcanonical, canonical, grand canonical, etc). Those being the most frequently considered in computer simulations are the canonical [with $N$, $V$, and $T$ being constant ($NVT$)] or the isothermal-isobaric ensemble [with $N$, $P$, and $T$ constants ($NPT$)]. From the ensemble, any physical property ($\mathcal{A}$) is computed, through its expected value [$\langle\mathcal{A}\rangle$],

$$\langle\mathcal{A}\rangle = \sum_{i=1}^{states} p_i\mathcal{A}_i, \tag{4.11}$$

where $p_i$ is the probability to find the system in the state $i$, and $\mathcal{A}_i$ is the ensemble in the state $i$. In Equation 4.11, the angle brackets denote average quantities. To estimate $p_i$ the Boltzmann Weighted average is used:

$$p_i = \frac{1}{Q}\exp\left(-E_i\beta\right), \tag{4.12}$$

where the denominator $Q$ $\left[=\sum_{i=1}^{states}\exp\left(-E_i\beta\right)\right]$ is the partition function, and $\beta$ $(=1/k_BT)$ is the Boltzmann factor. The partition function is also considered as the "cornerstone" of Statistical Mechanics because knowing it, we can derive the rest of physical quantities such as heat capacity, chemical potential, free energy, etc (McQuarrie and Simon, 1997). The Boltzmann factor is a variable that allows determining the relationship between the states of the system and the energy they present at a given temperature. Thus, low energy states will be more likely in a low temperature, while high energy states will contribute less in that temperature. On the other hand, in high temperatures, the low and high energy states will not be relatively biased (McQuarrie, 1976). Once it is defined the kind of ensemble (*i.e.* if $NVT$ or $NPT$ is chosen, for example), Equation 4.11 is integrated, as follows:

$$\langle\mathcal{A}\rangle = \int_{\mathbb{R}^N} d\vec{r}\left[\frac{1}{Q}\exp\left(-E\left\{\vec{r}^N\right\}\beta\right)\right]\mathcal{A}\left\{\vec{r}^N\right\}, \tag{4.13}$$

where $\left\{\vec{r}^N\right\}$ is the configuration of the $N$ particles of the system. Solving integrals analytically like those shown in Equation 4.13 can be very difficult. In general, numerical integration methods that are time-dependent (*e.g.* MD) or stochastically (*e.g.* MC) are more appropriate.

Classical MD and MC simulations were used in this study. Although these methods differ in how sampling occurs and what properties can be sampled, they are equivalent due to the ergodic hypothesis (McQuarrie, 1976). In practice, the choice among MC or MD depends on the convenience from available codes to compute the desired property and how efficient is the sampling for a particular molecular model (Cramer, 2013; Allen and Tildesley, 2017). This was largely explored by us in the next chapters. In this section, some aspects about how MD and MC simulations work will be briefly emphasized.

### 4.2.1 Molecular Dynamics

Before performing a MD calculation, it is necessary to relax the solvated system to a low energy state. During this process, bad contact problems among beads are solved until the system reaches a favorable position in the energetic landscape (local minimum) (Lemkul, 2018). For this, a proper algorithm can be used (*e.g.* steepest descent, conjugate gradient, among others) (Rapaport, 2004; Lemkul, 2018). For all the simulations carried out in this study, the steepest descent method was considered, which works by following the following steps:

1. The initial positions ($x_0$) are defined.
2. Calculate the potential energy of the system $[V(x_i)]$ in a particular position ($x_i$) and its gradient $[\nabla V(x_i)]$ (this shows in which direction the energy increases and decreases).
3. The atomic positions ($x_{i+1}$) are updated.
4. Convergence and the total force of the system are verified.
5. If the previous step satisfies convergence or the system reaches a preset total force value, the program stops, otherwise, it returns to step 2.

When our system reaches a favorable energy position after the minimization, a MD simulation can be started. Classical MD consists in solving Newton's equations of motion. The force $\left(\vec{F}\right)$ experienced by a particle $i$ in a given system can be calculated by means of the relation between the derivative of the potential energy $\left[dU\left\{\mathbf{r}^N(t)\right\}\right]$ and the acceleration,

$$\vec{F}_i(t) = \frac{dU\left\{\mathbf{r}^N(t)\right\}}{dr_i} = m_i \frac{d^2 r_i(t)}{dt^2}, \tag{4.14}$$

where $m_i$ is the mass of the particle $i$, $r$ is its position, and $t$ is the instant when the event happens. Notice that the first term incorporates the total potential energy like that one described in Equation 4.1. This differential equation is solved numerically using very small time steps ($\Delta t$) through some algorithm such as Verlet, leap-frog, and others (Bou-Rabee, 2014). Because the integration time needs to be shorter than the fastest frequency of motion (*i.e.* bond stretching), it is often set to $1-2$ and $10-100$ fs, depending on the degree of details included in the model (Andrew, 2001; Kmiecik et al., 2016). In the first run of the MD simulation after minimization, the initial positions of the system are defined (*e.g.* as shown in Figure 4.2). The velocities of each particle are estimated

assuming a determined temperature through a Maxwell-Boltzmann distribution:

$$p\left(v_i\right) = \sqrt{\frac{m_i}{2\pi k_B T} \exp\left(-\frac{m_i v_i^2}{2k_B T}\right)}. \tag{4.15}$$

where $v_i$ is the velocity of the particle $i$. The distribution can be used to determine the absolute temperature given by the total kinetic energy of the $N$particles:

$$E_{kin} = \frac{1}{2} N_{df} k_B T, \tag{4.16}$$

where $N_{df}$is the number of degrees of freedom. From this, we can calculate the associated force for each atom according to Equation 4.14. Summarizing, the MD algorithm works as follows[3]:

1. The initial positions are defined.
2. Velocities and momentum are attributed to each particle by means of Equation 4.15. This step is called "the heating step".
3. The forces acting on each particle are calculated according to Equation 4.14.
4. New positions and velocities are predicted.
5. Physical variables are computed (*e.g.* positions, velocities, energies, forces, temperature, pressure).
6. The process returns to step 3 until it reaches the total number of steps previously set.

In order to get an appropriate MD trajectory, a first round of simulation must be carried out. For this stage, we seek that the temperature or kinetic energy of the system be equilibrated. This is also known as the "equilibration phase" and can usually be performed under an $NVT$ ensemble (although other statistical ensembles could be employed). From the equilibration, it is also sought that the positions of the solvent, ions, and macromolecule have appropriate orientations for each other. After this phase, we can carry out a "production phase" where the pressure is controlled with an $NPT$ ensemble in order to reach a suitable density. Through the production phase, the trajectory is built of interest to subsequently compute all different dynamic properties described in Section 4.3.1.

These basic principles are implemented by the majority of simulation packages such as GROMACS (Abraham et al., 2015), NAMD (Phillips et al., 2005), or GROMOS[4]

[3]For further technical details, we recommend the studies by Rapaport (2004).

[4]Be careful not to confuse "GROMOS simulation package" with "GROMOS force field".

(Scott et al., 1999). In the present study, GROMACS v5.1.5 or later versions was used. The models described in Subsections 4.1.2 and 4.1.3 were solved by MD simulations (see also Chapters 8 and 9). MD was chosen because we are interested in calculating time-dependent properties such as correlated motions of protein parts or essential dynamics (see Subsection 4.3.1.2).

### 4.2.2 Metropolis Monte Carlo

Through MC simulations the system propagates via Markov chains (Satoh, 2010). This means that every time when a new state is randomly generated (*e.g.* by the rotation or translation of a molecule), it does not depend on the immediately preceding condition. However, as new states are obtained, they should be evaluated in order to verify if they are physically relevant. For a system with $N$ particles with initial positions $\{r_k\}$ and total energy $U(\{r_k\})$ (see Equation 4.8), the Metropolis algorithm can be implemented as follows:

1. From the current configuration $\{r_k\}$, a new configuration is chosen $\{r_l\}$ and its potential energy $U(\{r_l\})$ is calculated.

2. Each trial movement is accepted with the probability,

$$\min\left[1, \exp\left(-\beta \Delta U\right)\right].$$

3. During the MC simulation, the process returns to step 2 until convergence is reached.

Suitable values of acceptance rate are in the range of 30%−50% depending on the code efficiency to accept and/or reject configurations. In an ideal code with equally cpu costs for acceptance and rejections, if the acceptance rate is less than 50% it means that we are rejecting several configurations that may have favorable energies, while if the acceptance is greater than 50% it means that we are accepting more configurations with unfavorable energies. The acceptance rate is controlled by the displacement parameters for each simulated condition.

The MC simulations performed here assumed that the two proteins A and B can rotate, translate, and titrate as a function of pH. For the titration process, the Fast Proton Titration Scheme (FPTS) was used (Teixeira et al., 2010). This method has shown a high degree of precision and accuracy in previous benchmark studies (Teixeira et al., 2010; Barroso da Silva and Dias, 2017; Barroso da Silva and MacKernan, 2017; Barroso da Silva et al., 2018). All the advantages and disadvantages for the FPTS approach are described by Barroso da Silva and Dias (2017) and Barroso da Silva and MacKernan (2017). In short, FPTS assume that the exchange of protons between a protein and the

medium can be calculated through the computation of the difference of the potential energy $w_{\mathrm{TK}}$[5] (Barroso da Silva and MacKernan, 2017).

The FPTS method in a MC implementation works as follows (Teixeira et al., 2010):

1. From the protein coordinates, a titratable site is chosen randomly.

2. In the case of a *deprotonation* process, the proton charge is attempted to move from the site (*i.e.* an ionizable group) to the bulk solution, whereas in a *protonation* process exactly the opposite happens. In each MC trial, $\kappa$ should be adjusted in order to consider variations in the quantity of counterions in the solution. The total protein charge ($Z_P$) is also updated.

3. Each trial charge movement is accepted with the probability,

$$\min\left[1, \exp\left(-\beta \Delta w_{\mathrm{TK}}\right)\right],$$

4. During the MC simulation, the process returns to step 2 until convergence of the multiple titratable sites is reached.

When FPTS is employed, a semigrand canonical ensemble is implemented (Srivastava et al., 2017). In this statistical ensemble the chemical potential ($\mu$), volume ($V$), temperature ($T$), and the number of particles ($N$) are conserved. This ensemble should not be confused with the grand canonical ($\mu VT$), which does admit variation in $N$. The semigrand canonical ensemble was a convenient choice because the phase space for protein complexation processes can be easily explored, in addition to the intrinsic variation of protons. Furthermore, variables that can influence this type of phenomenon such as volume and temperature are exact when using the MC approach.

## 4.3 What is quantified in the computer simulations?

### 4.3.1 Physical quantities measured in the MD runs

MD allows exploring the conformational changes of a structure over time (Karplus and McCammon, 2002). Exploring structural conformational changes may help in understanding how proteins perform biological functions (Hollingsworth and Dror, 2018).The MD trajectory consists in a collection of frames for the studied biomolecule at a particular physicochemical condition. From this trajectory, several analyses can be carried out. Here, we describe the different measurements done in Chapters 8 and 9.

[5]This term is calculated assuming a Tanford-Kirkwood (TK) model.

#### 4.3.1.1 Canonical measures

The most frequent measures typically used to analyze MD trajectories are:

- **Root mean square deviation (RMSD):** The RMSD is the Euclidean distance in $3N$ configuration space as a function of the time step:

$$\mathrm{RMSD}\,(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{r}_i\,(t) - \mathbf{r}_i^{\mathrm{ref}}\right]^2}. \tag{4.17}$$

  Equation 4.17 tells that given the coordinates frame $\mathbf{r}_i$ at time $t$, their variation using as a base a reference position $\mathbf{r}_i^{\mathrm{ref}}$ can be estimated. This calculation can be done for any subset of atoms. It is usually performed for the $C_\alpha$ atoms or the backbone structure. The RMSD calculation gives us an initial insight into how the structure behaves over time. It is also frequently considered to check the stability of the simulation, namely when the system has reached an equilibrium.

- **Root mean square fluctuation (RMSF):** One can measure the fluctuations of a atom/bead/residue by calculating its square root of the variance around the mean position:

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{N_{steps}-1}\sum_{j=1}^{N_{steps}}\left\langle\left(\mathbf{r}_i - \langle\mathbf{r}_i\rangle\right)^2\right\rangle}, \tag{4.18}$$

  where $\mathbf{r}_i$ is the position of the atom/bead/residue $i$, $\langle\mathbf{r}_i\rangle$ is the position average, and $N_{steps}$ is the number of steps. When the RMSF is plotted as a function of sequence position, it is possible to distinguish rigid from more flexible regions in a biomolecular structure. This measurement is interesting because it can be compared with the $B-factor$ derived from crystallographic data[6]:

$$B - factor = \frac{8\pi^2}{3}\left(\mathrm{RMSF}_i\right)^2. \tag{4.19}$$

- **Radius of gyration** $(R_g)$**:** The compactness of a protein structure can be measured by $R_g$, described bellow:

$$R_g^2 = \frac{1}{M}\sum_{i=1}^{N} m_i\left(\mathbf{r}_i - \mathbf{R}\right)^2, \tag{4.20}$$

  where $M$ $\left(=\sum_{i=1}^{N} m_i\right)$ is the total mass, $\mathbf{r}_i$ is the coordinates frame of the system, and $\mathbf{R}$ $\left(=N^{-1}\sum_{i=1}^{N}\mathbf{r}_i\right)$ is the center of mass of the protein consisting of $N$ atoms/beads. The radius of gyration can provide information on the large

[6] See Appendix A.1 as an example.

conformational transitions (*e.g.* denaturation). However, long-time trajectories are necessary to have good sampling and detect these structural changes.

- **Solvent accessible surface area (SASA):** The surface area of a biomolecule that is accessible to the solvent can be calculated as suggested by Eisenhaber et al. (1995):

$$\mathrm{SASA} = 4\pi \sum_{i=1}^{N} r_i^2 \frac{m_{acc}\,(i)}{m},$$

where $m_{acc}\,(i)$ is the number of dots on an atom $i$ not occluded by neighboring beads and the summation is over atoms in the biomolecule. With SASA we have a measure of the amount of area that is interacting with the solvent and how this could be influencing certain dynamic aspects of the system.

#### 4.3.1.2 Statistical measures

The dimensionality of MD trajectories can be so big that emergent properties with relevant functionality are not noticeable by canonical analyses. For example, $\mathrm{NS1_{ZIKV\text{-}UG}}$ and $\mathrm{NS1_{ZIKV\text{-}BR}}$ being such similar proteins can exhibit discrepant structural patterns that can not be detected by conventional measurements (see Chapter 8). One way to solve that is by performing more sophisticated analyses where statistical tools allow us to filter the trajectory data to catch more informative patterns. For the sake of convenience, we highlight three main statistical calculations:

- **Principal component analysis (PCA):** PCA is a robust statistical tool where linear transformations are applied for dimensionality reduction of data sets of variables (Queen et al., 2002). From the point of view of MD simulations, PCA is applied to filter the atomic displacements capturing the most important movements or the *essential dynamics* (ED) for a given trajectory (David and Jacobs, 2014). When PCA is used, we start from the coordinates of the system. Then, a covariance matrix ($\mathbf{C}$) with dimension $3N$ via,

$$c_{ij} = \langle (x_i - \langle x_i \rangle)\,(x_j - \langle x_j \rangle) \rangle\,, \tag{4.21}$$

where $x_i$ and $x_j$ are the atomic coordinates and the brackets denote the ensemble average. The diagonalization of $\mathbf{C}$ is equivalent to solving the values associated with the variance or eigenvalues ($\lambda$). Then,

$$\mathbf{A}^{\mathrm{T}}\mathbf{C}\mathbf{A} = \lambda, \tag{4.22}$$

where $\mathbf{A}$ represents the eigenvectors. The trajectory can be projected onto an eigenvector that defines a given a "*principal component*" (PC) (Maisuradze et al.,

2009; David and Jacobs, 2014). Each eigenvalue captures a percentage of variance that is associated with its corresponding eigenvector defining a top set of PCs. These PCs are the "essential" modes related to particular motions with biological relevance. When the simulation is long enough a few PCs can catch a large part of the variance. For example, the first two PCs can capture 50% of the variance, but sometimes up to 20 PCs may be required (David and Jacobs, 2014). PCA is not the only method to estimate ED. Nevertheless, it has been shown as an efficient algorithm for the treatment of high-dimensionality data (Tribello and Gasparotto, 2019).

- **Dynamic cross-correlation map (DCCM):** Other features that may be relevant in a protein are the coupling of residues that move in concert (Harte et al., 1990; Ichiye and Karplus, 1991). To identify these correlated movements, the DCCM can be applied. Thus, the coefficient $C_{ij}$ for the pairs of beads, $i$ and $j$ are computed, like this:

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle \left\langle \Delta r_j^2 \right\rangle}}, \tag{4.23}$$

  where $\Delta r_i$ and $\Delta r_j$ are respectively the displacements from the mean position of the $i$-th and $j$-th beads determined from all configurations in the trajectory segment under analysis. When $C_{ij}$ is positive or negative, it means that we have a correlated or an anticorrelated motion, respectively (Ichiye and Karplus, 1991). DCCM is a measure for large scale movements and not for local fluctuations. Thus, DCCM is an important tool for analysis of biological functionality (Scarabelli and Grant, 2013).

- **Protein structure networks (PSN):** Each residue can be considered as a node for the construction of a weighted graph. These nodes are connected in a network if they are in contact within a cutoff in at least 75% of the trajectory frames (Sethi et al., 2009; Scarabelli and Grant, 2013). This analysis is so-called PSN, which starts from a clustering method, where each edge between the nodes $i$ and $j$ are weighed ($W_{ij}$) by their correlation value ($C_{ij}$):

$$W_{ij} = -\log\left(|C_{ij}|\right). \tag{4.24}$$

  For the clustering process, the Girvan-Newman method can be used and the nodes are grouped into communities with both high correlation and high physical proximity (*i.e.* modularity) (Girvan and Newman, 2002; Newman, 2006). As argued by Newman (2006), the quality of a community structure can be measured by its maximum modularity value ($Q^{GN}$), which is a variable of the probability difference for intra- or inter-community edges. In an ideal system, the maximum value

of $Q^{GN}$ can be 1, indicating a suitable community structure. Thus, when the network is divided into small communities, the *modularity score* is measured for each community division. A typical network in the real world can have a modularity score between 0.4 and 0.7 (Newman and Girvan, 2004). From that, other measures within the network such as centrality, shortest path, and suboptimal path can be done. These network features help us to interpret how the communities are being connected by central or alternative pathways. Besides, the critical number of edges or interactions estimated by PSN gives us tools to deduce allosteric signals (Sethi et al., 2009; Scarabelli and Grant, 2013).

### 4.3.2 Physical quantities measured in the MC runs

Electrostatics governs various aspects of protein functionality (Perutz, 1978). Reversible or irreversible protein associations are dependent on physical-chemical conditions that affect their electric charges (Barroso da Silva, 2000). For this reason, our study has a special emphasis on understanding the effect produced by the solution pH or the ionic strength in NS1 interactions. Using FPTS our major interest is to understand the main electrostatic properties that influence a protein-protein complexation process. The physicochemical quantities described below were frequently estimated in Chapters 5, 6, and 7.

- **Protein charge number** ($Z_P$)**:** This physicochemical quantity is the sum of all the charges of each titratable site ($z_{ai}$):

$$Z_P \equiv \sum_{i=1}^{N_P} z_{ai}, \tag{4.25}$$

  where $N_P$ is the number of titratable sites. In a pure Coulombic description, proteins with opposite or same $Z_P$'s can suffer attraction or repulsion, respectively (Israelachvili, 2011).

- **Charge regulation capacity** ($C_P$): The $C_P$ is a measure of the charge fluctuation for the ionizable groups on a protein due to the presence of a nearby electric field (this phenomenon is also known as "charge regulation mechanism") (Jönsson et al., 2007). This mechanism leads to an attraction and generally occurs when $C_P$ is high and the system is at low salt regimes. The charge regulation parameter can be given by:

$$C_P \equiv \left\langle Z_P^2 \right\rangle - \left\langle Z_P \right\rangle^2 . \tag{4.26}$$

  Equation 4.26 is proportional to:

$$\frac{\partial \left\langle Z_P \right\rangle}{\partial \mathrm{pH}}. \tag{4.27}$$

Using Equation 4.27 as a basis, note that as the charge variation is higher [*i.e.* the slope is high in a titration curve ($Z_P$ vs pH) because of a protonation or deprotonation processes are happening], the charge regulation mechanism is produced at the presence of the charge partner macromolecule.

- **Dipole moment number** ($\mu_P$)**:** The heterogeneous distribution of ionizable groups of a protein can produce an electric dipole moment. $\mu_P$ is defined as:

$$\mu_P = \sum_{i=1}^{N_P} r_{ai} z_{ai}, \tag{4.28}$$

where $r_{ai}$ is the amino acid $i$ position. A relatively high dipole can produce an attraction in the system.

- **Dissociation constant of the amino acid** (**p**$K_a$)**:** From $z_{ai}$ (*i.e.* the charge number of an ionizable residue) calculated at different pH conditions, it is possible to calculate the p$K_a$ of this residue assuming that the protonation is happening 50% of the time (Barroso da Silva and Dias, 2017). Figure 4.4 illustrates this idea. In this example, the p$K_a$ for a basic amino acid it will be $\backsim$11.6 (when $z_{ai} = 0.5$), while for an acid residue will be $\backsim$3.6 (when $z_{ai} = -0.5$). Note that we are referring to p$K_a$ and not p$K_0$ (the experimental input condition for each amino acid listed in Table 4.1). The p$K_a$ concept is the basis to estimate the electrostatic epitopes through the "PROCEEDp$K_a$" method (described in Chapter 6).

- **Free energy of interaction** [$\beta w(r)$]**:** During a MC simulation for protein-protein interactions the radial distribution function [$g(r)$] is calculated:

$$g(r) = \frac{hist(r)}{\rho n_{obs} \Delta V}, \tag{4.29}$$

where $hist(r)$ is the frequency of each protein-protein position at the separation distance $r$, $\rho$ is the particle density, $n_{obs}$ is the number of observations (or MC steps), and $\Delta V$ is the volume of a specific slice in the simulation box. From $g(r)$, we can calculate the potential of mean force [$\beta w(r)$] through the mechanical-statistical relation (McQuarrie, 1976):

$$\beta w(r) = -\ln g(r), \tag{4.30}$$

Equation 4.30 allows us to estimate the protein-protein free energy interaction. Also, from $w(r)$, it is possible to calculate the second virial coefficient $B_2$ (see Chapter 7). Finally, a positive or negative value of free energy means within a simple thermodynamical framework that the system is repulsive or attractive, respectively.

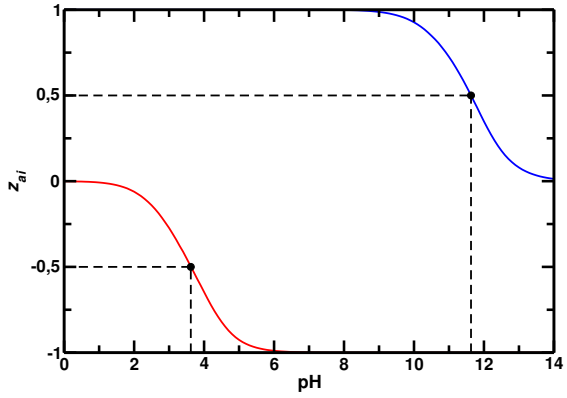



Figure 4.4: Example of a theoretical titration plot for an acid (red line) and basic residue (blue line). p$K_a$s are $\backsim$3.6 and $\backsim$11.6, respectivley.

## Chapter 5

# Preliminary insights about the NS1 electrostatic behavior

In this chapter, we are interested in evaluating several electrostatic properties and their impact on the three putative biological interfaces (*i.e.* NS1-NS1, NS1-Ab, and NS1-membrane). The study considered several NS1 from ZIKV variants with African and Asian origins. In parallel, through a colloidal-like framework, the free energy of the complexation processes for the same three main partners was examined.

The main question of this part of the study is:

*Are the electrostatic properties of NS1s from Africa and Asia different at the level of the three main biological interfaces?*

A summary of the molecular model used, the methods to solve them, and the measures calculated appear in Table 5.1.

Table 5.1: Molecular modeling and measurements carried out in Chapter 5.

| Level of detail | Solution | Measurements |
|---|---|---|
| DLVO | Analytical | Free energies |
| CG at the amino acid level | CpH MC simulations | $Z_P$, $C_P$, $\mu_P$, p$K_a$ |

Link paper: https://pubs.acs.org/doi/10.1021/acsomega.8b02081.

This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

ACS OMEGA

Article

Cite This: ACS Omega 2018, 3, 16212−16229

http://pubs.acs.org/journal/acsodf

# Insights into the ZIKV NS1 Virology from Different Strains through a Fine Analysis of Physicochemical Properties

Sergio A. Poveda-Cuevas,[†,‡,▽] Catherine Etchebest,[§,∥,⊥,▽] and Fernando L. Barroso da Silva*[,†,‡,▽,#]

[†]Programa Interunidades em Bioinformática, Universidade de São Paulo, São Paulo 05508-090, Brazil

[‡]Departamento de Física e Química, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo 14040-903, Brazil

[§]Institut National de la Transfusion Sanguine, Paris 75015, France

[∥]Biologie Intégrée du Globule Rouge, Equipe 2, Dynamique des Structures et des Interactions Moléculaires, Institut National de la Santé et de la Recherche Médicale, UMR_S 1134, Paris 75015, France

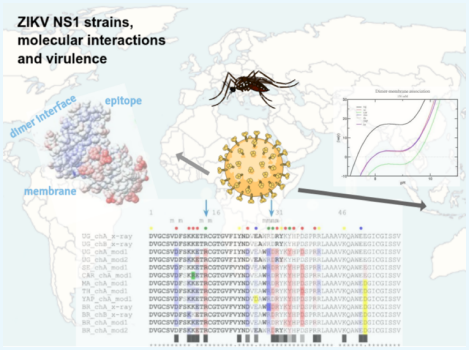
[⊥]Université Sorbonne Paris Cité and Université Paris Diderot, 75013 Paris, France

[▽]University of São Paulo and Université Sorbonne Paris Cité Joint International Laboratory in Structural Bioinformatics

[#]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, North Carolina 27606, United States

Supporting Information

**ABSTRACT:** The *flavivirus* genus has several organisms responsible for generating various diseases in humans. Recently, especially in tropical regions, Zika virus (ZIKV) has raised great health concerns due to the high number of cases affecting the area during the last years that has been accompanied by a rise in the cases of the Guillain−Barré syndrome and fetal and neonatal microcephaly. Diagnosis is still difficult since the clinical symptoms between ZIKV and other flaviviruses (e.g., dengue and yellow fever) are highly similar. The understanding of their common physicochemical properties that are pH-dependent and biomolecular interaction features and their differences sheds light on the relation strain-virulence and might suggest alternative strategies toward differential serological diagnostics and therapeutic intervention. Due to their immunogenicity, the primary focus of this study was on the ZIKV nonstructural proteins 1 (NS1). By means of computational studies and semiquantitative theoretical analyses, we calculated the main physicochemical properties of this protein from different strains that are directly responsible for the biomolecular interactions and, therefore, can be related to the differential infectivity of the strains. We also mapped the electrostatic differences at both the sequence and structural levels for the strains from Uganda to Brazil, which could suggest possible molecular mechanisms for the increase of the virulence of ZIKV in Brazil. Exploring the interfaces used by NS1 to self-associate in some different oligomeric states and interact with membranes and the antibody, we could map the strategy used by the ZIKV during its evolutionary process. This indicates possible molecular mechanisms that can be correlated with the different immunological responses. By comparing with the known antibody structure available for the West Nile virus, we demonstrated that this antibody would have difficulties to neutralize the NS1 from the Brazilian strain. The present study also opens up perspectives to computationally design high-specificity antibodies.




## INTRODUCTION

The *flavivirus* genus has several organisms responsible for generating diseases in humans, which often cause global health concerns [e.g., dengue virus (DENV), West Nile virus (WNV), and yellow fever virus (YFV)].[1−3] During the last years, the worldwide attention has moved to Zika virus (ZIKV) due to the large increase of infected people and affected areas, which has been accompanied by a rise in the cases of the Guillain−Barré syndrome and microcephaly.[4,5] A practical critical issue is that, in many cases, the ZIKV symptoms can be confused with other viruses, resulting in misdiagnoses and, posteriorly, inadequate treatments.[6,7] Consequently, the outbreak of ZIKV has produced a race to better diagnose it, as well as to treat and stop its epidemic growth. Yet, to date, there is no totally effective medication, easy and cheap diagnosis, or even treatment.

The complex life cycle of ZIKV and its capacity to be transported by several mosquito vectors aggravate the health issues.[8] Its spread is followed by an evolution of the strains together with distinct aggressive effects. The first isolation of

Received: August 17, 2018
Accepted: November 14, 2018
Published: November 29, 2018

Downloaded via UNIV OF SAO PAULO on July 10, 2019 at 14:24:43 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

the ZIKV occurred in Uganda (UG) in 1947, where the reported cases did not indicate the severe neurological disorders that the ZIKV can produce. Later, cases were found in Africa, Southeastern Asia, Oceania, and Americas.[9,10] From this, it has been established that there are two lineages of ZIKV, one Asian and another African,[11,12] as well as that the latter might be divided into two.[13] However, how each mutation directly contributes to the change in virulence remains unknown. As indicated in a recent review by Simonin and collaborators,[14] the high number of microcephaly cases observed for the Brazilian strain (Asian lineage) might be a result of a mild infection in contradiction to the common sense. The African lineage causes a much more severe infection inducing an early termination of pregnancy (higher mortality) without giving a chance to the development of congenital malformations.[15] Unfortunately, no molecular mechanisms have been proved to be responsible for these different pathogenicities. Therefore, the understanding of their common physicochemical properties and biomolecular interaction features sheds light on the relation strain-infectivity and might suggest alternative strategies toward an efficient serological diagnostics and possible therapeutic interventions.

ZIKV is a single-stranded RNA virus that encodes one polyprotein, which is later processed to generate three structural proteins (C, prM, and Env) and seven nonstructural (NS) proteins (NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5). Several of these proteins are directly involved in various molecular processes that define the structure or replication of the virus.[1−3] Here, we will focus on the NS1 protein due to its high biological relevance. Among other features, this protein is able to self-associate into a symmetric head-to-head homodimer, which increases its hydrophobic character in the early viral stages. The formed NS1 dimer interacts with the membrane of the endoplasmic reticulum (ER).[2,16−19] Later, the NS1 can have three possible destinations:[2,20] (1) it can function as a cofactor in the replication complex, (2) it can go toward the membrane surface to interact with different membrane receptors, or (3) it can go toward the extracellular space to interact with proteins of the immune system. Such versatile behavior of NS1 is exploited by the virus to carry out its development since this protein has the ability to evade the immune response of the host,[21,22] acting on several pathways of the complement system (e.g., see ref 23). It is known that all this diversity of interactions leads to a cascade of signaling, preventing the good metabolic functioning of the host.[2] Due to its involvement in the viral RNA replication, NS1 may also be an attractive target for drug developments.

pH is a trigger physicochemical parameter in several viral processes.[3,24−26] It is directly related to the macromolecular electrostatic properties and the molecular interactions.[27] In a previous study,[28] Song and coauthors compared electrostatic maps on the molecular surfaces of NS1 proteins from ZIKV, DENV, and WNV. They showed diverse electrostatic features among them at host interaction interfaces and suggested the possibility to relate the pathogenesis with the electrostatic behavior. Another study has shown that the dissociation of NS1 dimers occurs in acidic pH regimes, and more stable dimers prevail under physiological pH conditions.[29] All of these studies imply that the interactions of NS1 with any other charged molecule (including its self-association) and membranes of the host might be different for each flavivirus and even for each strain, producing a diverse of clinical manifestations. This also indicates the possibility to put NS1 as a potential candidate for the design of vaccines and as a biomarker for the detection of the virus in seropositive patients.[30−32]

All of these facts have motivated the present work where we extended the preliminary study of Song and collaborators[28] to explore the electrostatic properties of several ZIKV strains at several solution pH regimes. We propose that analyzing these properties at three key biological functional interfaces [(i) the self-association interface (homodimer interface), (ii) the predicted NS1 molecular region that interacts with the membranes, and (iii) a putative epitope region] could provide a way in the understanding of the differential infectious behavior of the ZIKV. This was achieved by means of numerical simulations and structural bioinformatics tools. Moreover, the quantification of biomolecular interactions is essential for understanding the molecular biology. In this direction, we performed semiquantitative estimations of the free energies of interactions of the ZIKV NS1 in some different oligomeric states at these biointerfaces. These thermodynamic measurements provide practical insights into some possible molecular mechanisms that can be related with the virology. Although the NS1 undergoes a complex maturation process and can be found in several oligomeric states from monomers that exist briefly (in the endoplasmic reticulum lumen) to hexamers (in late infection),[19,33−37] only the monomeric and dimeric forms are discussed here for several reasons. First, the monomer is the beginning of all processes and the only state where all biological interfaces can be compared on a common basis. Second, the idea that the high-molecular-weight form of NS1 is a homodimer as introduced by Winkler et al.[37] seems to be a consensus of a general feature of flavivirus infection. The subject has also been recently reviewed in different papers (e.g., ref 36), which puts the dimer in a central position. The homodimerization is a fundamental step to provide to NS1 the necessary characteristics to associate with membranes, which is central for the viral replication.[34] Third, the homodimer is the predominant form of intracellular NS1.[35] Fourth, the glycosylation of NS1 is not required for both the dimerization and the NS1−membrane association, which allows the use of a simpler theoretical approach.[38] Finally, the hexamer state is more obscure in the experiments where it has been discussed that they might even require the presence of stabilizing elements to maintain the three NS1 dimers complexed to form the hexameric state of NS1.[34] Tetrameric intermediates have also been reported.[33] The present analyses are carried out exploring global titration features as well as more local effects at the biofunctional interfaces at both the sequence and structural levels. It is also interesting to note that despite the small changes in the sequence due to the high sequence identity among the studied strains, the electrostatic properties are strongly affected by the pH, which also affects their biomolecular interactions with partners and, consequently, the molecular viral biology.

This paper is organized as follows. In the next section, we present the theoretical methods used in this work starting with the comparative modeling of the NS1 structures for each ZIKV strain, the molecular model, and numerical simulation details. This section describes the main physicochemical properties calculated, how the studied biointerfaces were defined based on either bioinformatical tools or crystallographic data, and how some relevant biomolecular interactions invoking simple colloidal-like ideas were estimated. A key concept is the p$K_a$ that is computed for all titratable residues and represented in



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212−16229

different color-coded maps for a comparison among the strains especially at the biointerfaces. In the subsequent Results and Discussion section, two groups of data are given. First, we focus on the global protein properties related to charges and dipoles and on the corresponded estimated free energies of interactions. Then, we present and discuss p$K_a$'s and their maps comparing the strains. This allows us to cover from a more basic physicochemical description in the first part to a more biochemical meaningful discussion in the second part. Finally, the main results of this paper are discussed and summarized.

## THEORETICAL METHODS

Computational approaches play an important role in modern virology contributing from studies of the virus assembly to insights into the dynamical and interactive properties of the structures of viruses and their components.[32,39] In this scenario, numerical simulations based on statistical mechanics methods allow to explore many thermodynamical properties and also the main features of the interaction of macromolecules providing free-energy derivatives as a function of the macromolecules separation at different environmental conditions. Here, we combine a fast constant-pH Monte Carlo (MC) scheme with canonical bioinformatic tools to investigate the ZIKV NS1 from different strains.

**Three-Dimensional (3D) Modeling.** Initially, we defined the most interesting ZIKV strains for this investigation based on previous phylogenetic studies, the history of ZIKV outbreaks, and the rise in the cases of microcephaly and the Guillain−Barré syndrome. The Uganda (African lineage) and Brazilian (Asian lineage) strains were assumed to be the two extreme and most important cases supported by the literature.[9,11,12] The sequences of Uganda (UG), Senegal (SE), Central African Republic (CAR), Malaysia (MA), Thailand (TH), Yap Island (YAP), and Brazil (BR) were selected from the study of Kochakarn and collaborators,[12] where phylogenetic reconstructions were carried out from ZIKV genomes from different regions of the world. These sequences are identified, respectively, with accession numbers AY632535, KF383117, KF268949, HQ234499, KU681081, EU545988, and KU365777, provided by the National Center for Biotechnology Information. Conceptual translations were made with the Basic Local Alignment Search Tool, using "blastx" with default parameters.[40] Crystal structure coordinates are available for the strains from Brazil and Uganda only. They were provided by the RCSB Protein Data Bank (PDB)[41] with PDB ids 5GS6 (resolution 2.85 Å) and 5K6K (resolution 1.89 Å).[17,18] These two structures are quite similar except for a few structural differences that are observed in some loop regions (1.9 Å root-mean-square deviation (RMSD) across all 352 pairs and 0.706 Å RMSD between 310 pruned atom pairs). The PDB files were edited before the calculations. All water molecules and heteroatoms were removed. Missing residues were added in the structures with the "UCSF Chimera 1.11.2" interface of the program "Modeller" for building all missing segments.[42] We will refer to these completed three-dimensional structures with the label "X-ray" in this text. The original PDB structures with chains A and B were used as templates in all comparative modeling reported here. There are no crystal structures available for the ZIKV NS1 proteins from the other strains (SE, CAR, MA, TH, and YAP). Their three-dimensional models in both oligomeric states (monomer and dimer) were built from the known structures using Modeller version 9.18.[43] For each variant and oligomeric state, we generated two different models using the completed structures of the PDB ids 5GS6 and 5K6K as templates. Accordingly, the use of two comparative models makes it possible to evaluate the impact of the structural fluctuations in all outcomes. For the same reason, comparative models for the UG and BR strains were also produced with the crossing structures (e.g., the BR sequence and the UG template were used together to generate the BR three-dimensional model). We shall refer to the resulting models obtained with the completed structures (the templates) from PDB ids 5GS6 and 5K6K, respectively, as 1 (for the BR strain) and 2 (for the UG strain). The subscripts "mod1" and "mod2" are used to label them in the next sections. Monomers were produced by the simple deletion of one given chain.

Finally, as suggested by Wang and coauthors,[19] we used the structure of the complex between the antibody Fab fragment 22NS1 and the incomplete structure of WNV (PDB id 4OII with a resolution of 3.0 Å)[37] to estimate the location of possible residues of ZIKV NS1 that have a chance to be part of epitopes and be responsible for the protein−antibody interaction. The residues 1−175 that correspond to the N-terminal domain of the NS1 protein were missing in this X-ray structure. By comparison at the sequence level of the NS1 proteins given by the PDB ids 4OII (WNV NS1), 5GS6 (ZIKV $NS1_{BR}$), and 5K6K (ZIKV $NS1_{UG}$), we removed from the PDB files 5GS6 and 5K6K the 175 residues from the N-terminal domain that were missing in the PDB file 4OII. The generated models were used only for the calculations related to the antibody analysis. We assumed that the absence of the N-terminal domain was not necessary for the interaction of NS1 with the antibody. This is reasonable since the structure without the N-terminal domain was found bound with the antibody at the crystal, indicating that the missing part is not essential for the complexation. The used antibody (available in the PDB id 4OII) is the Fab fragment 22NS1 as reported by Edeling et al.[44] The structures of the NS1 proteins without the antibody are superimposed and shown in Figure 1. It is interesting to point out that 13 residues on the WNV epitope out of the 21 residues are identical to the corresponding residues in ZIKV.[19]

**Coarse-Grained (CG) Modeling.** Coarse-grained models are very convenient and appropriate to be applied in the present study because they offer the possibility to explore the main physical features of a system with a reduced number of parameters and lower computational costs, which allows us to repeat the calculations on several different experimental conditions. Therefore, the proteins were modeled here as rigid bodies (i.e., bond lengths, angles, and dihedral angles are kept fixed) in an amino acid coarse-grained (CG) level of details according to the three-dimensional structures provided by the comparative modeling as described above. Following previously successful studies,[45,57] each protein atomistic coordinate is transformed into a collection of charged Lennard-Jones (LJ) spheres of radii ($R_i$) and valence $z_i$ mimicking amino acids. Amino acids sizes were taken from ref 45. For the sake of convenience, these values are also reported in Table S1. The valences of all ionizable residues were assigned by the fast proton titration scheme (FPTS)[46,47] according to the solution pH. Details of this titration scheme, its p$K_a$ benchmarks, and the discussions related to its approximations can be found elsewhere.[47] We note that this CG model focuses on the electrostatic features of the



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212−16229

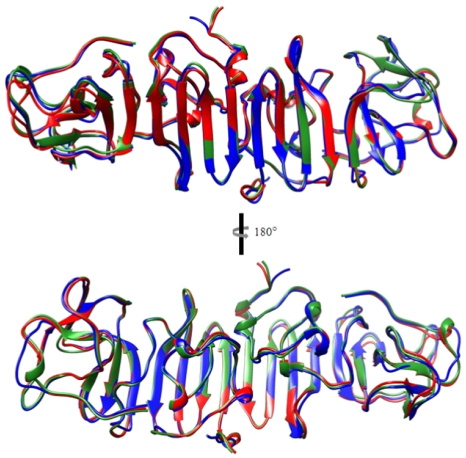



**Figure 1.** Comparison of different crystallographic $NS1_{176-352}$ structures from two flaviviruses. Superimposed structures in ribbon representation are the NS1 from WNV (blue), $ZIKV_{UG}$ (red), and $ZIKV_{BR}$ (green). The PDB ids are 4OII, 5K6K, and 5GS6, respectively.

biomolecular system and has been predicting p$K_a$'s with good accuracy at low computational costs.[27] This is highly important in the present study due to the elevated number of ionizable residues in the NS1 proteins, which increases the electrostatic coupling between them. As in the original FPTS paper,[46] the intrinsic acid dissociation constants (p$K_0$) were taken from the experimental work of Nozaki and Tanford.[48] These values together with the number of each amino acid in the ZIKV NS1 structures and their charge numbers in the deprotonated and protonated states are listed in Table S1.

Figure 2 gives a scheme of the CG model. Each protein is placed inside an electroneutral open cylindrical cell of radius

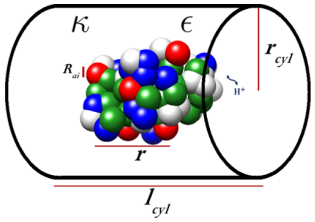



**Figure 2.** Schematic representation of model cell used in this study. For each titration, one protein was built up by a collection of charged LJ spheres of radii ($R_{ai}$), and valence $z_{ai}$ is placed in an electroneutral open cylindrical cell of radius $r_{cyl}$ and height $l_{cyl}$. The solvent is represented by its static dielectric constant $\epsilon$. Counterions and added salt particles are represented by Debye's term $\kappa$. Positively and negatively charged, neutral, and hydrophobic amino acids are represented in blue, red, white, and green, respectively.

$r_{cyl}$ and height $l_{cyl}$ with a static dielectric constant of 78.7 (assuming a temperature of 298 K). The shape of the simulation box is chosen only for convenience and has no effect on the outcomes as soon as the box is bigger than the protein size. Counterions and added salt particles were represented implicitly using a screening term, i.e., for two titratable residues $i$ and $j$, the screening is given by $[\exp(-\kappa r_{ij})]$, where $\kappa$ is the modified inverse Debye length length (as suggested in ref 81) and $r_{ij}$ is the interparticle separation distance. Therefore, the electrostatic interactions $[u_{el}(r_{ij})]$ between any two charged amino acids of valences $z_i$ and $z_j$ are given by

$$u_{el} = \frac{z_i z_j e^2}{4\pi\epsilon_0\epsilon r_{ij}} \exp(-\kappa r_{ij}) \quad (1)$$

where $\epsilon_0$ is the dielectric constant of the vacuum ($\epsilon_0 = 8.854 \times 10^{-12}$ $C^2$ $N^{-1}$ $m^{-2}$), $\epsilon_s$ is the dielectric constant of the medium (78.7 here), and $e = 1.602 \times 10^{-19}$ C is the elementary charge. See refs 27 and 47 for more details.

**Constant-pH Monte Carlo Simulations.** Monte Carlo (MC) methods are numerical integration tools that have become a useful procedure to compute thermodynamic averages in theoretical biochemistry and contribute to rationalize complex molecular mechanisms.[45,49,50] The CG model was solved via MC simulations using the Metropolis criteria in an NVT ensemble. Calculations were performed with a personalized version of the Faunus biomolecular simulation package,[49] where the FPTS is implemented to study pH-dependent processes.[46,47] In the first part of the work, to illustrate that the differences observed in the simulated p$K_a$'s do influence the biomolecular interactions, three physicochemical properties [the protein net charge number ($Z_p$), the charge regulation parameter ($C_p$), and the dipole moment number ($\mu_p$)] were evaluated during the Monte Carlo runs for each ZIKV strain (UG, SE, CAR, MA, TH, YAP, and BR) using each structure model (X-ray, mod1, and mod2) and oligomeric state (monomer and dimer). They are averaged quantities calculated from all molecular charges spatially distributed in the system sampled during the MC runs. $Z_p$, $C_p$, and $\mu_p$ were also computed for the antibody, the incomplete structure of WNV NS1 protein, and the corresponding truncated structures for $NS1_{ZIKV\ BR}$ and $NS1_{ZIKV\ UG}$. These quantities are equivalent to the ones obtained by the Henderson–Hasselbalch equation with the local electrostatic potential used by Podgornik and coauthors[51] and represent the lower multipoles moments and the charge fluctuations based on the details of charge distributions in these proteins and their acid–base equilibria. Their magnitudes will be used here to estimate different free energies of interactions within a simple colloidal-like framework. $C_p$ is the variance of the mean net protein charge, $C_p \equiv \langle Z^2\rangle - \langle Z\rangle^2 \propto -\frac{\partial\langle Z\rangle}{\partial \mathrm{pH}}$, which is an example of a linear response. Note that $z_i$ (a given charge in a titratable group $i$) is based on the total electrical potential at the coordinates of the titratable site $i$, and $\langle Z\rangle = \sum_1^N \langle z_i\rangle$ is measured after all $z_i$ values for the $N$ titratable sites were obtained during the MC runs. Salt concentration was kept constant at physiological conditions (150 mM) in all calculations. It was assumed that this is the ionic strength where the most important metabolic processes of the virus occur. A vast number of simulations were performed as a consequence of the numbers of strains, structural models, and oligomeric states. For each system (i.e., a given strain, structural model, and oligomeric state), a series of MC simulations with pH varying from 0 to 14 using a 0.1 interval were run to produce the theoretical titration plots and similar ones for the other pH-related properties. It was from this series of 140 independent simulations that the p$K_a$'s were computed as the pH where the residue is half-protonated as it is done in wet laboratory experiments.[52,53] For each pH condition, at least two production runs with $10^8$ MC steps were carried out after an equilibration phase of $10^6$ MC runs.



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212–16229

**Maps of the Electrostatic Properties.** Computed p$K_a$'s were converted into color-coded maps using the completed structure of the ZIKV $NS1_{UG}$ (seemingly older strain) as reference. For the comparisons at the monomeric state, chain A of the $NS1_{UG}$ was adopted as the reference. The p$K_a$ shifts were mapped at both the proteins sequences and their molecular surfaces. For the mapping at the sequence level, the alignments were done using the high-speed multiple sequence alignment web tool "MUSCLE" (Multiple Sequences Alignment by Log-Expectation).[54] Default parameters and the sequence conservation criteria were used. Positive shifts are represented in dark blue, whereas red is used for the negative ones. The p$K_a$ shifts can be due to the sequence (strain), structure, and mixed effects. To differentiate them, three p$K_a$ shifts were calculated for each case depending on the molecular model (1 or 2) used as input in the MC runs: (a) $\Delta pK_{a_1} = pK_{a_strain_mod1} - pK_{a_UG_xray}$, (b) $\Delta pK_{a_2} = pK_{a_strain_mod2} - pK_{a_UG_xray}$, and (c) $\Delta\Delta pK_a = \Delta pK_{a_1} - \Delta pK_{a_2}$. $\Delta pK_a$ values other than zero are candidates for interesting cases. However, to discriminate p$K_a$ shifts related to the sequence from those shifts due to the comparative modeling process, we analyzed the $\Delta\Delta pK_a$'s too. Only the cases where the signs of the p$K_a$ for the two models are the same (i.e., the variation of the p$K_a$ occurs in the same direction, which happens when $\Delta pK_{a_1} \times \Delta pK_{a_2} > 0$) are the relevant ones for the present analysis. These are the cases that can be used to quantify the effects from the evolutionary mutations. By simultaneously observing these two sets of p$K_a$ values (for models 1 and 2), we can find the five possible situations listed in Table 1. There are clear situations where the p$K_a$ shifts are due to the different sequences. For instance, this can be seen for condition 1 in this table. When $\Delta\Delta pK_a$ is equal to 0 and the product of the p$K_a$ shifts between the two models is greater than 0 (condition 1 in Table 1), there is a p$K_a$ shift and both models give the same shifts. So, there are no artifacts that were produced by the comparative modeling process. This is one of the most important cases. Conversely, we can find situations where there can be a mixture of the effects produced by the structure and by the primary sequence. These two features are contributing together and are indistinguishable. This is what happens for condition 4. There are shifts ($\Delta\Delta pK_a \neq 0$), but each model predicts the shift in the opposite direction ($\Delta pK_{a_1} \times \Delta pK_{a_2} \leq 0$). In this situation, the shifts are most likely due to the structural differences. Finally, since $NS1_{UG}$ is used as a reference, we have distinguished those amino acids where some mutation occurs in $NS1_{BR}$ (condition 5 in Table 1). For our purposes, the most interesting situations are given by conditions 1, 3, and 5. Additional colored geometric figures were included at the top of the alignments to guide the reader to the most interesting p$K_a$ shifts from the biological perspective. They were based on the shifts between the BR and UG strains. Table 1 summarizes all of these issues and lists the adopted criterion, the relevant cases, and the corresponding colored geometric figures. Alignments were also made from the sequences of the incomplete structures of the NS1 of WNV and the corresponding one from the ZIKV (UG and BR strains). For these comparisons, we have used the NS1 of WNV as reference. This analysis was simpler because all of the computed p$K_a$'s were done using the crystallographic structures in this case.

**Table 1. Criteria To Find the Biologically Relevant Cases with p$K_a$ Shifts**[a]

| | condition | biological relevance | graphical representation |
|---|---|---|---|
| 1 | $\Delta\Delta pK_a = 0$ and $\Delta pK_{a_1} \times \Delta pK_{a_2} > 0$ | relevant | circle |
| 2 | $\Delta\Delta pK_a = 0$ and $\Delta pK_{a_1} \times \Delta pK_{a_2} = 0$ | not relevant | triangle |
| 3 | $\Delta\Delta pK_a \neq 0$ and $\Delta pK_{a_1} \times \Delta pK_{a_2} > 0$ | relevant | square |
| 4 | $\Delta\Delta pK_a \neq 0$ and $\Delta pK_{a_1} \times \Delta pK_{a_2} \leq 0$ | not relevant | inverted triangle |
| 5 | mutation | relevant | diamond |

[a] $\Delta\Delta pK_a$ is calculated between the p$K_a$ shifts values from models 1 ($\Delta pK_{a_1} = pK_{a_strain_mod1} - pK_{a_UG_xray}$) and 2 ($\Delta pK_{a_2} = pK_{a_strain_mod2} - pK_{a_UG_xray}$) of $NS1_{BR}$, where $pK_{a_strain_mod1}$, $pK_{a_strain_mod2}$, and $pK_{a_UG_xray}$ refer to the computed p$K_a$'s obtained using the modeled protein structures, respectively, from the template from the BR strain, from the template from the UG strain, and from the completed UG crystallographic structure.

Complementarily, the p$K_a$ shifts were also projected at the three-dimensional structures. For the sake of consistency, the color-coded map was obtained by the $\Delta pK_a$'s between the UG and BR strains. These two strains represent the oldest and the newest mutations acquired during the ZIKV evolution. They also belong to different lineages, which makes them appealing for this comparison. We used the available crystallographic coordinates for this analysis, from where we also extracted the *B*-factors. For the estimation of the conserved regions, the "CONSURF" server[55] was employed with default parameters and 90 sequences as suggested by the server.

**Analytical Estimates.** Semiquantitative theoretical estimates of the free energy of interactions provided a practical way to rationalize the molecular processes and are often obtained from analytical equations based on simple Derjaguin−Landau−Verwey−Overbeek (DLVO) theory arguments.[45,56,57] Following a previous work,[45] the electrostatic free energy for the self-association of the ZIKV NS1 proteins $[A_{DVLO}(r_{ij})]$ can be approximately described in terms of electrostatically screened contributions as

$$A_{DVLO}(r_{ij}) \cong A_{ZZ}(r_{ij}) + A_{chreg}(r_{ij}) + A_{Z\mu}(r_{ij}) + A_{\mu\mu}(r_{ij}) \quad (2)$$

where, for the two macromolecules immersed in an electrolyte solution, these terms can approximately be written as

$$A_{ZZ}(r_{ij}) = \frac{l_B\langle Z\rangle\langle Z\rangle}{r_{ij}}\left\{\frac{\exp[-\kappa(r_{ij}-d)]}{(1+0.5\kappa d)^2}\right\} \quad (3)$$

$$A_{chreg}(r_{ij}) = -l_B^2\left(\frac{\langle C_p\rangle^2 + 2\langle C_p\rangle\langle Z\rangle^2}{2r_{ij}^2}\right)\left\{\frac{\exp[-\kappa(r_{ij}-d)]}{(1+\kappa d)}\right\}^2 \quad (4)$$

$$A_{Z\mu}(r_{ij}) = -\frac{1}{3}l_B^2\frac{\langle Z\rangle^2\langle\mu\rangle^2}{r_{ij}^4}S_{Z\mu} \quad (5)$$

$$A_{\mu\mu}(r_{ij}) = -\frac{9}{3}l_B^2\frac{\langle\mu\rangle^4}{r_{ij}^6}S_{\mu\mu} \quad (6)$$

The screened factors for the dipolar interactions are



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212−16229

$$S_{Z\mu} = \left\{ \frac{3(1+\kappa r_{ij})\exp[-\kappa(r_{ij}-d)]}{(1+0.5\kappa d)[2+\kappa d+0.25(\kappa d)^2+(1+0.5\kappa d)]} \right\}^2$$

and

$$S_{\mu\mu} = \frac{[(2+2\kappa r_{ij}+(\kappa r_{ij})^2)^2+2(1+\kappa r_{ij})^2]\exp[-2\kappa(r_{ij}-d)]}{[2+\kappa d+0.25(\kappa d)^2+(1+0.5\kappa d)]^4}$$

The size of the proteins in different oligomeric states was estimated from their full atomistic coordinates using the completed structures. The "ProteinVolume" server[58] was employed for this purpose. For the sake of convenience, on the basis of the calculated sizes, we assumed a separation distance $r_{ij}$ equal to 50 Å for the monomeric state and 100 Å for the dimeric state. The theoretical framework behind this set of equations is the Phillies model,[59−61] where anomalous properties of the screening environment are not considered. The limitations of this approximation are discussed in the literature.[51,62,63] A related point to consider here is that although more elaborated theoretical approaches might be important for protein−protein complexation in some cases, there is still an interest in general mechanisms such as the used free-energy equations using these simpler colloidal-like frameworks usually in good agreement with numerical simulations and that can contribute to rationalize more complex systems.[45,57,64,82] The variable $d$ was seen as an adjustable parameter here[65] and assumed to be equal to $r_{ij}$.[45] For the NS1−antibody association, only the charge−charge contribution was taken into account due to the severe screening from the other terms.[66] All results were expressed in $kT$ units, where $k$ is the Boltzmann constant ($1.3807 \times 10^{-23}$ J mol$^{-1}$ K$^{-1}$) and $T$ is the temperature (in kelvin).

A similar approach was used by Lund, Åkesson, and Jönsson to estimate the free energy of interaction [$\Delta A(R)$] between a protein (hisactophilin) and a charged surface mimicking a lipid membrane[67]

$$\beta\Delta A(R) = \frac{e\phi}{k_B T} Z_p - \frac{1}{2}\left(\frac{e\phi}{kT}\right)^2 C_p \quad (7)$$

where $\phi$ is the electrostatic surface potential generated by the charged membrane groups and $\beta$ is equal to $1/kT$. We have adopted $\phi = -95$ mV, which roughly corresponds to the potential calculated for ER membranes in neurons.[68]

**Biological Interfaces.** As already mentioned, we have focused on three selected biological interfaces of interactions: (1) dimeric (or self-association), (2) membrane, (3) and a putative epitope. They were determined by canonical structural bioinformatics tools. For the dimeric interface, the online server "PDBePISa"[69] was used, providing the residues directly involved in the NS1−NS1 interaction. In a similar way, the interface NS1−antibody as given by the PDB id 4OII was dissected by "PDBePISa". The choice of an experimental crystallographic structure to determine the epitopes as done here avoids all of the uncertainties commonly seen in epitope predictions by different theoretical methods.[32] The embedded residues in the membrane were estimated by the "PPM" server.[70] As the outputs from the "PPM" can be slightly dependent on the rotation, we have submitted rotated coordinates of the same structural model to have an inclusive protocol taking into account all amino acids that most likely interact with the membrane as given by each coordinate submitted to the "PPM" server. A few differences were observed between the rotated models. Depending on the coordinates, about one to three neighbor amino acids can be included or removed from the obtained prediction. Therefore, we have included all returned list from the server for the different runs with a set of rotated coordinates. The percentage of the exposure of the titratable amino acids was determined by the "PROPKA" program (version 3.1).[71] The outcomes were used to create a grayish color scale that is included in the alignment figures. The most exposed ionizable amino acids are represented in black in this scale, whereas white is used for buried residues. All programs were run with default parameters.

## RESULTS AND DISCUSSION

**Physicochemical Properties of the Proteins.** Electrostatic interactions are described as very important for virus in general[25,26,72,73] and also specifically for the NS1 protein.[17,28] The high number of titratable groups in NS1 (e.g., there are 34 GLUs in the monomeric structure) does indicate that these interactions should contribute to the molecular mechanisms. For this reason, we started the present study, quantifying the main physicochemical properties of the proteins that are directly responsible for the molecular interactions and, therefore, can be related to the virus molecular basis and, as such, contribute to the understanding of the differential infectivity of the strains. In this context, within a simple colloidal-like framework, the total charge ($Z_p$) of the system, the charge regulation parameter [or capacitance, ($C_p$)], and the dipole moment number ($\mu_p$) are key physical quantities that provide an important initial insight into the biomolecular interactions between the NS1 proteins and their partners. They directly affect the different physical forces responsible for protein associations (e.g., coulombic attraction for oppositely charged proteins,[57] dipolar attraction for macromolecules with high dipole moments,[45] and macromolecular attraction due to a charge regulation mechanism[50,64,66,74,82]). The global similarities and especially the differences among these physicochemical properties of the NS1 proteins from all of the studied strains might have a direct biological implication as discussed above. Charge−charge interactions are expected to be the most important one.[66] Hence, the values of $Z_p$, $C_p$, and $\mu_p$ at the physiological concentration of salt (150 mM) and different solution pH values were obtained from the numerical simulations with a single protein in the electrolyte solution. Calculations were carried out for each NS1 genetic variant of ZIKV, i.e., from Uganda (UG), Senegal (SE), Central African Republic (CAR), Malaysia (MA), Thailand (TH), Yap Island (YAP), and Brazil (BR), for the two oligomeric states (monomer and dimer). Three-dimensional (3D) structures were obtained directly from X-ray atomistic data when available or were built using homology modeling methods (see Theoretical Methods for details of the procedure). The high sequence identity (the smallest sequence identity between these strains is 96.9%; see Table 2) makes the procedure robust and the models reliable. Nevertheless, to evaluate the impact of the structural fluctuations brought by the building procedure, as already mentioned, we used two different models (noted models 1 and 2), based on two different templates (5GS6 and 5K6K), for each variant and oligomeric state.



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212−16229

**Table 2. Percentages (%) of Identity of Sequences for Different NS1 ZIKV Strains**

| | identity | | | | | | |
|---|---|---|---|---|---|---|---|
| | UG | SE | CAR | MA | TH | YAP | BR |
| UG | 100.0 | | | | | | |
| SE | 98.3 | 100.0 | | | | | |
| CAR | 98.0 | 98.0 | 100.0 | | | | |
| MA | 97.4 | 97.4 | 97.7 | 100.0 | | | |
| TH | 96.9 | 96.9 | 97.4 | 99.4 | 100.0 | | |
| YAP | 96.9 | 97.4 | 97.4 | 98.9 | 98.9 | 100.0 | |
| BR | 97.4 | 97.4 | 98.0 | 99.4 | 99.4 | 99.4 | 100.0 |

Figure 3 shows the computed average net charges and their charge regulation parameters as a function of pH at physiological ionic strength from pH 5 to 7.5. This is the pH range of important cell compartments.[75] The observed titration pattern of each strain is well characterized. Despite the small changes in the sequence, the total protein charge of each strain is differently affected by the pH. The $NS1_{UG}$ and $NS1_{CAR}$ strains are predicted to be two extreme cases, the $NS1_{UG}$ being the more negatively charged protein and the $NS1_{CAR}$ the more positively charged one. Three different groups of strains can be identified for the ZIKV NS1 proteins in terms of their titration behavior and isoelectric points (pI). For the monomeric state, (a) the UG strain with pI = 6.7 is apart from the other groups, (b) the SE, MA, TH, YAP, and BR strains with an intermediated value of pI (7.1) form a group with similar titration properties (the curves for MA, TH, YAP, and BR are seen on top of each other), and (c) the CAR with a pI shifted to the basic regime (pI = 7.4) defines another extreme case (see also Table S2). For the dimeric state, a similar behavior is observed ($pI_{UG} < pI_{SE,MA,TH,YAP,BR} < pI_{CAR}$). In this case, the corresponding pIs are 6.6, 7.1, and 7.3 (see also Table S3). Note that the pI values are all around physiological pH (~7.0) for both the monomeric and dimeric states. This facilitates, for instance, the NS1 self-association around the pI due to the relatively small protein net charges that result in weaker or almost null repulsive electrostatic interactions between the protein pairs. Since this pH condition is found in several cellular environments, it indicates that all ZIKV strains are able to take advantage of it. It is interesting to note that the intermediated group contains strains mostly from the Asian lineage (MA, TH, YAP, and BR strains). This is particularly more evident for the pH window between 5 and 6.5, where the SE strain (African lineage) is a bit outside the intermediated group. From the monomeric to the dimeric state, the difference between the UG strain and the others tends to increase. This means that the UG strain dimer tends to be less positively charged at the acid regime (pH < pI) and more negatively

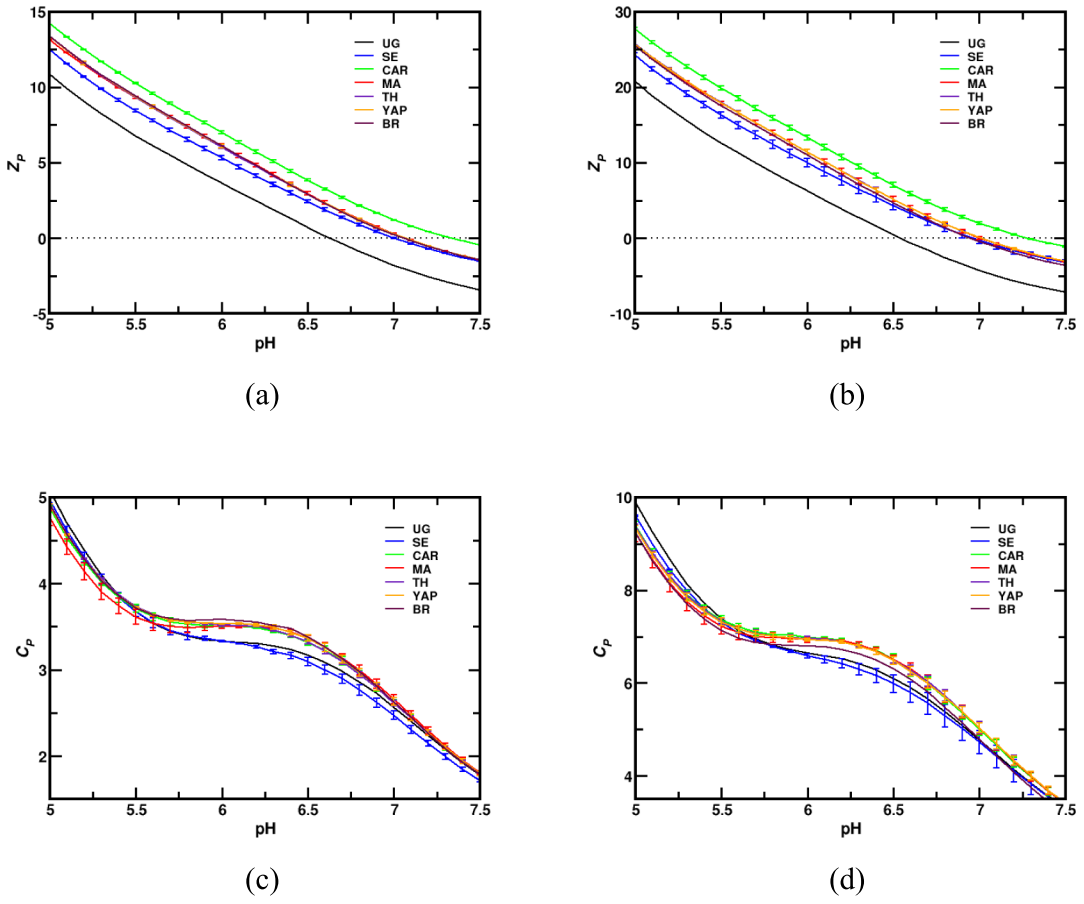



**Figure 3.** Physicochemical properties of different ZIKV NS1 strains (using model 1). (a, b) Simulated protein net charge number as a function of pH. (c, d) Simulated charge regulation parameter as a function of pH. Plots on the left- and right-hand sides are, respectively, for the monomers and the dimers. The estimated deviations calculated based on the averaged values for models 1 and 2 are also shown. Data from MC simulations at 150 mM of salt concentration. We note that some curves are on top of each other.



DOI: 10.1021/acsomega.8b02081
*ACS Omega* 2018, 3, 16212−16229

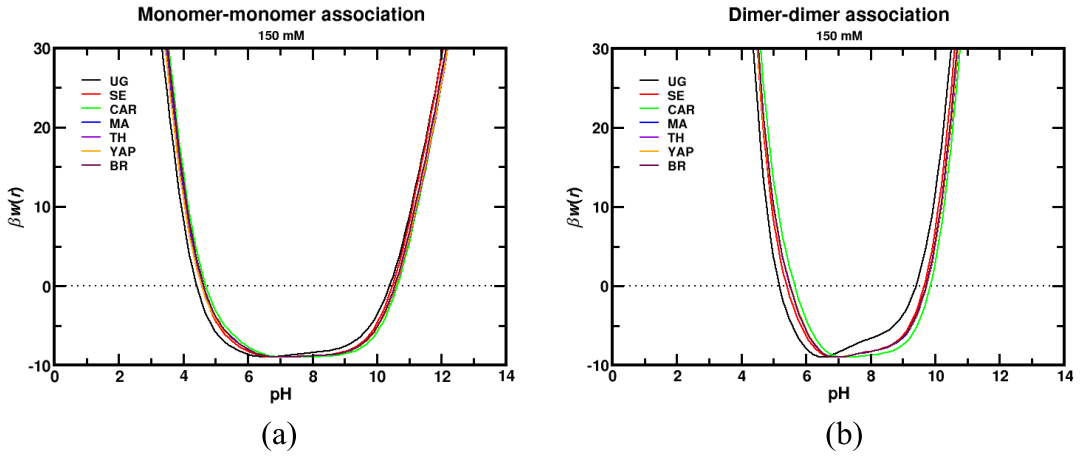



**Figure 4.** Homodimerization in two different types of oligomers for different NS1 of ZIKV. The electrostatic free energies of interactions are calculated as a function of pH. (a) Monomer state. (b) Dimer state. Data from the analytical equations using the parameters obtained from the MC simulations at 150 mM of salt concentration.

charged at the basic regime (pH > pI). Henceforth, it is expected that the $NS1_{UG}$ will interact differently from other charged molecules in the cell (e.g., for pH < 6.7, the monomeric form of $NS1_{UG}$ will be less attracted by a negatively charged membrane, whereas $NS1_{CAR}$ will be strongly attracted). This leads us to think that the electrostatic properties have an important evolutionary role defining different functionalities in highly conserved proteins such as NS1. Full titration plots obtained in our simulations are provided in the Supporting Information (Figure S1). From the very acidic to the basic pH conditions, the charge $Z_p$ approximately varies from +57 to −55 (for the monomeric state) and from +116 to −109 (for the dimeric state). In both high and low pH regimes, the dimers present a higher total protein charge than the monomers (see Figure S1). Structural differences due to the manner that the protein models were built are shown to be small.

Considering a pair of macromolecules, the second most relevant contribution in the interaction energy between them is often the charge regulation mechanism.[66] This induction interaction is especially important for the interaction of an approximately neutral protein with another (highly) charged macromolecule. Plots of the charge regulation parameter as a function of pH at physiological salt concentration are also shown in Figure 3. In the pH window of 5−7.5, it decreases from ca. 5 to 1.7, for the monomeric state, and from ca. 10 to 3.5, for the dimeric state. Since the attractive charge regulation mechanism depends on the product ($C_pZ_p^2$), this indicates that the charge regulation mechanism also decreases its importance for the homoassociation of the NS1 proteins as pH becomes closer to pI. Full plots of $C_p$ as a function of pH at 150 mM of NaCl are shown in Figure S1c,d. For the monomers, we observe that the highest peaks occur around pH values 3.6 ($C_p \approx 8.7$) and 10.8 ($C_p \approx 6.0$) (see Figure S1c). For the dimers, high peaks around pH values 3.6 ($C_p \approx 16.6$) and 12.5 ($C_p \approx 11.8$) are obtained as shown in Figure S1d. Between pH 5.5 and 7.0, we can observe that $C_p$ values are somewhat different for $NS1_{UG}$ and $NS1_{SE}$ compared to the other strains for both monomers and dimers. They tend to be smaller in this pH regime. Unlike what was seen for $Z_p$, we cannot clusterize the strains based on their values. There seem to be no strong differences between those of Africa and Asiatic origins in terms of their capacitance. Therefore, we can hypothesize that during the evolutionary process, this property has been conserved probably because it might be so significant for the biomolecular associations of NS1 and even would play no role in the differential virulence of ZIKV strains.

The other physicochemical property that can play a role in the molecular association is the dipole moment. However, the contribution of the charge−dipole term often vanishes for biomolecules at physiological salt condition. The dipole−dipole interactions also tend to be smaller than those of the other contributions.[66] Higher other moments would have a negligible contribution. For the sake of completeness, we evaluated $\mu_p$ for all of the strains. Results are presented in Figure S2 for both the monomeric and dimeric states at 150 mM of salt. There are significant differences depending on structural models because dipole moments are quite sensitive to the structural coordinates. Despite these, we can observe that each strain exhibits a specific curve pattern. This might indicate that although the total net charge is relatively maintained, there are local differences probably at the biointeractive interfaces. $NS1_{UG}$ has higher $\mu_p$ at all pH regimes. This can be explored in more detail in the future when theoretical methods that couple titration and conformational changes became feasible for larger proteins.

**NS1 Associations.** We have limited the present study on the investigation of the impact of each ZIKV strain in the interactions of NS1 with its three main partners: (a) itself in a dimerization process, (b) a membrane represented by a negative surface potential, and (c) an antibody. Following ref 66, we estimated the corresponding electrostatic free energy of interactions [$\beta w(r)$], applying the simulated $Z_p$, $C_p$, and $\mu_p$ values obtained at pH values in eqs 2−6. In this semi-quantitative analytical approach,[56] we have used the values of $Z_p$, $C_p$, and $\mu_p$ obtained from the MC simulations with model 1 at 150 mM of salt. On the basis of a simple thermodynamical criterion,[76] negative values of $\beta w(r)$ indicate an attractive regime, whereas $w(r) > 1$ corresponds to repulsion. Despite the intrinsic approximations, the central aspect here is to verify possible differences in the interactions between the NS1 protein for each strain and its different targets.

*NS1−NS1 Homooligomers.* We first examined the pH effects on the homoassociation process of the NS1 proteins for



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212−16229

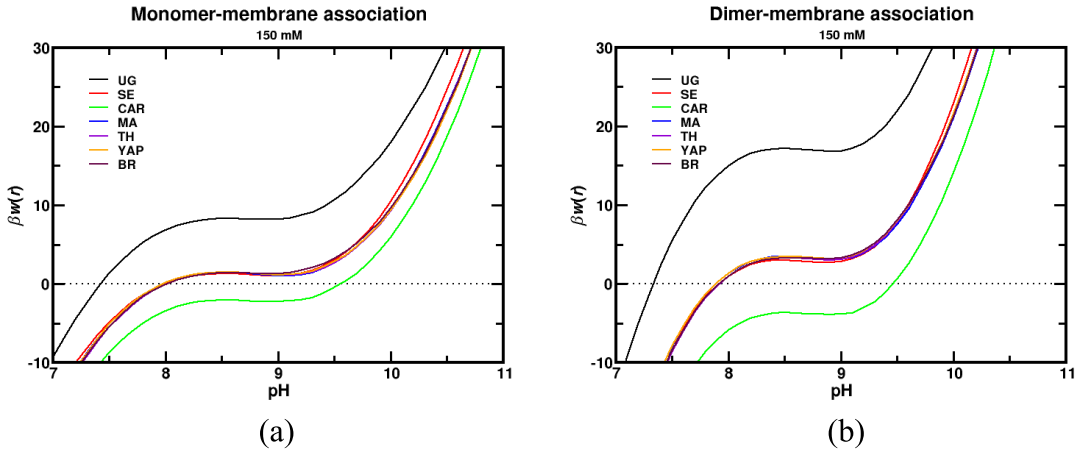



**Figure 5.** Protein−membrane interactions in two different types of oligomers for different NS1 of ZIKV. (a) Monomer state. (b) Dimer state. The electrostatic free energies of interactions between the complexes are calculated as a function of pH. Data from the analytical equations using the parameters obtained from the MC simulations at 150 mM of salt concentration.

all of the strains to form the homodimers from the monomeric units (NS1 + NS1 → NS1 dimer) as this is a very fundamental step to give NS1 the capacity to associate with membranes.[34] Another important association is the formation of hexamers from three homodimers (NS1 dimer + NS1 dimer + NS1 dimer → NS1 hexamer). For the sake of convenience, this last process was only partially investigated here in a simplified manner by means of a dimer−dimer interaction (NS1 dimer + NS1 dimer → NS1 tetramer) that might be relevant to the hexamer formation. The transition between the membrane-bound dimer (NS1 $\text{dimer}_{\text{bound}}$) to form a soluble trimer of dimers and then a hexamer is still unclear in the literature. Moreover, since tetrameric intermediates have also been reported,[33] it might be useful to have more information about this eventually intermediate step in the hexamer formation observed in late infection. Figure 4 shows the calculated electrostatic free energy of the interactions as given by eq 2 at physiological ionic strength, assuming $r$ = 50 Å. This corresponds to roughly the NS1 diameter (i.e., the center−center separation distance of two NS1 proteins at the contact). For the sake of convenience, all plots were shifted down by an arbitrary value of −9 $kT$ to mimic the addition of an estimated attractive van der Waals (vdW) contribution. In a typical Hamaker description often used in colloidal science, attractive numbers are usually in the range of 3−10 $kT$.[77] With this assumption, negative values of $\beta w(r)$ are observed at pH values between ca. 5 and 10, indicating that a homoassociation can happen at this relatively large pH window. Despite the simplicity of the model, this result agrees with the experimental data.[29] A smaller attractive vdW contribution will narrow this pH window without affecting the qualitative comparison between the strains, which is the central aspect of the present study. The picture is largely dominated by the charge−charge (coulombic) contributions. Both the charge regulation and dipolar interactions contribute with negligible numbers (less than 0.1$kT$ at maximum). As expected, the behavior of $\beta w(r)$ for each strain follows what was shown before for $Z_p$. The $\text{NS1}_{\text{UG}}$ is shifted to the acid pH regime, which permits the complexation to both start and be over at smaller pH values. All of the other strains have a similar profile with the $\text{NS1}_{\text{CAR}}$ being the farther case. The differences between the strains become clearer for the tetramerization process (Figure 4b). Thus, the observed differences between the $\text{NS1}_{\text{UG}}$ and the other strains in the self-association process might be related to their distinct biological responses as it is known that the formation of NS1 oligomers is very important for the ZIKV functionality.[2,28]

*NS1−Membrane Interaction.* We next investigated the pH effects on the interactions of the ZIKV NS1 protein from each strain with a general membrane model. Assuming that a surface membrane electrical potential equals −95 mV[68] and using in eq 7 the values of $Z_p$ obtained from the MC simulations, we estimated the electrostatic free energy of interactions for the protein−membrane association [$\beta\Delta A(R)$]. The results at 150 mM salt are shown in Figure 5 for both oligomeric states. The monomer−membrane case is included here for comparisons and the sake of completeness since the only biologically relevant system is the dimer−membrane. The same tendency for the $\text{NS1}_{\text{UG}}$ and $\text{NS1}_{\text{CAR}}$ to be extreme cases as observed for the NS1 dimer−dimer association is seen here too for both cases. The transition from monomer to dimer is reported to increase the hydrophobicity of NS1. Electrostatic attraction might contribute to this process as well. For instance, comparing the monomeric and dimeric cases, we can see that the discrepancy in $\beta w(r)$ between the different strains is increased for the dimers, i.e., the dimerization amplifies the biomolecular interactions in comparison with the monomers. The dimerization also slightly decreased the pH where a transition between an electrostatic attraction and a repulsive regime is observed [pH where $\beta w(r)$ = 0]. For the monomeric case, the $\text{NS1}_{\text{UG}}$ is attracted by the membrane until pH 7.4, whereas the $\text{NS1}_{\text{CAR}}$ is attracted by pH values up to 9.5. Conversely, for the dimer case, these pH values suffer a small decrease to 7.3 and 9.4 for, respectively, UG and CAR. Most of the NS1 proteins from the other strains ($\text{NS1}_{\text{SE}}$, $\text{NS1}_{\text{MA}}$, $\text{NS1}_{\text{TH}}$, $\text{NS1}_{\text{YAP}}$, and $\text{NS1}_{\text{BR}}$) can be grouped together in terms of their estimated interactive behavior. This group of proteins can be attracted by the membrane until an intermediate pH value (~8.0). At the pH of the ER lumen (pH 7.2), all proteins have an attractive behavior. CAR is the system where the strongest electrostatic attraction is analytically estimated at this pH value, followed by the group of strains mentioned above (SE, MA, TH, YAP, and BR), and finally by the UG strain.



DOI: 10.1021/acsomega.8b02081
*ACS Omega* 2018, 3, 16212−16229

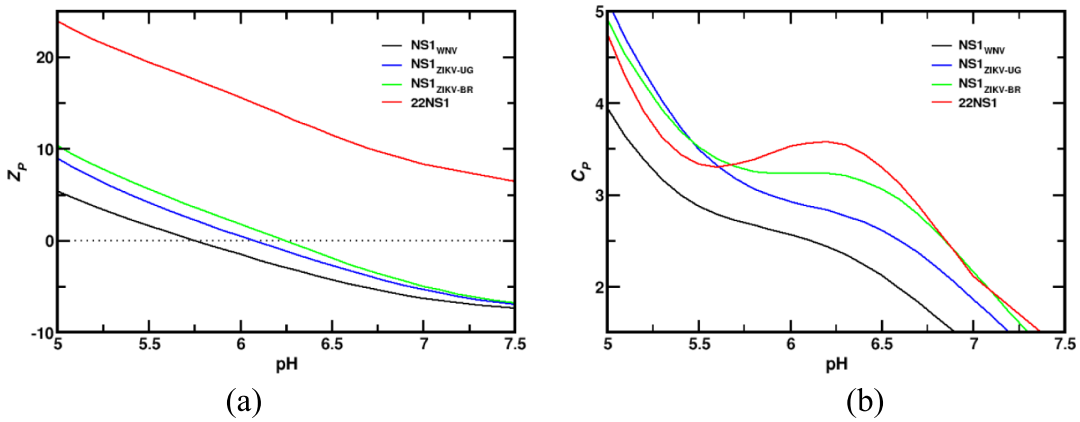



**Figure 6.** Physicochemical properties for different $NS1_{176-352}$ from two flaviviruses [WNV PDB id 4OII and ZIKV PDB id 5GS6 (BR) and 5K6K (UG)] and the Fab fragment 22NS1 (derived from 4OII). (a) Net protein charge number as a function of pH. (b) Charge regulation parameter as a function of pH. Plots have specifically focused on the typical physiological pH conditions for the virus stability and infectivity. Data from MC simulations at 150 mM of salt concentration.

Overall, we can see that the NS1 proteins from different strains do not have the same $\beta w(r)$ profile. Each strain produces an NS1 protein with specific molecular interactive properties, i.e., the magnitude of the attraction NS1 dimer−membrane is specific for each strain. It is legitimate to suppose that such differences on how NS1 interacts with the membrane should have an impact on the differential virulence of each strain among other factors. There should be a direct relation between the biomolecular interactions and the biological effects. The NS1−membrane association is a critical step to the assembly and release of the viral lipoprotein particle.[34] Akey and coauthors previously suggested the relation between the observed phenotype and the loss of effective interactions with transmembrane proteins of the replication complex.[35] On the basis of such arguments, a stronger affinity between the NS1 dimer and the membrane would promote a better association of this protein with intracellular membranes and consequently would favor the viral replication. Indeed, it has already been reported that African strains have a more aggressive effect on human neural stem cells compared to strains of Asian origin.[15] Moreover, it has been hypothesized that the strains from an African lineage ZIKV have a deadly effect on the early stages of the fetus, whereas the Asian lineage is less destructive, allowing the fetal development to continue, but then congenital malformations can appear at a later stage.[15,67] If these processes are related to the viral genome replication as it seems to be the case, it could be assumed that there is some kind of relationship between that and the results found here. As seen in Figure 5, for both the monomer and dimeric cases, it might be that African NS1 proteins (as $NS1_{CAR}$) have a higher affinity for negatively charged membranes since the residues involved in the interaction are positively charged. On the contrary, a less attractive effect at pH 7.2 should occur in the NS1 strains of Asiatic origin, due to the presence of more negative residues that prevent a strong affinity with the intracellular membranes. The hydrophobic characteristic of the NS1 homodimer might contribute to further enhance the NS1−membrane attraction.

*West Nile Virus NS1.* To go further and better understand the impact of sequence differences on the main physicochemical properties of NS1, we decided to examine a primary sequence, which has a sequence identity much lower than that found between the ZIKV strains. In this regard, the West Nile virus sequence (denoted WNV), for which a crystallographic structure of the NS1 complexed with an antibody is available, is a convenient case. Considering that the sequence identity between the available regions of the NS1 of WNV and ZIKV in the crystallographic structure is around 59%, we can examine how the amino acid composition affects these physicochemical properties in these proteins along the pH scale. Furthermore, the NS1−antibody complex can also be used as a "template" to estimate the interactions between the antibody and the NS1 from ZIKV. This does not necessarily mean that the WNV NS1−antibody can have a promiscuous behavior and also be able to bind the ZIKV NS1 *in vivo*. To our knowledge, there is no information about the possibility of polyspecificity or cross-reactions for the WNV−antibody. The theoretical calculations done here are simply to illustrate that each ZIKV strain would have a different electrostatic response for a given antibody molecule. This will be true even if the real ZIKV−antibody will be different from the WNV one as common sense suggests. In addition, it is interesting to point out that Wang and coauthors suggested that the WNV NS1 could be used in the development of antibodies against ZIKV.[19] Moreover, Rocha and coauthors[78] identified NS1−dengue epitopes that are recognized by two monoclonal antibodies (2H5 and 4H1BC) that also cross-reacted with NS1 ZIKV protein. In some cases, the sequence identity between the dengue epitope and the corresponding epitope in Zika can be less than 77% (see the Supporting Information of ref 78). Since our main goal was to evaluate the interaction properties of an epitope region with an antibody, in the absence of 3D structure of ZIKV with Rocha's antibodies, we turned out to the sole 3D complex that was indubitably characterized, namely, the evolutionary closely related WNV NS1 protein interacting with 22NS1. The crystallographic $NS1_{WNV}$ structure was missing the 1−175 N-terminal region. We have assumed that the missing amino acids would not affect the complexation because, even when absent in the known experimental structure, the oligomer could be formed in the crystal conditions. For comparison purpose, we reevaluated the electrostatic properties of $NS1_{ZIKV}$ restricting to the same region. The first procedure is to compute these properties for both the antibody and the $NS1_{WNV}$ and to compare them with the results for ZIKV. This


DOI: 10.1021/acsomega.8b02081
*ACS Omega* 2018, 3, 16212−16229

can be seen in Figure 6. The pI for $NS1_{WNV}$ (5.7) is smaller than what was previously computed for all $NS1_{ZIKV}$ strains. From pH 7.5 to the acid regime, the $NS1_{WNV}$ is always less charged than the ZIKV strains. The strain from the UG is the intermediate case, whereas the BR strain is the most positively charged molecule within this pH range. The same order is observed for the charge regulation parameter. The pI for the antibody is 8.7. From these titration data, it can be anticipated that the $NS1_{WNV}$ will form the strongest complex with the antibody, and the $NS1_{BR}$ the weaker complex. Due to its negligible contribution, $\mu_p$ was not reported here.

*NS1−Antibody.* Finally, using the C-terminus NS1 of WNV and ZIKV, we calculate the $\beta w(r)$ as a function of pH for these proteins with the antibody found in the crystallographic structure of WNV (see Figure 7). For $NS1_{ZIKV}$, we assumed

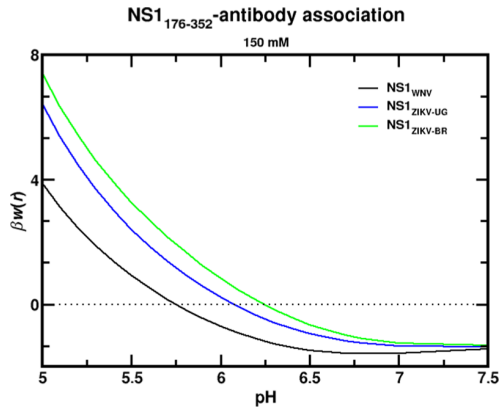



**Figure 7.** Protein−antibody interactions for different $NS1_{176-352}$ of two flaviviruses. The electrostatic free energies of interactions between the complexes are calculated as a function of pH. Data from the analytical equations using the parameters obtained from the MC simulations at 150 mM of salt concentration. The antibody Fab fragment 22NS1 used in these calculations is from the WNV as given by the PDB id 4OII.

that the epitope region is the same as that for $NS1_{WNV}$. We built the NS1 ZIKV−antibody complex by superimposing the common regions between $NS1_{WNV}$ and $NS1_{ZIKV}$. As expected, $\beta w(r)$ starts to be negative at a smaller pH for the WNV−antibody complex, confirming that this is the system with the stronger affinity followed by the $NS1_{ZIKV\ UG}$ and then the $NS1_{ZIKV\ BR}$. This result confirms that the ZIKV BR strain is less attracted by this specific antibody for WNV, which should imply a different immunological response. As expected based on the common sense for specificity for antibody−antigen interactions, an immune system that produces this antibody for WNV would have difficulties to neutralize the ZIKV $NS1_{BR}$. Our result clearly demonstrated that the electrostatics properties of the 22NS1 antibody have to be modified first before any other fine subtle residue changes. This result is of utmost importance as it significantly limits the number of residues combination to design a synthetic antibody.

**Mapping the Electrostatic Properties at the Sequence Level.** We described above the detected differences at the main global protein physicochemical properties. To refine this analysis, we also evaluated the differences in the electrostatic properties at the amino acid level due to the small variation on the sequence identity. This allows us to map the protein regions largely affected by the evolutionary mutations that could be related to the self-association of NS1 and the interactions with the membrane and the antibody. Consequently, these regions have different interaction affinities with their molecular partners and could be directly related to the patterns of differential infectivity found in different strains of ZIKV. In addition, regions where variations do not occur must be fundamental for the virus, and, as a consequence, their electrostatic conservation must be maintained. Computed $pK_a$ values of all titratable groups were employed in this mapping process. The $pK_a$ calculations were carried out for all available structures, i.e., the experimental ones as well as the 3D models. Thus, we can differentiate the shifts from the sequence itself from those that are due to the tertiary structure. For both the UG and BR strains, we performed the calculations with the two chains (A and B) that were accessible at the PDB. They are referred to as UG_chA_X-ray, UG_chB_X-ray, BR_chA_X-ray, and BR_chB_X-ray. For the other strains, the same notation was followed, replacing the "_X-ray" by the correspondent comparative structural models labels ("_mod1" and "_mod2" for the structures obtained from the templates from the BR and UG strains, respectively).

In fact, the interesting properties are the $pK_a$ shifts from a chosen reference system. The $pK_a$ shifts can be due to the sequence (strain), structure, and mixed effects. To distinguish impact from sequences to structures, three $pK_a$ shifts were calculated for each case depending on the molecular model, i.e., 1 or 2. In this study, we used the $NS1_{UG}$ as such reference. Accordingly, $\Delta pK_{a_1}$ (or 2) stands for the difference between $pK_a$ of a given amino acid in model 1 (or 2) of a given strain ($pK_a$_strain) compared to the same amino acid in X-ray structure of UG ($pK_a$_UG). We also evaluated $\Delta\Delta pK_a = \Delta pK_{a_1} - \Delta pK_{a_2}$, which reflects the fine influence of the structural building procedure. The ionic strength was kept constant at 150 mM during all simulation runs. Figure 8 presents the sequence alignments for all studied NS1 strains. At the bottom line, both the degree of conservation and the exposition to the solvent of the titratable amino acids are represented by the usual stars and the grayish little squares (where black means a very exposed amino acid, and white, a buried one). The results for the $\Delta pK_a$'s are displayed as a color-coded $\Delta pK_a$ scale, with blue being the positive shifts ($\Delta pK_a > 0$), and red the negative ones ($\Delta pK_a < 0$). The numerical data used to build up this color-coded $\Delta pK_a$ scale are reported in Table S4. Amino acids' one-letter codes highlighted in bold indicate those that are found at the NS1 self-association interfaces. The other important interfaces are indicated at the top line (above the amino acid numbering) by the letters "m" (for the interaction with the membrane) and "a" (for the NS1−antibody interface defined here based on the NS1 antigenic molecule known from the WNV NS1−Fab fragment 22NS1 from the PDB id 4OII). For the sake of convenience, we also introduced another notation at the third top line with a colored geometric figure. This is done to guide the reader to the most interesting $pK_a$ shifts from the biological perspective. Colored circles, triangles, squares, inverted triangles, and diamonds correspond to conditions 1, 2, 3, 4, and 5, respectively, as given in Table 1. This symbol comparison is done between the UG and BR strains. Sequence positions labeled with triangles indicate the $pK_a$ shifts that come from the differences in the three-dimensional structure. The most relevant cases to determine the shifts associated with the strains are both the ones where there was a mutation on a



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212−16229

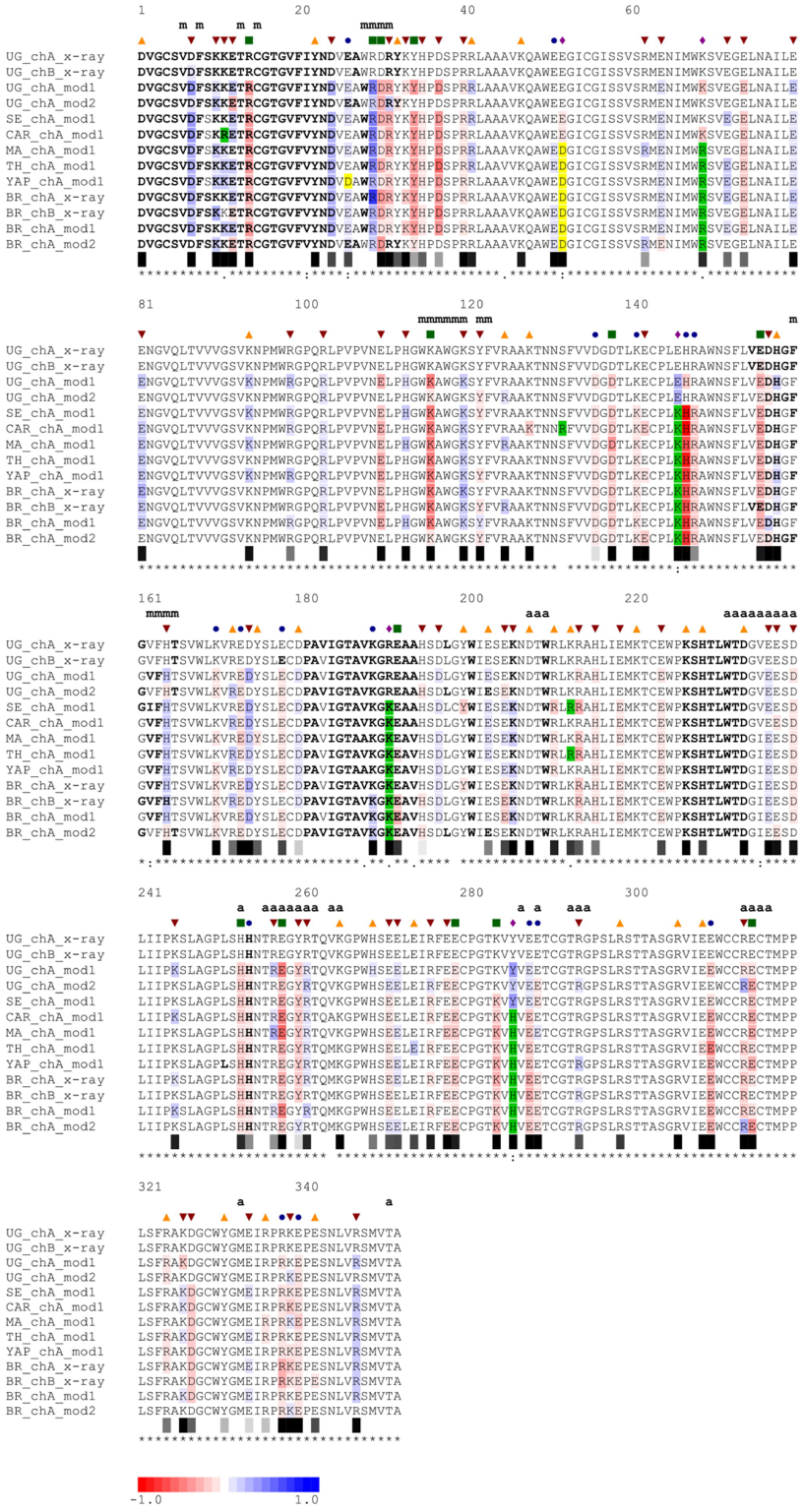



**Figure 8.** Alignment from different NS1 (monomers) of several geographic regions. When the amino acid suffers p$K_a$ variation, it is shaded with blue or red indicating, respectively, that it is more positive or more negative with regard to the reference ($NS1_{UG}$). Residues are shaded with yellow or green when a substitution per acidic or basic residue occurs with respect to the reference. Explanations about the symbols above the alignment are given in the text (see also Table 1, where the numerical criterion is provided to guide the use of these symbols). The letters "m" and "a" above the alignment are residues embedded in the membrane and epitope regions (this last region was defined based on the crystallographic structure of the complex WNV-Fab fragment 22NS1). Residues that belong to the dimeric interface are shown in bold. The bars below the alignment indicate exposure of charged residues on the protein surface, with darker areas indicating more exposure. The last line indicates the grade of conservation between proteins.



DOI: 10.1021/acsomega.8b02081
*ACS Omega* 2018, 3, 16212−16229

Table 3. Number of Ionizable Residues with p$K_a$ Shifts in Some Important Regions of $NS1_{BR}$ of ZIKV and Their Relations[a]

| | total | | | | dimer interface | | | | membrane interface | | | | epitope | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| residue | $N^T$ | $N'^T$ | $N'^T/N^T$ | $N'^V/N'^T$ | $N^D$ | $N'^D$ | $N'^D/N^D$ | $N'^D/N'^T$ | $N^M$ | $N'^M$ | $N'^M/N^M$ | $N'^M/N'^T$ | $N^E$ | $N'^E$ | $N'^E/N^E$ | $N'^E/N'^T$ |
| ARG | 25 | 4 | 0.16 | 0.50 | 1 | 1 | 1.00 | 0.25 | 2 | 1 | 0.50 | 0.25 | 3 | 0 | 0.00 | 0.00 |
| ASP | 16 | 3 | 0.19 | 0.33 | 5 | 0 | 0.00 | 0.00 | 1 | 1 | 1.00 | 0.33 | 3 | 0 | 0.00 | 0.00 |
| GLU | 34 | 13 | 0.38 | 0.31 | 2 | 1 | 0.50 | 0.08 | 0 | 0 | 0.00 | 0.00 | 5 | 3 | 0.60 | 0.60 |
| HIS | 12 | 2 | 0.17 | 1.00 | 3 | 1 | 0.33 | 0.50 | 1 | 0 | 0.00 | 0.00 | 1 | 1 | 1.00 | 1.00 |
| LYS | 22 | 6 | 0.27 | 0.50 | 5 | 2 | 0.40 | 0.33 | 2 | 1 | 0.50 | 0.17 | 0 | 0 | 0.00 | 0.00 |
| TYR | 8 | 1 | 0.13 | 0.00 | 1 | 0 | 0.00 | 0.00 | 1 | 0 | 0.00 | 0.00 | 1 | 0 | 0.00 | 0.00 |

[a] $N^T$ = total number of residues, $N'^T$ = number of residues with p$K_a$ shifts, $N'^V = N'^D + N'^M + N'^E$, $N^D$ = total number of residues in the dimeric interface, $N'^D$ = number of residues with p$K_a$ shifts in the dimeric interface, $N^M$ = total number of residues embedded in the membrane, $N'^M$ = number of residues with p$K_a$ shifts embedded in the membrane, $N^E$ = total number of residues in the antibody interface of NS1 WNV, $N'^E$ = number of residues with p$K_a$ shifts in the antibody interface of NS1 WNV.

specific sequence position (see the diamonds in Figure 8) and the cases where the p$K_a$ shifts for models 1 and 2 are in the same direction ($\Delta pK_{a_1} \times \Delta pK_{a_2} > 0$; see the circles and squares in Figure 8). For instance, the diamond at position 52 means that there is a mutation from the UG sequence (our reference in this study) to the BR strain. In this example, we can observe the effect of mutation E52D for the BR strain (and also for MA, TH, and YAP strains). At position 51, we can see the circle. In this case, E51 is conserved in the BR sequence (and all of the others), but in some strains, this residue experiences a different electrostatic environment, and positive p$K_a$ shifts were measured. From Table S4, we can see that both models (1 and 2) give exactly the same p$K_a$ shift ($pK_{a_1}$ mod1_br = 0.1 and $pK_{a_2}$ mod2_br = 0.1; see Table S4), indicating that the observed shifts are not an artifact from the structure. Similar cases where the p$K_a$ shifts are not exactly the same but keep the same signal can also be found. See the green square at position 14 in Figure 8, where $pK_{a_1}$ mod1_br = −0.4 and $pK_{a_2}$ mod2_br = −0.2 (values from Table S4). Actually, there are other cases like this one closer to the NS1 self-association interface (e.g., R29, E156, and E192). There are p$K_a$ shifts at the biological functional interfaces and also at other sequence positions with apparently no specific relation with them. A detailed description of each amino acid behavior with a description of the involved biofunctional interface is provided in Table S5. We can summarize the main observations from these analyses with the numerical quantification of all possible cases. These statistical data are given in Table 3. Interesting comments can be made using the data presented in this table. For instance, it becomes clear that TYR maintains a high degree of conservation in terms of the p$K_a$ shifts. There are seven TYRs that are not affected by the changes in the sequences from each strain, including the TYRs present at the functional interfaces (there is one TYR at each interface). The only TYR affected is not at these interfaces. The other titratable amino acids have situations where p$K_a$ shifts were observed in two functional interfaces. The total number of GLUs is 34, being the highest number of ionizable amino acids in the NS1 protein followed by 25 ARGs, 22 LYSs, 16 ASPs, and 12 HISs. From the 34 GLUs, 13 were affected by the mutations from the UG to the BR strains. Considering the total number of GLUs, this corresponds to a ratio equal to 0.38. It is also the highest relation between the number of affected cases of a given amino acid divided by its total number. Nevertheless, only parts of the affected GLUs are at the functional interfaces. There is one at the dimeric interface and three at the epitope. This amino acid is not presented at the membrane interface. HISs that are located at the functional interfaces have a high tendency to be sensitive to the mutations. We found three HISs at the dimeric interface, the membrane interface, and the epitope. HISs experienced p$K_a$ shifts at both the self-association and the epitope interfaces due to the evolutionary mutations: one out of three at the self-association interface and one out of one at the epitope. In short, we can say that (a) ARG (1), GLU (1), HIS (1), and LYS (2) are affected at the dimeric interface; (b) ARG (1), ASP (1), and LYS (1) at the membrane interface; and (c) GLU (3) and HIS (1) at the epitope. The number of occurrences of each amino acid at these interfaces was given between parentheses. It seems that these residues are essential for the functionality and survival of the virus since they are engaged in regions of crucial interactions.

**Mapping the Electrostatic Properties at the Structural Level.** The analysis at the sequence level can be seen as a rough estimation since the three-dimensional distribution of the amino acids charges has a clear effect on the strength of the molecular interactions that are distance-dependent properties. Even homologous proteins might have regions with different folds,[79] which have a direct effect on the protonation states of the titratable groups. Therefore, the p$K_a$ shifts reported above were also mapped at the protein surface following a similar color code. Exploring possible relations with other relevant properties, Figure 9 shows the p$K_a$ shifts regarding the same reference ($NS1_{UG}$) as a molecular surface color mapping. The threshold of $\Delta pK_a$ = 0.2 defined this color mapping to better highlight all affected molecular regions. In the same figure, we included a map for the degree of conservation and the *B*-factor of each residue. Because the *B*-factor is a thermodynamic measure of the structural flexibility and provides important information about the averaged protein dynamics properties, a comparison between this variable and p$K_a$ shifts is interesting to be investigated. Comparing the maps for the $\Delta pK_a$'s and *B*-factors, we can conclude that there is no clear relation between the protein flexibility and the $\Delta pK_a$'s for the $NS1_{BR}$. Conversely, there seem to be more p$K_a$ shifts at conserved regions of the protein. This makes sense in evolutionary terms since viruses can have some regions where the mutation rate is higher "to counter" the attacks from the hosts they infect.[73] However, we must emphasize that it is also interesting that independent of that degree of conservation, the variation of p$K_a$ can be seen between strains with a high sequence identity. In this way, each particular NS1 strain of ZIKV would have the ability to interact differently from their possible different partners to which they can be associated.



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212−16229

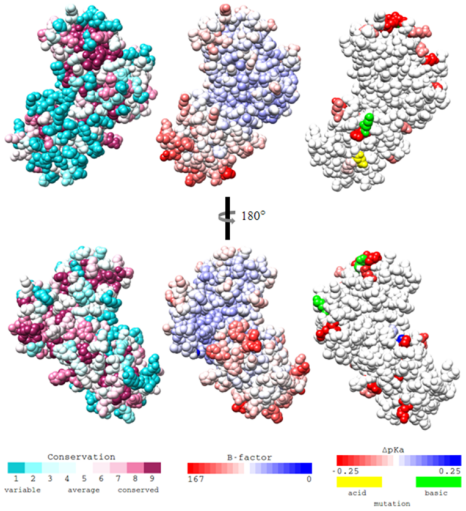



**Figure 9.** Tridimensional structures of the $NS1_{BR}$ monomer. Models from the left to the right show the degree of conservation, *B*-factor of each residue, and p$K_a$ variation regarding the reference ($NS1_{UG}$). See the text for details.

In Figure 10, the focus is on the comparison between the p$K_a$ shifts and the biological functional interfaces. For the sake

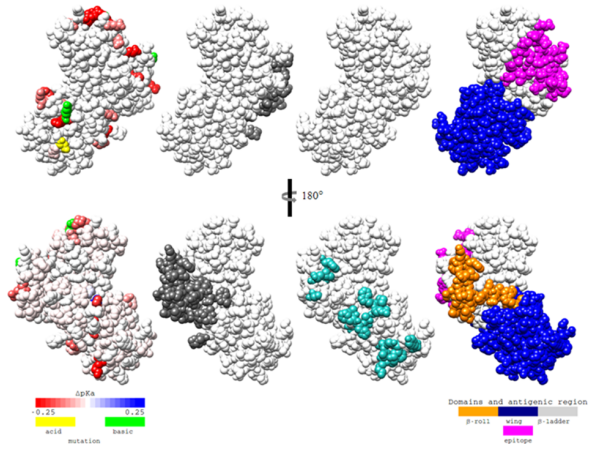



**Figure 10.** Tridimensional structures of the $NS1_{BR}$ monomer. Models from the left to right show p$K_a$ variation regarding the reference ($NS1_{UG}$), residues of the dimeric interface, residues embedded in the membrane, and domains of the NS1. Epitope region from NS1 of WNV (PDB id 4OII) is colored violet in the last image.

of convenience, the same molecular map of the $\Delta pK_a$'s previously seen in Figure 9 is reproduced again in this figure. From the left to right in this figure, we start with the p$K_a$ shifts color mapping, followed by the molecular surface representation of the dimeric interface and the residues that are embedded in the membrane. The last image on the right-hand side described the three distinct domains ($\beta$-roll, wing domain, and $\beta$-ladder domain) that are structurally important for NS1 together with the molecular regions that interact with the antibody. There are p$K_a$ shifts in all of these domains. In general, most of the p$K_a$ shifts occur at the dimeric interfaces and epitope. There are only a few amino acids affected at the region that directly interacts with the membrane: two with negative shifts, D30 ($\Delta pK_a \cong -0.3$) and K116 ($\Delta pK_a \cong -0.3$), and one with positive shift, R29 ($\Delta pK_a \cong 0.3$). It might be not surprising since it has been put forward that the interaction with the membrane would be driven by a hydrophobic mechanism at least for the WNV.[80] Nevertheless, such shifts explain the reduction in the free energy of interaction between NS1 and the membrane, as discussed in Figure 5. We have to bear in mind that electrostatic interactions are long-range interactions, and even charges at larger separation distances can contribute with the attraction between NS1 and the membrane. Thus, the presence of some titratable groups at the interface reinforces the evidence that electrostatic interactions are important to favor the approach of NS1 and the membrane, while the hydrophobic ones guide the final step of the anchoring of the NS1 into the membrane. The two mechanisms might work together for this biointerface. From the UG to the BR strain, most titratable residues experienced negative shifts both at the functionally important regions of the protein and in others without an apparent direct functional relation. We can hypothesize that these residues either contribute to the structural stability of NS1 or indirectly help the titration mechanisms of the other ionizable groups due to their high electrostatic coupling. Residues D52, K146, K191, and H286 seem to be affected by the neighborhood (see also Figure 8). It is clear that even the small different compositions of amino acids affect the electrostatics properties of NS1, therefore allowing it to interact in multiple ways depending on the strains.

Some earlier studies reported more negative electrostatic potential profiles at the loop surface on $NS1_{UG}$ of ZIKV compared to those on other flaviviruses.[17,28] This result is in agreement with our present outcomes. Furthermore, it seems as if the natural tendency of the future strains is to maintain more negatively charged interfaces while new geographical areas are conquered by the virus. It might be a strategy used by the ZIKV to trick the immune system. Focusing on the $NS1_{BR}$, we have listed the number of ionizable residues and the distribution of these residues in the three main functional interfaces in Table 3, with or without showing p$K_a$ shifts, as commented above. Among the titratable residues (117 in total), roughly a third is found in biological functional interface regions. Interestingly, more than 40% of residues with p$K_a$ shifts are found in these regions, which underlines the fine role of electrostatic features in key interactions. By the abundant number of GLUs present in the structure, it seems reasonable that the ZIKV has explored this feature. Although most GLU residues fall outside of the biological functional interfaces regions, we can observe how the presence of this amino acid is directly related to an electrostatically negative charges pattern in ZIKV reported by other studies.[17,28] It is important to point out that we have limited our analysis to known functional regions. This does not preclude the impact of these changes on interactions with putative other partners that still need to be identified. It should also be noted that some $\Delta pK_a$'s were positive as seen in residues E26 ($\Delta pK_a = 0.1$), R29 ($\Delta pK_a \cong 0.3$), E51 ($\Delta pK_a = 0.1$), and K189 ($\Delta pK_a = 0.2$), and mutations R69K ($\Delta pK_a = 1.7$) and E146K ($\Delta pK_a = 6.9$), as given in Table S4. In the antigenic region, the mutation Y286H exhibits a relatively high negative p$K_a$ shift ($\Delta pK_a = -3.2$). The highest positive shifts are observed for the mutations to LYS.

The results obtained to date suggest that the difference in the virulence is related to the p$K_a$ shifts in the three biological functional interfaces. For possible therapeutic interventions, the molecular interface between NS1 and the antibody is


DOI: 10.1021/acsomega.8b02081
*ACS Omega* 2018, 3, 16212−16229

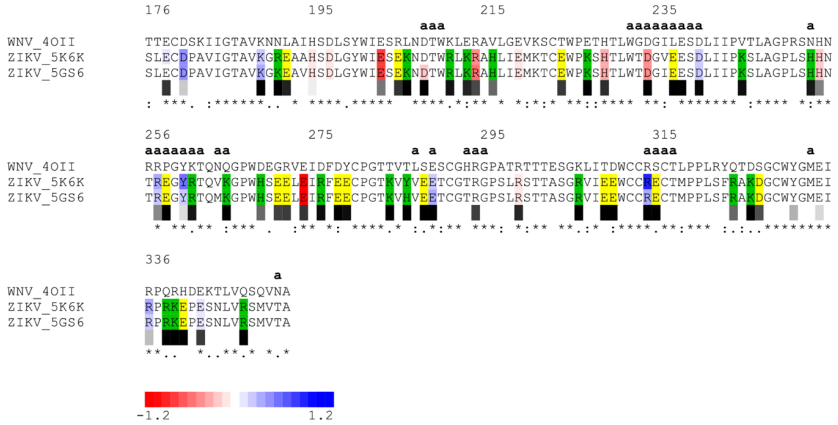



**Figure 11.** Alignment from different NS1 of two different flaviviruses. Calculations were done using the incomplete dimeric structure of WNV (PDB id 4OII) that interacts with the antibody (but removing this part). Incomplete structures for NS1 of ZIKV were used as well. Conventions are the same as in Figure 8.

highly interesting to be explored. Using the epitope region crystallographically known for WNV (only 185 residues from amino acids 176 to 352), we compared the $\Delta pK_a$ for WNV, $ZIKV_{UG}$, and $ZIKV_{BR}$ in Figure 11 (note that this alignment has a $pK_a$ shift scale from −1.2 to 1.2). The missing portion of the structure in the WNV−antibody complex was removed from the ZIKV strains structures too. All of these calculations were done in the apo form. Several $pK_a$ shifts are observed specific at the antigenic region (see cases with the "a"s at the top line of Figure 11). We can see that even conservative amino acids between the ZIKV and the WNV experienced $pK_a$ shifts (e.g., E178 and D180). There are cases where the shifts are larger for the $ZIKV_{BR}$ than for $ZIKV_{UG}$ and vice versa (e.g., E178). Nevertheless, the effective result of these mutations is to make the $ZIKV_{BR}$ more positively charged, as seen in Figure 6 and, consequently, less recognized by the antibody (see Figure 7). These regions affected by the protein sequence indicate a potential candidate for the development of vaccines.

## CONCLUSIONS

We have explored here the electrostatic properties of different strains of the ZIKV NS1 protein that are structurally conserved (in terms of the scaffold) but could have differential infectivities despite the small variations in the sequence level. This small variation in the sequences can have a considerable impact on the main physicochemical properties that are pH-dependent. This has been observed for global properties, such as the net protein charge numbers, and for more refined analyses based on $pK_a$'s and electrostatic maps. The mutations can affect neighboring amino acids and others that are spatially far but still electrostatically connected due to the long range of electrostatic interactions and the complex coupling between the titratable groups. Three different groups of strains could be identified in terms of their titration behavior and isoelectric points, the Uganda and the Central Africa Republic being the two extreme opposite cases. Most of the Asian strains tend to have a similar behavior.

The biomolecular interactions of the NS1 protein related to the three most important biological interfaces were investigated and demonstrated the differences between the strains that can be related to their distinct biological responses. For the self-association interaction, the UG strain is able to start its complexation at more acid pH regimes, whereas most of the studied strains need higher pH values. The differences between the strains increased from the monomeric to the dimeric states. Charge−charge interactions were confirmed as the main electrostatic-driven force for the complexation. Using the available structural data of the West Nile virus, we could demonstrate the difficulties that the same antibody would have to neutralize the NS1 from the Brazilian strain of the ZIKV. Another interesting observation is that it would have less difficulty to neutralize the strain from Uganda. In terms of the interaction with membranes, excluding the Uganda strain, most of the others have a similar behavior and can be attracted by the charged membrane until an intermediate pH value of ca. 8. All of these differences can be related to the differential clinical observed virulence. In fact, several $pK_a$ shifts were observed specifically at the three main studied biological interfaces reinforcing the idea that the NS1 electrostatic properties have a relation with the differential infectivity seen for this virus. It remains to be investigated how these observations are related to the hexamer transition, which was left for a future work. We have detected an important role for GLU residues and highlighted that several of its $pK_a$ shifts are found at the epitope. Furthermore, the antigenic region exhibits a negative $pK_a$ variation compared to the most ancestral strains. Since the NS1 protein has been proposed as a biomarker for flaviviruses and also as a potential vaccine target, the observed differences in these electrostatic properties maps indicate key issues for the development of antibody-based technologies and/or for the design of new medicines. The present study also opens up perspectives to computationally design high-specificity antibodies.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.8b02081.

Main physicochemical properties of different NS1 strains of ZIKV; dipole moment number of different NS1 strains of ZIKV (Figures S1 and S2); properties of coarse-grained amino acids; physicochemical properties of different ZIKV NS1 strains (monomeric state); physicochemical properties of different ZIKV NS1 strains (dimeric state); $\Delta\Delta pK_a$ calculated for two



DOI: 10.1021/acsomega.8b02081
ACS Omega 2018, 3, 16212−16229

models of NS1BR; and classification of the residues based on NS1BR (Tables S1−S5) (PDF)

## AUTHOR INFORMATION


### Corresponding Author

*E-mail: flbarroso@usp.br and flbarros@ncsu.edu.

### ORCID

Fernando L. Barroso da Silva: 0000-0003-2526-2085


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work has been supported in part by the "Fundação de Amparo à Pesquisa do Estado de São Paulo" [Fapesp 2015/16116-3 (F.L.B.d.S.)] and the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior"—CAPES (S.A.P.-C.). F.L.B.d.S. and S.A.P.-C. acknowledge the support of the University of São Paulo (USP) through the NAP-CatSinQ (Research Core in Catalysis and Chemical Synthesis) and the computing hours at Rice University through the international collaboration program with USP. F.L.B.d.S. and C.E. acknowledge the mobility support from the joint Universitè Sorbonne Paris Citè-USP Research program.

## ABBREVIATIONS

ZIKV, Zika virus; WNV, West Nile virus; NS1, nonstructural protein 1; FPTS, fast proton titration scheme; CG, coarse-grained; DVLO, Derjaguin−Landau−Verwey−Overbeek

## REFERENCES


(1) Cox, B. D.; Stanton, R. A.; Schinazi, R. F. Predicting Zika Virus: Structural Biology Challenges and Opportunities for Intervention. *Antiviral Chem. Chemother.* **2015**, *24*, 118−126.

(2) Muller, D. A.; Young, P. R. The Flavivirus NS1 Protein Molecular and Structural Biology, Immunology, Role in Pathogenesis and Application as a Diagnostic Biomarker. *Antiviral Res.* **2013**, *98*, 192−208.

(3) Lindenbach, B. D.; Thiel, H. J.; Rice, C. M. *Fields Virology*, 5th ed.; LR Publishers: Philadelphia, 2007.

(4) Mlakar, J.; Korva, M.; Tul, N.; Popović, M.; Poljšak-Prijatelj, M.; Mraz, J.; Kolenc, M.; Rus, K. R.; Vipotnik, T. V.; Vodušek, V. F.; et al. Zika Virus Associated with Microcephaly. *N. Engl. J. Med.* **2016**, *374*, 951−958.

(5) Petersen, E.; Wilson, M. E.; Touch, S.; McCloskey, B.; Mwaba, P.; Bates, M.; Dar, O.; Mattes, F.; Kidd, M.; Ippolito, G.; et al. Rapid Spread of Zika Virus in the Americas-Implications for Public Health Preparedness for Mass Gatherings at the 2016 Brazil Olympic Games. *Int. J. Infect. Dis.* **2016**, *44*, 11−15.

(6) Paixão, E. S.; Teixeira, M. G.; Rodrigues, L. C. Zika, Chikungunya and Dengue the Causes and Threats of New and Re-Emerging Arboviral Diseases. *BMJ Global Health* **2018**, *3*, No. e000530.

(7) Patterson, J.; Sammon, M.; Garg, M. Dengue, Zika and Chikungunya Emerging Arboviruses in the New World. *West. J. Emerg. Med.* **2016**, *17*, 671.

(8) Vasilakis, N.; Weaver, S. C. Flavivirus Transmission Focusing on Zika. *Curr. Opin. Virol.* **2017**, *22*, 30−35.

(9) Talero-Gutiérrez, C.; Rivera-Molina, A.; Pérez-Pavajeau, C.; Ossa-Ospina, I.; Santos-García, C.; Rojas-Anaya, M. C.; de-la-Torre, A. Zika Virus Epidemiology from Uganda to World Pandemic, an Update. *Epidemiol. Infect.* **2018**, *146*, 673−679.

(10) Beckham, J. D.; Pastula, D. M.; Massey, A.; Tyler, K. L. Zika Virus as an Emerging Global Pathogen Neurological Complications of Zika Virus. *JAMA Neurol.* **2016**, *73*, 875−879.

(11) Faye, O.; Freire, C. C. M.; Iamarino, A.; Faye, O.; de Oliveira, J. V. C.; Diallo, M.; Zanotto, P. M. A. Molecular Evolution of Zika Virus during Its Emergence in the 20th Century. *PLoS Neglected Trop. Dis.* **2014**, *8*, No. e2636.

(12) Kochakarn, T.; Kotanan, N.; Kümpornsin, K.; Loesbanluechai, D.; Thammasatta, M.; Auewarakul, P.; Wilairat, P.; Chookajorn, T. Comparative Genome Analysis between Southeast Asian and South American Zika Viruses. *Asian Pac. J. Trop. Med.* **2016**, *9*, 1048−1054.

(13) Shen, S.; Shi, J.; Wang, J.; Tang, S.; Wang, H.; Hu, Z.; Deng, F. Phylogenetic Analysis Revealed the Central Roles of Two African Countries in the Evolution and Worldwide Spread of Zika Virus. *Virol. Sin.* **2016**, *31*, 118−130.

(14) Simonin, Y.; van Riel, D.; de Perre, P.; Rockx, B.; Salinas, S. Differential Virulence between Asian and African Lineages of Zika Virus. *PLoS Neglected Trop. Dis.* **2017**, *11*, No. e0005821.

(15) Simonin, Y.; Loustalot, F.; Desmetz, C.; Foulongne, V.; Constant, O.; Fournier-Wirth, C.; Leon, F.; Molès, J.-P.; Goubaud, A.; Lemaitre, J.-M.; et al. Zika Virus Strains Potentially Display Different Infectious Profiles in Human Neural Cells. *EBioMedicine* **2016**, *12*, 161−169.

(16) Winkler, G.; Maxwell, S. E.; Ruemmler, C.; Stollar, V. Newly Synthesized Dengue-2 Virus Nonstructural Protein NS1 Is a Soluble Protein but Becomes Partially Hydrophobic and Membrane-Associated after Dimerization. *Virology* **1989**, *171*, 302−305.

(17) Brown, W. C.; Akey, D. L.; Konwerski, J. R.; Tarrasch, J. T.; Skiniotis, G.; Kuhn, R. J.; Smith, J. L. Extended Surface for Membrane Association in Zika Virus NS1 Structure. *Nat. Struct. Mol. Biol.* **2016**, *23*, 865−867.

(18) Xu, X.; Song, H.; Qi, J.; Liu, Y.; Wang, H.; Su, C.; Shi, Y.; Gao, G. F. Contribution of Intertwined Loop to Membrane Association Revealed by Zika Virus Full-Length NS1 Structure. *EMBO J.* **2016**, *35*, 2170−2178.

(19) Wang, D.; Chen, C.; Liu, S.; Zhou, H.; Yang, K.; Zhao, Q.; Ji, X.; Chen, C.; Xie, W.; Wang, Z.; et al. A Mutation Identified in Neonatal Microcephaly Destabilizes Zika Virus NS1 Assembly in Vitro. *Sci. Rep.* **2017**, *7*, No. 42580.

(20) Lindenbach, B. D.; Rice, C. M. Trans-Complementation of Yellow Fever Virus NS1 Reveals a Role in Early RNA Replication. *J. Virol.* **1997**, *71*, 9608−9617.

(21) Rastogi, M.; Sharma, N.; Singh, S. K. Flavivirus NS1 a Multifaceted Enigmatic Viral Protein. *Virol. J.* **2016**, *13*, 131.

(22) Chen, S.; Wu, Z.; Wang, M.; Cheng, A. Innate Immune Evasion Mediated by Flaviviridae Non-Structural Proteins. *Viruses* **2017**, *9*, 291.

(23) Conde, J. N.; Silva, E. M.; Barbosa, A. S.; Mohana-Borges, R. The Complement System in Flavivirus Infections. *Front. Microbiol.* **2017**, *8*, 213.

(24) Lefkowitz, E. J.; Dempsey, D. M.; Hendrickson, R. C.; Orton, R. J.; Siddell, S. G.; Smith, D. B. Virus Taxonomy: The Database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **2018**, *46*, D708−D717.

(25) Chaudhury, S.; Ripoll, D. R.; Wallqvist, A. Structure-Based pKa Prediction Provides a Thermodynamic Basis for the Role of Histidines in pH-Induced Conformational Transitions in Dengue Virus. *Biochem. Biophys. Rep.* **2015**, *4*, 375−385.

(26) Prakash, M. K.; Barducci, A.; Parrinello, M. Probing the Mechanism of pH-Induced Large-Scale Conformational Changes in Dengue Virus Envelope Protein Using Atomistic Simulations. *Biophys. J.* **2010**, *99*, 588−594.

(27) Barroso da Silva, F. L.; Dias, L. G. Development of Constant-pH Simulation Methods in Implicit Solvent and Applications in Biomolecular Systems. *Biophys. Rev.* **2017**, *9*, 699−728.

(28) Song, H.; Qi, J.; Haywood, J.; Shi, Y.; Gao, G. F. Zika Virus NS1 Structure Reveals Diversity of Electrostatic Surfaces among Flaviviruses. *Nat. Struct. Mol. Biol.* **2016**, *23*, 456−458.

(29) Falconar, A. K. I.; Young, P. R. Immunoaffinity Purification of Native Dimer Forms of the Flavivirus Non-Structural Glycoprotein, NS1. *J. Virol. Methods* **1990**, *30*, 323−332.




DOI: 10.1021/acsomega.8b02081
*ACS Omega* 2018, 3, 16212−16229

(30) Figueiredo, A.; Vieira, N. C. S.; Dos Santos, J. F.; Janegitz, B. C.; Aoki, S. M.; Junior, P. P.; Lovato, R. L.; Nogueira, M. L.; Zucolotto, V.; Guimaraes, F. E. G. Electrical Detection of Dengue Biomarker Using Egg Yolk Immunoglobulin as the Biological Recognition Element. *Sci. Rep.* **2015**, *5*, No. 7865.
(31) Cecchetto, J.; Fernandes, F. C. B.; Lopes, R.; Bueno, P. R. The Capacitive Sensing of NS1 Flavivirus Biomarker. *Biosens. Bioelectron.* **2017**, *87*, 949−956.
(32) Freire, M. C. L. C.; Pol-Fachin, L.; Coêlho, D. F.; Viana, I. F. T.; Magalhães, T.; Cordeiro, M. T.; Fischer, N.; Loeffler, F. F.; Jaenisch, T.; Franca, R. F.; et al. Mapping Putative B-Cell Zika Virus NS1 Epitopes Provides Molecular Basis for Anti-NS1 Antibody Discrimination between Zika and Dengue Viruses. *ACS Omega* **2017**, *2*, 3913−3920.
(33) Flamand, M.; Megret, F.; Mathieu, M.; Lepault, J.; Rey, F. A.; Deubel, V. Dengue Virus Type 1 Nonstructural Glycoprotein NS1 Is Secreted from Mammalian Cells as a Soluble Hexamer in a Glycosylation-Dependent Fashion. *J. Virol.* **1999**, *73*, 6104−6110.
(34) Gutsche, I.; Coulibaly, F.; Voss, J. E.; Salmon, J.; d'Alayer, J.; Ermonval, M.; Larquet, E.; Charneau, P.; Krey, T.; Mégret, F.; et al. Secreted Dengue Virus Nonstructural Protein NS1 Is an Atypical Barrel-Shaped High-Density Lipoprotein. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 8003−8008.
(35) Akey, D. L.; Brown, W. C.; Dutta, S.; Konwerski, J.; Jose, J.; Jurkiw, T. J.; DelProposto, J.; Ogata, C. M.; Skiniotis, G.; Kuhn, R. J.; et al. *Flavivirus* NS1 Structures Reveal Surfaces for Associations with Membranes and the Immune System. *Science* **2014**, *343*, 881−885.
(36) Zeidler, J. D.; Fernandes-Siqueira, L. O.; Barbosa, G. M.; Da Poian, A. T. Non-Canonical Roles of Dengue Virus Non-Structural Proteins. *Viruses* **2017**, *9*, 42.
(37) Winkler, G.; Randolph, V. B.; Cleaves, G. R.; Ryan, T. E.; Stollar, V. Evidence That the Mature Form of the *Flavivirus* Nonstructural Protein NS1 Is a Dimer. *Virology* **1988**, *162*, 187−196.
(38) Winkler, G.; Maxwell, S. E.; Ruemmler, C.; Stollar, V. Newly Synthesized Dengue-2 Virus Nonstructural Protein NS1 Is a Soluble Protein but Becomes Partially Hydrophobic and Membrane-Associated after Dimerization. *Virology* **1989**, *171*, 302−305.
(39) Darré, L.; Machado, M. R.; Brandner, A. F.; González, H. C.; Ferreira, S.; Pantano, S. SIRAH a Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics. *J. Chem. Theory Comput.* **2015**, *11*, 723−739.
(40) Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L. NCBI BLAST a Better Web Interface. *Nucleic Acids Res.* **2008**, *36*, W5−W9.
(41) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(42) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(43) Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M.; Pieper, U.; Sali, A. Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinf.* **2006**, *15*, 5.6.1−5.6.30.
(44) Edeling, M. A.; Diamond, M. S.; Fremont, D. H. Structural Basis of Flavivirus NS1 Assembly and Antibody Recognition. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 4285−4290.
(45) Barroso da Silva, F. L.; Pasquali, S.; Derreumaux, P.; Dias, L. G. Electrostatics Analysis of the Mutational and pH Effects of the N-Terminal Domain Self-Association of the Major Ampullate Spidroin. *Soft Matter* **2016**, *12*, 5600−5612.
(46) Teixeira, A. A. R.; Lund, M.; Barroso da Silva, F. L. Fast Proton Titration Scheme for Multiscale Modeling of Protein Solutions. *J. Chem. Theory Comput.* **2010**, *6*, 3259−3266.
(47) Barroso da Silva, F. L.; MacKernan, D. Benchmarking a Fast Proton Titration Scheme in Implicit Solvent for Biomolecular Simulations. *J. Chem. Theory Comput.* **2017**, *13*, 2915−2929.
(48) Nozaki, Y.; Tanford, C. [84] Examination of Titration Behavior. *Methods Enzymol.* **1967**, *11*, 715−734.
(49) Lund, M.; Trulsson, M.; Persson, B. Faunus An Object Oriented Framework for Molecular Simulation. *Source Code Biol. Med.* **2008**, *3*, 1.
(50) Barroso da Silva, F. L.; Derreumaux, P.; Pasquali, S. Fast Coarse-Grained Model for RNA Titration. *J. Chem. Phys.* **2017**, *146*, No. 035101.
(51) Božič, A. L.; Podgornik, R. pH Dependence of Charge Multipole Moments in Proteins. *Biophys. J.* **2017**, *113*, 1454−1465.
(52) Assadi-Porter, F. M.; Fillingame, R. H. Proton-Translocating Carboxyl of Subunit *c* of $F_1F_o$ $H^+$-ATP Synthase the Unique Environment Suggested by the pKa Determined by $^1$H NMR. *Biochemistry* **1995**, *34*, 16186−16193.
(53) Kaplan, H.; Stevenson, K. J.; Hartley, B. S. Competitive Labelling, a Method for Determining the Reactivity of Individual Groups in Proteins. The Amino Groups of Porcine Elastase. *Biochem. J.* **1971**, *124*, 289−299.
(54) Edgar, R. C. MUSCLE a Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinf.* **2004**, *5*, 113.
(55) Ashkenazy, H.; Abadi, S.; Martz, E.; Chay, O.; Mayrose, I.; Pupko, T.; Ben-Tal, N. ConSurf 2016 an Improved Methodology to Estimate and Visualize Evolutionary Conservation in Macromolecules. *Nucleic Acids Res.* **2016**, *44*, W344−W350.
(56) Barroso da Silva, F. L.; Boström, M.; Persson, C. Effect of Charge Regulation and Ion−Dipole Interactions on the Selectivity of Protein−Nanoparticle Binding. *Langmuir* **2014**, *30*, 4078−4083.
(57) Delboni, L. A.; Barroso da Silva, F. L. On the Complexation of Whey Proteins. *Food Hydrocolloids* **2016**, *55*, 89−99.
(58) Chen, C. R.; Makhatadze, G. I. ProteinVolume: Calculating Molecular van der Waals and Void Volumes in Proteins. *BMC Bioinf.* **2015**, *16*, 101.
(59) Phillies, G. D. J. Excess Chemical Potential of Dilute Solutions of Spherical Polyelectrolytes. *J. Chem. Phys.* **1974**, *60*, 2721−2731.
(60) Striolo, A.; Bratko, D.; Wu, J. Z.; Elvassore, N.; Blanch, H. W.; Prausnitz, J. M. Forces between Aqueous Nonuniformly Charged Colloids from Molecular Simulation. *J. Chem. Phys.* **2002**, *116*, 7733−7743.
(61) Tavares, F. W.; Bratko, D.; Striolo, A.; Blanch, H. W.; Prausnitz, J. M. Phase Behavior of Aqueous Solutions Containing Dipolar Proteins from Second-Order Perturbation Theory. *J. Chem. Phys.* **2004**, *120*, 9859−9869.
(62) Kjellander, R.; Ramirez, R. Yukawa Multipole Electrostatics and Nontrivial Coupling between Electrostatic and Dispersion Interactions in Electrolytes. *J. Phys.: Condens. Matter* **2008**, *20*, No. 494209.
(63) Božič, A. L.; Podgornik, R. Anomalous Multipole Expansion: Charge Regulation of Patchy Inhomogeneously Charged Spherical Particles. *J. Chem. Phys.* **2018**, *149*, No. 163307.
(64) Srivastava, D.; Santiso, E.; Gubbins, K.; Barroso da Silva, F. L. Computationally Mapping pKa Shifts Due to the Presence of a Polyelectrolyte Chain around Whey Proteins. *Langmuir* **2017**, *33*, 11417−11428.
(65) Lund, M. *Electrostatic Interactions In and Between Biomolecules*; University of Lund, 2006.
(66) Jönsson, B.; Lund, M.; Barroso da Silva, F. L. In *Electrostatics in Macromolecular Solutions*, Conference on Food Colloids 2006; 2007; Vol. *302*, pp 129−154.
(67) Lund, M.; Åkesson, T.; Jönsson, B. Enhanced Protein Adsorption Due to Charge Regulation. *Langmuir* **2005**, *21*, 8385−8388.
(68) Burdakov, D.; Petersen, O. H.; Verkhratsky, A. Intraluminal Calcium as a Primary Regulator of Endoplasmic Reticulum Function. *Cell Calcium* **2005**, *38*, 303−310.
(69) Krissinel, E.; Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **2007**, *372*, 774−797.
(70) Lomize, M. A.; Pogozheva, I. D.; Joo, H.; Mosberg, H. I.; Lomize, A. L. OPM Database and PPM Web Server Resources for Positioning of Proteins in Membranes. *Nucleic Acids Res.* **2012**, *40*, D370−D376.


DOI: 10.1021/acsomega.8b02081
*ACS Omega* 2018, 3, 16212−16229

(71) Rostkowski, M.; Olsson, M. H. M.; Søndergaard, C. R.; Jensen, J. H. Graphical Analysis of PH-Dependent Properties of Proteins Predicted Using PROPKA. *BMC Struct. Biol.* **2011**, *11*, 6.
(72) Hernando-Pérez, M.; Cartagena-Rivera, A. X.; Božič, A. L.; Carrillo, P. J. P.; Martín, C. S.; Mateu, M. G.; Raman, A.; Podgornik, R.; de Pablo, P. J. Quantitative Nanoscale Electrostatics of Viruses. *Nanoscale* **2015**, *7*, 17289−17298.
(73) Knipe, D. M.; Howley, P. *Fields Virology*, 6th ed.; Lippincott Williams and Wilkins, 2014.
(74) Barroso da Silva, F. L.; Jönsson, B. Polyelectrolyte−Protein Complexation Driven by Charge Regulation. *Soft Matter* **2009**, *5*, 2862−2868.
(75) Yu, I.-M.; Zhang, W.; Holdaway, H. A.; Li, L.; Kostyuchenko, V. A.; Chipman, P. R.; Kuhn, R. J.; Rossmann, M. G.; Chen, J. Structure of the Immature Dengue Virus at Low PH Primes Proteolytic Maturation. *Science* **2008**, *319*, 1834−1837.
(76) McQuarrie, D. *Statistical Mechanics*; Harper and Row, 1976.
(77) Lund, M.; Jönsson, B. A Mesoscopic Model for Protein-Protein Interactions in Solution. *Biophys. J.* **2003**, *85*, 2940−2947.
(78) Rocha, L. B.; dos Santos Alves, R. P.; Caetano, B. A.; Pereira, L. R.; Mitsunari, T.; Amorim, J. H.; Polatto, J. M.; Botosso, V. F.; Gallina, N. M. F.; Palacios, R.; et al. Epitope Sequences in Dengue Virus NS1 Protein Identified by Monoclonal Antibodies. *Antibodies* **2017**, *6*, 14.
(79) Chothia, C.; Lesk, A. M. The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J.* **1986**, *5*, 823−826.
(80) Youn, S.; Li, T.; McCune, B. T.; Edeling, M. A.; Fremont, D. H.; Cristea, I. M.; Diamond, M. S. Evidence for a Genetic and Physical Interaction between Nonstructural Proteins NS1 and NS4B That Modulates Replication of West Nile Virus. *J. Virol.* **2012**, *86*, 7360−7371.
(81) De Carvalho, S. J.; Ghiotto, R. C. T.; Barroso da Silva, F. L. Monte Carlo and Modified Tanford−Kirkwood Results for Macromolecular Electrostatics Calculations. *J. Phys. Chem. B* **2006**, *110* (17), 8832−8839.
(82) Duran, N. M.; Spelzini, D.; Wayllace, N.; Boeris, V.; Barroso da Silva, F. L. A combined experimental and molecular simulation study of factors influencing interaction of quinoa proteins−carrageenan. *Int. J. Biol. Macromol.* **2018**, *107*, 949−956.



DOI: 10.1021/acsomega.8b02081
*ACS Omega* 2018, 3, 16212−16229

## Chapter 6

# Electrostatic epitopes prediction through PROCEEDp$K_a$

From the results of Chapter 5, it was determined that p$K_a$'s in putative epitope regions of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ may play an important role in the differential recognition of Abs. Thus, for this part of the thesis, we wanted to establish whether the p$K_a$ can be a key electrostatic characteristic to predict antigenic regions. For this, the PROCEEDp$K_a$ method was developed. PROCEEDp$K_a$ was also compared with other tools available on the web for the prediction of B-cell epitopes. For the validation process of PROCEEDp$K_a$ it was used the known NS1-Ab heterocomplex of WNV (PDB id 4OII).

Once the PROCEEDp$K_a$ method was validated, we characterized the EE's in $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$.

The main question of this part of the study is:

*Do the proteins $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ present different EE's?*

A summary of the molecular model used, the method to solve it, and the measures calculated appear in Table 6.1.

Table 6.1: Molecular modeling and measurements carried out in Chapter 6.

| Level of detail | Solution | Measurements |
|---|---|---|
| CG model at the amino acid level | CpH MC simulations | p$K_a$, free energies |

Link paper: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.9b00895.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2020, 60, 944−963

pubs.acs.org/jcim

# Identification of Electrostatic Epitopes in Flavivirus by Computer Simulations: The PROCEEDpKa Method

Sergio A. Poveda-Cuevas,[†,‡,#] Catherine Etchebest,[§,∥,⊥,#] and Fernando L. Barroso da Silva[*,†,‡,#,¶]

[†]Universidade de São Paulo, Programa Interunidades em Bioinformática, Rua do Matão, 1010, BR, 05508-090 São Paulo, São Paulo, Brazil

[‡]Universidade de São Paulo, Departamento de Ciências Biomoleculares, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Av. Café, s/no−Campus da USP, BR, 14040-903 Ribeirão Preto, São Paulo, Brazil

[§]Université de Paris, Biologie Intégrée du Globule Rouge, UMR_S1134, BIGR, INSERM, F-75015 Paris, France

[∥]Equipe 2, Dynamique des Structures et des Interactions Moléculaires, Université Paris Diderot−Paris 7, INTS, 6 Rue Alexandre Cabanel, 75015 Paris, France

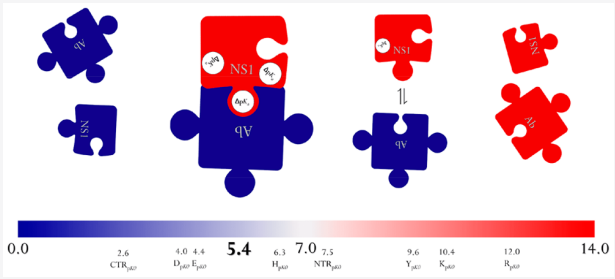
[⊥]Laboratoire d'Excellence GR-Ex, Paris, France

[#]University of São Paulo-Université Sorbonne Paris Cité International Laboratory in Structural Bioinformatics, Av. do Café, s/no−FCFRP, Bloco B, BR, 14040-903 Ribeirão Preto, São Paulo, Brazil

[¶]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States

*S* *Supporting Information*

**ABSTRACT:** Viruses are enthusiastically studied due to the great impact that these organisms can have on human health. Computational approaches can contribute offering tools that can shed light on important molecular mechanisms that help to design new diagnostic procedures. Several cellular processes between the immune-host system and the pathogenic organism are dependent on specific intermolecular interactions. In this study, we evaluated theoretical approaches to understand some properties of the antigen−antibody interactions considering the titratable properties of all ionizable residues of the nonstructural viral protein 1 (NS1) of the West Nile virus (WNV) and the Zika virus (ZIKV). Constant-pH Monte Carlo simulations were performed to estimate electrostatic properties such as the p$K_a$ shifts ($\Delta pK_a$). We proposed an alternative criterion for the discrimination of antigenic residues based on $\Delta pK_a$s. Our outcomes were analyzed by an evaluation of the sensitivity and specificity through a receiver operating characteristic (ROC). As a starting point, we used the known crystallographic structure for the complex of $NS1_{WNV(176-352)}$ and the specific antibody 22NS1 (PDB ID 4OII) to differentiate the residues belonging to that interface. With an optimal threshold for the absolute value of the p$K_a$ shifts, we found that is possible to predict antigenic epitopes reproducing the interfaces as defined by the X-ray structure. After this validation, we evaluated theoretical predictions based on protein−protein (PP) complexation simulations. From them, we observe amino acids with an antigenic potential and defined the optimum threshold that was applied for two strains of ZIKV (i.e., Uganda and Brazil). Several ionizable residues with antigenic capacity were identified. This is favorably related to some studies that show the high immunogenicity of secreted NS1. This approach opens up an important discussion about what are termed here "electrostatic epitopes" and how they work as an important reference in the paratope−epitope interaction for viral systems.




## INTRODUCTION

Studies in *flaviviruses* have received special global attention in recent decades due to the diseases that these kinds of viruses can elicit, especially in different tropical and subtropical regions of the world.[1−3] Among the members of this viral genus, we found the Dengue virus (DENV), West Nile virus (WNV), Japanese encephalitis virus (JEV), Yellow fever virus (YFV), Zika virus (ZIKV), and others.[4] This is a topic of intense research and special interest for Latin American countries due to the recent outbreak of ZIKV especially in Brazil and the repeated occurrence of seasons with a high number of cases.[5−13] Despite the diversity that these species in the genus can have, the structure of their genomes is very similar. Flaviviruses are single-stranded RNA of 10.7 kilobases

**Special Issue:** Molecular Simulation in Latin America: Coming of Age

**Received:** September 30, 2019
**Published:** November 27, 2019

Downloaded via UNIV OF SAO PAULO on March 2, 2020 at 18:15:00 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.


© 2019 American Chemical Society

DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

with positive-sense that translate to a polyprotein. Virus- and host-encoded proteases manage the polyprotein generating three structural (C, prM, and Env) and seven nonstructural (NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5) biomolecules in the lumen of the endoplasmic reticulum.[4,14,15] Within infected cells, NS1 plays a multifunctional and antagonistic role. For instance, in the cytosol, the dimeric NS1 integrated into a membrane (mNS1) works as a cofactor of the replication complex. However, nonclassical secretion pathways are also used by the NS1 (sNS1), allowing it to interact with the immune system in the extracellular space. Because NS1 protein can be secreted independently of virions, it has been proposed as an important candidate for the detection, treatment, and discrimination of flaviviruses.[15] NS1 works as an efficient biomarker because it is secreted at high levels (an infected patient may have up to 50 $\mu$g/mL).[16] It has also been shown *in vivo* that immunization can be elicited by the administration of monoclonal antibodies (mAb) anti-NS1. This raises an important point about how host proteins operate against a pathogen such as NS1 and what intermolecular interactions are the base for these molecular mechanisms. Therefore, we focused the present study on NS1.

The recognition processes between biomolecules involve different physical intermolecular interactions. Among them, electrostatic interactions are playing a key role in contributing significantly to the binding specificity or the complex stability as exemplified in different studies, e.g., refs 17−21. The molecular interface antibody−antigen has a peculiar electrostatic pattern (richer in titratable groups) that is different than other general PP interfaces.[22] This also indicates that the canonical view of the "lock and key" concept which assumes short-range interactions (with a clear focus on the protein surface) as the more relevant ones urge to be reviewed due to the long-range nature of electrostatic interactions.

Despite the fact that antibody−antigen recognition is an intensive research field since immunology reactions are of major importance in health care, electrostatic effects for this important biological interface have not been largely explored. Specificity for antibody−antigen interactions is a topic that can be even more complex (see, for instance, refs 10,20−28). A good paper is the review work done by Regenmortel,[29] where he started saying that "the concept of antibody specificity is widely used in immunology, although it has rarely been defined satisfactorily". He also explains the concept of "antibody polyspecificity" (as the "ability of one antibody to bind a large variety of epitopes in different antigens"). Also in these cases, electrostatics is playing a key role as it was very recently shown by Yoshida et al.[30] who demonstrated how it was possible to recover antigen−antibody binding by designing variants that compensate electrostatic interactions. Interestingly, a similar conclusion was obtained a few years before by Yoshida et al.[30] who also highlighted the importance, in the specificity process, of the *transition state* along the reaction pathway.

Additionally, Ab dependent enhancement due to subneutralizing cross-reactive antibodies generated by previous flavivirus infections (mainly Dengue infection) also plays a role in the debilitating consequences of this flavivirus infection.[31] This scenario is especially complicated due to the frequent subclinical manifestations of both viruses;[32,33] the high similarity of symptoms, which often leads to misdiagnosis; and the possibility of transfusion and sexual transmissions of ZIKV.[34,35] The differential diagnostic between DENV infection and ZIKV infection is performed by detecting the viral genome in bodily fluids through RT-PCR (reverse transcriptase-polymerase chain reaction) and PRNT (plaque reduction neutralization test), which prove to be expensive and labor-intensive for routine diagnostics.[36−38] Therefore, there is an urgent need to decipher its molecular basis and to develop reliable and low-cost diagnostic tools to discriminate between both infections and allow for an optimized allocation of public health resources and better management of patient care.

In this challenging field, three main questions are generally addressed with *in silico* methods. The first one concerns the location of the epitope on a protein targeted by a given antibody (Ab), the second relates on finding how the two proteins interact, and third, once knowing a complex how to optimize the complex in terms of stability and specificity. The first two points have been addressed for a long time and are mainly based on docking procedures associated with more or less elaborated scoring functions. Antigen−antibody recognition is frequently the system of choice to elaborate and validate new docking and scoring procedures. In this case, the epitope region, i.e., the antigen-binding region, needs to be identified, while Ab complementary determining regions (CDR, L1, L2, and L3), the most famous key regions for binding, are generally targeted. The last point, which is the most recent one, is also the most challenging but with a high biomedical impact. It requires to search for alternative Ab sequences with an enhanced stability and a highest specificity for a given target epitope, avoiding or favoring cross-reactivity. In this aim, a number of different softwares has been developed (e.g., refs 39−42 and others cited below). Although it is often argued that they are more and more efficient, they are based on similar hypotheses and *per se* they share common features. Indeed, most of them focus on Ab complementary determining regions (CDR, L1, L2, and L3) and commonly take advantage of sequences, position-sequence-substitution matrices (PSSM), and structures information deposited in databases. The main differences between them rely on the way the data are used, classified, and eventually exploited to train machine-learning procedures like in a very recent example that uses deep-learning methods,[43,44] but the results were not as successful as expected. This information is generally combined with docking approaches associated with different energy functions (physical-based or statistical-based), large sampling procedure in the sequence and structure space, clustering of a large set of candidates and ranking of these candidates (for a review, see refs 45−47). The most popular methods are, e.g., PIGS,[48] SabPred,[49] RosettaAntibody,[50] AbodyBuilder,[51] OptCDR and OptMAVEN,[52,53] AbDesign,[54] or RosettaAntibodyDesign,[55] or mCSM-AB.[56] Interestingly, Nimrod et al.[57] have proposed a strategy that is slightly different and closer to the one we present in this work. Actually, rather than optimizing a full 3D complex, the method consists in focusing on optimal residue−residue interactions between an epitope zone and any Ab showing complementary features like shape or charge with respect to the epitope surface. This seems to be a promising approach to enlarge the number of new designed antibodies (Abs). A related idea has been the basis of a very recent and interesting study conducted by Jesperen et al.[58]

Overall, despite some relative success, improvements are still needed. For instance, epitope prediction bioinformatics tools (e.g., ElliPro,[39] SEPPA,[40] DiscoTope,[42] EPCES,[59] etc.; see refs 60 and 61 for a review) are based on the available structural data obtained at *very specific* physical-chemical experimental conditions which might even be quite different than the *in vivo*



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

ones. The main output information is the potentially antigenic propensity of the amino acids which is *not* precise for a given Ab. Although docking procedures can deal with a specific Ab and are very efficient to reproduce known crystallographic macromolecular complexes, they suffer from similar premises ignoring that physical-chemical parameters as the pH and ionic strength can drastically affect the complexation process. Often the electrostatic contributions are diminished in these approaches because the experimental data used to calibrate the models is usually obtained from crystallization where the conditions for proteins are usually very harsh with high salt concentrations leading to strong electrostatic screening. Under these crystallization conditions, the influence of electrostatic interactions is expected to be only marginal and far from the physiological environment. Moreover, focusing on the binding site region, it is assumed that they should always form a (stable) complex, and it also ignored the whole association process, i.e., the pathway for the two molecules to come together and eventually associate. This pathway can contain energetic barriers and even other minima that would prevent the association in real conditions.

Accordingly, in the present work, we choose to go further assuming three main hypotheses. The first one is that electrostatic properties are of major importance in driving interactions between macromolecules. Indeed, they strongly depend on environmental conditions like pH or salt concentration that may significantly vary in different biological cellular compartments or in biocontainers. Despite their well-recognized role, these fundamental interactions are often neglected in large complexes exploration or at least over-simplified, because their calculations are time-consuming when environmental conditions are accounted for.[62] It is, in particular, the case when more detailed and sophisticated methodologies, like molecular dynamics simulations in the presence of an explicit solvent, are employed.[63,64] Frequently, they neglect a complete description of the pH effects and/or suffer from slow convergence,[62,64−66] which might be critical for an epitope−paratope interface that has more titratable groups than other PP interfaces. Importantly, since electrostatic interactions prevail at a long-range distance, they can definitely drive complexation processes and govern the final complex state. Titratable amino acids that are at the interior of the proteins can also contribute to these interactions due to the electrical coupling between the ionizable sites. Hence, focusing only on the final stage when putative pairing partners are supposed to be in close contact without considering the full conformational pathway would lead to misleading results. The second hypothesis assumes that flexibility/dynamics can be omitted in the preliminary study as they should have a strong effect only at short separation distances. In fact, most of the bioinformatic tools assume the same premise.[39,40,60] Elucidation of the dynamical pathways can be of great interest to identify key positions that might be further targeted to improve the stability of the complex or the specificity of the recognition process after this initial investigation. Finally, in the case of proteins, the last hypothesis supposes that in different cases, the sequence-structure relationship is robust enough to introduce mutations or modifications while preserving the overall fold of the partners.

As one support to these hypotheses, we have shown in a previous study[11] that this interface epitope−paratope for the ZIKV NS1 has characteristic electrostatic features depending on the strain. This initial work indicated the possibility to map key amino acids that could lead to the development of new biomolecular diagnostics for flavivirus capable to differentiate the ZIKV from others.

In this present work, we took a closer look at key residues involved in the host−pathogen PP interaction of two flaviviruses (WNV and ZIKV) controlled by pH. These amino acids were identified during fast constant-pH Monte Carlo simulations specially devised by our lab[62,65,67] for biomolecules and also by some canonical bioinformatic tools.[39−42] All discussion is based on a physical-chemical framework where the p$K_a$s of titratable groups can be shifted by the presence of other charged objects at their closer vicinity.[68,69,86] We have also limited the present study to only one Ab (the Fab fragment 22NS1). This Ab is experimentally found solely complexed with the incomplete $NS1_{176-352}$ structure of WNV. As such, the theoretical tests with ZIKV NS1 are waiting for validation by laboratory experiments.

This paper is organized as follows. In the first section, we present the theoretical foundation including the molecular models and the simulation details. Next, we demonstrate how the p$K_a$s shifts can be a powerful tool to probe the host−pathogen interactions as *a proof-of-concept*. Within this conceptual framework, we restricted our work here to a single and specific Ab. We also demonstrate how the thermodynamic pathway of the biomolecular interactions can be used to recover the proper biological interface of the epitope−paratope, and to define the "electrostatic epitopes" (EE) concept in this broader definition that we would like to bring into the discussion. The predictive properties of the available servers for the flavivirus antigenic determinant are examined as well. In the last section, we compare the EE for WNV and two strains of the ZIKV, the Brazilian and Ugandan, to discuss their possible cross-reactivity.

## THEORETICAL METHODS

Computational methods have been widely used to probe the thermodynamic, dynamics, and interactive properties of proteins and nucleic acids in general.[4,1314,17−21,62,63,93] They are also playing an important role in modern Virology shedding light on the understanding of molecular mechanisms and predicting possible antigenic sites.[4,11−13] Here, we combine canonical structural bioinformatics tools and a fast constant-pH Monte Carlo (MC) scheme for protein−protein studies to evaluate their potential for the identification of key amino acids for the host−pathogen interaction in flaviviruses.

**Molecular Systems and Their Structural Modeling.** An experimental referential system is necessary to test and validate the theoretical predictions. Experimental characterization of antibodies against NS1 flaviviruses is often done by enzyme-linked immunosorbent assay (ELISA) without providing the structural information on the hybridomas.[70,71] This makes it difficult to use their outcomes for the validation of the theoretical methods. Conversely, X-ray crystallographic studies of proteins complexed with Abs have been of great value to enhance our knowledge of antigenic specificity from an experimental point of view.[72−74] The unique crystallographic data available at the RCSB Protein Data Bank (PDB) for a flavivirus NS1 protein complexed with an Ab concern the incomplete structure of WNV with the antibody Fab fragment 22NS1 obtained at a solution pH 5.4.[72] In the X-ray structure, the $NS1_{WNV}$ protein exists as a dimer formed by the interaction between two chains, labeled A and B, respectively. Residues 1−175, which corresponds to the $\beta$-roll and wing domains of the



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

NS1 protein, are missing in this X-ray structure for both chains. This indicates that these absent amino acids are dispensable for the complexation $NS1_{WNV}$-Fab fragment 22NS1 at least at the physical chemical environment given by the crystallographic conditions. Figure 1 shows this structure emphasizing the epitope−paratope interface and the ionizable amino acids.

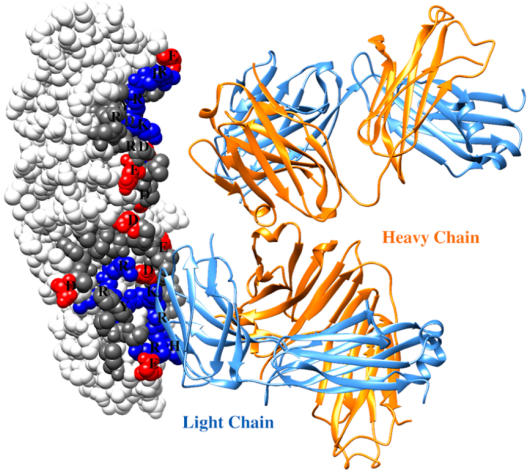



**Figure 1.** Crystal structure of $NS1_{WNV}$−Ab complex (PDB ID 4OII). The incomplete $NS1_{WNV}$ protein (missing the N-terminal domain) is shown in a molecular representation using spheres for its atoms. Atoms that are part of the epitopes are red (acid amino acids), blue (base amino acids), and gray (nontitrating amino acids) colored. The antibody Fab fragment 22NS1 is shown as a ribbon. For a better visualization of the interface, the Ab structure was translated as ∼12 Å.

Besides this experimental reference system, we include the prediction of possible epitopes for ZIKV NS1. The most interesting ZIKV strains for this investigation were defined based on our previous study.[11] Considering the evolutive features, the Ugandan (African lineage) and Brazilian (Asian lineage) strains were assumed to be the two extremes and most important cases in terms of health impact supported by the literature.[75,76] Crystal structures coordinates are available for them with PDB IDs 5K6K and 5GS6,[77,78] respectively. These two structures are quite similar except for a few structural differences that are observed in some loop regions (1.9 Å RMSD across all 352 pairs and 0.706 Å RMSD between 310 pruned atom pairs). The PDB files were edited before the calculations. All water molecules and heteroatoms were removed. We will refer to these completed three-dimensional structures in this text with the label "BR" and "UG", respectively, for the Brazilian and Ugandan strains. Additional calculations were performed removing residues 1−175 from these completed ZIKV structures to compare with the $NS1_{WNV}$ that lacks this N-terminal domain in the crystallographic data. The subscript "176−352" was added to their labels in the next sections to indicate when the incomplete structure was used.

In order to evaluate the effect of the missing residues in $NS1_{WNV(176-352)}$, the more complete PDB ID 4O6D structure[1] was used. Missing regions (108−128 chains A and B) structures were built using the "UCSF Chimera 1.11.2" interface of the program "Modeller" with default parameters.[79] The PDB IDs 5GS6 and 4O6D were used as structural templates.

**Coarse-Grained Modeling.** Cost-effective coarse-grained models are very convenient and appropriate to be applied in the present study because they offer the possibility to explore the main physical features of a system with a reduced number of parameters and lower computational costs. This allows us to repeat the calculations on several different experimental conditions. It gives us the opportunity to work at an intermediate theoretical level between the quite expensive atomistic molecular simulation schemes and common bioinformatic approaches with the advantage of being able to fully consider the pH effects. Therefore, the proteins (NS1 and the antibody Fab fragment 22NS1) were modeled here as rigid bodies (i.e., bond lengths, angles, and dihedral angles are kept fixed) in an amino acid coarse-grained (CG) level of detail according to the three-dimensional structures provided by the X-ray structure (for the incomplete $NS1_{WNV(176-352)}$ and the Fab fragment 22NS1) and the comparative modeling (for $NS1_{WNV}$, $NS1_{ZIKV-BR}$, $NS1_{ZIKV-UG}$, $NS1_{ZIKV-BR(176-352)}$, and $NS1_{ZIKV-UG(176-352)}$) as described above. Following previously successful studies,[17,18,20,69,80,81] each protein atomistic coordinate is transformed into a collection of charged Lennard-Jones (LJ) spheres of radii ($R_i$) and valence $z_i$ mimicking amino acids. Amino acid sizes were taken from ref 20. The valences of all ionizable residues were assigned by the fast proton titration scheme (FPTS)[65,67] according to the solution pH. Details of this titration scheme, its p$K_a$ benchmarks, advantages, and the discussions related to its approximations can be found elsewhere.[62,65,67,82] We note that this CG model focuses on the electrostatic features of the biomolecular system and has been predicting p$K_a$s with good accuracy at low computational costs.[65,82] This is highly important in the present study due to the elevated number of ionizable residues in the NS1 proteins (216 titratable residues for $NS1_{WNV}$) which increases the electrostatic coupling between them. As in the original FPTS paper,[67] the intrinsic acid dissociation constants (p$K_0$) were taken from the experimental work of Nozaki and Tanford.[83] Details as the number of each amino acid in the $NS1_{ZIKV}$ structures and their charge numbers in the deprotonated and protonated states are listed in ref 11.

The same CG model was employed for the titration calculations both for macromolecules at the crystallographic conditions (as given by the PDB coordinates) and for them during the complexation process (as given by the simulations where the two macromolecules can translate, rotate, and titrate inside the simulation box). Figure 2 gives a scheme of this CG model for the complexation studies. Each protein (NS1 and the Ab) is placed inside an electroneutral open cylindrical cell of radius $r_{cyl}$ and height $l_{cyl}$ with a static dielectric constant of 78.7 (assuming a temperature of 298 K). One additional advantage of cell models as this one is that all electrostatic interactions within the cell can be exactly treated in the sense that no cutoff scheme or potential truncation need to be included.[62] Counterions and added salt particles were represented implicitly using a screening term, i.e., for two titratable residues $i$ and $j$, the screening is given by $[\exp(-\kappa r_{ij})]$, where $\kappa$ is the modified inverse Debye length, and $r_{ij}$ is the interparticle separation distance. Although excellent agreement was found with explicitly salt models,[66] this implicit salt scheme is more suitable for the description of monovalent salt systems where ion−ion correlations do not play an important role. This is appropriate here where salt was assumed to be NaCl (and there are no multivalent ions present in the system). In the case of studying the titration properties


DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

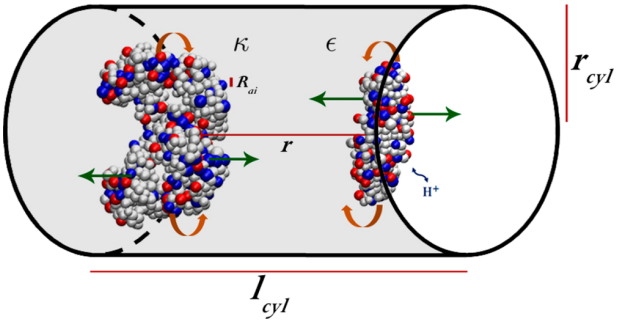



**Figure 2.** A sketch of the simulation model system for the constant-pH Monte Carlo simulations. Two macromolecules represented by a collection of charged Lennard–Jones spheres of radii $R_i$ and valences mimicking amino acids are surrounded by counterions and added salt, implicitly described by the modified inverse Debye length $\kappa$. The solvent is represented by its static dielectric constant $\varepsilon$. Positive and negatively charged protein amino acids are represented in blue and red, respectively. The macromolecules's centers of mass are separated by a distance $r$. The cylindrical simulation box is defined by the length $l_{cyl}$ and radius $r_{cyl}$.

of a single macromolecule in the apo (no Ab bound), the simulation cell model simplifies by the removal of the second molecule from the simulation box. Two macromolecules (the NS1 and the Ab) are always present in the simulation box in the holo (Ab bound) form at a fixed conformation (as given by the X-ray data). In both cases, the electrostatic interactions $[u_{el}(r_{ij})]$ between any two charged amino acids of valences $z_i$ and $z_j$ are given by

$$u_{el} = \frac{z_i z_j e^2}{4\pi\epsilon_0\epsilon_s r_{ij}} \exp(-\kappa r_{ij}) \tag{1}$$

where $\epsilon_0$ is the dielectric constant of the vacuum ($\epsilon_0 = 8.854 \times 10^{-12}$ C$^2$/N m$^2$), $\epsilon_s$ is the dielectric constant of the medium (78.7 here), and $e = 1.602 \times 10^{-19}$ C is the elementary charge. The values of $z_i$ and $z_j$ are defined by the FPTS and will vary during the simulation sampling as a function of pH and the environment. See refs 65 and 67 for more details.

For the macromolecular complexation studies, other interactions (van der Waals interactions, hydrophobic effect, and excluded volume repulsion) must also be included.[17,18] To account for these physical interactions, a simple Lennard–Jones potential between the beads (amino acids) is used. Therefore, for any two charged or neutral beads $i$ and $j$, $u_{vdw}(r_{ij})$ is given by

$$u_{vdw} = 4\epsilon_{LJ}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2}$$

where $\epsilon_{LJ}$ is, following refs 18, 20, and 80, assumed to be universal for any system and equals 0.124 kJ/mol. This should correspond to a Hamaker constant of ~9 $k_BT$ ($k_B = 1.380 \times 10^{-23}$ m$^2$ kg s$^{-2}$ K$^{-1}$ is the Boltzmann constant, and $T$ is the temperature in Kelvin) for amino acid pairs.[18,20,80] The choice of $\epsilon_{LJ}$ is somehow arbitrary, and it regulates the strength of the attractive forces in the system.[17] For this reason, we also tested the effect of stronger attractions ($\epsilon_{LJ}$ is equal to 0.1734 kJ/mol) for the hydrophobic amino acids in an attempt to improve the modeling of the macromolecular hydrophobic moments[84] and to guide a correct docking orientation at short separation distances. Nonspecific contributions from the hydrophobic effect are already partially incorporated in the model by the different sizes of the beads.[18] The terms $\sigma_{ij}$ were obtained as the sum of these beads' sizes ($\sigma_{ij} = R_i + R_j$).

Combining eqs 1 and 2, the total system's interaction energy for a given configuration $[U(\{r_k\})]$ can be written as

$$U(\{r_k\}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(u^{el}(r_{ij}) + u^{vdw}(r_{ij})) \tag{3}$$

where $\{r_k\}$ are amino acid positions and $N$ is the total number of beads.

**Mapping p$K_a$ Shifts Due to the Presence of Other Molecules.** The central physical chemistry framework that defines our approach is based on the simple acid–base equilibrium properties. It is well-known that any electrical charged body in the vicinity of a titratable group will shift its equilibrium. This is the identical physical mechanism related to the explanation of the reason for pepsin, commonly used for the digestion of immunoglobulins, to have an isoelectric point of 1.4.[85] For instance, an isolated arginine (p$K_0$ = 12.0) has an averaged charge number of 0.5 at pH 12. In the presence of a monovalent positively charged molecule, its p$K_a$ is shifted to 11.3 ($\Delta$p$K_a$ = −0.7) and the averaged charged number is reduced to 0.2 at pH 12. This is due to the fact that the existence of this positive charge closer to arginine favors the arginine deprotonation. Conversely, replacing the positive charge by a monovalent negative one, the protonated state is now favored, which grants the proton to stay longer at arginine. Consequently, the p$K_a$ is now shifted in the opposite direction up to 12.7 ($\Delta$p$K_a$ = +0.7). Likewise, the averaged charge number is increased to 0.8 at pH 12 under this condition. Therefore, knowing the p$K_a$ of an ionizable residue is a practical means to probe molecular interactions in its neighborhood as demonstrated before.[86]

In fact, when an Ab molecule comes nearer or binds to the virus protein, its charges can induce these changes in the p$K_a$ values of the nearby titratable amino acids and vice versa. Mapping the observed p$K_a$ shifts can reveal the existence of preferential binding spots.[86] This is the physical phenomenon that we explored here to map the "electrostatic epitopes" of the studied flaviviruses. Through the comparison of computed p$K_a$ values for the *apo* and *holo* forms, we could identify the amino acids of NS1 most affected by the presence of the Ab. Due to the long-range of electrostatic interactions and the electrical coupling between the titratable groups, this will include ionizable amino acids that are both at the molecular interface (the epitope as defined by the classical biological view) and buried inside the protein structure (the broader definition that we support in this paper).

**Constant-pH Monte Carlo Simulations.** The MC simulations allowed us to explore the ionizable and thermodynamic properties and also the main features of the interactions of NS1-Ab providing free energy derivatives as a function of the macromolecules' separation at different environmental conditions. For our purposes, several sets of MC simulations in the NVT ensemble were performed:

(a) ***Set A***. A single protein $NS1_{WNV(176-352)}$ in the dimeric state (chains A and B as provided by the crystallographic structure PDB ID 4OII removing the Ab) in NaCl solution. The protein is kept rigid in the simulation box. These simulations were used to calculate the p$K_a$s of all titratable groups of the *apo* form of the $NS1_{WNV(176-352)}$ dimer


DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

Table 1. Set of MC Simulations for the Three Dimeric NS1 Proteins of WNV (PDB ID 4OII), $ZIKV_{UG}$ (PDB ID 5K6K), and $ZIKV_{BR}$ (PDB ID 5GS6)[a]

| set | system | PDB, resolution | pH[b] | form | MC movements | MC steps[c] | replicates | main information | label |
|---|---|---|---|---|---|---|---|---|---|
| A | $NS1_{WNV(176-352)}$ | 4OII, 3 Å | 5.4 | *apo* | titration | $10^8$ | 5 | $pK_a$ | X-ray |
| B | $NS1_{WNV(176-352)}$-Ab | 4OII, 3 Å | | *holo* | titration | $10^8$ | 5 | $pK_a$ | X-ray |
| C | $NS1_{WNV(176-352)}$-Ab | 4OII, 3 Å | | *holo* | titration, translation and rotation[d] | $10^7$ | 5 | $pK_a$, $\beta w(r)$ | complex |
| D | two $NS1_{ZIKV-(176-352)}$ | 5K6K, 5GS6, 1.89 Å, 2.85 Å | 8.5, 8.5 | *apo* | titration | $10^8$ | 5 | $pK_a$ | X-ray |
| E | two $NS1_{ZIKV-(176-352)}$-Ab | 5K6K, 5GS6 | | *holo* | titration, translation and rotation[d] | $10^7$ | 5 | $pK_a$, $\beta w(r)$ | complex |
| F | $NS1_{WNV(176-352)}$-Ab | 4OII 3 Å | 5.4 | *holo* | titration, translation and rotation[d] | $25 \times 10^9$ | 1 | $\beta w(r)$ | complex |
| G | $NS1_{WNV}$ | 4O6D+5GS6, 2.6 Å + 2.85 Å | 5.5 + 8.5 | *apo* | titration | $10^8$ | 5 | $pK_a$ | X-ray complete |
| H | $NS1_{WNV}$-Ab | 4O6D+5GS6, 2.6 Å + 2.85 Å | | *holo* | titration, translation and rotation[d] | $10^7$ | 5 | $pK_a$, $\beta w(r)$ | complex |

[a]All simulations were done with $\varepsilon_{LJ}$, $T$, $\epsilon$, and $c_s$ equal to 0.124 kJ/mol, 298 K, 78.7, and 150 mM (of NaCl), respectively. See the text for more details. [b]Experimental pH. The MC Simulations were performed at several other pH conditions. See the text. [c]After the proper equilibration of the systems[87] (that was defined by monitoring the evolution during the MC runs of all the quantified properties). [d]This is a full protein-protein complexation study at CpH.

[$pK_{a_WNV(176-352)}$(X-ray)]. This is the reference state of the $\Delta pK_a$ that we used to map the EE of the $NS1_{WNV(176-352)}$ for this specific Ab.

(b) ***Set B***. The complex $NS1_{WNV(176-352)}$-Ab as provided by the crystallographic structure PDB ID 4OII in NaCl solution. The $NS1_{WNV(176-352)}$ was in the dimeric state as in Set A. The same quantities computed in Set A are computed for the complex [$pK_{a_WNV(176-352)\text{-}Ab}$(X-ray)]. Calculations were carried out with the *holo* form of the $NS1_{WNV(176-352)}$ dimer. These data are used to determine titratable amino acids affected by the presence of the Fab fragment 22NS1.

(c) ***Set C***. The protein $NS1_{WNV(176-352)}$ dimer initially placed apart from the Ab protein in NaCl solution. The structures of each protein were the same as those used in Sets A and B, respectively. Such a system allows studying the association mechanism of the two macromolecules and the impact of Ab on the charged $NS1_{WNV(176-352)}$ dimer residues, when they approach each other. Both macromolecules were allowed to translate, rotate, and titrate. Simulations provide (i) the computed free energy of interactions as a function of the molecules separation distance [$\beta w(r)$] considering all electrostatics mechanisms at a given pH and (ii) $pK_a$s for the $NS1_{WNV(176-352)}$ dimer during the protein-protein complexation process with Ab [$pK_{a_WNV(176-352)\text{-}Ab}$(complex)]. These $pK_a$s data are used to determine what titratable amino acids are affected by the visits of the Fab fragment 22NS1. This type of simulation is essential in practical applications when the crystallographic structure for the complex is not experimentally available.

(d) ***Set D***. Same as Set A but for the incomplete structures of the $NS1_{ZIKV\text{-}BR(176-352)}$ (PDB ID 5GS6) and $NS1_{ZIKV\text{-}UG(176-352)}$ (PDB ID 5K6K) dimers. The main results were $pK_{a_ZIKV\text{-}BR(176-352)}$(X-ray) and $pK_{a_ZIKV\text{-}UG(176-352)}$(X-ray), respectively. These quantities are the reference state of the $\Delta pK_a$ that we used to map the EE of the ZIKV.

(e) ***Set E***. Same as Set C but for the pairs $NS1_{ZIKV\text{-}UG(176-352)}$-Ab and $NS1_{ZIKV\text{-}BR(176-352)}$-Ab. In both cases, the Ab Fab fragment 22NS1 structure was extracted from PDB ID 4OII. These CpH MC simulations provided the values for $pK_{a_ZIKV\text{-}BR(176-352)\text{-}Ab}$(complex) and $pK_{a_ZIKV\text{-}UG(176-352)\text{-}Ab}$(complex).

(f) ***Set F***. Same as Set C with the purpose to obtain $\beta w(r)$ that is used to evaluate the pH effects on the association pathway of the $NS1_{WNV(176-352)}$ and the Ab.

(g) ***Set G***. Same as Set A but for the complete structure $NS1_{WNV}$. The main result was $pK_{a_WNV}$($X\text{-ray}_{complete}$).

(h) ***Set H***. Same as Set C but for the complete structure $NS1_{WNV}$ and the Fab fragment 22NS1. The principal output extracted from these CpH MC simulations was $pK_{a_WNV\text{-}Ab}$($complex_{complete}$). For the sake of the reader, a comparison between these simulations sets is summarized in Table 1. All calculations performed with the rigid and unmovable crystallographic structures are referenced with the extension "*X*-ray" [e.g., $pK_{a_ZIKV\text{-}BR(176-352)}$(X-ray)], while results from the simulations of the protein-protein complexation process where both macromolecules can also translate and rotate are specified by the word "complex" [e.g., $pK_{a_ZIKV\text{-}BR(176-352)\text{-}Ab}$(complex)].

All the $pK_a$ values were calculated from the averaged charge numbers of each titratable amino acid ($z_{ai}$) sampled at a pH range from 0 to 14 every 0.1 using the FPTS.[67] Following the procedure described in ref 65, we obtained from these $z_{ai}$ values the corresponding $pK_a$s at the pH condition when the titratable group is half protonated. After equilibration, the number of MC steps used in Sets A, B, D, and G (titrations for rigid macromolecules) was $10^8$ while for Sets C, E, and H (titrations during the simulation of the complexation) was $10^7$; and for Set F, we employed longer runs ($2.5 \times 10^9$ MC steps) due to statistical reasons for the estimation of $\beta w(r)$.[18] These are massive simulations and very costly even at the CG representation. For all simulation, at least five replicates were carried out to estimate the standard deviations. Each system was placed in an open cylinder simulation box with proper periodic boundary conditions as shown in Figure 2. A personalized version of the Faunus biomolecular simulation package,[88] where the FPTS is implemented to study pH-dependent processes was employed to perform all the constant-pH simulations.

For the sake of convenience, the simulation box was adjusted for each of these sets. For titrations with rigid macromolecules (Sets A, B, D, and G), the most important condition is to ensure that all the macromolecules are inside



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

the cell that mimics the electrolyte solution. Respecting this constraint, any box size bigger than that has no effect on the outcomes. We used a quite larger box of 300 Å of length and radius. On the other hand, for the complexation studies (Set F), we found that for measuring $\beta w(r)$, it was necessary to keep the cylindric length at 300 Å. Conversely, for the p$K_a$ values measured during the complexation process, shorter boxes could be used with 150 Å of length and radius to improve the sampling efficiency of this particular quantity (for Sets C, E, and H). This number of 150 Å was defined based on the previous simulation runs with the bigger box size after observing the behavior of $\beta w(r)$. A smaller box is a benefit to increase the p$K_a$ values since the molecules tend to be at a shorter separation distance without significantly altering the p$K_a$ shifts from which we based our conclusions. The estimation of $\beta w(r)$ as described in Set F for $NS1_{WNV(176-352)}$ and the Ab was performed at a solution pH values equal to 3, 5 (close to the pH 5.4 of the crystallization conditions), and 11. For each simulation, the PP radial distribution function $[g(r)]$ between the center of mass of the two proteins was calculated with a bin size of 1 Å. Then, the potential of the mean force $[\beta w(r) = -\ln g(r)]$ was obtained as an estimation of the free energy of complexation, see refs 17 and 18. The aqueous solution static dielectric constant ($\epsilon_s$), temperature, and the salt concentration ($c_s$) were, respectively, 78.7, 298 K, and 150 mM (of NaCl) in all simulations.

**Epitope Determined by Experimental Data.** As a reference for the epitope definition, we determined the amino acids that are experimentally found in contact with each other at the $NS1_{WNV(176-352)}$-Ab interface based on changes in the solvent-accessible surface area (SASA). This is a good criterion to define the amino acids that are part of the antigenic determinant with a reasonable probability and without invoking any other criterion and/or energetic function. It is solely based on the only known experimental crystallographic structure for a complex, namely, PDB file 4OII. The calculations were done using NACCESS with default parameters, *i.e.*, 1.4 Å for the solvent probe radius.[89] In practice, residue surface accessibility (RSA) was evaluated for the *isolated* $NS1_{WNV(176-352)}$ in the complex X-ray structure and compared to RSA of $NS1_{WNV(176-352)}$ complexed with Ab. Any amino acid of $NS1_{WNV(176-352)}$ that exhibits some change between the two situations is considered as constituting the antigenic or epitope region while change for Ab will define the paratope zone. As such defined, this will be the reference for all methods used in this work.

The performance of each theoretical method was evaluated by standard statistical concepts (in a binary classifier context) using the numbers of true positives, true negatives, false positives, and false negatives when comparing with the reference from the SASA analysis. Calculations of the sensitivity, specificity, accuracy, precision, and balanced accuracy were done to estimate them. These measurements are also the most commonly used to estimate the performance of algorithms for the prediction of epitopes.[39,90]

Briefly, residues that are predicted as antigenic, i.e., defining an epitope, and are located in the experimental X-ray interface are true positives (TP). Those that are predicted as antigenic but are not located in the X-ray interface are false positives (FP). True negative (TN) residues are those that are not predicted antigenic and not found in the X-ray interface, and finally, false negatives (FN) are residues located in the interface but not predicted as antigenic. From this definition, we can directly calculate the main descriptors:

$$\text{(a) sensitivity (or True Positive Rate, TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{(b) specificity (or True Negative Rate, TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{(c) precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{(d) accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}}$$

$$\text{(e) balanced accuracy (BACC)} = \frac{\text{TPR} + \text{TNR}}{2}$$

The results were displayed as the receiver operating characteristic (ROC) plots. The ROC evaluation considers a comparison between TPR and FPR, i.e., the sensitivity against (1 − specificity) calculated as a function of a key parameter, i.e., a threshold. For each method, the relevant threshold was tested by varying by step of 0.001. For the sake of convenience, the area under the curve (AUC) was also computed. All theoretical predictions were evaluated using these statistical quantities. The best thresholds for each case were also defined from ROC analyses and measurements of the performance in comparison with the SASA data. Some analyses were done taking into account only the titratable residues. We refer to them as $SASA_{ionizable}$.

**Epitope Predictions by Bioinformatic Tools.** To map the antigenic regions by bioinformatic tools, we employed several (free) Web servers for the prediction of conformational or linear B-cells epitopes. For the present purposes, we selected five popular different algorithms: (a) "PDBePISA" (PISA), which was included in the predictor set because, besides the accessibility, it uses additional empirical parameters to predict assemblies based on free binding energies and consequently, the corresponding interface;[91] (b) DiscoTope, that uses a statistical, spatial, and surface exposure evaluation of the amino acids;[42] (c) ElliPro, where antigenicity, solvent accessibility, and flexibility of the protein structure are considered;[39] (d) SEPPA2 that takes into account the unit-patch of residue triangle within neighboring distance, the accessible surface area, the localization of the protein antigen, and the kind of species;[40] and (e) SEPPA3 that is a recently updated version of SEPPA2 but now including parameters related to the glycosylation in its score function.[92] We performed calculations both with their default parameters and with other values to determine their accuracy as a function of these input parameters.

These tools are supposed to be "universal" predictors, *i.e.*, not specific for one particular Ab as the Fab fragment 22NS1. Thus, in principle, they can predict the antigenic regions that has a potential to interact with *any* Ab. Hence, it is expected that their predictions would find the epitopes as observed at the complex given by the crystallographic structure. This would let us access their prediction capacity for flaviviruses.

Initially, we used the thresholds as defined by default for each algorithm (the values used here were, respectively, −3.7, 0.5, 0.09, and 0.089 for DiscoTope, ElliPro, SEPPA2, and SEPPA3). Calculations with ElliPro were performed both with the maximum distance setup to 6 Å (the default value) and


DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

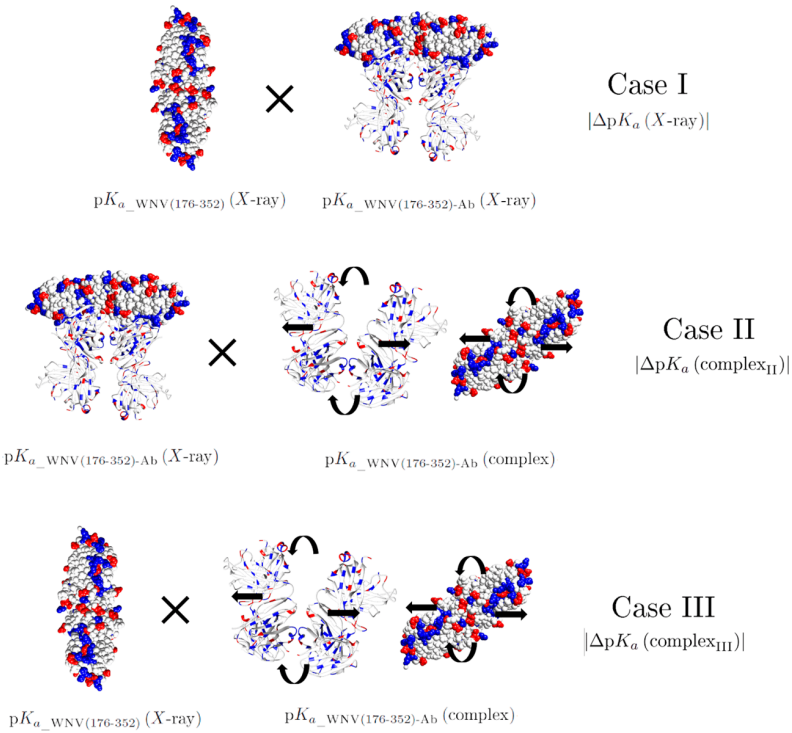



**Figure 3.** Scheme illustrating all the comparisons of the p$K_a$ shifts performed in this work. Systems with no arrows drawn on it can only titrate. The arrows indicate the MC simulations of the complexation process where the two macromolecules are also free to translate and rotate inside the cylindrical simulation box. See the text for the details.

simultaneously varying the two parameters available from their server.[39] In this case, the maximum distance was varied from 4 to 8 Å by steps of 1 Å together with a reference score value varying from 0.5 to 1.0 every 0.1. The clustered data based on distances as provided by the server for the conformational epitopes were selected for the predictions from this tool. Linear epitopes could be in principle used for this algorithm as well. However, we restricted the calculations to conformational epitopes in order to make these predictions comparable with the other servers. For SEPPA2 and SEPPA3, the antigen protein subcellular localization selected was "Secreted" and the species of the host was "Homo". Then, we investigated the optimal thresholds using a ROC analysis. With the best threshold for each program based on the SASA definition of the true epitopes as a reference, we built alignments as in our previous work[11] to visualize more easily the predicted antigenic regions of each program. As argued by Yang and Yu,[61] a consensus approach using two or more programs tends to improve the prediction rates. Here, we defined the epitopes regions as the set of amino acids where the predictions given by DiscoTope and SEPPA3 match. This consensus was chosen because we found the best performance for these two servers when the thresholds were optimized. Due to the scope of the present work, as we have a greater interest in the ionizable residues of the antigen−antibody interface, we applied the same methodology but using only the titratable amino acids in our references ($SASA_{ionizable}$, see above).

**Estimates of the Antigenic Regions by p$K_a$ Shifts: The PROCEEDp$K_a$ Method.** The theoretical physical fundamentals to support the use of p$K_a$ shifts for the epitope prediction were given above. Three cases with different complementary aims were analyzed here and for the sake of clarity, they are illustrated in Figure 3. Cases I and II are initial validation tests while Case III is a real prediction test. Predictions obtained from this electrostatic method are referenced as "PROCEEDp-$K_a$" (**PR**ediction **O**f ele**C**trostatic **E**pitopes bas**ED** on **p*K*a** shifts).

These cases are (a) **Case I**. We compare the p$K_a$ results obtained from simulation Sets A and B to calculate $\Delta pK_a$(X-ray) as the difference between $pK_{a_WNV(176-352)\text{-}Ab}$(X-ray) and $pK_{a_WNV(176-352)}$(X-ray). Doing this comparison, we identified all titratable amino acids from $NS1_{WNV}$ that were perturbed by the presence of the Ab Fab fragment 22NS1. The affected ionizable residues by this *specific* Ab are defined here as EE. The subscript "I" will be used together with the abbreviation EE (i.e., $EE_I$) to specify from which case the EE was obtained from. This is the first aim of such analysis, *i.e.*, to map the EE for a known host−pathogen PP system. The second is to confirm the capacity to use these p$K_a$ shifts to reproduce the experimental observations as provided by the SASA analysis. In other words, we want to evaluate to which extent $|\Delta pK_a$(X-ray)| can be used to identify titratable amino acids that are classified as epitopes in the known X-ray structure for the complex. Accordingly, different thresholds for $\Delta pK_a$(X-ray) were tested to define an ideal condition for the two following cases.

(b) **Case II.** While Case I is useful to validate the present approach, it has limited practical use for general applications in other systems, since it requires the knowledge of the X-ray structure of the complex, which is not available in most of the situations. A possible alternative is to study the PP association process as described in Set C. The simulations done in Set C only require the 3D structures of the individual molecules (NS1 and Ab) and no need for the structure of the complex.



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

Table 2. Performance of Several Canonical Bioinformatic Predictor Tools of Antigenic Regions of B-Cell[a]

| | measure | PISA[b] | DiscoTope | ElliPro | SEPPA2 | SEPPA3 | Consensus |
|---|---|---|---|---|---|---|---|
| (a) | sensitivity | | 0.121 | 0.485 | 0.182 | 0.818 | |
| | specificity | | 0.934 | 0.410 | 0.833 | 0.618 | |
| | accuracy | | 0.782 | 0.424 | 0.712 | 0.655 | |
| | precision | | 0.296 | 0.158 | 0.200 | 0.329 | |
| | balanced accuracy | | 0.523 | 0.447 | 0.508 | 0.718 | |
| (b) | sensitivity | 1.000 | 0.667 | 0.182 | 0.500 | 0.818 | 0.606 |
| | specificity | 0.840 | 0.764 | 0.951 | 0.743 | 0.622 | 0.809 |
| | accuracy | 0.870 | 0.746 | 0.808 | 0.698 | 0.658 | 0.771 |
| | precision | 0.589 | 0.393 | 0.462 | 0.308 | 0.331 | 0.421 |
| | balanced accuracy | 0.855 | 0.716 | 0.567 | 0.622 | 0.720 | 0.790 |
| (a) | default | | −3.700 | 0.500 (6)[c] | 0.090 | 0.089 | |
| (b) | best threshold | | −6.940 | 0.700 (5)[c] | 0.049 | 0.093 | |

[a]The SASA data is used as a reference. Data obtained with (a) default and (b) optimized parameters (best thresholds). [b]The PISA server does not offer an option to change the default parameters except for the processing mode. All PISA calculations were done with the "auto" mode. [c]Parentheses value refers to the maximum distance.

On the contrary, the complexation mechanism should be revealed by such simulations. Nevertheless, it is compulsory to first demonstrate that these simulations of the complexation are able to correctly reproduce the crystallographic PP interface. This is the main purpose of this case II. To test it, $|\Delta pK_a(\text{complex}_{II})|$ as equal to the difference between $pK_{a_WNV(176-352)\text{-}Ab}(\text{complex}) - pK_{aWNV(176-352)\text{-}Ab}(\text{X-ray})$ was calculated. These quantities were obtained from the set of simulations C and B, respectively (see Table 1). The minimum value of $|\Delta pK_a(\text{X-ray})|$ for an amino acid to be classified as an EE is *a priori* unknown. We have defined it based on analyses performed with Case I assuring that $|\Delta pK_a(\text{X-ray})|$ was larger than the estimated standard deviations for the calculated $pK_a$s during the MC simulation runs. This set of EE are referred to as $EE_{II}$ in this text. A comparison between $EE_I$ and $EE_{II}$ can confirm if the MC simulations of the PP complexation can reproduce the same interfaces as revealed by the crystallographic structure of the complex.

(c) **Case III**. This is analogous to Case II, but now the aim is the prediction of the EE for practical studies. Case III represents one of the most interesting assessment because we do not need to know the experimental interface antigen−antibody. Indeed, the comparison here is made between $pK_a$ calculations from the *individual* NS1 and those obtained for NS1 from the complexation simulations. This means that the EE is predicted by the calculation of $\Delta pK_a(\text{complex}_{III})$ as the difference given by $pK_{a_WNV(176-352)\text{-}Ab}(\text{complex})$ and the $pK_{a_WNV(176-352)}(\text{X-ray})$ (see sets of simulations C and A in Table 1). Different thresholds were also investigated for this case as mentioned above. Since the available experimental structure that we have used is for the incomplete dimeric NS1 of WNV, the relevant role of the missing residues and the long-range effects of the electrostatic interactions was also discussed. Our analysis will be confronted with the experimental results given by $SASA_{ionizable}$ as a reference.

The definition of the minimum value of $|\Delta pK_a|$ to optimize the thresholds for each case mentioned above followed the same procedure to define the accuracy as done for the bioinformatic tools through a ROC analysis.

**Estimation of Epitopes on ZIKV by $pK_a$ Shifts for Two Other Strains.** We extended the predictions to the ZIKV comparing also the Brazilian and Ugandan strains. The EE for $NS1_{ZIKV(176-352)}$ from these strains were determined by the values of $|\Delta pK_a(\text{complex}_{III})|$ greater than the best threshold defined through the ROC analysis of Case III. Here, for the Ugandan strain, the differences given by $pK_{a_ZIKV\text{-}UG(176-352)\text{-}Ab}(\text{complex})$ (derived from Set E) and $pK_{a_ZIKV\text{-}UG(176-352)}(\text{X-ray})$(derived from Set D) were used. An identical calculation was done for the Brazilian strain with its respective values of $pK_a$s.

**Estimation of the Impact of $\beta$-Roll and Wing Domains on $pK_a$ Shifts.** The differences between the $pK_a$ values of the *apo* and *holo* forms of the complete structures of $NS1_{WNV}$ were carried out to estimate the $|\Delta pK_a(\text{complex}_{III})|$ of this system as well (please see above regarding $pK_a$s for Sets G and H of the simulations in Table 1). In all situations, the EEs were considered taking into account the best threshold derived from Case III of the last subsection. Analysis of EE for these three flaviviruses systems was done at the sequence and three-dimensional level.

## RESULTS AND DISCUSSION

**Prediction of Antigenic Regions by Canonical Bioinformatic Tools.** Epitope prediction is a known challenging task. Most of the tools developed so far are trained on data sets that do not necessarily include particular information about flaviviruses. Since our purpose is to discuss and evaluate a new, wider, and more efficient theoretical approach to predict antigenic epitope regions for flaviviruses, we must critically analyze the efficiency of the available tools first. Thus, rather than reporting published benchmarking results, we calculated the efficiency of the different commonly used bioinformatic tools for identifying antigenic sites on the NS1 protein of WNV by comparing with the known crystallographic complex $NS1_{WNV(176-452)}$-Ab deposited in the PDB file ID 4OII. Accordingly, as a first reference target, we considered *all* residues of the PP interface identified by the SASA procedure, *i.e.*, ionizable and nonionizable residues, resulting in a total number of 66 amino acids and then we restricted to ionizable residues, *i.e*, the electrostatic epitope (EE) that encompasses 24 residues. Therefore, the critical analyses were carried out following both the classical definition of epitopes and exploring the EE concept presented here.



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

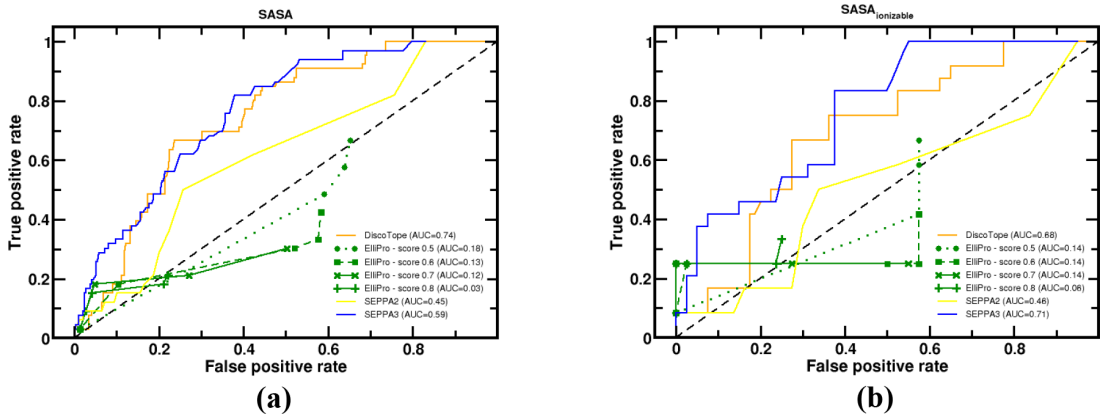



**Figure 4.** Performance of different canonical bioinformatic tools for the prediction of antigenic residues of B-cell epitopes. Data for $NS1_{WNV(176-352)}$. (a) Receiver operating characteristic curves obtained based on the comparison between SASA as a reference and each tool. (b) Same as part a but using $SASA_{ionizable}$(only ionizable amino acids) as a reference.

As usually done, the performance of the different algorithms was evaluated based on the values of TPR (or sensitivity) and FPR considering all residues. The corresponding results with additional statistical measures obtained using the default and optimized parameters are summarized in Table 2. Predictions from PISA and the consensus policy are also reported. More detailed results are given in Figure S1 and Table S1.

The reliability of the predictions is an important aspect to be considered when the performance of the algorithms is discussed. Generally speaking, it is desired to have a sensitivity and a specificity greater than 0.5. This guarantees that the estimates are not either due to chance or due to a high number of FP and FN. As is often the case, the better sensitivity is reached at the expense of an increase in the number of false positives (i.e., a reduced specificity) and vice versa. Using the default parameters, DiscoTope and SEPPA2 show poor sensitivity but high specificity. This means that they efficiently identify residues that do *not* belong to the interface but fail to characterize residues that belong to the host−pathogen PP interface. In contrast, SEPPA3 is able to find the amino acids that belong to the epitope region, but this method is less efficient to distinguish those that do not participate in this biological interface. ElliPro poorly performs in terms of sensitivity and specificity. DiscoTope, SEPPA2, and SEPPA3 performed well in terms of the accuracy, while ElliPro presented values less accurate than the random behavior (0.5). All servers performed poorly in terms of the precision. Most of the values for their precision measurements are below 0.4. The highest observed value was found for SEPPA3 (0.329). In addition, the BACC, which is an average between sensitivity and specificity, is also optimal for SEPPA3 (0.718) when it is compared with the other tools at the default conditions.

Another useful way to assess the predictive characteristics of these tools is using ROC plots. These plots measure the balance between the true positive and false positive rates of a predictor also allowing to compare it with a random guess. Such analysis has already been previously used for epitope predictors.[90,92] Figure 4a illustrates the ROC curves for all the studied bioinformatic prediction servers. We observed that the best performance is seen for the DiscoTope and SEPPA3 servers. DiscoTope has the AUC value (0.74) closer to 1 (the perfect classifier). The second-best result was obtained by SEPPA3 with AUC equals to 0.59 which is slightly higher than a random prediction (AUC = 0.5). Conversely, an insufficiently performance was obtained by ElliPro and SEPPA2 (the curves have AUCs less than 0.5). It is important to mention that although AUC is relatively high for DiscoTope and SEPPA3, their FPRs are also comparatively increased. However, it should be noted that the poor performance found for ElliPro and SEPPA2 might be due to the fact that we are restricted to a unique reference (i.e., the SASA data of the well-known crystallographic of $NS1_{WNV(176-352)}$ - PDB id 4OII for a *specific* Ab).

The optimized parameters for each epitope predictor algorithm (see values reported as part b in Table 2) were defined through the ROC analysis. They were obtained from the analyses of the input parameters that resulted in the best prediction, *i.e.*, the situations when TPR and FPR have the closest values to 1 and 0, respectively. In practice, it is expected that the predictor should correctly estimate the antigenic regions through the default parameters proposed for each algorithm. Nevertheless, these present results confirm the complexity of epitopes predictions and indicate that the available tools are not providing the desired outcomes for flaviviruses. A quick overview shows that improvements can be achieved in the TPR values for DiscoTope, SEPPA2, and SEPPA3 when the optimized input parameters are used. With respect to the accuracy, from the default to the optimized cases, the values improve for ElliPro (from 0.424 to 0.808) and SEPPA3 (from 0.655 to 0.658), while for DiscoTope and SEPPA2 the opposite is seen (from 0.782 to 0.746 and from 0.712 to 0.698, respectively). In terms of precision and BACC, an improvement is detected for all studied algorithms. Surprisingly, the consensus approach (i.e., those residues estimated that simultaneously match for DiscoTope and SEPPA3) shows higher performance, *e.g.*, BACC equals 0.790, proving to be an effective means to determine epitopes as previously suggested.[92] Furthermore, the PISA web server reveals high performance in terms of the statistical descriptors calculated for this purpose, *e.g.*, BACC equals 0.855. It is not surprising since it is the method that includes the most experimental information from the system itself. The difference with the pure SASA results is due to the inclusion of residues whose accessibility is not modified by the presence of the Ab,


DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

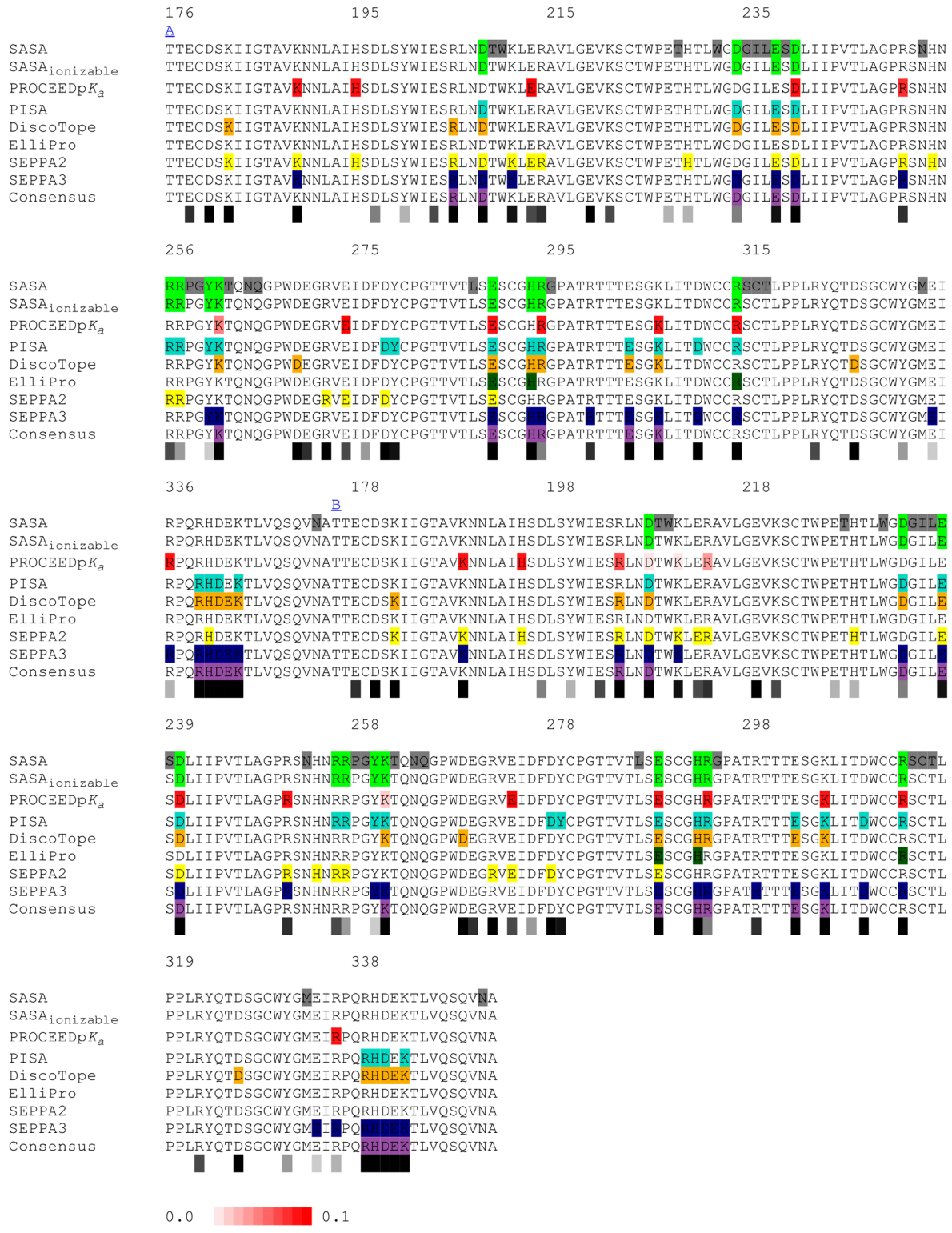



**Figure 5.** Primary sequence of the dimeric NS1$_{WNV(176-352)}$ with the estimated ionizable amino acids of the antigenic regions predicted by several bioinformatic tools using the best thresholds for each one. The used experimental reference was SASA$_{ionizable}$ which comprises 24 residues (only including ionizable amino acids). The $|\Delta pK_a|$s are estimated from Case III where the EE is determined if 0.007 as a threshold [$|\Delta pK_a(complex_{III})| > 0.007$]. The reference used for all algorithms was SASA$_{ionizable}$ which comprises 24 ionizable groups. The results for the $\Delta pK_a$s are displayed as a color-coded $\Delta pK_a$ scale, with white being zero ($|\Delta pK_a| = 0$) and red the maximum variation observed ($|\Delta pK_a| > 0$). The scale is shown at the bottom of this figure. All other details are the same as in Figure S1. The main statistical information for each tool is given in Table S1.



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

Table 3. Performance of Canonical Bioinformatic Tools to Predict Antigenic Regions of B-Cell and the Estimated EE[a]

| measure | $\lvert\Delta pK_a\ (complex_{III})\rvert$ | PISA | DiscoTope | ElliPro | SEPPA2 | SEPPA3 | Consensus |
|---|---|---|---|---|---|---|---|
| threshold | 0.007 | | −6.850 | 0.700 (5)[b] | 0.049 | 0.093 | |
| sensitivity | 0.500 | 1.000 | 0.667 | 0.250 | 0.500 | 0.833 | 0.667 |
| specificity | 0.800 | 0.775 | 0.725 | 1.000 | 0.662 | 0.625 | 0.800 |
| accuracy | 0.721 | 0.827 | 0.702 | 0.827 | 0.625 | 0.673 | 0.769 |
| precision | 0.407 | 0.571 | 0.410 | 1.000 | 0.308 | 0.400 | 0.500 |
| balanced accuracy | 0.650 | 0.888 | 0.696 | 0.625 | 0.581 | 0.729 | 0.734 |

[a]Data for $NS1_{WNV(176-352)}$. EE were estimated $\lvert\Delta pK_a(complex_{III})\rvert$ (using $pK_a$ shifts data from Sets A and C, see Table 1). $SASA_{ionizable}$ is used as the reference. Data correspond to the best threshold for each algorithm. [b]Parentheses value refers to the maximum distance.

but their energetic score functions are considered important to their contribution.

Yet, it remains open to be investigated if these parameters are universal for all proteins from flaviviruses. In any case, this set of results indicates that there is still room for improvement to get better epitope predictions.

**Putative Ionizable Residues at the Interface Obtained by Canonical Bioinformatic Tools.** The analyses reported in the previous section allow us to determine how the most common available tools predict antigenic regions for $NS1_{WNV(176-352)}$. They revealed the intrinsic difficulties of these methodologies and the relatively good performance of PISA, DiscoTope, SEPPA3 and the consensus approach for the Fab fragment 22NS1, when the input parameters can be tuned. However, since our central discussion is the EE, it is necessary to re-evaluate the predictions specifically for the ionizable residues of the PP interface of NS1. There are 24 amino acids that can be protonated/deprotonated in the antigen−antibody interface. Therefore, we investigated in this section the effect of restricting the analysis only to titratable amino acids using $SASA_{ionizable}$ instead of SASA as used in the previous subsection. In Figure 5, the primary sequence of protein NS1 is plotted together with the estimated ionizable amino acids of the antigenic regions as defined by $SASA_{ionizable}$ and the predicted (ionizable) ones by DiscoTope, ElliPro, SEPPA2, SEPPA3, and a consensus policy. The predicted results correspond to the best thresholds. Results from PISA are given at this figure too. The values of TP, FN, TN, and FP are reported in Table S1. We repeated the same protocol using the ROC analysis as done before. The only difference is the replacement of the reference data by $SASA_{ionizable}$. This will be a test on the predictive power of these popular algorithms to identify electrostatic important residues at the host−pathogen PP interface. Figure 4b shows the ROC curves obtained for the servers using $SASA_{ionizable}$ as a reference. It can be seen that DiscoTope and SEPPA3 show the best performance compared to the others (AUC are, respectively, 0.68 and 0.71). Compared to the situation when all amino acids are considered, the predictive abilities tend to reduce. This data also confirms that DiscoTope and SEPPA3 are the most appropriate servers to define a consensus policy even when only ionizable residues are considered. The data from PISA (not included in this figure) was slightly better than SEPPA3 in terms of the TP (24 for PISA instead of 20 measured for SEPPA3) and FP (18 for PISA instead of 30 measured for SEPPA3). This is not a surprise, obviously, because this tool is strongly dependent on the known X-ray structure (as we discussed above). Conversely, PISA still has some FP (see, for example, the lines colored cyan for residues 278, 279, 303, 306, and 310 from chain A in Figure 5) because more amino acids are wrongly included due to its additional energetic contribution.[91] This extra contribution on the top of the surface accessibility analysis is what prevents PISA to reach a value equal to 1 in terms of specificity (see Table 3 and the discussion below).

From the ROC analyses, we obtained the best thresholds for each program using $SASA_{ionizable}$(see Table 3). Generally, all the statistical measures for the different tools show minor changes when only the ionizable residues are considered at the interface compared to the total residues, except for PISA. The best values of BACC for the predictor B-cell epitope servers were found for DiscoTope and SEPPA3 (see Table 3). When $SASA_{ionizable}$ is considered, DiscoTope and SEPPA3 predict a high number of TP (20 and 16, respectively, for them). In this comparison, the sensitivity and specificity are modestly improved for SEPPA3 (from 0.622 to 0.662). On its turn, DiscoTope has a bit more difficulty to predict the ionizable residues involved in the antigenic region. Its specificity decreases from 0.764 to 0.725. ElliPro increases its sensitivity from 0.182 to 0.250 although it is still behind the other tools for these predictions. For SEPPA2, we did not observe any change in the sensitivity (the value 0.5 was conserved in both cases). The specificity improved markedly for ElliPro using $SASA_{ionizable}$, while for SEPPA2 it worsened. In terms of accuracy, some servers perform better for the exclusive identification of titratable amino acids. For instance, ElliPro had a value equal to 0.808 (using SASA), but now its accuracy is 0.827 (using $SASA_{ionizable}$). The opposite was observed for DiscoTope that had the accuracy value using SASA equal to 0.746, and now it is 0.702 (using $SASA_{ionizable}$). In terms of precision, all servers showed an improvement in their performances with the new thresholds. Values are below 0.5 for most of the algorithms with the exception of ElliPro that now has a precision equal to 1.

**Testing the Predictive Power of the $pK_a$ Shifts.** In this subsection, we want to answer four important questions that are the base for the detection of EE in proteins of flaviviruses using the $pK_a$ shift ideas (i.e., the PROCEEDp$K_a$ method): (1) can we observe the complexation of $NS1_{WNV(176-352)}$ with the Fab fragment in a constant-pH MC simulation?, (2) can an analysis of $pK_a$ shift the identification of EE for a known system?, (3) can we reproduce the $pK_a$ shifts observed for a known crystallographic structure of the complex $NS1_{WNV(176-352)}$-Ab also in a constant-pH MC simulation of the complexation process?, and (4) how good are the EE predictions based on $pK_a$ shifts?

We started investigating the association pathway by means of the constant-pH MC simulations (described as Set F, see Table 1). The complexation process between the $NS1_{WNV(176-352)}$ and the Ab Fab fragment 22NS1 was investigated by means of free energy derivatives. The calculated free energy of interactions as given by the potentials



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

of mean force $[\beta w(r)]$ for the three studied pH conditions (pH 3, 5, and 11) at physiological ionic strength are given in Figure 6. These are very expensive simulations due to the free energy barriers and the need to populate all the used histogram bins during the sampling.[17,18] Closer to the crystallographic conditions (pH 5.4), there is a large minimum plateau between 30 and 60 Å which indicates a complexation based on a simple thermodynamical criterion $[\beta w(r) < 0]$. At pH 11, there is a small tendency for attractive interactions, while the system is repulsive at pH 3 $[\beta w(r) > 0]$. This data describes the association pathway and also confirms that it is possible to observe the complexation of these macromolecules in a MC simulation as seen before for other studied systems.[17−19,80,81] The appearance of this kind of plateau at the attractive regime (pH 5) reveals that the two macromolecules are able to explore different separation distances with equivalent probability. This is probably due to their irregular molecular shapes. Moreover, these results coincide with the free energy profiles obtained from a semianalytical theory description used in our previous work.[11] In spite of some deviations, the relevant role of electrostatic interactions for the formation of complexes antigen−antibody is again confirmed.

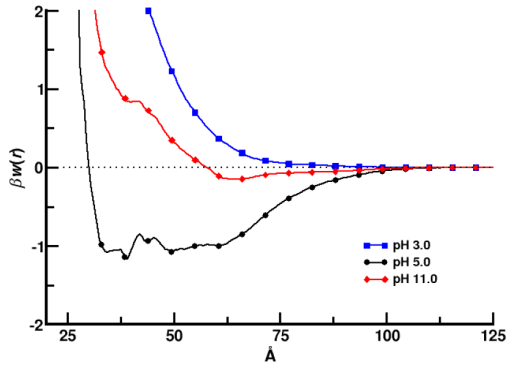



**Figure 6.** Simulated free energy of interactions $[\beta w(r)]$ between the centers of mass of the $NS1_{WNV(176-452)}$ and the antibody 22NS1 at different solution pH conditions. Salt concentration was fixed at 150 mM. The structure of each macromolecule was extracted from the PDB ID 4OII. Simulations started with the two molecules placed at random orientation and separation distance.

We next examined if the analysis of the $pK_a$ shifts were able to correctly identify the EE for the $NS1_{WNV(176-452)}$ and the antibody fragment 22NS1. This was described above as Case I (see also Figure 3). The titratable amino acids identified as EE

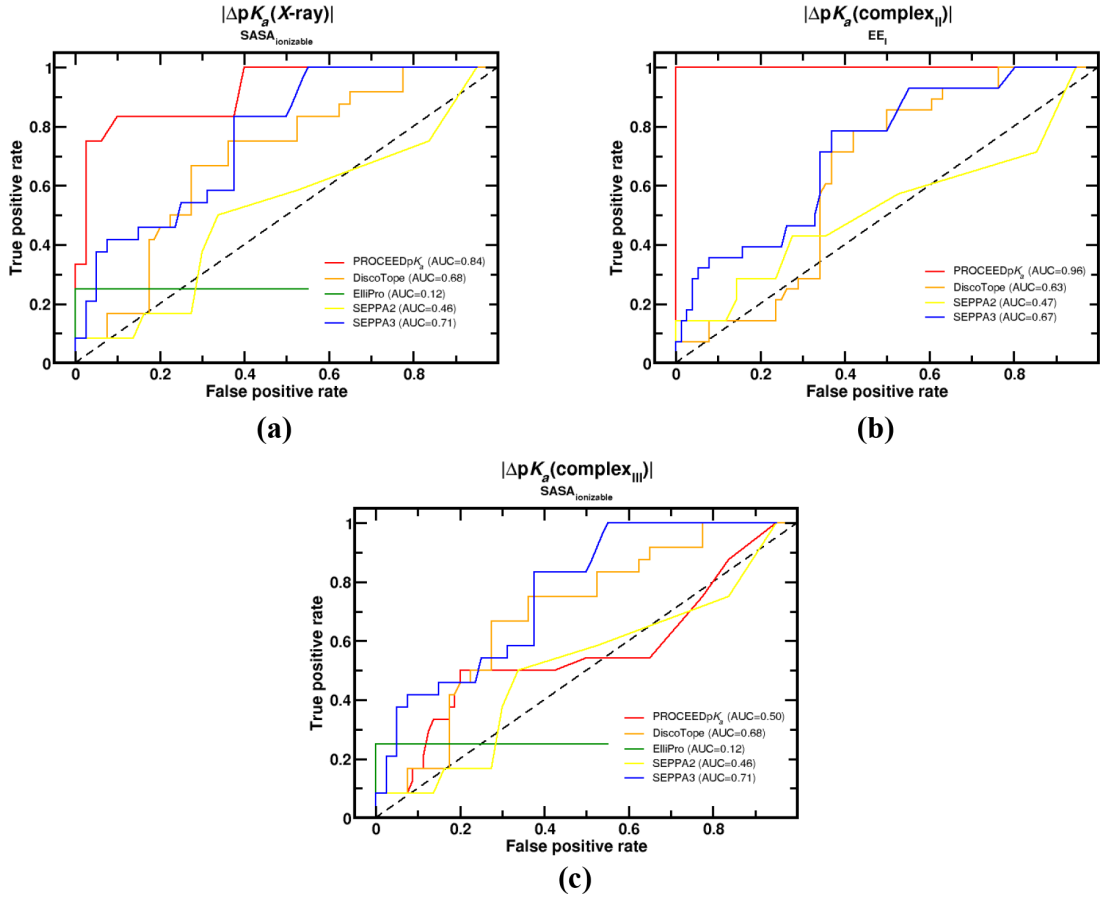



**Figure 7.** Performance of the electrostatic method and other theoretical tools for the prediction of antigenic residues of B-cell epitopes. Data for $NS1_{WNV(176-352)}$. The electrostatic method based on the $pK_a$ shifts analysis is represented in red curves in all the plots. (a) Receiver operating characteristic curves obtained based on the comparison between $SASA_{ionizable}$ as a reference and each other approach. The $pK_a$ shifts analysis was derived from Case I. (b) The same as part a but assuming the $EE_I$ as the reference. The $pK_a$ shifts analysis was derived from Case II. (c) The same as part a but using the electrostatic data from Case III. Calculations of the area under the curves (AUC) are given for all predictions. Titles and subtitles in each plot are given to guide the reader to identify the case and the used reference, respectively.



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

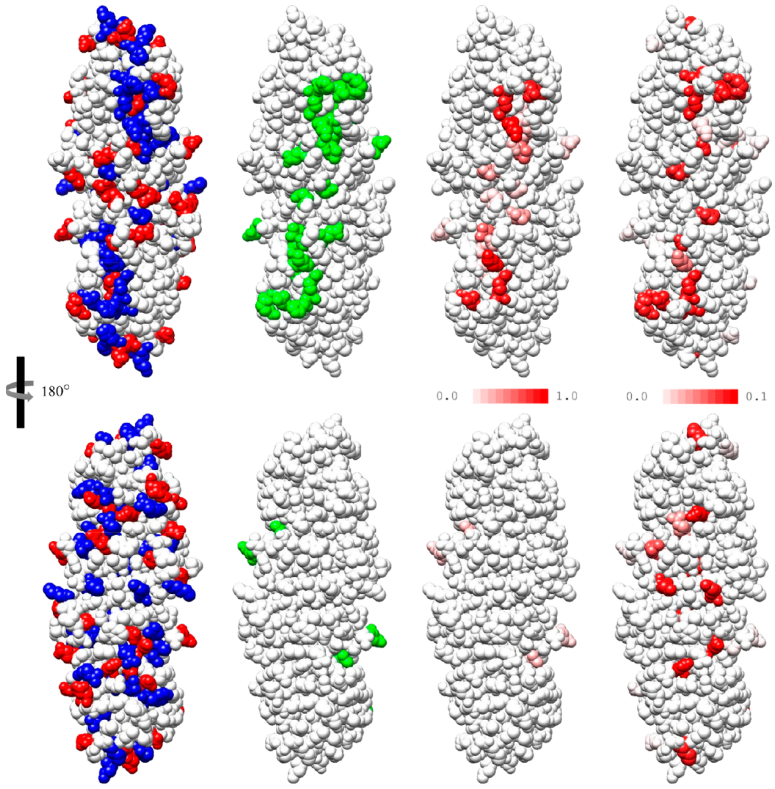



**Figure 8.** Estimates of the EE by p$K_a$ shifts as a descriptor for the protein $NS1_{WNV(176–352)}$ (PDB ID 4OII). From left to right are the ionizable residues, $SASA_{ionizable}$ interface, data from Case I [threshold defined by $|\Delta pK_a(\text{X-ray})| > 0.1$] and Case III [threshold defined by $|\Delta pK_a(\text{complex}_{III})| > 0.007$], respectively. Images in the top correspond to the antibody interface.

at the optimized threshold are displayed as a color-coded $\Delta pK_a$ scale at a three-dimensional level at Figure 8 (look at the third structural model from left to right). The $\Delta pK_a$s were obtained from the comparison of p$K_a$ values of the *apo* ($NS1_{WNV(176–352)}$) and *holo* ($NS1_{WNV(176–352)}$-Ab) forms. The averages and estimated standard deviations of the p$K_a$ for each kind of ionizable group are reported in Table S2. The tested thresholds were always greater than the estimated standard deviations.

By visual inspection, it can be seen in Figures 5 and 8 that the outcomes are closer to the results reported by $SASA_{ionizable}$ but with a tendency to include in the epitopes more amino acids not detected before. Sometimes ionizable residues that are neighbors of the EE obtained by $SASA_{ionizable}$ were also predicted as part of the EE. These are natural features of this p$K_a$ approach that fully explores the long-range nature of the electrostatic interactions and the electrostatic coupling between titratable sites. When the Ab approaches the molecular surface of NS1, its charges will perturb not only the superficial ionizable groups but also other residues that can titrate and are inside the protein. On top of that, the perturbation of one titratable site can also indirectly perturb others in its vicinity like a "chain effect". On its turn, titratable groups of the NS1 are also going to perturb the ionization of the Ab. All the effects of this electrostatic phenomenon will be added together and indiscriminately in the computed p*K*a shifts. As the chosen reference for these analyses, the $SASA_{ionizable}$ data only considers superficial residues of the host–pathogen PP interface and does not account for the contribution of buried residues, and this will negatively impact the predictive power of this approach to reproduce the EE defined by $SASA_{ionizable}$. However, the proposed approach here using the p$K_a$ shifts is much broader and is the only one able to reveal *all* titratable groups that feel the presence of the Ab and can play an important role for their association process. Nevertheless, even the inclusion of the extra amino acids does not prevent this approach to predict well the EE as it can be seen at Figure 7a. Being obtained from what we called Case I in the Theoretical Methods section, we shall call these titratable residues as $EE_I$.

In Figure 7a, the ROC plot for the electrostatic method testing different thresholds for the p$K_a$ shifts is given together with the curves for the other tools discussed above. The data for the other bioinformatic predictors are the same as seen in Figure 4b. It is repeated here for the sake of comparison. It becomes evident that this electrostatic approach (AUC = 0.84) can well predict EE with a small advantage compared to the other methods. For instance, SEPPA3 and DiscoTope have AUC values equal to 0.71 and 0.68, respectively. Of course, in this particular case, we have an ideal situation. The electrostatic method in this test had the benefit of being able to use the exact coordinates of the crystallographic structure (as described for Case I in the Theoretical Methods section). Yet, the main result from this subsection is the fact that the p$K_a$ shifts can indeed be used to identify the EE at least provided the spatial positions and orientations of the two macromolecules are known.

For practical applications, the experimental structure of the protein-Ab complex is likely to be unknown. This implies that a computational complexation study must be performed in



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944–963

most of the situations. Therefore, the third test is to verify if we can reproduce the p$K_a$ shifts observed for a known crystallographic structure of the complex $NS1_{WNV(176-352)}$-Ab (as seen for $EE_I$) in a constant-pH MC simulation of the complexation process. Figure 8 shows the data for this comparison. It can be seen that in fact the $\Delta pK_a$s from the ionizable groups can work as a descriptor of the PP interface. When using the $EE_I$ as the reference instead of the $SASA_{ionizable}$, we put the comparison on the same level. With this referential, the AUC value is equal to 0.96 for the ROC curve (Figure 7b) validating that the same electrostatic perturbation is seen at both situations.

With the best threshold for $|\Delta pK_a(\text{X-ray})|$ (i.e., 0.1 in Table 4), a high performance is detected of this approach for all measures (the values are always greater than 0.5). It means that all residues with a $|\Delta pK_a(\text{X-ray})|$ greater than 0.1 have the potential to be an EE. We observed even much better results in Case II compared to the bioinformatic tools than what was seen for the previous comparison (Case I, see Figure 7a). We can conclude that the two cases (Cases I and II) have a very good match between their EE. They have almost the same $SASA_{ionizable}$ patches. This confirms that the MC simulation of the complexation process exploring its pathway can reproduce the results obtained with the experimental crystallographic structure of the complex. This also implies that the EEs obtained in the complexation simulation are equivalent to the $EE_I$.

**Table 4. Performance of the PROCEEDp$K_a$ Electrostatic Method**[a]

| measure | $\|\Delta pK_a$ (X-ray)$\|$ | $\|\Delta pK_a$ (complex$_{II}$)$\|$ |
|---|---|---|
| best threshold | 0.100 | 0.100 |
| sensitivity | 0.833 | 1.000 |
| specificity | 0.900 | 1.000 |
| accuracy | 0.885 | 1.000 |
| precision | 0.714 | 1.000 |
| balanced accuracy | 0.867 | 1.000 |

[a]Data obtained from $|\Delta pK_a(\text{X-ray})|$ (using p$K_a$ data Sets A and B, see Table 1) and $|\Delta pK_a(\text{complex}_{II})|$ (using p$K_a$ data Sets B and C, see Table 1). This data corresponds to Cases I and II, respectively, for the system $NS1_{WNV(176-352)}$-Ab. $SASA_{ionizable}$ and $EE_I$ were used as references for them, respectively.

Table 4 summarizes the best thresholds obtained for Cases I and II together with the corresponding main statistical descriptors. This data reinforces that is possible to use the p$K_a$ shifts (BAC = 0.867) as a descriptor of EE and that the complexation MC simulations can correctly map *all* titratable amino acids affected by the presence of the Ab as in the crystallographic structure (BAC = 1.000).

Additionally, we challenged the electrostatic method to verify its capacity to reproduce the EE given by $SASA_{ionizable}$ referential. The purpose of this analysis is to compare the predictions from the MC simulations of the complexation process (and the resulting p$K_a$ changes) with the EE defined by the crystallographic data of the complex. This corresponds to case III in the methodology section, in which the two proteins are free to rotate, translate, and titrate. Thus, we are assuming an unknown interface for case III. In fact, the MC simulation will reveal it during the sampling. For this critical test, we used the p$K_a$s from simulation Sets A and C (see Table 1) to calculate $|\Delta pK_a(\text{complex}_{III})|$. These $\Delta pK_a$s are smaller compared to Case I because the two proteins are freely exploring the space during the MC trial movements, see Table S2. It is expected that the number of titratable residues affected by the presence of Ab when exploring the complexation pathway (during the MC sampling) should be higher than what is observed at the rigid crystallographic conformation. This comparison in terms of a ROC plot is shown in Figure 7c. As anticipated, the performance of the electrostatic method decreases (AUC = 0.50) under this condition. Among the bioinformatics tools that solely lie on the 3D structure of the antigen, SEPPA3 and DiscoTope are better than all others. They have their AUCs equal to 0.71 and 0.68, respectively. These three methods are better binary classifiers for EEs than the other bioinformatic tools [SEPPA2 (AUC = 0.46) and ElliPro (AUC = 0.12)] assuming the $SASA_{ionizable}$ referential. Conversely, it should be remembered that the electrostatic method is more appropriate to map *all* titratable groups related to the interaction with the Ab (see Table 4).

Interestingly, we can note a higher specificity for the PROCEEDp$K_a$ when compared to the servers. In terms of accuracy, our electrostatic predictor performs well and, in

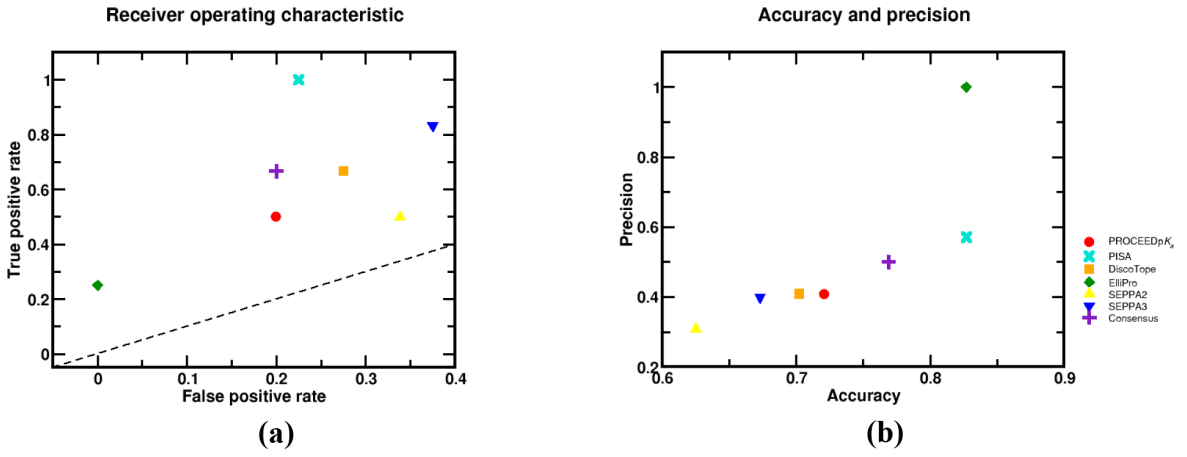



**Figure 9.** Plots of the (a) receiver operating characteristic analysis and (b) the precision against accuracy for the most important theoretical methods. Data for the system $NS1_{WNV(176-352)}$. $SASA_{ionizable}$ is used as a reference. Plots are showing the best thresholds for each algorithm (data is summarized in Table 4).



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

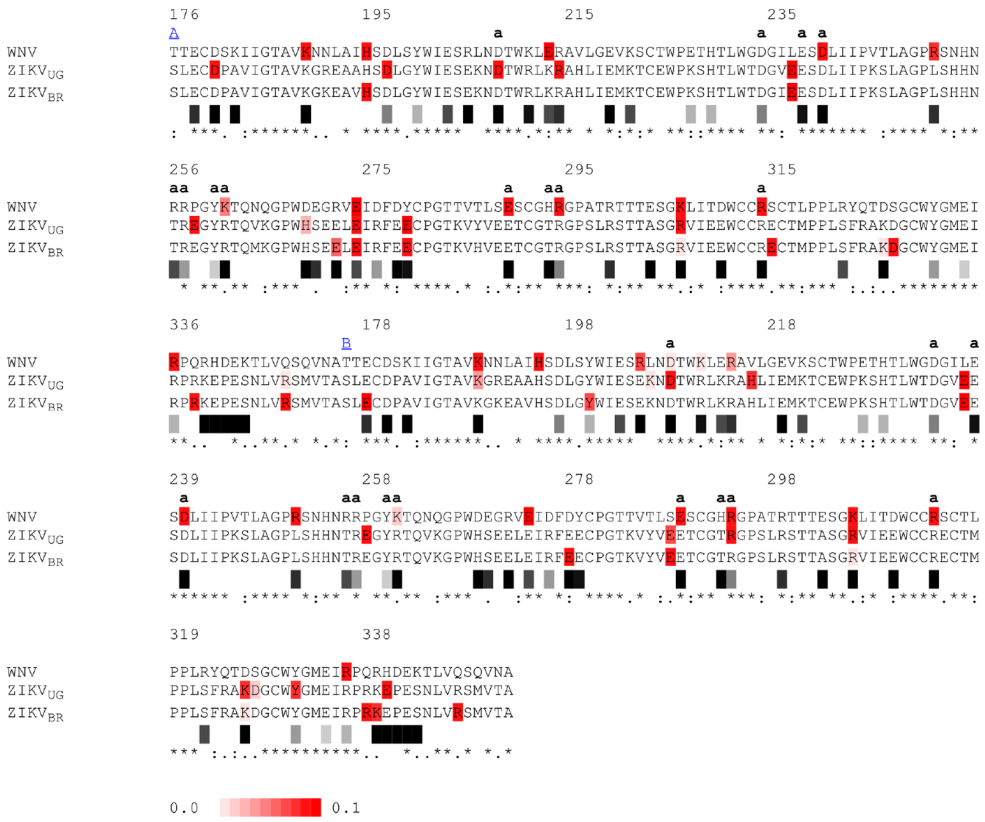



**Figure 10.** Primary sequences of the dimeric NS1$_{WNV(176–352)}$, NS1$_{ZIKV\text{-}UG(176–352)}$, and NS1$_{ZIKV\text{-}BR(176–352)}$ with the estimated antigenic regions by the electrostatic method. Data obtained using the threshold $|\Delta pK_a(\text{complex}_{III}|$>0.007. The p$K_a$s were computed at simulation Sets C, D, and E. The letters "a" on the top of the alignment means the SASA$_{ionizable}$ for NS1$_{WNV(176–352)}$. Symbols below are the degree of amino acid conservation. The same color scale from the previous figures is used here.

terms of precision, our data is comparable to the best bioinformatic servers, DiscoTope and SEPPA3. A summary of all these measures is presented in Figure 9. Using the best thresholds for each algorithm, a lower FPR is given for our data (see Figure 9a). The FPR values of our data are almost in line with the consensus. In terms of TPR, using the p$K_a$ shifts the data is enough to fall in a reliable position in the plot. Observing the plot of accuracy and precision (see Figure 9b), our data are best performed in terms of accuracy when compared to the bioinformatic prediction tools (with the exception of ElliPro). Furthermore, in terms of precision, the data from $|\Delta pK_a(\text{complex}_{III}|$ are competing with DiscoTope but is less efficient when compared to ElliPro. The consensus proves to be an effective strategy for the optimization of predictive tools when the performance is measured by accuracy and precision, for instance.

The data for Case III using $|\Delta pK_a(\text{complex}_{III})| > 0.007$ are graphically represented at the sequence and three-dimensional levels in Figures 5 and 8, respectively. The sequence analysis discussed above revealed how the amino acids E208, E240, K261, E289, R294, and R314 of the two protein chains are satisfactorily predicted. Most of them have a $|\Delta pK_a(\text{complex}_{III})| \approx 0.08$ (with the exception of the residue E208 that have a $\Delta pK_a$ close to the threshold and K261 in chain B with $|\Delta pK_a(\text{complex}_{III})| \approx 0.01$). An apparent symmetry between the two chains can be observed at both structural levels. The remaining ionizable residues fall outside of SASA$_{ionizable}$(FP = 16 amino acids). In general, most of the residues are on the surface or close to it as shown in Figure 8. Nevertheless, there are some amino acids buried in the protein interior that suffers an electrostatic perturbation like H195 in the two chains of NS1$_{WNV(176–352)}$ (compare the black bars of the other ionizable residues in the alignment in Figure 5, for the H195s. The bar is white indicating that they are internal amino acids). This effect can occur due to the long-range characteristics of the electrostatic interactions and the electrostatic coupling between the titratable sites as argued above. At the surface level (see Figure 8), there are several titratable groups distributed all over the protein surface. These residues are colored in blue (basic amino acids) and red (acid amino acids) in Figure 8. The SASA$_{ionizable}$ representation indicates in green the titratable groups defined as EE. The EE$_I$ predicted interface is included in this host–pathogen PP interface. There is virtually no perturbation (represented in red) in the opposite biological interface. Contrarily, when both macromolecules are free to translate and rotate exploring the association pathway (Case III), they can eventually visit and come closer to the opposite side of the PP interface and consequently alter the electrostatics properties of any titratable residues (see Figure 8, right and bottom, opposite face to the PP interface), which may contribute in increasing the number of false positive. In the present situation, we must remember that the used NS1$_{WNV}$ structure in these calculations is missing the N-terminus region. In the complete structure, the $\beta$-roll and a good part of the wing domains of NS1, constituting the N-terminus, occupy this PP opposite side. Thus, when this



DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944–963

missing structural part is included, the ROC analysis for the simulation Set H (using $SASA_{ionizable}$ as reference) reveals that the AUC increases (0.58) compared to AUC in Case III which is 0.5. This indicates, in fact, that the missing structural part influences the predictions.

Additional tests were also performed aiming to improve the accuracy of the electrostatic method. For instance, MC simulations with $\varepsilon_{LJ}$ equal to 0.1734 kJ/mol resulted in an AUC value equal to 0.48. For this reason, we kept the commonly used value of 0.124 kJ/mol[18,20] in all calculations.

**Electrostatic Epitopes for ZIKV NS1 by p$K_a$ Shifts.** So far, we have established that it is possible to determine epitope regions for $NS1_{WNV}$ by p$K_a$ shifts using the PROCEEDp$K_a$ approach. Using an absolute value of p$K_a$ greater than 0.007, ionizable residues can be identified as an EE on $NS1_{WNV(176-352)}$. Our electrostatic method is sensitive to the used Ab structure and its ionizable groups that can produce the electrical perturbations on the titratable groups of the antigen. For ZIKV, there is no NS1-Ab structure available. Although there is no experimental proof that this Fab fragment 22NS1 Ab can interact with the ZIKV NS1 protein, this is an interesting system to explore as argued before,[11] which might also bring light on putative cross-reactivity mechanisms. Thus, for this part of the study, we assumed that this specific Ab is able to interact with ZIKV and searched for its ZIKV EE recognition site. Moreover, the comparison between $NS1_{ZIKV\text{-}UG(176-352)}$ and $NS1_{ZIKV\text{-}BR(176-352)}$ is valid in any situation and can shed light in the understanding of their molecular mechanisms.[11]

The sequences alignments and the 3D representations of the $NS1_{WNV(176-352)}$ and the $NS1_{ZIKV(176-352)}$ of the two ZIKV strains are shown in Figures 10 and 11, respectively. We initially observe that the degree of conservation of each NS1 flaviviruses affects the potential EEs that they may have. Therefore, each system exhibits a different pattern of antigenic ionizable residues. Even observable differences between the two strains of ZIKV with an identity greater than 97% show a profile of EE not similar. Although there is no experimental guarantee that the 22NS1 can recognize ZIKV, the p$K_a$ shifts supports the hypothesis mentioned in our previous study.[11] Thus, we observed again how electrostatic properties impact in a transcendental way the differential interactions that these two proteins $NS1_{ZIKV}$ can have with a partner (as an Ab).

The total number of titratable amino acids that fulfill the criterion of $|\Delta pK_a(complex_{III})| > 0.007$ for $NS1_{WNV(176-352)}$, $NS1_{ZIKV\text{-}UG(176-352)}$, and $NS1_{ZIKV\text{-}BR(176-352)}$ is, respectively, 28 (ASP = 4, ARG = 10, GLU = 5, HIS = 2, LYS = 7), 24 (ASP = 5, ARG = 5, GLU = 8, HIS = 2, LYS = 3, TYR = 1), and 21 (ASP = 1, ARG = 6, GLU = 9, HIS = 1, LYS = 3, TYR = 1). The second protein with most predicted EEs is $NS1_{ZIKV\text{-}UG(176-352)}$, with $NS1_{ZIKV\text{-}BR(176-352)}$, being the one with less ionizable amino acids involved in the putative antigenic region. This could imply that the Brazilian strain would be less detected by this specific Ab. Based on our previous results using a semiquantitative theory approach,[11] the same order was observed for the attraction in the free energy of interactions for these NS1 proteins and this Ab. All these outcomes confirm the possibility for the cross-reactivity between these NS1 proteins with the Fab fragment 22NS1.

Finally, for the complete dimeric $NS1_{WNV}$ protein, we see how the presence of the missing part of the structure significantly influences the detection of EEs using the electrostatic method compared to the incomplete version

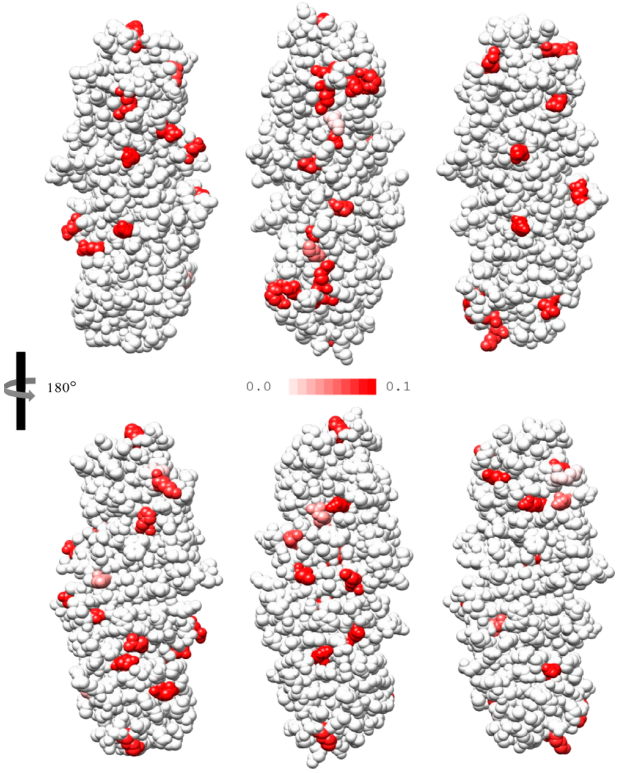



**Figure 11.** From left to right estimates of the EE by p$K_a$ shifts as a descriptor for proteins $NS1_{WNV(176-352)}$ (PDB ID 4OII), $NS1_{ZIKV\text{-}UG(176-352)}$ (PDB ID 5K6K), and $NS1_{ZIKV\text{-}BR(176-352)}$ (PDB ID 5GS6). EE were computed using a threshold: $|\Delta pK_a(complex_{III})| > 0.007$. Images in the top correspond to the antibody interface.

$NS1_{WNV(176-352)}$ (Figure S2). However, this also reveals possible target regions that could work for the recognition of the human immune system for future clinical and practical applications. As a final point, compared to the classical bioinformatics tools, the procedure we propose looks efficient not only to predict the epitope region but also the way the antibody might interact.

## CONCLUSIONS

In the present study, we have discussed the difficulties that the most commonly used bioinformatic tools can have to predict B-cell epitopes for flaviviruses. We have also introduced a broader view of epitopes from an electrostatic point of view where inner titratable groups can also participate in the interplay of interactions with Abs. These "electrostatic epitopes" were described for the NS1 protein of WNV and its Fab fragment of 22NS1 at different situations.

By means of statistical binary classifiers, we highlight that, in fact, p$K_a$ shifts can work as an alternative in the understanding and prediction of epitope regions especially the electrostatic epitopes for flaviviruses. This undoubtedly shows a novel way of addressing problems related to prediction algorithms. Our PROCEEDp$K_a$ descriptor based on p$K_a$ shifts was tested on $NS1_{WNV}$ and applied on $NS1_{ZIKV}$ proteins from two strains. An additional advantage was the possibility to test *specific* Abs. This allowed us to confirm the possibility of cross-reactivity and to find a higher antigenic capacity for the Ugandan strain compared to the Brazilian one. This result shows a high correlation with previous studies, opening a new path for the


DOI: 10.1021/acs.jcim.9b00895
J. Chem. Inf. Model. 2020, 60, 944−963

development of diagnostic strategies where electrostatic properties are considered to describe more realistically the nature and to contribute to obtain better predictive tools.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00895.

Additional details of the evaluation of efficiency and performance of different bioinformatics tools in predicting NS1 epitope and absolute values of the p$K_a$ shifts for each type of ionizable residue (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

*E-mail: fernando@fcfrp.usp.br.

### ORCID

Fernando L. Barroso da Silva: 0000-0003-2526-2085


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work has been supported in part by the "Fundação de Amparo à Pesquisa do Estado de São Paulo" [Grants Fapesp 2015/16116-3 and 2019/06139-7 (F.L.B.d.S.)], the Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) (F.L.B.d.S.), and the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior"—CAPES (S.A.P.-C.). F.L.B.d.S. and S.A.P.-C. acknowledge the support of the University of São Paulo (USP) through the NAP-CatSinQ (Research Core in Catalysis and Chemical Synthesis) and the computing hours at HPC/USP. F.L.B.d.S. is deeply thankful to the large use of computing hours at The Swedish National Infrastructure for Computing (Grant SNIC 2019/1-32). F.L.B.d.S. and C.E. acknowledge the mobility support from the joint Universitè Sorbonne Paris Citè-USP Research Program and the Print USP/Capes (process 88887.370314/2019).

## ■ REFERENCES


(1) Akey, D. L.; Brown, W. C.; Dutta, S.; Konwerski, J.; Jose, J.; Jurkiw, T. J.; DelProposto, J.; Ogata, C. M.; Skiniotis, G.; Kuhn, R. J.; Smith, J. L. Flavivirus NS1 Structures Reveal Surfaces for Associations with Membranes and the Immune System. *Science* **2014**, *343*, 881−885.

(2) Cohen, J. Where Has All the Zika Gone? *Science* **2017**, *357*, 631−632.

(3) Heinz, F. X.; Stiasny, K. The Antigenic Structure of Zika Virus and Its Relation to Other Flaviviruses: Implications for Infection and Immunoprophylaxis. *Microbiol. Mol. Biol. Rev.* **2017**, *81*, e00055-16.

(4) Lindenbach, B. D.; Rice, C. M. Molecular Biology of Flaviviruses. *Adv. Virus Res.* **2003**, *59*, 23−61.

(5) Ortiz-Baez, A. S.; Cunha, M. M. D. P.; Vedovello, D.; Colombo, T. E.; Nogueira, M. L.; Villabona-Arenas, C. J.; Zanotto, P. M. de A. Origin, Tempo, and Mode of the Spread of DENV-4 Genotype IIB across the State of São Paulo, Brazil during the 2012−2013 Outbreak. *Mem. Inst. Oswaldo Cruz* **2019**, *114*, e180251.

(6) Villabona-Arenas, C. J.; de Oliveira, J. L.; Capra, C. d. S.; Balarini, K.; Loureiro, M.; Fonseca, C. R. T. P.; Passos, S. D.; Zanotto, P. M. d. A. de A. Detection Of Four Dengue Serotypes Suggests Rise In Hyperendemicity In Urban Centers Of Brazil. *PLoS Neglected Trop. Dis.* **2014**, *8*, No. e2620.

(7) Ekins, S.; Perryman, A. L.; Horta Andrade, C. OpenZika: An IBM World Community Grid Project to Accelerate Zika Virus Drug Discovery. *PLoS Neglected Trop. Dis.* **2016**, *10*, No. e0005023.

(8) Mottin, M.; Braga, R. C.; da Silva, R. A.; Silva, J. H. M. da; Perryman, A. L.; Ekins, S.; Andrade, C. H. Molecular Dynamics Simulations of Zika Virus NS3 Helicase: Insights into RNA Binding Site Activity. *Biochem. Biophys. Res. Commun.* **2017**, *492*, 643−651.

(9) Mesci, P.; Macia, A.; LaRock, C. N.; Tejwani, L.; Fernandes, I. R.; Suarez, N. A.; de A. Zanotto, P. M.; Beltrão-Braga, P. C. B.; Nizet, V.; Muotri, A. R. Modeling Neuro-Immune Interactions during Zika Virus Infection. *Hum. Mol. Genet.* **2018**, *27*, 41−52.

(10) Rocha, L.; Alves, R.; Caetano, B.; Pereira, L.; Mitsunari, T.; Amorim, J.; Polatto, J.; Botosso, V.; Gallina, N.; Palacios, R.; Precioso, A.; Granato, C.; Oliveira, D.; Silveira, V.; Luz, D.; Ferreira, L.; Piazza, R. Epitope Sequences in Dengue Virus NS1 Protein Identified by Monoclonal Antibodies. *Antibodies* **2017**, *6*, 14.

(11) Poveda-Cuevas, S. A.; Etchebest, C.; Barroso da Silva, F. L. Insights into the ZIKV NS1 Virology from Different Strains through a Fine Analysis of Physicochemical Properties. *ACS Omega* **2018**, *3*, 16212−16229.

(12) Freire, M. C. L. C.; Pol-Fachin, L.; Coelho, D. F.; Viana, I. F. T.; Magalhaes, T.; Cordeiro, M. T.; Fischer, N.; Loeffler, F. F.; Jaenisch, T.; Franca, R. F.; Marques, E. T. A.; Lins, R. D. Mapping Putative B-Cell Zika Virus NS1 Epitopes Provides Molecular Basis for Anti-NS1 Antibody Discrimination between Zika and Dengue Viruses. *ACS Omega* **2017**, *2*, 3913−3920.

(13) Machado, M. R.; González, H. C.; Pantano, S. MD Simulations of Viruslike Particles with Supra CG Solvation Affordable to Desktop Computers. *J. Chem. Theory Comput.* **2017**, *13*, 5106−5116.

(14) Muller, D. A.; Young, P. R. The Flavivirus NS1 Protein: Molecular and Structural Biology, Immunology, Role in Pathogenesis and Application as a Diagnostic Biomarker. *Antiviral Res.* **2013**, *98*, 192−208.

(15) Watterson, D.; Modhiran, N.; Young, P. R. The Many Faces of the Flavivirus NS1 Protein Offer a Multitude of Options for Inhibitor Design. *Antiviral Res.* **2016**, *130*, 7−18.

(16) Young, P. R.; Hilditch, P. A.; Bletchly, C.; Halloran, W. An Antigen Capture Enzyme-Linked Immunosorbent Assay Reveals High Levels of the Dengue Virus Protein NS1 in the Sera of Infected Patients. *J. Clin. Microbiol.* **2000**, *38*, 1053−1057.

(17) Barroso da Silva, F. L.; Pasquali, S.; Derreumaux, P.; Dias, L. G. Electrostatics Analysis of the Mutational and pH Effects of the N-Terminal Domain Self-Association of the Major Ampullate Spidroin. *Soft Matter* **2016**, *12*, 5600−5612.

(18) Delboni, L.; Barroso da Silva, F. L. On the Complexation of Whey Proteins. *Food Hydrocolloids* **2016**, *55*, 89−99.

(19) Barroso da Silva, F. L.; Derreumaux, P.; Pasquali, S. Protein-RNA Complexation Driven by the Charge Regulation Mechanism. *Biochem. Biophys. Res. Commun.* **2018**, *498*, 264−273.

(20) Persson, B. A; Lund, M.; Forsman, J.; Chatterton, D. E.W.; Akesson, T. Torbjörn Akesson. Molecular Evidence of Stereo-Specific Lactoferrin Dimers in Solution. *Biophys. Chem.* **2010**, *151*, 187−189.

(21) Wade, R. C.; Gabdoulline, R. G.; Lüdemann, S. K.; Lounnas, V. Electrostatic Steering and Ionic Tethering in Enzyme-Ligand Binding: Insights from Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 5942−5949.

(22) Ramaraj, T.; Angel, T.; Dratz, E. A.; Jesaitis, A. J.; Mumey, B. Antigen-Antibody Interface Properties: Composition, Residue Interactions, and Features of 53 Non-Redundant Structures. *Biochim. Biophys. Acta, Proteins Proteomics* **2012**, *1824*, 520−532.

(23) O'Kennedy, R.; Fitzgerald, S.; Murphy, C. Don't Blame It All on Antibodies - The Need for Exhaustive Characterisation, Appropriate Handling, and Addressing the Issues That Affect Specificity. *TrAC, Trends Anal. Chem.* **2017**, *89*, 53−59.

(24) Saper, C. B. A Guide to the Perplexed on the Specificity of Antibodies. *J. Histochem. Cytochem.* **2009**, *57*, 1−5.

(25) Berzofsky, J. A.; Schechter, A. N. The Concepts of Crossreactivity and Specificity in Immunology. *Mol. Immunol.* **1981**, *18*, 751−763.




DOI: 10.1021/acs.jcim.9b00895
*J. Chem. Inf. Model.* 2020, 60, 944−963

(26) Hansen, S.; Hotop, S.-K.; Faye, O.; Ndiaye, O.; Bohlken-Fascher, S.; Pessoa, R.; Hufert, F.; Stahl-Hennig, C.; Frank, R.; Cherny, C.-P.; Schmidt-Chanasit, J.; Sanabani, S.S.; Sall, A.; Niedrig, M.; Bronstrup, M.; Fritz, H.-J.; Wahed, A. A. E. Elucidation of Specific and Cross Reactive Zika Virus Antibody Target Regions. *Int. J. Infect. Dis.* **2019**, *79*, 33.
(27) Reverberi, R.; Reverberi, L. Factors Affecting the Antigen-Antibody Reaction. *Blood Transfus.* **2007**, *5*, 227−240.
(28) Jiang, S.; Du, L. Advances in the Research and Development of Therapeutic Antibodies against the Zika Virus. *Cell. Mol. Immunol.* **2019**, *16* (1), 96−97.
(29) Van Regenmortel, M. H. V. Specificity, Polyspecificity, and Heterospecificity of Antibody-Antigen Recognition. *J. Mol. Recognit.* **2014**, *27*, 627−639.
(30) Yoshida, K.; Kuroda, D.; Kiyoshi, M.; Nakakido, M.; Nagatoishi, S.; Soga, S.; Shirai, H.; Tsumoto, K. Exploring Designability of Electrostatic Complementarity at an Antigen-Antibody Interface Directed by Mutagenesis, Biophysical Analysis, and Molecular Dynamics Simulations. *Sci. Rep.* **2019**, *9*, 4482.
(31) Dejnirattisai, W.; Supasa, P.; Wongwiwat, W.; Rouvinski, A.; Barba-Spaeth, G.; Duangchinda, T.; Sakuntabhai, A.; Cao-Lormeau, V.-M.; Malasit, P.; Rey, F. A.; Mongkolsapaya, J.; Screaton, G. R. Dengue Virus Sero-Cross-Reactivity Drives Antibody-Dependent Enhancement of Infection with Zika Virus. *Nat. Immunol.* **2016**, *17*, 1102−1108.
(32) Malone, R. W.; Homan, J.; Callahan, M. V.; Glasspool-Malone, J.; Damodaran, L.; Schneider, A. D. B.; Zimler, R.; Talton, J.; Cobb, R. R.; Ruzic, I.; Smith-Gagen, J.; Janies, D.; Wilson, J. Zika Virus Medical Countermeasure Development Challenges. *PLoS Neglected Trop. Dis.* **2016**, *10*, No. e0004530.
(33) Henchal, E. A.; Putnak, J. R. The Dengue Viruses. *Clin. Microbiol. Rev.* **1990**, *3*, 376−396.
(34) Kim, C. R.; Counotte, M.; Bernstein, K.; Deal, C.; Mayaud, P.; Low, N.; Broutet, N. Investigating the Sexual Transmission of Zika Virus. *Lancet Glob. Health* **2018**, *6*, No. e24.
(35) Tai, W.; Voronin, D.; Chen, J.; Bao, W.; Kessler, D. A.; Shaz, B.; Jiang, S.; Yazdanbakhsh, K.; Du, L. Transfusion-Transmitted Zika Virus Infection in Pregnant Mice Leads to Broad Tissue Tropism With Severe Placental Damage and Fetal Demise. *Front. Microbiol.* **2019**, *10*, 29.
(36) Musso, D.; Roche, C.; Nhan, T.-X.; Robin, E.; Teissier, A.; Cao-Lormeau, V.-M. Detection of Zika Virus in Saliva. *J. Clin. Virol.* **2015**, *68*, 53−55.
(37) Mansuy, J.-M.; Lhomme, S.; Cazabat, M.; Pasquier, C.; Martin-Blondel, G.; Izopet, J. Detection of Zika, Dengue and Chikungunya Viruses Using Single-Reaction Multiplex Real-Time RT-PCR. *Diagn. Microbiol. Infect. Dis.* **2018**, *92*, 284−287.
(38) Sharp, T. M.; Fischer, M.; Muñoz-Jordán, J. L.; Paz-Bailey, G.; Staples, J. E.; Gregory, C. J.; Waterman, S. H. Dengue and Zika Virus Diagnostic Testing for Patients with a Clinically Compatible Illness and Risk for Infection with Both Viruses. *MMWR Recomm. Rep.* **2019**, *68*, 1−10.
(39) Ponomarenko, J.; Bui, H.-H.; Li, W.; Fusseder, N.; Bourne, P. E.; Sette, A.; Peters, B. ElliPro a New Structure-Based Tool for the Prediction of Antibody Epitopes. *BMC Bioinf.* **2008**, *9*, 514.
(40) Sun, J.; Wu, D.; Xu, T.; Wang, X.; Xu, X.; Tao, L.; Li, Y. X.; Cao, Z. W. SEPPA: A Computational Server for Spatial Epitope Prediction of Protein Antigens. *Nucleic Acids Res.* **2009**, *37* (suppl_2), W612−W616.
(41) Fiorucci, S.; Zacharias, M. Computational Antigenic Epitope Prediction by Calculating Electrostatic Desolvation Penalties of Protein Surfaces. *Methods Mol. Biol.* **2014**, *1184*, 365−374.
(42) Haste Andersen, P.; Nielsen, M.; Lund, O. Prediction of Residues in Discontinuous B-Cell Epitopes Using Protein 3D Structures. *Protein Sci.* **2006**, *15*, 2558−2567.
(43) Wang, J.; Cao, H.; Zhang, J. Z. H.; Qi, Y. Computational Protein Design with Deep Learning Neural Networks. *Sci. Rep.* **2018**, *8*, 6349.
(44) Long, X.; Jeliazkov, J. R.; Gray, J. J. Non-H3 CDR Template Selection in Antibody Modeling through Machine Learning. *PeerJ* **2019**, *7*, No. e6179.
(45) Zhao, J.; Nussinov, R.; Wu, W.-J.; Ma, B. In Silico Methods in Antibody Design. *Antibodies* **2018**, *7*, 22.
(46) Sormanni, P.; Aprile, F. A.; Vendruscolo, M. Third Generation Antibody Discovery Methods: *In Silico* Rational Design. *Chem. Soc. Rev.* **2018**, *47*, 9137−9157.
(47) Fischman, S.; Ofran, Y. Computational Design of Antibodies. *Curr. Opin. Struct. Biol.* **2018**, *51*, 156−162.
(48) Marcatili, P.; Olimpieri, P. P.; Chailyan, A.; Tramontano, A. Antibody Modeling Using the Prediction of Immunoglobulin Structure (PIGS) Web Server [Corrected]. *Nat. Protoc.* **2014**, *9*, 2771−2783.
(49) Dunbar, J.; Krawczyk, K.; Leem, J.; Marks, C.; Nowak, J.; Regep, C.; Georges, G.; Kelm, S.; Popovic, B.; Deane, C. M. SAbPred: A Structure-Based Antibody Prediction Server. *Nucleic Acids Res.* **2016**, *44* (W1), W474−W478.
(50) Weitzner, B. D.; Jeliazkov, J. R.; Lyskov, S.; Marze, N.; Kuroda, D.; Frick, R.; Adolf-Bryfogle, J.; Biswas, N.; Dunbrack, R. L.; Gray, J. J. Modeling and Docking Antibody Structures with Rosetta. *Nat. Protoc.* **2017**, *12*, 401−416.
(51) Leem, J.; Dunbar, J.; Georges, G.; Shi, J.; Deane, C. M. A Body Builder: Automated Antibody Structure Prediction with Data-Driven Accuracy Estimation. *mAbs* **2016**, *8*, 1259−1268.
(52) Pantazes, R. J.; Maranas, C. D. OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng., Des. Sel.* **2010**, *23*, 849−858.
(53) Li, T.; Pantazes, R. J.; Maranas, C. D. OptMAVEn - A New Framework for the de novo Design of Antibody Variable Region Models Targeting Specific Antigen Epitopes. *PLoS One* **2014**, *9*, e105954.
(54) 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. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 10900−10905.
(55) Adolf-Bryfogle, J.; Kalyuzhniy, O.; Kubitz, M.; Weitzner, B. D.; Hu, X.; Adachi, Y.; Schief, W. R.; Dunbrack, R. L. RosettaAntibodyDesign (RAbD): A General Framework for Computational Antibody Design. *PLoS Comput. Biol.* **2018**, *14*, e1006112.
(56) Pires, D. E.; Ascher, D. B.; Blundell, T. L. MCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. *Bioinformatics* **2014**, *30*, 335−342.
(57) Nimrod, G.; Fischman, S.; Austin, M.; Herman, A.; Keyes, F.; Leiderman, O.; Hargreaves, D.; Strajbl, M.; Breed, J.; Klompus, S.; Minton, K.; Spooner, J.; Buchanan, A.; Vaughan, T. J.; Ofran, Y. Computational Design of Epitope-Specific Functional Antibodies. *Cell Rep.* **2018**, *25*, 2121−2131.
(58) Jespersen, M. C.; Mahajan, S.; Peters, B.; Nielsen, M.; Marcatili, P. Antibody Specific B-Cell Epitope Predictions: Leveraging Information From Antibody-Antigen Protein Complexes. *Front. Immunol.* **2019**, *10*, 298.
(59) Liang, S.; Zheng, D.; Zhang, C.; Zacharias, M. Prediction of Antigenic Epitopes on Protein Surfaces by Consensus Scoring. *BMC Bioinf.* **2009**, *10*, 302.
(60) Van Regenmortel, M. H. V. Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity. *Methods* **1996**, *9*, 465−472.
(61) Yang, X.; Yu, X. An Introduction to Epitope Prediction Methods and Software. *Rev. Med. Virol.* **2009**, *19*, 77−96.
(62) Barroso da Silva, F. L.; Dias, L. G. Development of Constant-pH Simulation Methods in Implicit Solvent and Applications in Biomolecular Systems. *Biophys. Rev.* **2017**, *9*, 699−728.
(63) Barroso da Silva, F. L.; Sterpone, F.; Derreumaux, P. OPEP6: A New Constant pH Molecular Dynamics Simulation Scheme with OPEP Coarse-Grained Force Field. *J. Chem. Theory Comput.* **2019**, *15*, 3875−3888.


DOI: 10.1021/acs.jcim.9b00895
*J. Chem. Inf. Model.* 2020, 60, 944−963

(64) Chen, W.; Wallace, J. A.; Yue, Z.; Shen, J. K. Introducing Titratable Water to All-Atom Molecular Dynamics at Constant pH. *Biophys. J.* **2013**, *105*, L15−L17.
(65) Barroso da Silva, F. L. B. da; MacKernan, D. Benchmarking a Fast Proton Titration Scheme in Implicit Solvent for Biomolecular Simulations. *J. Chem. Theory Comput.* **2017**, *13*, 2915−2929.
(66) Chen, Y.; Roux, B. Constant pH Hybrid Nonequilibrium Molecular Dynamics-Monte Carlo Simulation Method. *J. Chem. Theory Comput.* **2015**, *11*, 3919−3931.
(67) Teixeira, A. A. R.; Lund, M.; Barroso da Silva, F. L. Fast Proton Titration Scheme for Multiscale Modeling of Protein Solutions. *J. Chem. Theory Comput.* **2010**, *6*, 3259−3266.
(68) Kurut, A.; Henriques, J.; Forsman, J.; Skepö, M.; Lund, M. Role of Histidine for Charge Regulation of Unstructured Peptides at Interfaces and in Bulk. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 657−667.
(69) Kurut, A.; Lund, M. Solution Electrostatics beyond pH: A Coarse Grained Approach to Ion Specific Interactions between Macromolecules. *Faraday Discuss.* **2013**, *160*, 271−278.
(70) Johnson, A. J.; Martin, D. A.; Karabatsos, N.; Roehrig, J. T. Detection of Anti-Arboviral Immunoglobulin G by Using a Monoclonal Antibody-Based Capture Enzyme-Linked Immunosorbent Assay. *J. Clin. Microbiol.* **2000**, *38*, 1827−1831.
(71) Martin, D. A.; Muth, D. A.; Brown, T.; Johnson, A. J.; Karabatsos, N.; Roehrig, J. T. Standardization of Immunoglobulin M Capture Enzyme-Linked Immunosorbent Assays for Routine Diagnosis of Arboviral Infections. *J. Clin. Microbiol.* **2000**, *38*, 1823−1826.
(72) Edeling, M. A.; Diamond, M. S.; Fremont, D. H. Structural Basis of Flavivirus NS1 Assembly and Antibody Recognition. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 4285−4290.
(73) Toride King, M.; Brooks, C. L. Epitope Mapping of Antibody-Antigen Interactions with X-Ray Crystallography. *Methods Mol. Biol.* **2018**, *1785*, 13−27.
(74) Loyau, J.; Didelot, G.; Malinge, P.; Ravn, U.; Magistrelli, G.; Depoisier, J.-F.; Pontini, G.; Poitevin, Y.; Kosco-Vilbois, M.; Fischer, N.; Thore, S.; Rousseau, F. Robust Antibody-Antigen Complexes Prediction Generated by Combining Sequence Analyses, Mutagenesis, In Vitro Evolution, X-Ray Crystallography and In Silico Docking. *J. Mol. Biol.* **2015**, *427*, 2647−2662.
(75) Simonin, Y.; Loustalot, F.; Desmetz, C.; Foulongne, V.; Constant, O.; Fournier-Wirth, C.; Leon, F.; Moles, J.-P.; Goubaud, A.; Lemaitre, J.-M.; Maquart, M.; Leparc-Goffart, I.; Briant, L.; Nagot, N.; Van de Perre, P.; Salinas, S. Zika Virus Strains Potentially Display Different Infectious Profiles in Human Neural Cells. *EBioMedicine* **2016**, *12*, 161−169.
(76) Simonin, Y.; van Riel, D.; Van de Perre, P.; Rockx, B.; Salinas, S. Differential Virulence between Asian and African Lineages of Zika Virus. *PLoS Neglected Trop. Dis.* **2017**, *11*, No. e0005821.
(77) Xu, X.; Song, H.; Qi, J.; Liu, Y.; Wang, H.; Su, C.; Shi, Y.; Gao, G. F. Contribution of Intertwined Loop to Membrane Association Revealed by Zika Virus Full-Length NS1 Structure. *EMBO J.* **2016**, *35*, 2170−2178.
(78) Brown, W. C.; Akey, D. L.; Konwerski, J. R.; Tarrasch, J. T.; Skiniotis, G.; Kuhn, R. J.; Smith, J. L. Extended Surface for Membrane Association in Zika Virus NS1 Structure. *Nat. Struct. Mol. Biol.* **2016**, *23*, 865−867.
(79) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(80) Kurut, A.; Persson, B. A.; Åkesson, T.; Forsman, J.; Lund, M. Anisotropic Interactions in Protein Mixtures: Self Assembly and Phase Behavior in Aqueous Solution. *J. Phys. Chem. Lett.* **2012**, *3*, 731−734.
(81) Mendonça, D. C.; Macedo, J. N.; Itri, R.; Guimarães, S. L.; Barroso da Silva, F. L.; Cassago, A.; Garratt, R. C.; Portugal, R.; Araujo, A. P. U. A Revised Order of Subunits in Mammalian Septin Complexes. *Cytoskeleton* **2019**, *76*, 457−466.
(82) Barroso da Silva, F. L.; Derreumaux, P.; Pasquali, S. Fast Coarse-Grained Model for RNA Titration. *J. Chem. Phys.* **2017**, *146*, 035101.
(83) Nozaki, Y.; Tanford, C. Examination of Titratation Behavior. *Methods Enzymol.* **1967**, *11*, 715−734.
(84) Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C.; Wilcox, W. Hydrophobic Moments and Protein Structure. *Faraday Symp. Chem. Soc.* **1982**, *17*, 109−120.
(85) Knowles, J. R. On the Mechanism of Action of Pepsin. *Philos. Trans. R. Soc., B* **1970**, *257*, 135−146.
(86) Srivastava, D.; Santiso, E.; Gubbins, K.; Barroso da Silva, F. L. Computationally Mapping pKa Shifts Due to the Presence of a Polyelectrolyte Chain around Whey Proteins. *Langmuir* **2017**, *33*, 11417−11428.
(87) Binder, K. Applications of the Monte Carlo Methods to Statistical Physics. *Rep. Prog. Phys.* **1997**, *60*, 487−559.
(88) Lund, M.; Trulsson, M.; Persson, B. Faunus: An Object Oriented Framework for Molecular Simulation. *Source Code Biol. Med.* **2008**, *3*, 1.
(89) Hubbard, S. J.; Thornton, J. M. *NACCESS, Computer Program*; Department of Biochemistry and Molecular Biology, University College: London, 1993.
(90) Borley, D. W.; Mahapatra, M.; Paton, D. J.; Esnouf, R. M.; Stuart, D. I.; Fry, E. E. Evaluation and Use of *In-Silico* Structure-Based Epitope Prediction with Foot-and-Mouth Disease Virus. *PLoS One* **2013**, *8*, No. e61122.
(91) Krissinel, E.; Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **2007**, *372*, 774−797.
(92) Zhou, C.; Chen, Z.; Zhang, L.; Yan, D.; Mao, T.; Tang, K.; Qiu, T.; Cao, Z. SEPPA 3.0—Enhanced Spatial Epitope Prediction Enabling Glycoprotein Antigens. *Nucleic Acids Res.* **2019**, *47* (W1), W388−W394.
(93) Prudkin-Silva, C.; Perez, O. E.; Martinez, K. D.; Mendonça, D. C.; Barroso da Silva, F. L. Combined Experimental and Molecular Simulation Study of Insulin-Chitosan Complexation Driven by Electrostatic Interactions. *J. Chem. Inf. Model.* **2019**, DOI: 10.1021/acs.jcim.9b00814.


DOI: 10.1021/acs.jcim.9b00895
*J. Chem. Inf. Model.* 2020, 60, 944−963

## Chapter 7

# NS1 self-association

In Chapter 5, the dimerization and tetramerization process of several $NS1_{ZIKV}$ from two phylogenetic origins was examined via DLVO. However, greater structural details and pH effects could influence the complexation phenomena of these proteins. In this Chapter, we specifically study the NS1 self-association through free energies using a CG model at the level of the amino acid considering various salt and pH concentrations. It is important to mention that the NS1s studied belong to four flaviviruses: ZIKV-UG, ZIKV-BR, DENV2, and WNV.

The main question of this part of the study is:

*Are the free energies of interaction for dimerizations and tetramerizations different at the intraspecific and interspecific levels?*

A summary of the molecular model used, the method to solve them, and the measures calculated appear in Table 7.1.

Table 7.1: Molecular modeling and measurements carried out in Chapter 7.

| Level of detail | Solution | Measurements |
|---|---|---|
| CG model at the amino acid level | CpH MC simulations | $Z_P$, $C_P$, $\mu_P$, free energies |

Link m/s (*BioRxiv*): https://www.biorxiv.org/content/10.1101/2022.03.15.484429v1.

*Manuscript submitted to the Virus Research journal (m/s ID: VIRUS-D-22-00203) on 15/03/2022.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Self-association features of NS1 proteins from different flaviviruses

Sergio A. Poveda-Cuevas [a,b,e] , Catherine Etchebest [d], Fernando L. Barroso da Silva [a,b,c,e*]

[a]Universidade de São Paulo, Programa Interunidades em Bioinformática, Rua do Matão, 1010, BR-05508-090 São Paulo, São Paulo, Brazil.

[b]Universidade de São Paulo, Departamento de Ciências Biomoleculares, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Av. do Café, s/no-Campus da USP, BR-14040-903 Ribeirão Preto, São Paulo, Brazil.

[c] Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27695, United States.

[d]Université Paris Cité, Biologie Intégrée du Globule Rouge, Equipe 2, INSERM, F-75015 Paris, France.

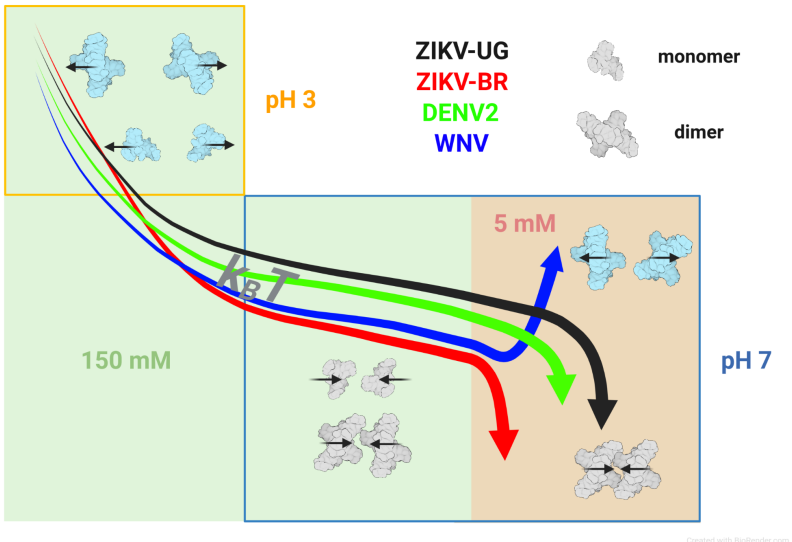
[e]University of São Paulo and Université de Paris International Laboratory in Structural Bioinformatics, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Av. do Café, s/no-Campus da USP, Bloco B, BR-14040-903 Ribeirão Preto, São Paulo, Brazil.

* Corresponding Author
E-mail: flbarroso@usp.br

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**ABSTRACT**

Flaviviruses comprise a large group of arboviral species that are distributed in several countries of the tropics, neotropics, and some temperate zones. Since they can produce neurological pathologies or vascular damage, there has been intense research seeking better diagnosis and treatments for their infections in the last decades. The flavivirus NS1 protein is a relevant clinical target because it is involved in viral replication, immune evasion, and virulence. Being a key factor in endothelial and tissue-specific modulation, NS1 has been largely studied to understand the molecular mechanisms exploited by the virus to reprogram host cells. A central part of the viral maturation processes is the NS1 oligomerization because many stages rely on these protein-protein assemblies. In the present study, the self-associations of NS1 proteins from Zika, Dengue, and West Nile viruses are examined through constant-pH coarse-grained biophysical simulations. Free energies of interactions were estimated for different oligomeric states and pH conditions. Our results show that these proteins can form both dimers and tetramers under conditions near physiological pH even without the presence of lipids. Moreover, pH plays an important role mainly controlling the regimes where van der Waals interactions govern their association. Finally, despite the similarity at the sequence level, we found that each flavivirus has a well-characteristic protein-protein interaction profile. These specific features can provide new hints for the development of binders both for better diagnostic tools and the formulation of new therapeutic drugs.




**Keywords:** Protein-protein interaction, free energy, non-structural protein 1, pH, Monte Carlo simulations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## 1. INTRODUCTION

The World Health Organization has reported in recent years viruses such as influenza H1N1, Chikungunya, Ebola, Lassa fever, MERS-CoV, Zika (ZIKV), Dengue (DENV), or West Nile (WNV) with epidemic or pandemic capacity (Trovato et al., 2020). The most recent outbreak by SARS-CoV-2 responsible for the Covid-19 pandemic, for example, has produced more than 452 million infections and more than 6 million deaths to date (World Health Organization, 2022). This ongoing crisis placed on high alert the health and economic worldwide systems, and reinforced the damage that these biological entities could have on a global scale. Through mutation processes, any virus can develop new variants resulting in discrepant pathological effects (Fields, 2014). In most cases, probably due to selective pressure, mutations can increase the viral mortality [as seen, for example, for Covid-19 or hemorrhagic dengue fever (DHF) by SARS-CoV-2 and DENV respectively (Gubler, 1998; Hu et al., 2021)] or can produce congenital damage [*e.g.*, microcephaly by ZIKV (Mlakar et al., 2016)] The concern may be even greater considering the space-time coexistence of different viral species (Beck et al., 2013; Bhatt et al., 2013; Plourde and Bloch, 2016). For instance, in Brazil, up to 10 different species of flaviviruses have been reported in the last decades (Figueiredo, 2000). Many of these species affect the health of urban populations and there are no totally efficient serological tests due to the heterotypic immune responses associated with flaviviruses (Musso and Desprès, 2020).

The *Flavivirus* genus is composed of +ssRNA viruses with a genome of ~11 kb that translates into a polyprotein. Through viral and host proteases the polyprotein is cleaved producing three structural proteins (prM, Env, and C) and seven non-structural (NS) proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5) (Neufeldt et al., 2018). The NS1 protein is involved in regulatory functions and different viral processes, such as replication, virulence, and pathogenicity. Consequently, NS1 is one of the most relevant clinical targets for flavivirus differentiation (Muller and Young, 2013; Watterson et al., 2016). It has also been suggested as a target for vaccine development due to the low risk of antibody-dependent enhancement (ADE) mechanism (Li et al., 2018; Sridhar et al., 2018; Wessel et al., 2021). Nevertheless, NS1 is well

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

known to produce autoantibodies, which can promote inflammation and/or apoptosis into the host (Muller and Young, 2013). The first stage of NS1 in the endoplasmic reticulum (ER) shows a tendency to form dimeric structures with more liposoluble characteristics compared to its monomeric state (Winkler et al., 1989). The dimeric NS1 can interact with membranes (mNS1) (Noisakran et al., 2008), where it later participates as a cofactor in early events of viral RNA replication, virion assembly, and particle release (Scaturro et al., 2015). On the other hand, another portion of mNS1 is trafficked through the Golgi apparatus where it is secreted as a hexamer (sNS1) to be ready to interact with the host's immune system (Flamand et al., 1999; Gutsche et al., 2011; Watterson et al., 2016). Yet, the NS1 dimer can also do so the same thing, showing that hexamerization is not a strict requirement for secretion (Watterson et al., 2016). Furthermore, the transition of a dimer to a hexamer implies the formation of tetrameric intermediates which are essential for the formation of sNS1 (Flamand et al., 1999). This viral toxin exhibits a high versatility that could be influencing the differential tropism among viral strains or species (*i.e.*, at the intraspecific or interspecific level, respectively). Inflammatory processes during DHF by DENV, the overcoming of the blood-brain or placental barriers (by the neurotrophic WNV or ZIKV, respectively) imply the presence of sNS1 as an important step in immune activation or endothelial damage (Basu and Chaturvedi, 2008; Jurado et al., 2016; Neal, 2014; Puerta-Guardo et al., 2019; Watterson et al., 2016).

Some experimental and computational studies have focused on understanding the oligomerization processes of NS1 proteins as a possible therapeutic strategy. Site-directed mutagenesis assays have shown that there are an attenuation and virulence reduction for DENV, ZIKV, Kunjin, and Murray Valley encephalitis viruses when the NS1 dimerization is affected (Clark et al., 2007; Hall et al., 1999; Lin et al., 1998; Wang et al., 2017). Electrophoresis studies by Falconar and Young (1990) found that dimerization for DENV2 occurs for pH regimes greater than 3 and less than 7.4, while tetramerizations were reported at pH 8 and 8.5, for the NS1 of DENV2 and DENV1, respectively (Flamand et al., 1999; Muller et al., 2012). Nevertheless, no data is yet available to access the effects from their differences at the protein sequence level. Also, it is not clear if the presence of lipids is a fundamental need (Winkler et al., 1989). To

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

deeply understand how flaviviruses work, it is essential to dissect this molecular mechanism for different species and strains.

Computer simulations are an interesting approach to unveil the phenomena that control the NS1 self-association. It allows testing specific steps at the molecular level under a well-controlled condition and without any risk of contamination and/or external interference typically at lower costs. Individual aspects of the interplay of several factors can be one by one tested. Molecular dynamics studies, for instance, have revealed that the β-hairpins of the β-roll domain are crucial for the interchain fit and stability of the NS1 dimers (Gonçalves et al., 2019; Menezes et al., 2020; Poveda-Cuevas et al., 2021). It has also been shown that the monomer-monomer association is guided by salt bridges or hydrogen bonds (Gonçalves et al., 2019), validating the importance of electrostatics in these biomolecular interactions. This also implies that pH is a thermodynamic variable that can trigger the association process (Poveda-Cuevas et al., 2020, 2018; Qiao and Olvera de la Cruz, 2020). Indeed, it is worth mentioning that the role of electrostatic interactions in complexation processes is not limited to flaviviruses. The phenomena produced by these interactions have been studied in other viral genders such as *Betacoronavirus* (Giron et al., 2021, 2020), *Chlorovirus* (Xian et al., 2019), *Bromovirus* (Suci et al., 2005), etc.

There is a differential electrostatic profile among flaviviruses despite their structural and sequence conservation (20%–40% identity and 60%–80% similarity) (Brown et al., 2016; Freire et al., 2017; Song et al., 2016; Xu et al., 2016). Interestingly, this also happens on Ugandan (UG) and Brazilian (BR) ZIKV variants from African and Asian origins, respectively. As previously shown via constant-pH (CpH) Monte Carlo (MC) simulations, the $pK_a$s of ionizable groups from the $NS1_{ZIKV\text{-}UG}$ suffer a displacement (or $\Delta pK_a$) when compared with values from strains phylogenetically newer (*i.e.*, $NS1_{ZIKV\text{-}BR}$) (Poveda-Cuevas et al., 2018). In spite of their similar sequences, titratable residues are affected in a specific way for each NS1 protein at their three-dimensional configuration (Poveda-Cuevas et al., 2020, 2018). As it can be anticipated, such physical feature has consequences on three biological interface regions responsible for the

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

*i*) oligomerization, *ii*) membrane binding, and *iii*) antibody interactions (Poveda-Cuevas et al., 2018).

The present work aims to study the self-association of four NS1 flaviviruses: a) $NS1_{ZIKV\text{-}UG}$, b) $NS1_{ZIKV\text{-}BR}$, c) $NS1_{DENV}$ (specifically variant 2, $NS1_{DENV2}$), and d) $NS1_{WNV}$ at different pH conditions. Through CpH coarse-grained (CG) models solved via MC simulations, free energy derivatives are calculated in order to characterize/describe, dimerization and tetramerization processes of the NS1 protein. The hexamerization process is not studied here due to its multifactorial contributions: different types of biomolecules such as lipids and sugars could be necessary to stabilize the three dimers NS1, auxiliary proteins could manage the hexameric formation, or the hexamers are generated during a lipid droplet process nascent in the membrane of the Golgi apparatus (Gutsche et al., 2011; Somnuke et al., 2011). Thus, this would increase the complexity of the models far from the computational capacities. Further reasons for the choice of only these two simpler oligomeric states (*i.e.*, dimers and tetramers) are deeply discussed in previous work (Poveda-Cuevas et al., 2018).

The present paper is organized as follows. In the next section, we describe the methodological aspects starting with the molecular models. Next, the numerical simulations used to solve the model and details of the different sets of calculations are described together with the main physical properties sampled. Subsequently, in the Results and Discussion section, the CG model is validated by comparing the theoretical outcomes with available electrophoresis and crystallography data. In the next step, the dimerization and tetramerization processes for the four NS1 proteins under relevant pHs and salt conditions were evaluated. Special attention was given to the self-association at the inter- and intraspecific levels. The last section aims to understand the physicochemical reasons that govern the NS1 oligomerization and the discrepancies among the flaviviruses. Lastly, how these factors could have a relationship with the biological aspects observed for each viral species is discussed.

## 2. METHODOLOGY

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 2.1. Structural modeling

Atomistic coordinates for the three-dimensional structures from $NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$ deposited in the RCSB Protein Data Bank (PDB) (Berman et al., 2000) were used as input for the simulations. X-ray structures are available at the PDB and identified with ids 5K6K (1.89Å, pH 8.5) (Brown et al., 2016), 5GS6 (2.85Å, pH 8.5) (Xu et al., 2016), 4O6B (3.00Å, pH 5.5), and 4O6D (2.60Å, pH 5.5) (Akey et al., 2014) for $NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$, respectively. The experimental resolution and pH conditions for each of these coordinates are the numbers given in parenthesis. The two $NS1_{ZIKV}$ protein models employed here were selected based on our previous study (Poveda-Cuevas et al., 2018). Missing residues in the loop regions were completed through an appropriate homology modeling protocol (Poveda-Cuevas et al., 2021, 2018). The crystallographic structures of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ are homodimers where one chain is fully resolved in the experiment and the other is not. Hence, the complete chain was used as a template to model the second chain with missing residues. The modeling was done through the "Modeller" program that is integrated into UCSF Chimera 1.11.2 (Pettersen et al., 2004). For each $NS1_{ZIKV}$ five models were built and the structure with the best DOPE was selected (Shen and Sali, 2006). Afterwards, the PDB files were edited before the simulation by deleting all water molecules and heteroatoms.

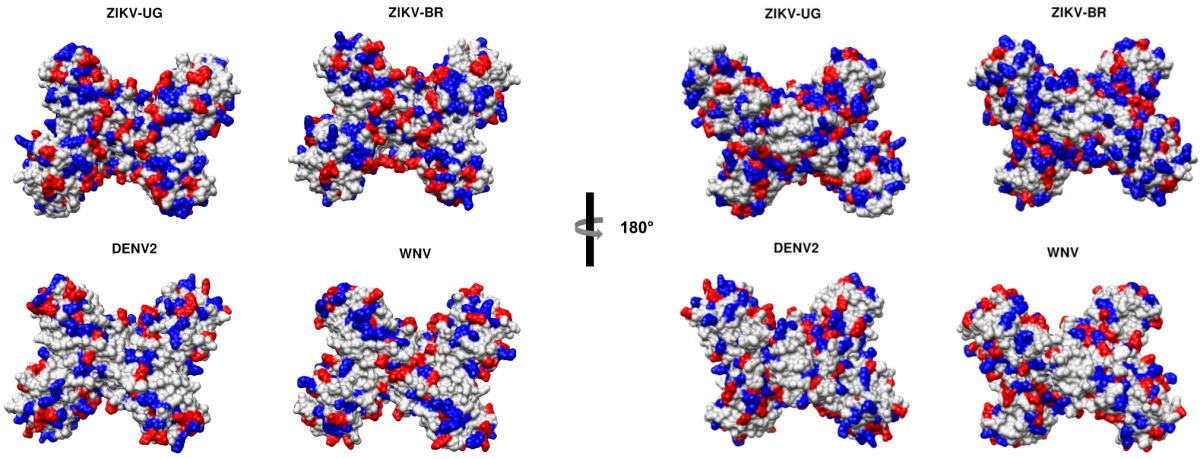



**Figure 1.** Molecular representation of the NS1 proteins (dimers) for the flaviviruses ZIKV-UG (PDB id 5K6K), ZIKV-BR (PDB id 5GS6), DENV2 (PDB id 4O6B), and WNV (PDB id 4O6D)

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in a surface representation. Positive and negatively charged amino acids are represented in blue and red, respectively.

The dimers $NS1_{DENV2}$ (PDB id 4O6B) and $NS1_{WNV}$ (PDB id 4O6D) also have missing residues in loop regions. The missing regions of the structure 4O6B (DENV2) are $8\text{-}10_{chain_A}$, $107\text{-}130_{chain_B}$, $108\text{-}128_{chain_A}$, $159\text{-}165_{chain_A}$, $162\text{-}163_{chain_B}$, and $350\text{-}352_{chain_A/B}$. Regarding the crystal given by the PDB id 4O6D (WNV), the missing parts are $108\text{-}128_{chain_A}$ and $108\text{-}128_{chain_B}$. Because both chains in these two structures are incomplete we can not follow the same strategy as implemented for the two $NS1_{ZIKV}$. For this part of the study, a more advanced homology modeling protocol was applied. This process involved multiple templates (Larsson et al., 2008). Thereby, we start from the query sequence of the dimeric $NS1_{DENV2}$ with the respective 704 amino acids. Later, three templates in the dimeric state were selected: *a)* the coordinates for $NS1_{DENV2}$ given by the PDB id 4O6B (with missing residues but 100% of identity sequence when compared to the query sequence and excluding gaps), *b)* the $NS1_{ZIKV\text{-}UG}$ (without any missing amino acids), and *c)* the $NS1_{ZIKV\text{-}BR}$ (without any missing residues). The two structures of items *b* and *c* have some differences, especially in some loop regions (see Figure S1) (Poveda-Cuevas et al., 2021). Since both atomic arrangements of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ are probable, they should be used simultaneously in order to increase the chance of obtaining structures for $NS1_{DENV2}$ with more favorable energies (Fernandez-Fuentes et al., 2007; Larsson et al., 2008). As a first stage in the homology modeling, the protein sequences of the three templates were extracted to make alignments with the query sequence. For this process, the python scripts "salign.py" and "align2d_mult.py" (available in https://salilab.org/modeller/) were employed. From the multiple alignment sequence, ten 3D structure models with complete parts were built with the python script "model_mult.py" (also available in the previous web URL). The 3D structural model with the best DOPE score was used for our simulations (Shen and Sali, 2006). Regarding $NS1_{WNV}$, the same strategy was followed as for $NS1_{DENV2}$. However, the structure of item *a)* was replaced by the uncompleted crystal coordinates given by PDB id 4O6D. All scripts were executed with Modeller 9.24 (Webb and Sali, 2014) and default parameters.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The molecular representation of all NS1 proteins after all modeling steps above described is given in Figure 1. Note the differences in the location and spatial distribution of the titratable residues for each system. Monomers were prepared by the simple deletion of one given chain. Chain A was selected for most models except for $NS1_{ZIKV\text{-}UG}$ where chain B was chosen. All final atomistic coordinates are available for download as Supplementary Materials (https://zenodo.org/record/5608521). The monomeric and dimeric states are referenced in this work with the subscripts "m" and "d", respectively. After this initial preparation, the atomic coordinates were converted into the simplified CG molecular model as described in the next subsection.

### 2.2. Structural analysis

Additional calculations were done in order to compare structural features of the four $NS1_{flavivirus}$: *a)* all possible pairs of the whole dimer structures were overlapped and computations of the root mean square deviation (RMSD) using the alpha carbons ($C_\alpha$) were performed, *b)* from the overlapped dimers of item *a*, the RMSDs per residue were estimated, *c)* the percentage of identity among the NS1 protein sequences were also calculated, and *d)* volumes of the NS1 proteins in the monomeric and dimeric states were estimated. The items *a* to *d* were done in Chimera (Pettersen et al., 2004). For item *b* sequences were aligned with the web tool MUSCLE (Edgar, 2004) and visualized with CPRISMA program 1.0 (Poveda-Cuevas, 2021) Regarding item *d*, the "ProteinVolume server" was employed (Chen and Makhatadze, 2015).

### 2.3. Coarse-grained model

The NS1 proteins are relatively large objects which imply several interacting sites in a computer simulation with available classical force fields. For instance, the $NS1_{DENV2}$ homodimer has 704 amino acids (5604 atoms). The need to repeat the calculations for four different biological systems ($NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$) at two oligomeric states (dimers and tetramers) and at several pH and salt regimes makes simplified molecular models quite appealing for this study. Moreover, the aim is also to investigate the driving forces behind self-association. This requires estimating the free energy of interactions [$\beta w(r)$, where $\beta$ is equal to $1/k_BT$,

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$k_B = 1.380 \times 10^{-23}$ $m^2$ kg $s^{-2}$ $K^{-1}$ is the Boltzmann constant, and $T$ is the temperature in Kelvin] as a function of separation distances ($r$) between the centers of the proteins. It is well known that such measurements are very expensive in terms of computational time. Longer simulation runs are needed to ensure both that the most important parts of the configurational space have been properly sampled and that the histogram bins used to estimate $\beta w(r)$ are well-populated within a reasonable number of observations to be statistically meaningful (Barroso da Silva et al., 2006; Engkvist and Karlström, 1996). These reasons together motivated us to invoke a constant-pH (CpH) coarse-grained (CG) model especially devised before to explore protein-protein interactions (Da Silva et al., 2016; Delboni and Da Silva, 2016; Giron et al., 2020; Mendonça et al., 2019). Such a simplified biophysical model with a reduced number of degrees of freedom and input parameters has already proved to be advantageous and suitable for this kind of purpose at a lower (not necessarily cheaper) and feasible computational cost (Da Silva and Dias, 2017). Accordingly, amino acids are converted into a set of charged Lennard-Jones (LJ) spheres (beads) with radii $R_{ai}$ and valences $z_{ai}$. The resulting mesoscopic structures are treated as rigid bodies (*i.e.*, bond lengths, angles, and dihedral angles are kept fixed). The size and density of each of these charged LJ spheres were defined following the parameters given by Persson et al. (2010). The valences of the titratable residues are calculated as a function of pH during the simulation runs with the fast proton titration scheme (FPTS) (Da Silva and Dias, 2017; Da Silva and MacKernan, 2017; Teixeira et al., 2010). Intra and interchain charged amino acids can interfere in the acid-base properties of any other ionizable group as expected. Non-ionizable residues are not affected by the FPTS scheme. They are kept neutral at all times.

During the simulation runs the NS1 proteins are placed in an electroneutral cylinder with radius $r_{cyl}$ (all systems = 350 Å) and length $l_{cyl}$ (monomer-monomer = 200 Å, and dimer-dimer = 350 Å), where they can titrate, rotate in any direction, and translate back and forward along one axis. This strategy has been recently named FORTE (**F**ast c**O**arse-grained p**R**otein-pro**T**ein mod**E**l) (Neamtu et al., 2022). A detailed discussion of this theoretical approach, its pros and cons were given in previous works (Da Silva and Dias, 2017; Da Silva and MacKernan, 2017; Teixeira et

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

al., 2010). Lipids and sugars are not present in the model. For the dimerization process, two single monomers were placed in the simulation box. In an analogous manner, two dimers were present in the simulation box to study the tetramerization process. Water molecules are effectively modeled by a static dielectric constant $\epsilon_s$ equals to 78.7 (at 298 K). The main effect of counterions and added salt particles is given by an electrostatic screening term $[exp(-\kappa r_{ij})]$, where κ is the inverse modified Debye length [more details are given in the refs. (Da Silva and Dias, 2017; Da Silva and MacKernan, 2017)], and $r_{ij}$ is the inter amino acid separation distance (Teixeira et al., 2010). A sketch representing the model is shown in Figure 2.

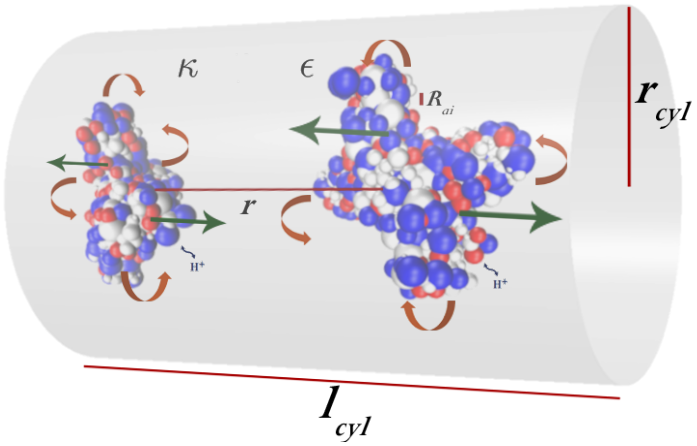



**Figure 2.** Schematic representation of the model cell used in this study. Two proteins build up by a collection of charged LJ spheres of radii ($R_{ai}$) and valence $z_{ai}$ mimicking amino acids. The simulation "box" is an electroneutral open cylindrical cell of radius $r_{cyl}$ and height $l_{cyl}$. The solvent is represented by its static dielectric constant $\epsilon_s$. Counter-ions and added salt particles are represented by Debye's term κ. Acidic, basic, and neutral residues are represented in blue, red, and white respectively. See the text for more details.

The effective configurational Hamiltonian of the system for a given configuration $[U(\{r_k\})]$ is calculated as,

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$U(\{r_k\}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(u^{el}(r_{ij}) + u^{vdw}(r_{ij})), \quad [1]$$

where $\{r_k\}$ are the positions of the amino acids, $N$ is the total number of amino acids, $[u^{el}(r_{ij})]$ is the Coulombic contribution, $[u^{vdw}(r_{ij})]$ is the LJ term, and $r_{ij}$ is the separation distance between amino acids $i$ and $j$.

The potential energy $u^{el}(r_{ij})$ describes the electrostatic interaction between any two ionizable groups of valences $z_i$ and $z_j$:

$$u^{el} = \frac{z_i z_j e^2}{4\pi\epsilon_0\epsilon_s r_{ij}} exp(-\kappa r_{ij}), \quad [2]$$

where $\epsilon_0$ is the dielectric constant of the vacuum ($= 8.854 \times 10^{-12}$ C$^2$/Nm$^2$) and $e$ is the elementary charge ($= 1.602 \times 10^{-19}$ C).

Using different radii for the amino-acids, the van der Waals (vdw) interactions partially incorporating the non-specific contributions of the hydrophobic effect and the Pauli exclusion principle are effectively calculated with the $u^{vdw}(r_{ij})$ potential energy as given by,

$$u^{vdw} = 4\varepsilon_{LJ}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \quad [3]$$

where $\varepsilon_{LJ}$ is an adjustable parameter that defines the depth of the potential well. It can be tuned up to favor the complexation or down to result in a lower attraction depending on the simulated

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

system. Here, we assumed a more or less universal value for protein studies where $\varepsilon_{LJ}$ is equal to 0.124 kJ/mol as suggested by Persson and co-authors (Persson et al., 2010). This should correspond to a Hamaker constant of ca. 9 $k_BT$ for an amino acid pair (Lund and Jönsson, 2003). The variable $\sigma_{ij}$ [$= R_{ai} + R_{aj}$)/2] is related to the size of beads ($R_{ai}$ and $R_{aj}$). These values for each amino acid were taken from the study by Da Silva et al. (2016). Larger amino acids should strongly attract others which results in a reasonable description of macromolecular hydrophobic moments (Eisenberg et al., 1982; Lund and Jönsson, 2003).

### 2.4. Constant-pH Monte Carlo simulations

The molecular systems were simulated at different pH solutions and salt concentrations by Monte Carlo (MC) simulations using the Metropolis algorithm (Metropolis et al., 1953). Free energy derivatives were computed as a function of the separation distance for both monomer-monomer and dimer-dimer complexations. For the sake of convenience, several sets of simulations in the NVT ensemble were performed for all viruses/strains:

1) **Set A:** A single NS1 protein (monomeric state) in NaCl solution (150 mM). This set was used to provide the main physical-chemical properties of the proteins;
2) **Set B:** Same as set A for the dimeric state;
3) **Set C:** Two NS1 monomers in salt solution (150 mM) during the dimerization processes ($NS1_{flavivirus_m} + NS1_{flavivirus_m} \rightarrow NS1_{flavivirus_mm}$). Here, both macromolecules can translate, rotate, and titrate to sample the complexation process. These simulations were employed to provide $\beta w(r)$ for the dimerization. Note that the resulting dimer was labeled with the subscript "mm";
4) **Set D:** Same as set C for the tetramerization processes ($NS1_{flavivirus_d} + NS1_{flavivirus_d} \rightarrow NS1_{flavivirus_dd}$). $NS1_{flavivirus_d}$ is the dimer as given by the experimental structure. Note that here the resulting tetrameter was labeled with the subscript "dd";
5) **Set E:** Same as set D for a hypothetical system with all amino acids assumed to be constantly neutral during the simulation run. This artificial system was solely used to

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

access the free energy contributions exclusively from the van der Waals terms excluding the electrostatic interactions;

6) **Set F:** Same as set D for lower ionic strength (5 mM).

For sets A and B, the equilibration and production phases were performed using $10^6$ and $10^8$ MC cycles, respectively. Only titration was included as an MC movement in these runs. The FPTS was employed at a different pH condition from 0.1 to 14 every 0.1 pH unit to compute the corresponding protein net charge number ($Z_P$), the charge regulation capacity ($C_P$), and the dipole moment number ($\mu_P$) as previously proposed (Delboni and Da Silva, 2016; Jonsson and Lund, 2007; Lunkad et al., 2022; Poveda-Cuevas et al., 2018). For the other simulation sets where the focus was on the complexation process, much longer and quite expensive (in terms of CPU time) simulations were necessary. At least 1.0 x $10^{10}$ and 2.0 x $10^{10}$ MC steps were carried out for the dimerization and tetramerization processes, respectively, after the equilibration phase. Due to the high computational cost of these simulations, a bit shorter runs were executed for sets E and F. In these runs (*i.e.* E and F), 2.6 x $10^8$ MC and 9.0 x $10^8$ steps were employed, respectively. In all sets, three replicates were carried out for each simulated system. They were also used to estimate mean values and standard deviations. All MC simulations were done through a modified version of the biomolecular simulation package Faunus (Lund et al., 2008).

The central aspects of the protein-protein complexation process were covered by sets C and D. Three different pH solutions were explored: 3, 7, and 8.5. These values correspond to the experimental conditions where dissociation and association were reported for $NS1_{DENV2}$ (Falconar and Young, 1990). Previously electrophoresis assays and X-ray data have shown that the dimerization occurs at a pH interval from 3.5 to 8.5 (Akey et al., 2014; Brown et al., 2016; Falconar and Young, 1990; Flamand et al., 1999; Song et al., 2016; Xu et al., 2016). At pH 3, monomers were observed instead of the dimers (Falconar and Young, 1990). Few experimental information is available for the tetramerization. There are only two reported cases at pH 8 and 8.5 for $NS1_{DENV2}$ and $NS1_{DENV1}$, respectively (Flamand et al., 1999; Muller et al., 2012). Therefore, the choice of the pH regimes in this work was twofold: *a*) to validate the model

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

capacity to reproduce experimental data (for $NS1_{DENV2}$), and *b*) to apply it to other conditions and viral systems where there is a lack of experimental data (*e.g.* pH 7 and $NS1_{ZIKV\text{-}UG_dd}$). To uncover the impact of the specific protein sequences, it is important to compare the behavior of NS1 proteins from $NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$ on exactly the same experimental conditions. Other pH conditions were not simulated due to the already high CPU costs of these three selected conditions. For each single MC step, the computing time was 0.06 seconds (processor Intel(R) Xeon(R) CPU E7- 2870 @ 2.40GHz) for a dimer-dimer complexation study. This number grows too fast when multiplied by the needed ca 3 x $10^{10}$ steps and the replicates together additional tests per case and condition.

The last two sets (E and F) were complementary runs to provide the energy basis for the self-association of these NS1 proteins. Simulations with all ionizable groups artificially forced to keep zero their charges are used in set F to quantify the vdw contributions on each flavivirus NS1 system (in the absence of electrostatic interactions). Finally, set E was carried out at pH 7 to assess the role of a low salt condition (5 mM) on the tetramerization process.

### 2.5. Free energy of interactions

The main result from sets C to F is $\beta w(r)$. During each simulation run, it was recorded in histograms the frequency of each separation distance between the center of the proteins observed. The sampling was performed with a bin size equal to 1 Å. These histograms were converted to radial distribution functions [$g(r)$]. Through a well-established mechanical-statistical relation, $g(r)$ is related to the two-particle potential of mean force [$\beta w(r) = -\ln g(r)$] that was used to calculate the corresponding free energy of interactions (McQuarrie, 1976).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 2.6. Second virial coefficient

When a pair of proteins are immersed in an electrolyte solution, several physical interactions are acting on the system (*e.g.* repulsion due to the Pauli-exclusion principle, vdw, hydrophobic "interactions" and multipole-multipole electrostatic interactions). Practical estimation of their importance can be done involving the calculation of second virial coefficients ($B_2$). Such quantity can also be obtained by means of light scattering (Velev et al., 1998), chromatography (Tessier et al., 2002), or membrane osmometry experiments (Pjura et al., 2000). Theoretically, $B_2$ (often given in mol ml/g$^2$) can be estimated following the canonical partition $\left[B_2 = B_{q+vdw} + B_{hs}\right]$, where $B_{q+vdw}$ represents the electrostatics and vdw contributions, and $B_{hs}$ represents the excluded-volume (hard-sphere) term. Mathematically,

$$B_{q+vdw} = -\frac{2\pi N_A}{M^2}\int_{\sigma}^{\infty}\left[exp\left(-\,w(r)/k_BT\right) - 1\right]r^2 dr, \quad [4]$$

$$B_{hs} = \frac{N_A}{3M^2}\pi\sigma_B^3, \quad [5]$$

where $M$ is the molecular weight (in g/mol) of proteins, $\sigma_B$ is the sum of radii of the two proteins (in Å) and $N_A$ is the Avogadro number ($6.022\times10^{23}$ mol$^{-1}$). Equations 4 and 5 are numerically solved through the Simpson rule from $\beta w(r)$ given by the MC simulation data and simple analytical math, respectively. Values for $B_{hs}$ for the NS1 protein systems in the monomeric and dimeric state for each NS1$_{\text{flavivirus}}$ are compiled in Table S1. This data does not depend on the simulation results as seen in Eq. 5. A positive $B_2$ (total term) indicates an attractive behavior, whereas a negative value means a repulsion between the proteins (George and Wilson, 1994). The dependence on $r^2$ in Equation 4 makes $B_{q+vdw}$ largely affected by the statistical noises of $w(r)$. Much longer runs would be needed for a precise quantitative prediction of $B_2$. Instead, due to the prohibitive CPU costs we focuses on the qualitative trends given by the calculated $B_2$. The

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

molecular weights (*Mw*) of the studied proteins were estimated in the web server tool "Compute pI/Mw" (Wilkins et al., 1999). The *Mw*'s in the monomeric states are 40078.36, 40097.47, 39939.47, and 39749.8 g/mol, for respectively, $NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$. For all $NS1_{flavivirus}$ in the dimeric state these, *Mw*'s were doubled.

## 3. RESULTS AND DISCUSSION

The systems studied in the present paper cover different ranges of viral time-scale evolution, a short-range for the two ZIKV strains (Gong et al., 2016), and a long-range between ZIKV, DENV2, and WNV (Moureau et al., 2015; Pettersson and Fiz-Palacios, 2014). This is illustrated by the identity percentages between the four $NS1_{flavivirus}$ as reported in Table S2. The $NS1_{ZIKV}$ strains present 97% identity, while the remaining pairs have an average identity percentage equivalent to 55.2% when compared. The flavivirus NS1 protein is a multifaceted molecule often referred to as a promiscuous biological system due to its ability to interact with different partners (Muller and Young, 2013; Watterson et al., 2016). This can be advantageous from the point of view of viruses since the cellular machinery is integrated by several macromolecules and biological media that should be altered to produce the host-cell programming. Electrostatic forces have an impact on these interactions as it has already been seen in both experiments and computer simulations (Brown et al., 2016; Poveda-Cuevas et al., 2020, 2018; Song et al., 2016; Xu et al., 2016). A considerably high number of titratable sites (see Figure 1 and S1) is observed for both monomers and dimers. Table S3 lists the number of ionizable residues for each flavivirus NS1 protein at the monomeric state. Note that these four viral systems have a different number of titratable groups. The two $NS1_{ZIKV}$ have a similar number of basic (66 and 67, for $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$, respectively) and acidic residues (51 and 50, for $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$, respectively) because they are phylogenetically closer while $NS1_{DENV2}$ and $NS1_{WNV}$ exhibit a larger difference because they belong to other flavivirus species. In terms of their three-dimensional structures, RMSD calculations among the $C_\alpha$ for all possible pairs of NS1 dimers show some structural differences among them. The most and least similar NS1 dimers correspond to the pairs $NS1_{ZIKV\text{-}UG}$-$NS1_{DENV2}$ (0.492 Å) and $NS1_{ZIKV\text{-}BR}$-$NS1_{WNV}$ (0.765 Å),

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

respectively (see Table S4). A detailed view when comparing the $C_\alpha$s of the NS1 structures per residue shows that regions with the greatest structural variations as measured by the RMSD comprise the loop zones between residues 115-122 (see Figure S1). These two features (*i.e.* number/composition of titratable residues and a specific three-dimensional structure for each NS1 protein with a different spatial arrangement of the ionizable residues) should have an impact on electrostatic properties and the biomolecular interactions of these macromolecules. Consequently, physical-chemical variables such as pH are important in the molecular mechanisms that control the formation of NS1 complexes and impact the virus life cycle.

Free energies of interactions [$\beta w(r)$] and second virial coefficients ($B_2$) were employed to desiccate the self-association features of the $NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$ at different pH conditions: 3 (acidic regime), 7 (neutral/physiological regime), and 8.5 (basic regime). These free energies were computed assuming an ionic strength of 150 mM of NaCl. On the basis of these physicochemical conditions, this section focuses four aspects: *i*) the validation of our theoretical model using the available experimental information, *ii*) a report of the specificities of these viruses for both dimerization ($NS1_{flavivirus_m} + NS1_{flavivirus_m} \rightarrow NS1_{flavivirus_mm}$) and tetramerization ($NS1_{flavivirus_d} + NS1_{flavivirus_d} \rightarrow NS1_{flavivirus_dd}$) processes, *iii*) the examination of the discrepancies at the inter- (using the four NS1 flavivirus proteins) or intraspecific level (focusing in the two $NS1_{ZIKV}$ proteins), and *iv*) the analysis of the physical forces that govern the self-association and their possible biological implications. The $\beta w(r)$ for both $NS1_{flavivirus_mm}$ and $NS1_{flavivirus_dd}$ for all four flaviviruses were calculated through CpH MC. Figure 3 displays the outcomes of these simulations for all studied pH conditions and the two oligomeric states.

### 3.1. Validation of the computational model

A common procedure to validate computational simulations of biomolecules is to compare specific physical quantities predicted by the models with experimental data.(Cragnolini et al., 2015) The reproducibility of the experimental information by the theoretical model allows establishing its predictive capacity for subsequent novel applications (*e.g.*, the prediction of the tetramerizations features that are not yet experimentally available in the literature). Although the

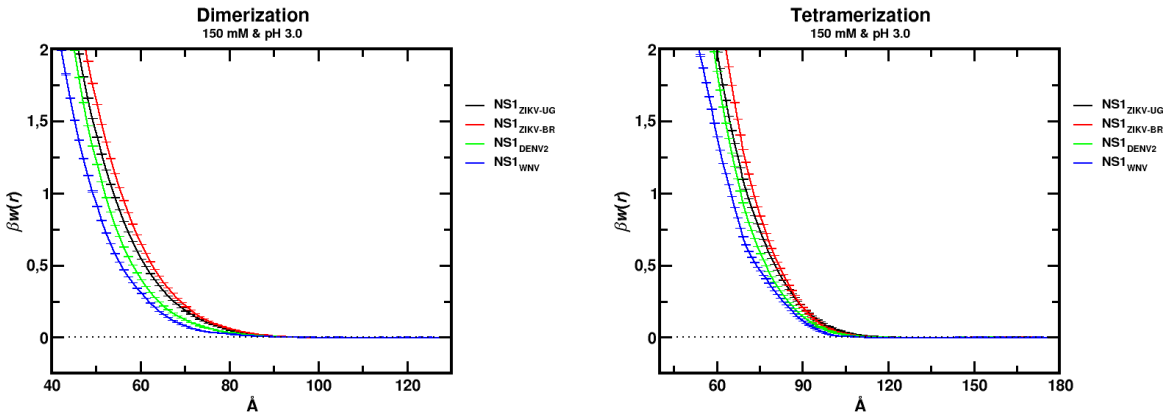
bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model has been validated for other systems (Da Silva et al., 2016; Delboni and Da Silva, 2016; Giron et al., 2021, 2020; Neamtu et al., 2022), for this purpose and NS1 proteins, we have mainly used electrophoresis data by Falconar and Young (1990) for dimerizations of $NS1_{DENV2}$ (*i.e.*, $NS1_{DENV2_mm}$). Data is available for the pH range of 2.2 to 7.4. It was found in this lab experiment that for situations above pH ~3 the system forms stable dimers, whereas below pH ~3 a monomeric state prevails. No electrophoresis data is available for the other three NS1 proteins and/or pH conditions. Some additional information can be extracted from structural data. For instance, combining the crystallography conditions where crystals for $NS1_{ZIKV\text{-}UG_mm}$ and $NS1_{ZIKV\text{-}BR_mm}$ (Brown et al., 2016; Xu et al., 2016) were obtained with electrophoresis data for $NS1_{DENV1_dd}$ and $NS1_{DENV2_dd}$ (Flamand et al., 1999; Muller et al., 2012) indicates that there are dimers or tetramers above pH 8. This was tested in our biophysical simulations. The $\beta w(r)$ profiles for the dimerization process of $NS1_{DENV2_mm}$ were calculated at pH values of 3, 7, and 8.5 (see, respectively, the green curves in Figures 3a, 3c, and 3e). Following a conventional thermodynamic criterion to identify repulsive or attractive systems [*i.e.*, when $\beta w(r) > 0$ and $\beta w(r) < 0$, respectively] (McQuarrie, 1976), we observed that monomers of $NS1_{DENV2_mm}$ form homodimers at pH 7 and 8.5, while they are unable to associate at pH 3 in agreement with the results of the electrophoretic experiment. For the $NS1_{ZIKV\text{-}UG_mm}$ and $NS1_{ZIKV\text{-}BR_mm}$ at pH 8.5 (see black and red curves in Figure 3e), our model predicts attractive profiles for both viral proteins. The same was found for $NS1_{DENV2_dd}$ that experiences an attraction at pH 8.5 confirming it can self-associate at this condition (see Figure 3f). For all cases, our theoretical model is in good agreement with the available experimental data predicting that dimers and monomers can be found at the correct solution pH value. This allows us to further explore other new conditions in order to comprehend more about the NS1 self-association of these four viral systems (in particular the tetramerization process). It is important to mention that the model does not incorporate conformational changes that can occur under extreme pH conditions (*i.e.*, unfolding at acidic or basic regimes far from the protein isoelectric point). Nevertheless, no discrepancy was observed for the predictions at the acidic pH in comparison with the experiments.

### 3.2. Differences at the oligomerization level

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, we analyzed and compared the NS1 dimerization and tetramerization processes at the same physicochemical conditions for the four viral systems. We found that regardless of the type of virus, strain, or oligomerization state, the four systems are all repulsive at pH 3 as indicated by their estimated free energy of interactions ($\beta w(r) > 0$ - see Figures 3a and 3b). At this pH, the tetramerization is even more unfavorable compared to dimerization on the range of separation distances examined. Conversely, an attractive regime [$\beta w(r) < 0$] is observed for all NS1 proteins at pH 7 for both the dimerization (see Figure 3c) and tetramerization (see Figure 3d) processes. A higher well-depth is measured for the tetramerization in comparison with the dimer formation. For instance, using the minima values of the $\beta w(r)$ potential well as a parameter for this analysis, the attraction is 4-fold larger for $NS1_{ZIKV\text{-}BR_dd}$ than $NS1_{ZIKV\text{-}BR_mm}$ [$\beta w(r)_{min_dd}$ = -2.19±0.03 $k_BT$ at 44.2 Å and $\beta w(r)_{min_mm}$ = -0.575±0.002 $k_BT$ at 52.2 Å, for $NS1_{ZIKV\text{-}BR_dd}$ and $NS1_{ZIKV\text{-}BR_mm}$, respectively]. This behavior suggests an important contribution from the van der Waals interactions for these complexes. The theoretical predictions for $\beta w(r)$ at pH 7 and 8.5 are quite similar. These data can be seen in Figures 3e and 3f, respectively. The main difference is the slight tendency for the protein-protein association to be less attractive at pH 8.5 due to the increased Coulombic repulsion at this more basic condition [*e.g.*, $Z_P$s are -4.3 (pH 7) and -10.4 (pH 8.5) for $NS1_{ZIKV\text{-}UG_d}$].




**(a)** **(b)**

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

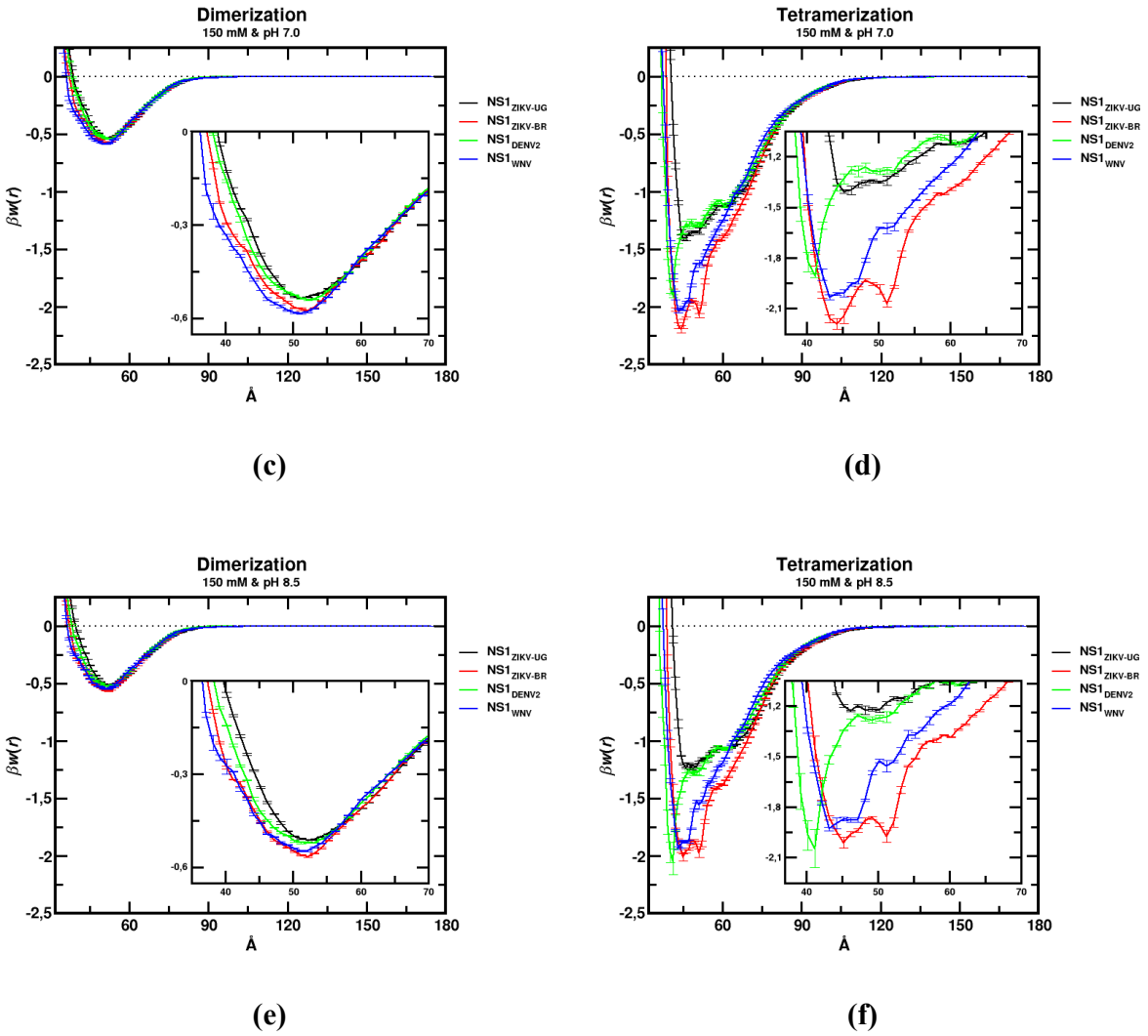



**Figure 3**. Free energy of interactions as a function of the center-center separation distance for NS1 homoassociation. Data for the dimerization and the tetramerization processes of four flaviviruses at different pH solutions and 150 mM of NaCl. The plots show the average data with the estimated standard deviations based on three replicate runs for each studied system. These data correspond to the results from simulation sets C ($NS1_{flavivirus_mm}$) and D ($NS1_{flavivirus_dd}$).

A complementary analysis can be done by means of second virial coefficients. They can be roughly estimated from the computed $\beta w(r)$ as shown in Equation 4. The quantity $B_2$ is interpreted similarly as for $\beta w(r)$, that is, when a system undergoes an attractive regime, $B_2 < 0$,

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

while for a repulsive condition, $B_2 > 0$. Figure 4 shows $B_2$ as a function of pH for all studied systems (inter and intraspecific levels), oligomeric states, and physical-chemical conditions. The data for the formation of dimers and tetramers are given by the dashed and solid curves, respectively. It can also be easily seen in this plot how tetramerizations are more attractive than dimerizations for any pH solution. All solid lines (tetramerization) are always below the dashed ones (dimerization) for all systems. Even at pH 3, lower repulsive features are observed for the dimer-dimer interactions compared to the monomer-monomer association, in contrast to the data represented in Figure 3a and Figure 3b. For instance, $B_2$ decreases from 1.90±0.02 ml mol x $10^{-4}/g^2$ (for the dimer formation) to 1.022±0.006 ml mol x $10^{-4}/g^2$ (for the tetramer formation) at pH 3 for $NS1_{WNV}$. This change might be explained by a smaller repulsive excluded-volume contribution for the dimer-dimer association compared to monomer-monomer interaction (see Table S1).

### 3.3. Differences at the inter and intraspecific levels

For the understanding of the specific molecular mechanisms that the NS1 protein might obey for each viral system, we compared their associative characteristics. In Figures 3a and 3b, we start with the comparison at pH 3 for their self-associations. For both oligomeric states, the four viral systems can be sorted out from the more repulsive behavior to the less one based on the minima values of their free energies of interactions: $\beta w(r)_{\text{ZIKV-BR}} > \beta w(r)_{\text{ZIKV-UG}} > \beta w(r)_{\text{DENV2}} > \beta w(r)_{\text{WNV}}$. At physiological pH, the monomer-monomer association follows a different order: $\beta w(r)_{\text{min_ZIKV-UG_mm}} > \beta w(r)_{\text{min_DENV2_mm}} > \beta w(r)_{\text{min_ZIKV-BR_mm}} > \beta w(r)_{\text{min_WNV_mm}}$ as seen in Figure 3c. The NS1 protein from the WNV is the system with the highest NS1-NS1 binding affinity. Yet, for the tetramer formation, there is an inversion between $NS1_{\text{WNV_dd}}$ and $NS1_{\text{ZIKV-BR_dd}}$: $\beta w(r)_{\text{min_ZIKV-UG_dd}} > \beta w(r)_{\text{min_DENV2_dd}} > \beta w(r)_{\text{min_WNV_dd}} > \beta w(r)_{\text{min_ZIKV-BR_dd}}$ (see Figure 3d). The same trend is seen at pH 8.5 as illustrated in Figures 3e and 3f for the dimerization and the tetramerization processes, respectively. This result also implies that each system will have a different pH condition for the transition between a repulsive to an attractive regime. The connecting lines are drawn to guide the eyes in Figure 4. The minima values of the well-depth for these free energies of interactions are found at different separation distances. This is a direct consequence of the different sizes that each system has (see Table S5). Some NS1 proteins can

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

come closer to their selfpartners than others. For instance, it is observed that the dimers of $NS1_{DENV_dd}$ present a high frequency of interaction in a short distance range (notice in Figures 3d and 3f how the green curves are narrower at ~40 Å).

The main features of this analysis can be summarized with the $B_2$ data as given in Figure 4. For instance, for the formation of tetramers, from the less attractive regime to the more one, the systems are following this order: $B_{2_NS1_ZIKV\text{-}UG_dd} > B_{2_NS1_DENV2_dd} > B_{2_NS1_WNV_dd} > B_{2_NS1_ZIKV\text{-}BR_dd}$. This can be seen at both pH 7 and 8.5 where the systems are under an attractive regime condition. In all the simulated conditions, lipids and sugars were absent from the system indicating that they are not necessary for these self-associations.

It is interesting to see how more phylogenetically related proteins such as $NS1_{ZIKV_UG}$ and $NS1_{ZIKV\text{-}BR}$ (97% of sequence identity among them) behave in a discrepant way. For example, at pH 7 (see Figure 3d), the smaller value for $\beta w(r)$ was found for $NS1_{ZIKV\text{-}BR_dd}$ (~2 $k_BT$ at $r$~44.5 Å) when compared to $NS1_{ZIKV\text{-}UG_dd}$ (~1.4 $k_BT$ at $r$~45.5 Å). This means a difference of $0.6k_BT$. Conversely, the same comparison for $NS1_{ZIKV\text{-}BR_dd}$ and $NS1_{WNV_dd}$ gives a smaller difference of $0.2k_BT$ despite their lower sequence identity (56%). This shows an interplay of the contributions determined at the linear sequence and three-dimensional structural levels. It also reveals an important diversity in the molecular mechanisms of these viral proteins with direct consequences on their virulences.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

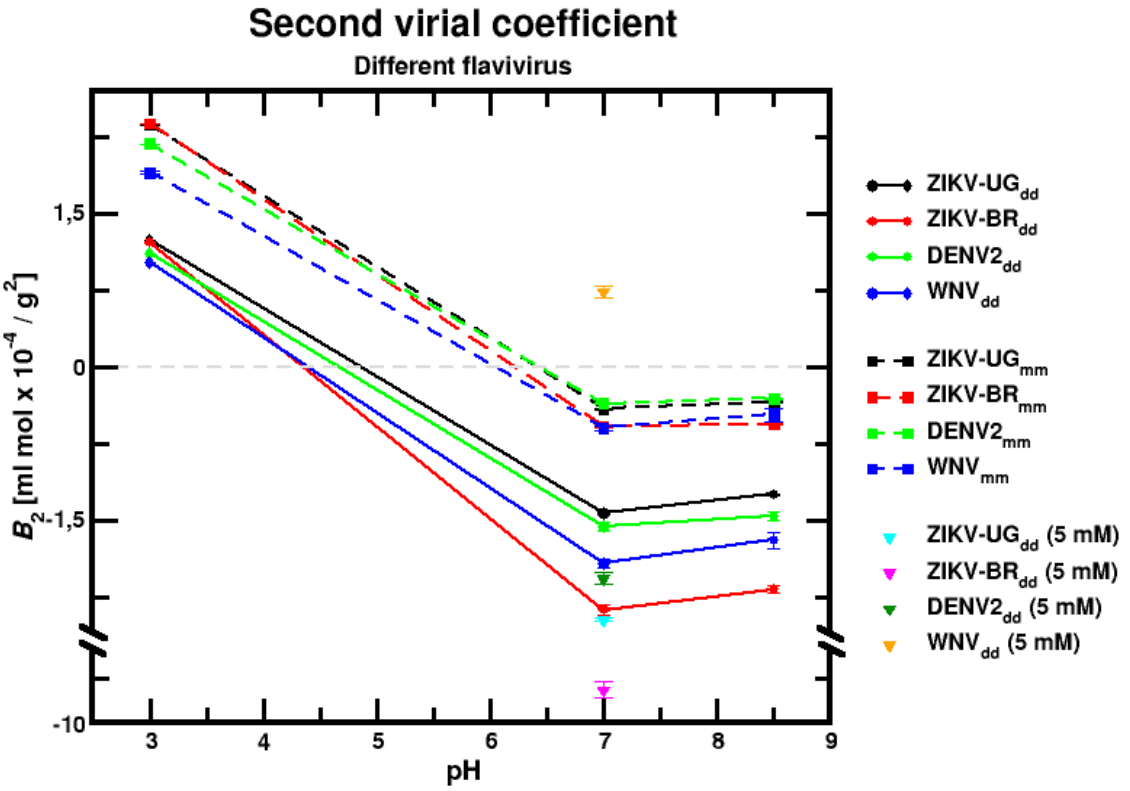



**Figure 4.** Second virial coefficient $B_2$ as a function of pH for several flavivirus NS1 and two oligomeric states at different pH solutions. The plots show an average behavior estimated on three replicate runs. Dashed and continuous lines are used to guide the eyes for the formation of dimers and tetramers, respectively. All symbols connected by these lines represent data obtained at 150mM of NaCl. Symbols not connected by lines represent data obtained at a low salt condition (5mM of NaCl).

### 3.4. Physicochemical features

The evaluation of the main physicochemical aspects that govern these systems can shed light to understand the physical reasons for the differences discussed above (Da Silva et al., 2016; Delboni and Da Silva, 2016; Giron et al., 2021, 2020; Poveda-Cuevas et al., 2018). Therefore, we next characterized the main physical-chemical quantities ($Z_P$, $C_P$, and $\mu_P$) of $NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$ for both monomeric and dimeric states at physiological salt

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

concentration and pH values from 0 to 14. Based on a previous study (Poveda-Cuevas et al., 2018), the total protein charge number was identified as the central outcome for the present work due to the electrostatic screening of the high order terms. This physical quantity seems especially important in viral proteins (Giron et al., 2021; Poveda-Cuevas et al., 2020, 2018; Yoshida et al., 2019). Table 1 provided the main data for pH conditions of 3, 7, and 8.5. For the sake of completeness, full titration plots for the two oligomeric states and the four NS1 flaviviruses are also shown in Figure S2. Monomers can be highly charged at very acidic and basic regimes with $Z_P$ ranging from ca. 53 to -53 (see Figure S2a). These protein net charges are relatively doubled in the dimeric state due to the presence of the second chain (see Figure S2b) that contributes to the charge regulation. Regarding the $C_P$s of monomers (see Figure S2c) and dimers (see Figure S2d), we observed two peaks at roughly pH 3.7 and 11.6. Note that the $C_P$ values are 2-fold large for NS1s in the dimeric state when compared to monomers (*e.g.*, the $C_P$s are 7.5 and 14.4 at pH 4.0 for, respectively, $NS1_{DENV_m}$ and $NS1_{DENV_d}$). This means that the second monomer in the dimeric NS1 is contributing to increasing $C_P$. As explained before by Jonsson and Lund (2007) the reason for these two peaks has a direct relationship with the number of titratable sites and their p$K_a$s values. Data for $\mu_P$ is given for the monomers and dimers, respectively, in Figures S2e and S2f. At pHs conditions closer to 7, $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ in the dimeric states have a particular characteristic presenting higher dipole values than the remaining $NS1_{flavivirus}$. These observations indicate that charge regulation and dipole interactions (Barroso da Silva et al., 2006; Da Silva and Jönsson, 2009; Lunkad et al., 2022) can play a more important and specific role for each viral system when the salt concentration is sufficiently low.

We noticed a correlation between the total protein net charges and the minima values of $\beta w(r)$ for both $NS1_{flavivirus_mm}$ and $NS1_{flavivirus_dd}$ at pH 3. That is, as the protein net charge is increased the protein-protein free energy of interaction becomes more repulsive. For instance, compare the blue ($NS1_{WNV}$), green ($NS1_{DENV2}$), black ($NS1_{ZIKV\text{-}UG}$), and red ($NS1_{ZIKV\text{-}BR}$) curves given in Figures 3a and 3b with the protein net charges given in Table 1. The protein net charges follow the same order ($Z_{P_WNV} < Z_{P_DENV2} < Z_{P_ZIKV\text{-}UG} < Z_{P_ZIKV\text{-}BR}$) indicating that charge-charge interactions dominate this situation in both the monomer and dimers states. At physiological or

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

basic pHs, this behavior is not clearly observed, namely, we do not see that $NS1_{WNV_dd}$ as the most repulsive system among the remaining NS1 despite having the higher protein net charge. This may be an indication that van der Waals forces are dominating the self-association at these pH solutions. Electrostatic interactions are probably screened out by physiological salt conditions which allow the vdw interactions to overcome them at these pH conditions. Non-Coulombic contributions such as those arising from the multipole-multipole interactions and the charge regulation mechanism are even more severely affected by salt (Barroso da Silva et al., 2006). Thus, in certain conditions, it has been experimentally shown that the oligomerization processes of some viral proteins with the same net charge can be driven by vdw interactions or the hydrophobic effect (Kegel and van der Schoot, 2004; Thuman-Commike et al., 1998).

The contributions of the vdw interactions were further explored by means of an artificial man-created model where electrostatic interactions were removed from the system. That is, all the charges of the titratable groups were artificially kept equal to zero during dimer-dimer complexation runs at physiological salt concentration. The corresponding $\beta w(r)$ plots for this situation are given in Figure S3 (red lines). For the sake of comparison, $\beta w(r)$ for the same system at pH 7 are included in this figure as solid black lines. The results suggest that vdw forces dominate the complexation process at this pH condition in contrast to what was observed before at the acidic pH condition. The protein net charges are not sufficiently high to produce a repulsion as it is observed at pH 3. Moreover, since the vdw contributions have a direct relationship with the dimension of the systems (as could be seen in Equation 3), they found differences between the two oligomeric states are largely explained by the protein sizes (the volumes of the monomers and dimers are equivalent to ~4.8 x $10^4$Å$^3$ and ~9.7 x $10^4$Å$^3$, respectively). Some specific amino acids can also prevent the proteins from coming closer. For example, residues 101-135 are localized in the wing domain and have high thermal fluctuation (see Figure S1). They could easily contribute to a steric hindrance during the protein-protein assembly. Previous computational studies have also shown that these flexible regions can play a particular role for a specific $NS1_{flavivirus}$ (Gonçalves et al., 2019; Menezes et al., 2020; Poveda-Cuevas et al., 2021).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Table 1.** Protein charge number ($Z_P$) for $NS1_{ZIKV\text{-}UG}$ (PDB id 5K6K), $NS1_{ZIKV\text{-}BR}$ (PDB id 5GS6), $NS1_{DENV2}$ (PDB id 4O6B), and $NS1_{WNV}$ (PDB id 4O6D). Data correspond to monomers (m) and dimers (d) at pH 3 (acid regime), 7 (neutral regime), and 8.5 (basic regime). The pIs for both oligomeric states are $6.7_{pI_m}/6.6_{pI_d}$, $7.1_{pI_m}/7.0_{pI_d}$, $6.8_{pI_m}/6.8_{pI_d}$, and $5.8_{pI_m}/5.8_{pI_d}$, for respectively $NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$. Data correspond to simulation sets A and B.

| pH | ZIKV-UG | | ZIKV-BR | | DENV2 | | WNV | |
|---|---|---|---|---|---|---|---|---|
| | $Z_P$_m | $Z_P$_d | $Z_P$_m | $Z_P$_d | $Z_P$_m | $Z_P$_d | $Z_P$_m | $Z_P$_d |
| 3 | 43.7 | 83.5 | 45.1 | 86.2 | 40.0 | 77.6 | 38.2 | 73.7 |
| 7 | -1.8 | -4.3 | 0.2 | -0.5 | -0.9 | -2.0 | -5.2 | -10.6 |
| 8.5 | -5.1 | -10.4 | -3.2 | -6.9 | -3.6 | -7.2 | -7.7 | -15.5 |

The understanding of the balance between electrostatics and vdw interactions at different conditions can be complemented by investigating a low salt condition where the electrostatic screening is reduced. Calculations were performed at 5mM of NaCl and pH 7 for the dimer-dimer homoassociation (see Figure S4). Most systems ($NS1_{ZIKV\text{-}UG_dd}$, $NS1_{ZIKV\text{-}BR_dd}$, and $NS1_{DENV2_dd}$) are more attractive at the low salt concentration regime: a) $NS1_{ZIKV\text{-}UG_dd}$ ( ca. $-1.8_{5mM} < -1.4_{150mM}$ $k_BT$), b) $NS1_{ZIKV\text{-}BR_dd}$ (ca $-4.2_{5mM} < -2_{150mM}$ $k_BT$), and c) $NS1_{DENV2_dd}$ (ca $-2.2_{5mM} < -1.9_{150mM}$ $k_BT$). Conversely, $NS1_{WNV_dd}$ shows a less attractive characteristic (ca $-1.5_{5mM} > -2.0_{150mM}$ $k_BT$). These trends were already included in Figure 4 showing the same behavior.

### 3.5. Biological relevance

NS1 is known both to interact with the immune system (Biering et al., 2021; Edeling et al., 2014; Hertz et al., 2017) and to increase the permeability of human endothelial cells from the lung, dermis, umbilical vein, brain, and liver (Puerta-Guardo et al., 2019). As shown through *in vitro* and *in vivo* experiments by Puerta-Guardo et al., (2019) the NS1 from a particular flavivirus is

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tissue-specific which can consequently produce particular physiopathology. Assuming physiological conditions of pH and salt (*i.e.*, regimes where several biological processes happen), our computational outcomes show that at the molecular scale, the free energy of the NS1 oligomerization from four flaviviruses present meaningful discrepancies.

Given that each flavivirus has different stabilities (especially when tetramers are formed), we can hypothesize that the differential virulence has a direct contribution from the amino acid composition of each NS1 that, at the same time, impacts their structures, the vdw (to a great extent particularly at the physiological condition) and electrostatic interactions during the complexation process. Surprisingly, these discrepancies are not limited to the *inter*specific level, but also to the intraspecific level, showing that the versatility of NS1 can be extended to flavivirus variants. Furthermore, it is reasonable to assume that the oligomeric stability is crucial for subsequent interactions with other biomolecules that must occur. Thus, each flavivirus having a differential free energy profile could have a relation with the particular tropism expressed by each one or the specific organ affinity found by laboratory assays (Biering et al., 2021; Puerta-Guardo et al., 2019).

As mentioned earlier, regions 101-135 seem to be important for the biological function of $NS1_{flavivirus}$. Previous studies have shown that this region is immunodominant and may work to inhibit the NS1 activity and pathogenicity via antibodies (Lai et al., 2017). Although these protein regions may be related to interactions with other macromolecules such as GAGs or membranes (Ci et al., 2021, 2020; Puerta-Guardo et al., 2019), we suppose that they may also be directly involved in the protein-protein NS1 assembly. This implies that these regions can work as a starting point to selectively inhibit oligomer association of different flavivirus species or strains. Finally, the detection of hotspots [*i.e.* regions or residues that can contribute significantly to protein-protein interaction as observed in other studies (Gonçalves et al., 2019; Persson et al., 2010)] for NS1 oligomerization gives important clues for the vaccine development, drugs, or therapies against flavivirus infections.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## 4. CONCLUSIONS

The dimerization and tetramerization processes for $NS1_{ZIKV\text{-}UG}$, $NS1_{ZIKV\text{-}BR}$, $NS1_{DENV2}$, and $NS1_{WNV}$ were examined through free energy derivatives at three different pH solutions (that is, 3, 7, and 8.5) and physiological salt concentration. As an essential step, the theoretical modeling was validated by comparing the available experimental data and free energy measurements for dimerizations and tetramerizations processes. Our outcomes confirmed the reliability of the used CG model also for this biological system. Complexation data for other flaviviruses and different experimental conditions that were lacking in the literature were provided. Our results show that these NS1 proteins can form both dimers and tetramers under conditions near physiological pH. The dimer-dimer interactions were found more stable than the monomer-monomer association as expected.

pH plays an important role mainly controlling the regimes where van der Waals interactions govern the NS1 self-associations. Two regimes could be detected that are common for all studied NS1 systems: *a)* a repulsive one, at pH 3, dominated by repulsive charge-charge interactions, and *b)* an attractive condition, at pHs 7 and 8.5, due to the vdw interactions to be able to overcome the electrostatic contributions. The number of additional peaks around the minima values of the free energies also suggests a complex energy landscape with steric contributions by the structural features of each system.

In previous studies, a probable mechanism that could explain the tetramerization and eventual hexamerization of NS1 has been hypothesized, where the interaction NS1-membrane is an essential step to produce stable assemblies among NS1 dimers during the secretion process (Gutsche et al., 2011). Nevertheless, despite the reduced number of parameters of our computational model, we have shown that the dimer-dimer interaction of NS1 can also occur in the *absence* of lipid and sugar molecules. Stable tetramers were formed in a pure electrolyte medium. Moreover, it was hypothesized that tetramers have a higher probability to interact with the immune system when compared to dimers or monomers because their greater stability facilitates the recognition of immunoglobulins.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, this set of information gives new clues about the molecular mechanisms that drive the self-association of NS1 proteins. This could be beneficial for the development of drugs based on oligomerization inhibition. This study also provides a complementary view on the discussion related to flavivirus virulence and differential cellular tropism expressed at the inter and intraspecific levels. Despite the similarity at the protein sequence level, we found that each flavivirus has a well-characteristic NS1 protein-protein interaction profile.

## 5. ACKNOWLEDGMENTS

This work has been supported in part by the "Fundação de Amparo à Pesquisa do Estado de São Paulo" [Fapesp 2015/16116-3 and 2020/07158-2 (F.L.B.d.S.)], the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [CNPq 305393/2020-0 (F.L.B.d.S.)], the Swedish National Infrastructure for Computing [SNIC 2020/14-38 and SNIC 2021/3-6 (F.L.B.d.S.)] and the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior"—CAPES (S.A.P.C.). F.L.B.d.S. and S.A.P.C. acknowledge the support of the University of São Paulo (USP) through the NAP-CatSinQ (Research Core in Catalysis and Chemical Synthesis) and the computing hours at HPC/USP. The authors also would like to thank CINES (Centre Informatique National de l'Enseignement Supérieur) for computational resources (project: A0060710878). F.L.B.d.S. and C.E. acknowledge the mobility support from the joint Université Sorbonne Paris Cité-USP Research program. The authors also gratefully acknowledge the computing time granted by the John von Neumann Institute for Computing (NIC) and provided on the supercomputer JURECA at Jülich Supercomputing Centre (JSC) - project C210400511.

**Author Contributions**: S.A.P.C., F.L.B.d.S., and C.E., performed simulations, analyzed data, and wrote the paper. F.L.B.d.S. and C.E. designed and supervised the research.

## 6. BIBLIOGRAPHY

Akey, D.L., Brown, W.C., Dutta, S., Konwerski, J., Jose, J., Jurkiw, T.J., DelProposto, J., Ogata, C.M., Skiniotis, G., Kuhn, R.J., Smith, J.L., 2014. Flavivirus NS1 structures reveal surfaces for

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associations with membranes and the immune system. Science 343, 881–885.

Barroso da Silva, F.L., Lund, M., Jönsson, B., Åkesson, T., 2006. On the Complexation of Proteins and Polyelectrolytes. J. Phys. Chem. B 110, 4459–4464. https://doi.org/10.1021/jp054880l

Basu, A., Chaturvedi, U.C., 2008. Vascular endothelium: the battlefield of dengue viruses. FEMS Immunol. Med. Microbiol. 53, 287–299. https://doi.org/10.1111/j.1574-695X.2008.00420.x

Beck, C., Jimenez-Clavero, M.A., Leblond, A., Durand, B., Nowotny, N., Leparc-Goffart, I., Zientara, S., Jourdain, E., Lecollinet, S., 2013. Flaviviruses in Europe: complex circulation patterns and their consequences for the diagnosis and control of West Nile disease. Int. J. Environ. Res. Public. Health 10, 6049–6083.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242. https://doi.org/10.1093/nar/28.1.235

Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George, D.B., Jaenisch, T., Wint, G.R.W., Simmons, C.P., Scott, T.W., Farrar, J.J., Hay, S.I., 2013. The global distribution and burden of dengue. Nature 496, 504–507. https://doi.org/10.1038/nature12060

Biering, S.B., Akey, D.L., Wong, M.P., Brown, W.C., Lo, N.T.N., Puerta-Guardo, H., Tramontini Gomes de Sousa, F., Wang, C., Konwerski, J.R., Espinosa, D.A., Bockhaus, N.J., Glasner, D.R., Li, J., Blanc, S.F., Juan, E.Y., Elledge, S.J., Mina, M.J., Beatty, P.R., Smith, J.L., Harris, E., 2021. Structural basis for antibody inhibition of flavivirus NS1–triggered endothelial dysfunction. Science 371, 194–200. https://doi.org/10.1126/science.abc0476

Brown, W.C., Akey, D.L., Konwerski, J.R., Tarrasch, J.T., Skiniotis, G., Kuhn, R.J., Smith, J.L., 2016. Extended surface for membrane association in Zika virus NS1 structure. Nat. Struct. Mol. Biol. 23, 865–867. https://doi.org/10.1038/nsmb.3268

Chen, C.R., Makhatadze, G.I., 2015. ProteinVolume: calculating molecular van der Waals and void volumes in proteins. BMC Bioinformatics 16, 101. https://doi.org/10.1186/s12859-015-0531-2

Ci, Y., Liu, Z.-Y., Zhang, N.-N., Niu, Y., Yang, Y., Xu, C., Yang, W., Qin, C.-F., Shi, L., 2020. Zika NS1–induced ER remodeling is essential for viral replication. J. Cell Biol. 219, e201903062. https://doi.org/10.1083/jcb.201903062

Ci, Y., Yang, Y., Xu, C., Qin, C.-F., Shi, L., 2021. Electrostatic Interaction Between NS1 and Negatively Charged Lipids Contributes to Flavivirus Replication Organelles Formation. Front. Microbiol. 12, 641059. https://doi.org/10.3389/fmicb.2021.641059

Clark, D.C., Lobigs, M., Lee, E., Howard, M.J., Clark, K., Blitvich, B.J., Hall, R.A., 2007. In situ reactions of monoclonal antibodies with a viable mutant of Murray Valley encephalitis virus reveal an absence of dimeric NS1 protein. J. Gen. Virol. 88, 1175–1183.

Cragnolini, T., Derreumaux, P., Pasquali, S., 2015. Ab initio RNA folding. J. Phys. Condens. Matter 27, 233102. https://doi.org/10.1088/0953-8984/27/23/233102

Da Silva, F.L.B., Dias, L.G., 2017. Development of constant-pH simulation methods in implicit solvent and applications in biomolecular systems. Biophys. Rev. 9, 699–728. https://doi.org/10.1007/s12551-017-0311-5

Da Silva, F.L.B., Jönsson, B., 2009. Polyelectrolyte–protein complexation driven by charge regulation. Soft Matter 5, 2862–2868. https://doi.org/10.1039/B902039J

Da Silva, F.L.B., MacKernan, D., 2017. Benchmarking a fast proton titration scheme in implicit solvent for biomolecular simulations. J. Chem. Theory Comput. 13, 2915–2929. https://doi.org/10.1021/acs.jctc.6b01114

Da Silva, F.L.B., Pasquali, S., Derreumaux, P., Dias, L.G., 2016. Electrostatics analysis of the mutational and pH effects of the N-terminal domain self–association of the major ampullate spidroin. Soft Matter 12, 5600–5612. https://doi.org/10.1039/c6sm00860g

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Delboni, L.A., Da Silva, F.L.B., 2016. On the complexation of whey proteins. Food Hydrocoll. 55, 89–99. https://doi.org/10.1016/j.foodhyd.2015.11.010

Edeling, M.A., Diamond, M.S., Fremont, D.H., 2014. Structural basis of Flavivirus NS1 assembly and antibody recognition. Proc. Natl. Acad. Sci. 111, 4285–4290. https://doi.org/10.1073/pnas.1322036111

Edgar, R.C., 2004. MUSCLE a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113. https://doi.org/10.1186/1471-2105-5-113

Eisenberg, D., Weiss, R.M., Terwilliger, T.C., Wilcox, W., 1982. Hydrophobic moments and protein structure. Faraday Symp. Chem. Soc. 17, 109. https://doi.org/10.1039/fs9821700109

Engkvist, O., Karlström, G., 1996. A method to calculate the probability distribution for systems with large energy barriers. Chem. Phys. 213, 63–76. https://doi.org/10.1016/S0301-0104(96)00247-9

Falconar, A.K., Young, P.R., 1990. Immunoaffinity purification of native dimer forms of the Flavivirus non-structural glycoprotein, NS1. J. Virol. Methods 30, 323–332. https://doi.org/10.1016/0166-0934(90)90075-Q

Fernandez-Fuentes, N., Rai, B.K., Madrid-Aliste, C.J., Eduardo Fajardo, J., Fiser, A., 2007. Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. Bioinformatics 23, 2558–2565. https://doi.org/10.1093/bioinformatics/btm377

Fields, B.N., 2014. Fields virology, 6th ed. Lippincott Williams and Wilkins.

Figueiredo, L.T.M., 2000. The Brazilian flaviviruses. Microbes Infect. 2, 1643–1649. https://doi.org/10.1016/S1286-4579(00)01320-4

Flamand, M., Megret, F., Mathieu, M., Lepault, J., Rey, F.A., Deubel, V., 1999. Dengue Virus Type 1 Nonstructural Glycoprotein NS1 Is Secreted from Mammalian Cells as a Soluble Hexamer in a Glycosylation-Dependent Fashion. J. Virol. 73, 6104–6110. https://doi.org/10.1128/JVI.73.7.6104-6110.1999

Freire, M.C., Pol-Fachin, L., Coêlho, D.F., Viana, I.F., Magalhães, T., Cordeiro, M.T., Fischer, N., Loeffler, F.F., Jaenisch, T., Franca, R.F., Marques, E.T.A., Lins, R.D., 2017. Mapping putative B–cell Zika virus NS1 epitopes provides molecular basis for anti-NS1 antibody discrimination between Zika and Dengue viruses. ACS Omega 2, 3913–3920. https://doi.org/10.1021/acsomega.7b00608

George, A., Wilson, W.W., 1994. Predicting protein crystallization from a dilute solution property. Acta Crystallogr. D Biol. Crystallogr. 50, 361–365. https://doi.org/10.1107/S0907444994001216

Giron, C.C., Laaksonen, A., Barroso da Silva, F.L., 2021. Up State of the SARS-COV-2 Spike Homotrimer Favors an Increased Virulence for New Variants. Front. Med. Technol. 3, 694347. https://doi.org/10.3389/fmedt.2021.694347

Giron, C.C., Laaksonen, A., Barroso da Silva, F.L., 2020. On the interactions of the receptor-binding domain of SARS-CoV-1 and SARS-CoV-2 spike proteins with monoclonal antibodies and the receptor ACE2. Virus Res.

Gonçalves, R.L., de Lima Menezes, G., Sussuchi, L., Moreli, M.L., Mottin, M., Andrade, C.H., Pereira, M., da Silva, R.A., 2019. Dynamic behavior of Dengue and Zika viruses NS1 protein reveals monomer–monomer interaction mechanisms and insights to rational drug design. J. Biomol. Struct. Dyn. 1–11.

Gong, Z., Gao, Y., Han, G.-Z., 2016. Zika virus: two or three lineages? Trends Microbiol. 24, 521–522.

Gutsche, I., Coulibaly, F., Voss, J.E., Salmon, J., d'Alayer, J., Ermonval, M., Larquet, E., Charneau, P., Krey, T., Mégret, F., 2011. Secreted dengue virus nonstructural protein NS1 is an atypical barrel-shaped high-density lipoprotein. Proc. Natl. Acad. Sci. 108, 8003–8008.

Hall, R.A., Khromykh, A.A., Mackenzie, J.M., Scherret, J.H., Khromykh, T.I., Mackenzie, J.S., 1999. Loss of dimerisation of the nonstructural protein NS1 of Kunjin virus delays viral replication and reduces virulence in mice, but still allows secretion of NS1. Virology 264, 66–75.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hertz, T., Beatty, P.R., MacMillen, Z., Killingbeck, S.S., Wang, C., Harris, E., 2017. Antibody Epitopes Identified in Critical Regions of Dengue Virus Nonstructural 1 Protein in Mouse Vaccination and Natural Human Infections. J. Immunol. 198, 4025–4035. https://doi.org/10.4049/jimmunol.1700029

Jonsson, B., Lund, M., 2007. Electrostatics in Macromolecular Solution, in: Food Colloids: Self-Assembly and Material Science. pp. 129–154.

Jurado, K.A., Simoni, M.K., Tang, Z., Uraki, R., Hwang, J., Householder, S., Wu, M., Lindenbach, B.D., Abrahams, V.M., Guller, S., others, 2016. Zika virus productively infects primary human placenta-specific macrophages. JCI Insight 1.

Kegel, W.K., van der Schoot, P., 2004. Competing Hydrophobic and Screened-Coulomb Interactions in Hepatitis B Virus Capsid Assembly. Biophys. J. 86, 3905–3913. https://doi.org/10.1529/biophysj.104.040055

Lai, Y.-C., Chuang, Y.-C., Liu, C.-C., Ho, T.-S., Lin, Y.-S., Anderson, R., Yeh, T.-M., 2017. Antibodies Against Modified NS1 Wing Domain Peptide Protect Against Dengue Virus Infection. Sci. Rep. 7, 6975. https://doi.org/10.1038/s41598-017-07308-3

Larsson, P., Wallner, B., Lindahl, E., Elofsson, A., 2008. Using multiple templates to improve quality of homology models in automated homology modeling. Protein Sci. 17, 990–1002. https://doi.org/10.1110/ps.073344908

Li, A., Yu, J., Lu, M., Ma, Y., Attia, Z., Shan, C., Xue, M., Liang, X., Craig, K., Makadiya, N., He, J.J., Jennings, R., Shi, P.-Y., Peeples, M.E., Liu, S.-L., Boyaka, P.N., Li, J., 2018. A Zika virus vaccine expressing premembrane-envelope-NS1 polyprotein. Nat. Commun. 9, 3067. https://doi.org/10.1038/s41467-018-05276-4

Lin, B., Pryor, M.J., Davidson, A.D., Wright, P.J., Gualano, R.C., 1998. Growth restriction of dengue virus type 2 by site-specific mutagenesis of virus-encoded glycoproteins. J. Gen. Virol. 79, 2631–2639. https://doi.org/10.1099/0022-1317-79-11-2631

Lund, M., Jönsson, B., 2003. A mesoscopic model for protein–protein interactions in solution. Biophys. J. 85, 2940–2947. https://doi.org/10.1016/S0006-3495(03)74714-6

Lund, M., Trulsson, M., Persson, B., 2008. Faunus: an object oriented framework for molecular simulation. Source Code Biol. Med. 3, 1.

Lunkad, R., Barroso da Silva, F.L., Košovan, P., 2022. Both Charge-Regulation and Charge-Patch Distribution Can Drive Adsorption on the Wrong Side of the Isoelectric Point. J. Am. Chem. Soc. 144, 1813–1825. https://doi.org/10.1021/jacs.1c11676

McQuarrie, D., 1976. Statistical mechanics. Harper and Row.

Mendonça, D.C., Macedo, J.N., Guimarães, S.L., Barroso da Silva, F.L., Cassago, A., Garratt, R.C., Portugal, R.V., Araujo, A.P.U., 2019. A revised order of subunits in mammalian septin complexes. Cytoskeleton 76, 457–466. https://doi.org/10.1002/cm.21569

Menezes, G. de L., Gonçalves, R.L., Lima, R.M., Pereira, M., Moreli, M.L., da Silva, R.A., 2020. Mutation of critical residues reveals insights of yellow fever virus nonstructural protein 1 (NS1) stability and its formation. J. Biomol. Struct. Dyn. 38, 5277–5286. https://doi.org/10.1080/07391102.2019.1695673

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., 1953. Equation of State Calculations by Fast Computing Machines. J. Chem. Phys 21, 7.

Mlakar, J., Korva, M., Tul, N., Popović, M., Poljšak-Prijatelj, M., Mraz, J., Kolenc, M., Resman Rus, K., Vesnaver Vipotnik, T., Fabjan Vodušek, V., 2016. Zika virus associated with microcephaly. N. Engl. J. Med. 2016, 951–958. https://doi.org/10.1056/NEJMoa1600651

Moureau, G., Cook, S., Lemey, P., Nougairede, A., Forrester, N.L., Khasnatinov, M., Charrel, R.N., Firth, A.E., Gould, E.A., de Lamballerie, X., 2015. New Insights into Flavivirus Evolution, Taxonomy and Biogeographic History, Extended by Analysis of Canonical and Alternative Coding

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequences. PLOS ONE 10, e0117849. https://doi.org/10.1371/journal.pone.0117849

Muller, D.A., Landsberg, M.J., Bletchly, C., Rothnagel, R., Waddington, L., Hankamer, B., Young, P.R., 2012. Structure of the dengue virus glycoprotein non-structural protein 1 by electron microscopy and single-particle analysis. J. Gen. Virol. 93, 771–779. https://doi.org/10.1099/vir.0.039321-0

Muller, D.A., Young, P.R., 2013. The Flavivirus NS1 protein: molecular and structural biology, immunology, role in pathogenesis and application as a diagnostic biomarker. Antiviral Res. 98, 192–208. https://doi.org/10.1016/j.antiviral.2013.03.008

Musso, D., Desprès, P., 2020. Serological Diagnosis of Flavivirus-Associated Human Infections. Diagnostics 10, 302. https://doi.org/10.3390/diagnostics10050302

Neal, J.W., 2014. Flaviviruses are neurotropic, but how do they invade the CNS? J. Infect. 69, 203–215. https://doi.org/10.1016/j.jinf.2014.05.010

Neamtu, A., Mocci, F., Laaksonen, A., Barroso da Silva, F.L., 2022. Towards an optimal monoclonal antibody with higher binding affinity to the receptor-binding domain of SARS-CoV-2 spike proteins from different variants. bioRxiv 2022.01.04.474958. https://doi.org/10.1101/2022.01.04.474958

Neufeldt, C.J., Cortese, M., Acosta, E.G., Bartenschlager, R., 2018. Rewiring cellular networks by members of the Flaviviridae family. Nat. Rev. Microbiol. 16, 125–142.

Noisakran, S., Dechtawewat, T., Avirutnan, P., Kinoshita, T., Siripanyaphinyo, U., Puttikhunt, C., Kasinrerk, W., Malasit, P., Sittisombut, N., 2008. Association of dengue virus NS1 protein with lipid rafts. J. Gen. Virol. 89, 2492–2500.

Persson, B.A., Lund, M., Forsman, J., Chatterton, D.E., Akesson, T., 2010. Molecular evidence of stereo–specific Lactoferrin dimers in solution. Biophys. Chem. 151, 187–189.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612.

Pettersson, J.H.-O., Fiz-Palacios, O., 2014. Dating the origin of the genus Flavivirus in the light of Beringian biogeography. J. Gen. Virol. 95, 1969–1982. https://doi.org/10.1099/vir.0.065227-0

Pjura, P.E., Lenhoff, A.M., Leonard, S.A., Gittis, A.G., 2000. Protein crystallization by design: chymotrypsinogen without precipitants 1 1Edited by I. A. Wilson. J. Mol. Biol. 300, 235–239. https://doi.org/10.1006/jmbi.2000.3851

Plourde, A.R., Bloch, E.M., 2016. A literature review of Zika virus. Emerg. Infect. Dis. 22, 1185.

Poveda-Cuevas, S.A., 2021. CPRISMA (Coloring Proteins by Inputs and Sets of Multiple Alignments) software. https://github.com/seapovedac/cprisma.

Poveda-Cuevas, S.A., Barroso da Silva, F.L., Etchebest, C., 2021. How the Strain Origin of Zika Virus NS1 Protein Impacts Its Dynamics and Implications to Their Differential Virulence. J. Chem. Inf. Model. 61, 1516–1530. https://doi.org/10.1021/acs.jcim.0c01377

Poveda-Cuevas, S.A., Etchebest, C., Barroso da Silva, F.L., 2020. Identification of Electrostatic Epitopes in Flavivirus by Computer Simulations: The PROCEEDpKa Method. J. Chem. Inf. Model. 60, 944–963.

Poveda-Cuevas, S.A., Etchebest, C., Barroso da Silva, F.L., 2018. Insights into the ZIKV NS1 Virology from Different Strains through a Fine Analysis of Physicochemical Properties. ACS Omega 3, 16212–16229.

Puerta-Guardo, H., Glasner, D.R., Espinosa, D.A., Biering, S.B., Patana, M., Ratnasiri, K., Wang, C., Beatty, P.R., Harris, E., 2019. Flavivirus NS1 triggers tissue-specific vascular endothelial dysfunction reflecting disease tropism. Cell Rep. 26, 1598–1613.

Qiao, B., Olvera de la Cruz, M., 2020. Enhanced Binding of SARS-CoV-2 Spike Protein to Receptor by Distal Polybasic Cleavage Sites. ACS Nano 14, 10616–10623. https://doi.org/10.1021/acsnano.0c04798

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Scaturro, P., Cortese, M., Chatel-Chaix, L., Fischl, W., Bartenschlager, R., 2015. Dengue Virus Non-structural Protein 1 Modulates Infectious Particle Production via Interaction with the Structural Proteins. PLOS Pathog. 11, e1005277. https://doi.org/10.1371/journal.ppat.1005277

Shen, M., Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. Protein Sci. 15, 2507–2524. https://doi.org/10.1110/ps.062416606

Somnuke, P., Hauhart, R.E., Atkinson, J.P., Diamond, M.S., Avirutnan, P., 2011. N–linked glycosylation of dengue virus NS1 protein modulates secretion, cell–surface expression, hexamer stability, and interactions with human complement. Virology 413, 253–264.

Song, H., Qi, J., Haywood, J., Shi, Y., Gao, G.F., 2016. Zika virus NS1 structure reveals diversity of electrostatic surfaces among flaviviruses. Nat. Struct. Mol. Biol. 23, 456–458. https://doi.org/10.1038/nsmb.3213

Sridhar, S., Luedtke, A., Langevin, E., Zhu, M., Bonaparte, M., Machabert, T., Savarino, S., Zambrano, B., Moureau, A., Khromava, A., Moodie, Z., Westling, T., Mascareñas, C., Frago, C., Cortés, M., Chansinghakul, D., Noriega, F., Bouckenooghe, A., Chen, J., Ng, S.-P., Gilbert, P.B., Gurunathan, S., DiazGranados, C.A., 2018. Effect of Dengue Serostatus on Dengue Vaccine Safety and Efficacy. N. Engl. J. Med. 379, 327–340. https://doi.org/10.1056/NEJMoa1800820

Suci, P.A., Klem, M.T., Douglas, T., Young, M., 2005. Influence of Electrostatic Interactions on the Surface Adsorption of a Viral Protein Cage. Langmuir 21, 8686–8693. https://doi.org/10.1021/la050217c

Teixeira, A.A.R., Lund, M., Da Silva, F.L.B., 2010. Fast proton titration scheme for multiscale modeling of protein solutions. J. Chem. Theory Comput. 6, 3259–3266. https://doi.org/10.1021/ct1003093

Tessier, P.M., Lenhoff, A.M., Sandler, S.I., 2002. Rapid Measurement of Protein Osmotic Second Virial Coefficients by Self-Interaction Chromatography. Biophys. J. 82, 1620–1631. https://doi.org/10.1016/S0006-3495(02)75513-6

Thuman-Commike, P.A., Greene, B., Malinski, J.A., King, J., Chiu, W., 1998. Role of the Scaffolding Protein in P22 Procapsid Size Determination Suggested by T=4 and T=7 Procapsid Structures. Biophys. J. 74, 559–568. https://doi.org/10.1016/S0006-3495(98)77814-2

Trovato, M., Sartorius, R., D'Apice, L., Manco, R., De Berardinis, P., 2020. Viral Emerging Diseases: Challenges in Developing Vaccination Strategies. Front. Immunol. 11, 2130. https://doi.org/10.3389/fimmu.2020.02130

Velev, O.D., Kaler, E.W., Lenhoff, A.M., 1998. Protein Interactions in Solution Characterized by Light and Neutron Scattering: Comparison of Lysozyme and Chymotrypsinogen. Biophys. J. 75, 2682–2697. https://doi.org/10.1016/S0006-3495(98)77713-6

Wang, D., Chen, Cheng, Liu, S., Zhou, H., Yang, K., Zhao, Q., Ji, X., Chen, Chen, Xie, W., Wang, Z., 2017. A mutation identified in neonatal Microcephaly destabilizes Zika virus NS1 assembly in vitro. Sci. Rep. 7, 42580.

Watterson, D., Modhiran, N., Young, P.R., 2016. The many faces of the flavivirus NS1 protein offer a multitude of options for inhibitor design. Antiviral Res. 130, 7–18.

Webb, B., Sali, A., 2014. Protein structure modeling with MODELLER. Protein Struct. Predict. 1–15.

Wessel, A.W., Doyle, M.P., Engdahl, T.B., Rodriguez, J., Crowe, J.E., Diamond, M.S., 2021. Human Monoclonal Antibodies against NS1 Protein Protect against Lethal West Nile Virus Infection. mBio 12, e02440-21. https://doi.org/10.1128/mBio.02440-21

Wilkins, M., Gasteiger, E., Bairoch, A.S., J-C, W., KL, A., 1999. Protein identification and analysis tools in the ExPASy server 112, 531–552.

Winkler, G., Maxwell, S.E., Ruemmler, C., Stollar, V., 1989. Newly synthesized Dengue–2 virus nonstructural protein NS1 is a soluble protein but becomes partially hydrophobic and membrane–associated after dimerization. Virology 171, 302–305. https://doi.org/10.1016/0042-6822(89)90544-8

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.15.484429; this version posted March 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

World Health Organization, 2022. WHO [WWW Document]. URL https://www.who.int/emergencies/diseases/novel-coronavirus-2019

Xian, Y., Karki, C.B., Silva, S.M., Li, L., Xiao, C., 2019. The Roles of Electrostatic Interactions in Capsid Assembly Mechanisms of Giant Viruses. Int. J. Mol. Sci. 20, 1876. https://doi.org/10.3390/ijms20081876

Xu, X., Song, H., Qi, J., Liu, Y., Wang, H., Su, C., Shi, Y., Gao, G.F., 2016. Contribution of intertwined loop to membrane association revealed by Zika virus full–length NS1 structure. EMBO J. 35, 2170–2178. https://doi.org/10.15252/embj.201695290

Yoshida, K., Kuroda, D., Kiyoshi, M., Nakakido, M., Nagatoishi, S., Soga, S., Shirai, H., Tsumoto, K., 2019. Exploring designability of electrostatic complementarity at an antigen-antibody interface directed by mutagenesis, biophysical analysis, and molecular dynamics simulations. Sci. Rep. 9, 4482. https://doi.org/10.1038/s41598-019-40461-5

## Chapter 8

# Preliminary insights about the NS1 dynamical behavior

The focus of this chapter was to evaluate the dynamical behavior of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ in an electrolyte solution without any partner. A greater emphasis was given to the evaluation of the dynamic and structural properties of the two $NS1_{ZIKV}$ through canonical and statistical measures[1] (see Subsection 4.3.1). This part of the investigation complements the results shown in Chapter 5 where only electrostatic properties were considered without any type of protein conformational change.

The main question of this part of the study is:

*Are the MD of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ discrepant?*

A summary of the molecular model used, the method to solve them, and the measures calculated appear in Table 8.1.

Table 8.1: Molecular modeling and measurements carried out in Chapter 8.

| Level of detail | Solution | Measurements |
|---|---|---|
| AA | Classical MD | Canonical + PCA |
| CG model at the 4to1 level | | Canonical + Statistical |

Link paper: https://pubs.acs.org/doi/10.1021/acs.jcim.0c01377.

[1]It is important to mention that proteins employed here do not have sugars in their 3D structures. However, sugar molecules seem to do not have a meaningful impact at least for the dynamic properties evaluated here (see Appendix B).

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

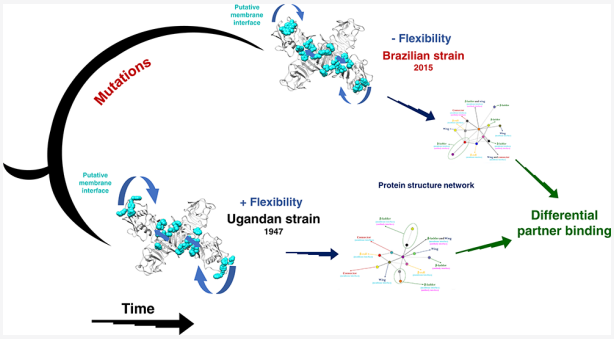
# How the Strain Origin of Zika Virus NS1 Protein Impacts Its Dynamics and Implications to Their Differential Virulence

Sergio A. Poveda-Cuevas, Fernando Luís Barroso da Silva, and Catherine Etchebest*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 1516−1530 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Viruses can impact and affect human populations in a severe way. The appropriate differentiation among several species or strains of viruses is one of the biggest challenges for virology and infectiology studies. The detection of measurables-quantified discrepancies allows for more accurate clinical diagnoses and treatments for viral diseases. In the present study, we have used a computational approach to explore the dynamical properties of the nonstructural protein 1 from two strains of Zika virus. Our results show that despite a high sequence similarity, the two viral proteins from different origins can exhibit significant dissimilar structural dynamics, which complement their reported differential virulence. The present study opens up new ways in the understanding of the infectivity for these biological entities.




Downloaded via UNIV OF SAO PAULO on March 24, 2021 at 14:57:19 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

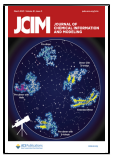
## INTRODUCTION

The recent SARS-CoV-2 pandemic has undoubtedly highlighted the impact of viral infections on the health of the population worldwide, associated with major consequences on economics.[1] Besides coronaviruses, flaviviruses, like dengue (DENV), yellow fever, or West Nile viruses, to name a few, are continuously a health threat.[2] The viruses have received renewed attention in the last past years because of the emergence of new species with a severe pathogenicity like the recent outbreaks of Zika or due to the increase of their geographical spread, for example, DENV with cases identified in the south of Europe.[3,4] Climate changes are strongly believed to be responsible for this expansion. The very recent review by Pierson and Diamond convincingly emphasizes the importance of this genus of viruses and their impact on human health.[5]

Despite several attempts, there are still no therapeutics solutions for these flaviviruses, and the main way to protect against them involves avoiding bites from the infectious vehicles, mainly mosquitoes of the *Aedes* genus.[6,7] Moreover, the similarity of symptoms consecutive to the infection makes difficult the clinical diagnostics of the specific disease. Classic laboratory tests (e.g., RT-PCR or serological tests) are thus required to precisely identify the infectious agent.[8] However, even with such elaborate tests, difficulties remain because of the high similarity of the genome structures and genome products of these different species.[9,10] Indeed, the flavivirus genome is a single-stranded positive-sense RNA (10.7 kb) that encodes a polyprotein precursor. The virion contains three structural proteins, the capsid protein (C), the envelope protein (Env), and the premembrane (prM) or membrane (M) protein in immature and mature virions, respectively. Besides, seven nonstructural (NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5) biomolecules are synthesized in the lumen of the endoplasmic reticulum of the host cells.[11,12] The sequence similarity between these proteins from different species is astonishingly high, in sharp contrast with the large diversity of virulence at the inter- (i.e., among several members of a genus) or intraspecific level (i.e., among strains of a unique virus species).[13] Accordingly, this may lead to a possible cross-reactivity of antibodies that might not clearly distinguish between Zika virus (ZIKV) and one of the four DENV serotypes, for instance.[14,15] This may result in misdiagnosing and prescribing inappropriate treatments.

Several viral targets have been studied for treatment against ZIKV. Among the structural proteins, Env is the most promising one for the inactivation of flaviviruses.[16,17] Some studies suggest that Env would be the best target for human ZIKV or DENV-neutralizing antibodies, proving to be a key element for the development of effective treatments.[16,18] To a lesser extent, NS proteins are also important for the treatment or identification of flaviviruses.[19,20] However, from the point of view of how these viruses come to exhibit a diverse virulence pattern, NS proteins turn out to be more relevant in

**Received:** November 27, 2020
**Published:** March 3, 2021




© 2021 American Chemical Society

https://dx.doi.org/10.1021/acs.jcim.0c01377
*J. Chem. Inf. Model.* 2021, 61, 1516−1530

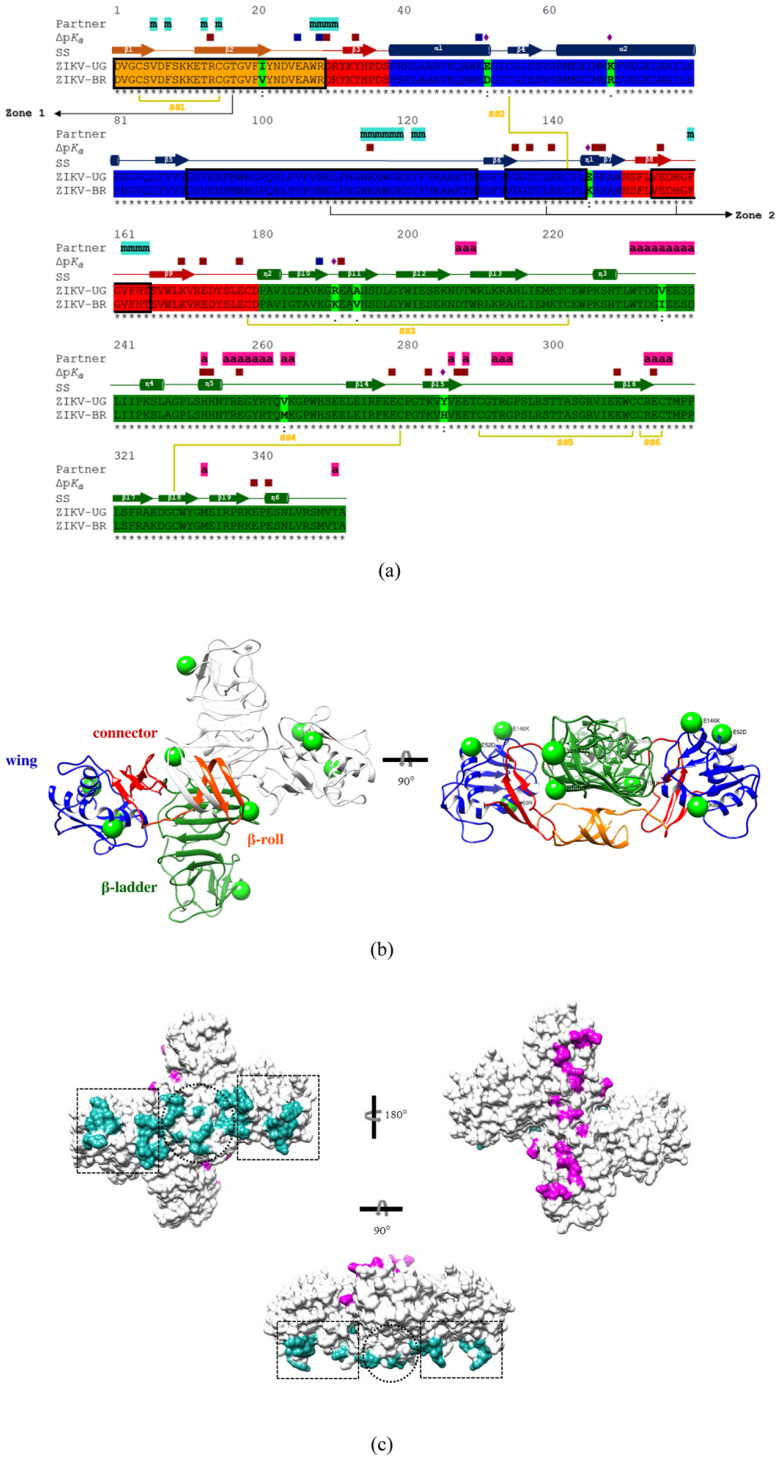



**Figure 1.** ZIKV NS protein 1. (a) Protein sequence alignment between $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ (sequences provided by NCBI and identified as AY632535 and KU365777, respectively). $\beta$-Roll, wing, and $\beta$-ladder domains are colored in orange, blue, and green, respectively.[19] The connector subdomain is colored in red. $\beta$-Strands, $\alpha$-helix, and coil structures are shown as arrows, cylinders, and lines, respectively. Regions with light green color highlight the mutations between the strains. Disulfide bridges (or SS bonds among cysteine residues) are represented by the yellow lines. $\Delta pK_a$ values are as given in ref 36, and putative regions of interaction with the membrane and antibody (defined in ref 36) are shown. Black rectangles refer to zones 1 and 2 (for more details, see the Results and Discussion section). (b) Domain organization of NS1 using ribbon's representation for the X-ray crystal structure of the ZIKV $NS1_{BR}$ dimer (PDB id 5GS6). The color pattern is the same as given in Figure 1a. The light green spheres represent the location of sequence changes between $NS1_{UG}$ and $NS1_{BR}$. (c) Putative regions of interaction with the membrane are colored in cyan and those with antibody in violet. It is to be noted that the 90° rotation about the $x$-axis for the dimeric NS1 represents the estimated position of interaction with the membrane. Dashed circles and rectangles refer to zones 1 and 2, respectively (for more details, see the Results and Discussion section).



https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

this regard.[21] NS1 is of particular interest because of its involvement in various key viral processes (e.g., viral replication, immune evasion, and pathogenesis).[11,12] In addition to being secreted at high levels, it displays significant differences in surface properties between ZIKV and DENV. Therefore, it has been proposed as a valuable biomarker for the distinction of flaviviruses.[9,22] Moreover, in the case of ZIKV, more aggressive virulence was found for African strains than for Asian viruses.[23−25] In fact, the biological partnerships of NS1 proteins and their behavior should be examined in priority because both infectivity and virulence are linked to the capacity of the viral strains of crossing membrane barriers, interacting with host protein partners, and interfering with their functions as well as of escaping from the immune host systems [that is, antibodies (Abs)]; all of them rely on the NS1-related mechanisms.[11,26,27]

In addition, during these different infectious processes, Zika NS1, denoted $NS1_{ZIKV}$ here, should adopt different oligomeric states, for example, dimer or hexamer, to be effective.[11,12,28] Any perturbation of the NS1 protein self-association could influence the viral future events. For example, a single mutation can impact the dimerization of $NS1_{ZIKV}$ by disrupting an elaborated hydrogen-bonding network across the dimer interface.[29] Site-directed mutagenesis assays on NS1 were shown to affect dimerization, which correlates with the slow growth of the Kunjin and Murray Valley encephalitis flaviviruses.[30,31] A particular molecular mechanism was reported by classical molecular dynamics (MD) simulations, giving a probable path in the understanding of these effects.[32] On the other hand, membrane associations (specifically, with lipid rafts) of the dimeric NS1 may be affected by the absence of specific sequences in the C-terminal region.[33] The change in the capacity to interact with lipid rafts was also shown to affect infectivity for other viruses.[34,35] Many of these processes are in general strongly interrelated and depend in particular on the membrane assembly (e.g., hexamer formation, cell surface recognition, and so on). Accordingly, they should be examined in concert. In summary, modification of the ability of NS1 to associate or recognize one of these biological partners could have a direct relationship with the differential virulence of the ZIKV strains from two phylogenetic origins.

From a therapeutics point of view, understanding how small sequence variations have a drastic impact is of major importance in improving the specificity of serological tools and even in designing appropriate therapeutics solutions. This is the reason why we have focused our recent studies on NS1[36,37] and explored in-depth the consequence of sequence variation on its physical−chemical properties and, per se, its recognition and binding capacity. Initially, we have concentrated on electrostatic properties because they act at a long-range distance and are particularly sensitive to environment changes like the salt concentration and pH, which in turn would affect the NS1 association properties with its three main in vivo biological partners (i.e., other NS1 monomers, antibodies, and the cellular membrane).[36,37] By employing constant-pH (CpH) Monte Carlo (MC) simulations, we found that the $pK_a$ shifts ($\Delta pK_a$) on the three NS1 interfaces are significantly discrepant between strains from Uganda (UG) and Brazil (BR), with African and Asian origins, respectively.[36] Subsequently, more detailed analysis confirmed that the NS1−Ab interface of these two strains can exhibit different electrostatic epitope regions, despite having high sequence identities (~97%).[37] Although it was already stated that discrepancies may occur from the electrostatic point of view or by observing pH-dependent processes,[9,22,36,37] it is also important to examine another fundamental biological property, namely, the dynamics of the protein. Actually, this is supported by several recent studies that examined NS1 proteins from DENV and ZIKV by using constant-charge MD simulations, where interesting local differences were found for the monomer−monomer interface, despite an extremely well-preserved fold.[38] Whether these differences at the dynamical level hold good at the intraspecific level, specifically for the ZIKV proteins $NS1_{UG}$ and $NS1_{BR}$, is the challenging question that we have addressed in detail in the present work. With this aim, we used both all-atoms (AA) and coarse-grained (CG) classical MD simulations that have already proved to be quite efficient in several previous computational studies of other viral proteins.[39−41] It is worth mentioning that CG MD simulations permit to reach a significant sampling of the conformational space and strengthen the conclusions accordingly.

Therefore, we first explored the dynamics of two oligomeric states, that is, monomer and dimer, in order to elucidate the role of each individual chain in the dimeric structure. Next, we evaluated the impact of the sequence changes (Figure 1a,b) on the dynamics of the two NS1s strains and concentrated our analyses more specifically on their functional interfaces, that is, NS1−NS1, NS1−membrane, and NS1−Ab (Figure 1c). We carried out canonical analyses and essential dynamics (ED) analyses to better detect putative discrepancies. A comparison of the dynamical network of the two strains finally reveals interesting features that could provide clues regarding the differential virulence observed between the two strains.

## RESULTS AND DISCUSSION

In a previous study,[36] we have shown that using the fast proton titration scheme (FPTS) approach,[42−44] we can reveal how slight sequence modifications in the NS1 monomer from different strains can affect electrostatic properties in regions far from the mutation sites and, per se, impact putative interactions between the NS1 with partners. In the present work, focusing on another fundamental aspect (i.e., the protein dynamics), we have examined the impact of the small sequence variations on the dynamics properties of $NS1_{ZIKV}$ protein from Uganda and Brazil strains and their possible relations with the biological features/mechanisms/observations. Accordingly, we have performed constant-charge MD simulations on the corresponding $NS1_{ZIKV}$ 3D structures of each strain. The $NS1_{ZIKV}$ dimer is considered as the functional unit, but the monomeric form has already been observed.[38] Therefore, we explored both oligomeric states in order to better decipher the role of the monomer unit in the dimer dynamics and also the role of the sequence itself by the dimeric analysis. Besides the overall dynamics features, we also focused on functional regions that we previously examined in-depth, namely, the monomer−monomer interface, the epitope sites, and the membrane binding region.[36,37] The results are detailed in the subsequent sections that are organized as follows: first, the monomer dynamic properties obtained from AA MD simulations are compared for each strain, and then, the dimer properties are examined in comparison with monomers. Afterward, results achieved with a CG force field, which improves conformational sampling, are detailed, and a similar comparison is conducted. For the sake of


https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

Table 1. Average Structural Properties Sampled by the MD Simulations for $NS1_{UG}$ and $NS1_{BR}$[a]

| | CHARMM36m | | | | MARTINI ElNeDyn | |
|---|---|---|---|---|---|---|
| property | $NS1_{UG}^{monomer}$ | $NS1_{BR}^{monomer}$ | $NS1_{UG}^{dimer}$ | $NS1_{BR}^{dimer}$ | $NS1_{UG}^{dimer}$ | $NS1_{BR}^{dimer}$ |
| RMSD (nm) | 0.5 ± 0.1 | 0.5 ± 0.1 | 0.28 ± 0.05 | 0.25 ± 0.06 | 0.38 ± 0.08 | 0.36 ± 0.04 |
| $RMSD_{chain}$ (nm)[b] | | | 0.23 ± 0.06 | 0.20 ± 0.04 | 0.25 ± 0.03 | 0.29 ± 0.04 |
| $RMSD_{no\text{-}flex}$ (nm)[c] | 0.24 ± 0.08 | 0.17 ± 0.02 | 0.11 ± 0.03 | 0.12 ± 0.02 | | |
| RMSF (nm) | 0.2 ± 0.2 | 0.2 ± 0.3 | 0.10 ± 0.09 | 0.10 ± 0.06 | 0.13 ± 0.06 | 0.12 ± 0.06 |
| $R_g$ (nm) | 2.35 ± 0.03 | 2.32 ± 0.02 | 3.12 ± 0.02 | 3.09 ± 0.02 | **3.10 ± 0.02** | **3.03 ± 0.02** |
| $R_{g_no\text{-}flex}$ (nm)[c] | 2.36 ± 0.04 | 2.32 ± 0.02 | 2.19 ± 0.01 | 2.21 ± 0.01 | | |
| SASA ($nm^2$) | 186 ± 4 | 184 ± 4 | 333 ± 5 | 329 ± 6 | 411 ± 7 | 409 ± 6 |
| $SASA_{no\text{-}flex}$ ($nm^2$)[c] | **171 ± 4** | **163 ± 3** | 146 ± 2 | 147 ± 2 | | |
| average distance (nm) | | | 4.10 ± 0.04 | 4.10 ± 0.05 | 4.11 ± 0.07 | 3.97 ± 0.07 |
| number of contacts[d] | | | 661 ± 30 | 610 ± 36 | 112 ± 9 | 108 ± 10 |
| OF[e] | | | 0.04 ± 0.15 | 0.04 ± 0.15 | 0.07 ± 0.20 | 0.06 ± 0.19 |

[a]Time of the trajectories corresponds to 200 ns and 10 $\mu$s for the sampling with CHARMM36m[45] and MARTINI ElNeDyn,[46,47] respectively. [b]Property calculated from trajectories for chains A and B separately. [c]Property calculated from AA trajectories without flexible parts, namely, residues 36−348 for monomers and 35−110 and 167−348 for dimers. [d]Number of contacts between two chains. [e]OF for residues of chain A using thresholds less than 0.5 and 0.52 for the AA and the CG models, respectively. See the text for the meaning of the other quantities and the data highlighted in bold.

synthesis and clarity, the average structural properties are listed in Table 1.

**Main Structural Properties for Monomers.** Figure 2 displays the standard structural properties computed from the MD trajectories generated with the AA representation model. The root mean square deviation (RMSD), radius of gyration ($R_g$), surface accessible surface area (SASA), and root mean square fluctuation (RMSF) are given for both the $NS1_{UG}$ and $NS1_{BR}$ monomers and averaged over the three replicates. For each monomer, most of the properties reach equilibrium at least on the last 100 ns of the simulations. For instance, a plateau of 0.5 nm was found after 70 ns of sampling for the RMSD (see Figure 2a). These substantial large RMSD values (>3 Å) would indicate that significant conformational changes might occur in the monomeric structure. Yet, changes are in fact well localized because the radius of gyration (see Figure 2b) and the solvent surface accessible area (see Figure 2c) remain rather stable over the simulation time. The regions of local changes are highlighted by the RMSF profile (see Figure 2d), with two main regions associated with high values responsible for the large RMSD seen values. The first one located (zone 1) in the $\beta$-roll domain is associated with motions of the loop segment (residues 19−33) that links the $\beta$-hairpin motif to the $\beta$3 strand of the connector subdomain, as shown in Figure 1a,c. The second region (zone 2) involves residues 91−130, 135−145, and 155−165, which correspond, respectively, to the very long intertwined loop of the wing domain, to the region loop linking the $\beta$6 strands with the $\eta$1-helix/$\beta$7 strand in the wing domain, and finally to the loop between $\beta$8 and $\beta$9 strands in the connector subdomain (see Figure 2d).[10] The quantitative impact of these regions is also confirmed by the measures in which they have been discarded and labeled $RMSD_{no\text{-}flex}$, $R_{g_no\text{-}flex}$, and $SASA_{no\text{-}flex}$ (see Table 1). Most of these quantities are reduced to values normally expected for thermal fluctuations.

Interestingly, both zones 1 and 2 are functionally important; zone 2 is associated with the interaction of NS1 with the membrane (compare cyan rectangles in Figure 2e), while zone 1 is involved both in the interaction with the membrane and in the oligomerization process.[28] The flexibility helps in optimizing conformational adjustments throughout complex formation.[38]

It is worth mentioning that each monomer is stabilized by several critical disulfide bridges, directly related to the protein folding as demonstrated by mutagenesis studies.[11,12,48−50] These SS bonds are distributed over the three-dimensional structure [C4−C15 (SS1), C55−C143 (SS2), C179−C223 (SS3), C280−C329 (SS4), C291−C312 (SS5), and C313−C316 (SS6), see Figure 1a][28] and bring additional stability, for example, SS1 located between the two $\beta$-strands of the $\beta$-hairpins motifs at the N-terminus domain. Others like SS2, SS3, and SS4 involve relatively distant pair of cysteine residues and thus contribute to the overall structural integrity. As expected, these regions exhibit low fluctuations.

**Main Structural Properties for Dimers.** Within dimers, which do not involve any covalent bond between monomers, regardless of the property examined, the values for each monomeric unit are significantly lower compared to the monomer alone, for example, the RMSD does not exceed 3 Å. Clearly, the conformational changes observed in Figure 2a and the flexibility shown in Figure 2d are attenuated when a second chain is present, imposing some natural constraints. However, care should be taken before definitely concluding because the slope of the RMSD curve continues to increase, which would indicate that the equilibrium is not fully reached even over the 200 ns of the simulation. Nevertheless, on this timescale, SASA and $R_g$ remain rather constant and the RMSF profiles confirm the reduction of flexibility. This is observed in particular for zone 1 (involved in the dimer/membrane interface) but also for zone 2 (the membrane binding region). Indeed, the interactions between the two chains in the dimeric state reduce the fluctuations of the $\beta$-hairpins in the $\beta$-roll domain and those of the intertwined loop in the wing domain. Similar outcomes were obtained in previous studies.[38] Yet, the two loops located between residues 135−145 and 155−165 retain a similar significant flexibility extent independently on the oligomeric state.

Overall, the presence of a second chain does not significantly influence the monomer secondary structure (SS) pattern (see Figure S1), whereas it has a notable impact on the dynamics of loops located in functional regions even far away from the interface.

**Influence of Strains on Dynamical Properties.** The comparison of the structural properties for each strain does



https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

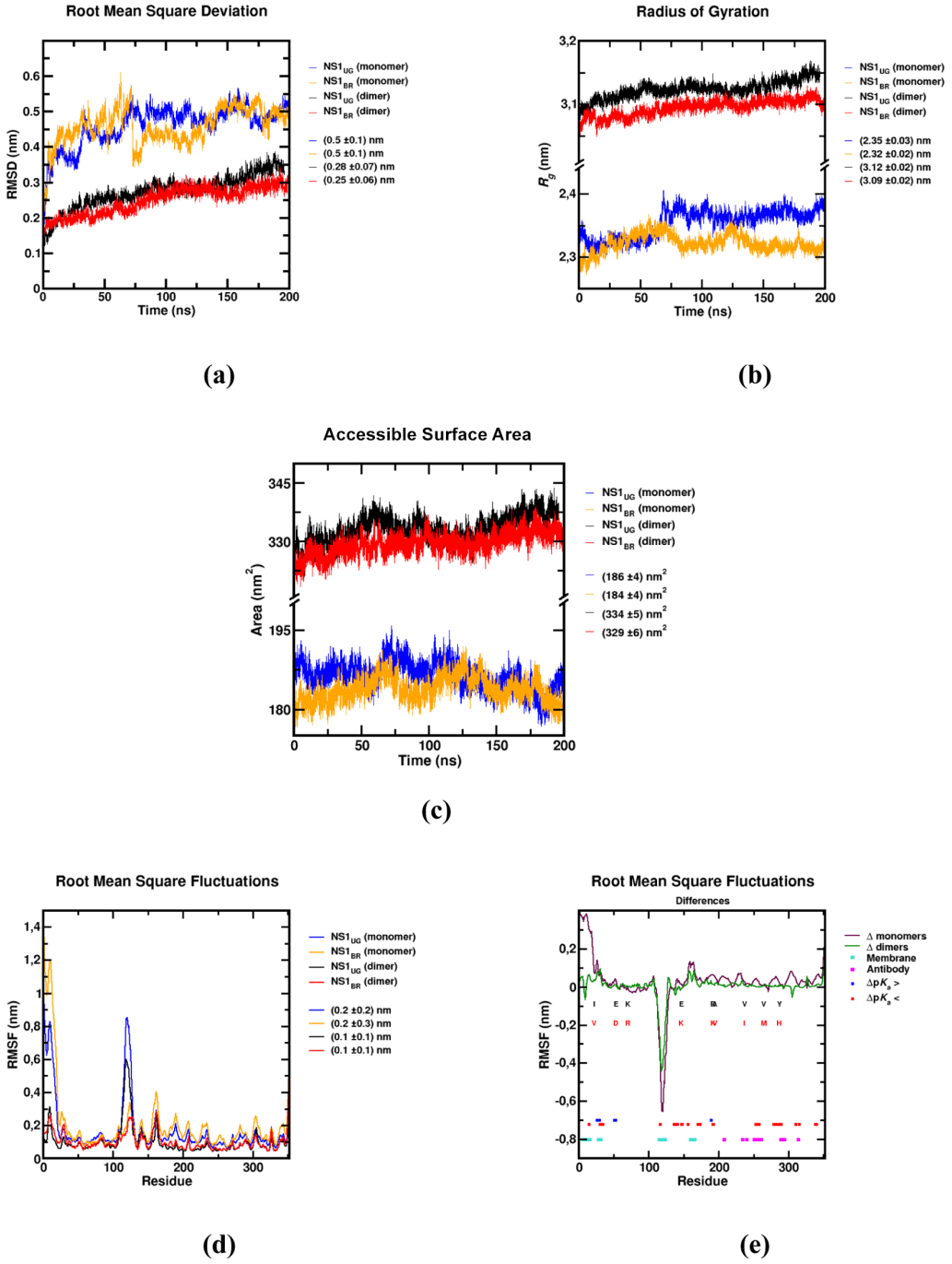



**Figure 2.** Averaged structural properties for the proteins $NS1_{UG}$ (PDB id 5K6K) and $NS1_{BR}$ (PDB id 5GS6) at the AA representation for their monomers and dimers. (a) RMSD. (b) Radius of gyration. (c) Solvent accessible surface area; (a−c) sampled as a function of time (200 ns). (d) RMSFs as a function of sequence. (e) Difference $RMSF_{NS1\text{-}BR} - RMSF_{NS1\text{-}UG}$ for monomers (maroon curve) and dimers (dark green curve) as a function of sequence. Sequence changes between the two strains, $\Delta pK_a$ values as given by ref 36, and putative regions of interaction with the membrane and antibody (defined in ref 36) are shown below the curves in plot (e). Mean values for all quantities are given at each panel.

not highlight significant differences between $NS1_{UG}$ and $NS1_{BR}$ monomers (see Table 1) alone or involved in the dimer at least during the observation time of 200 ns. The values obtained for the different properties and calculated over the different replicates remain within the standard deviation. It is also true for features relative to monomer−monomer interactions [e.g., average distance or occupancy fraction (OF) given in Table 1]. Previous studies revealed similar conclusions at the interspecific flavivirus level, where $NS1_{ZIKV\text{-}UG}$ and $NS1_{DENV}$ presented quite similar RMSD and $R_g$ profiles.[38,51]

Notwithstanding this remark, at the level of the monomer, the $\beta$-roll domain of $NS1_{BR}$ presents significantly larger fluctuations compared to $NS1_{UG}$. It is the reverse for the intertwined loop of the wing domain (Figure 2e), with $NS1_{UG}$ being much more dynamical in this region. Within the dimer,


https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

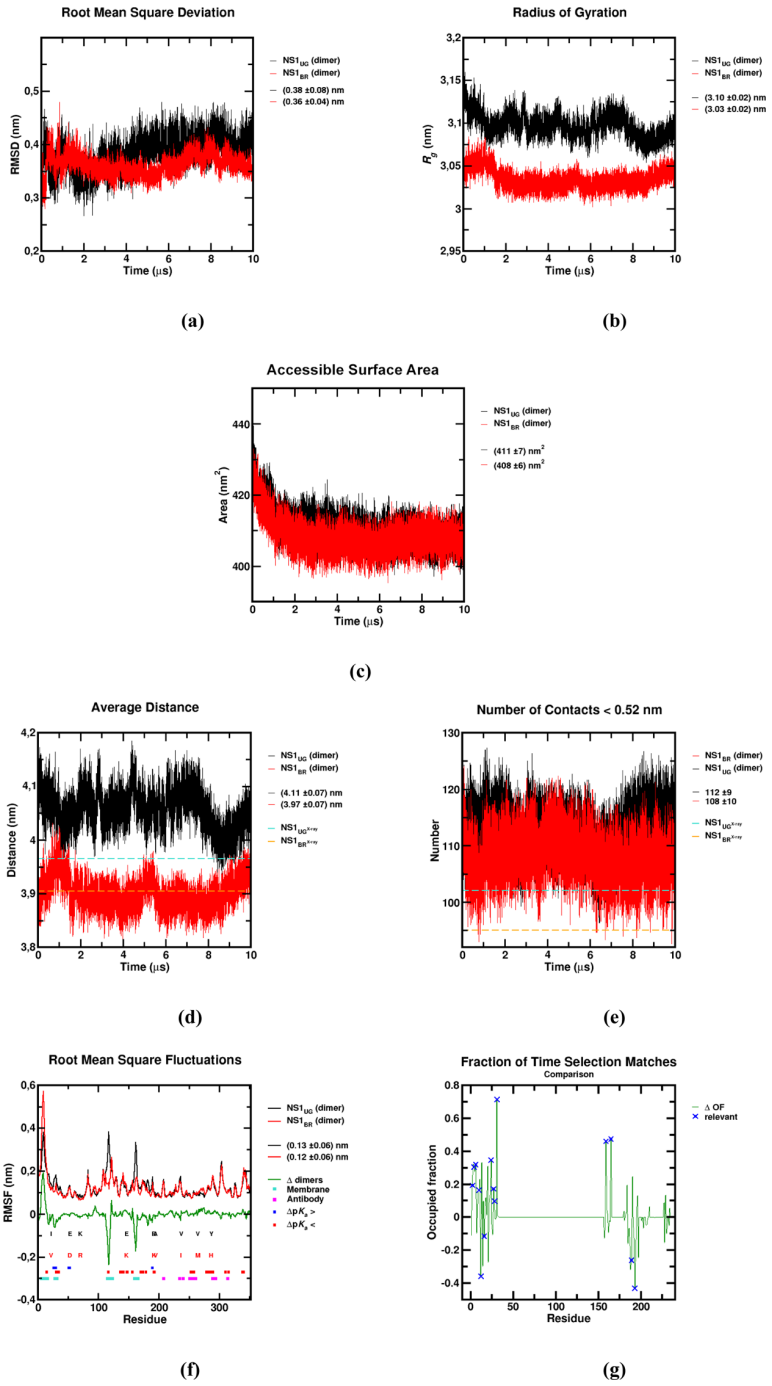



**Figure 3.** Averaged structural properties for the proteins $NS1_{UG}$ (PDB id 5K6K) and $NS1_{BR}$ (PDB id 5GS6) at CG representation for their dimers. (a) RMSD. (b) Radius of gyration. (c) Solvent accessible surface area. (d) Average distance. (e) Number of contacts. (f) RMSFs as a function of the number of residues. (g) OF of the residues. Plots (a–e) show variations as a function of time (10 $\mu$s). Plots (d,e,g) were calculated taking into account the chain A as a reference. Plot (f) shows the difference $RMSF_{NS1\text{-}BR}-RMSF_{NS1\text{-}UG}$ for dimers (dark green curve). Sequence changes between the two strains, $\Delta pK_a$'s as given by ref 36, and putative regions of interaction with the membrane and Ab (defined in ref 36) are shown below the curves in plot (f). The difference $OF_{NS1\text{-}UG}-OF_{NS1\text{-}BR}$ ($\Delta$OF) is shown in plot (g) as a green dark line. The relevant values $\Delta$OF are represented with blue cross symbols. Mean values for all quantities are given at each panel.

even being alleviated, this difference still remains. It must be noticed that the variations observed are not necessarily directly related to the exact position where the sequence changes (see Figure 2e, where the sequence variations are also indicated for the sake of comparison). Yet, interestingly, the largest negative peak is very close to a significant difference in electrostatic property variation (see $\Delta pK_a<$ in red in Figure 2e), echoing our recent findings relative to the great relevance of electrostatic properties in the detection of discrepancies between strains.[36]



https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

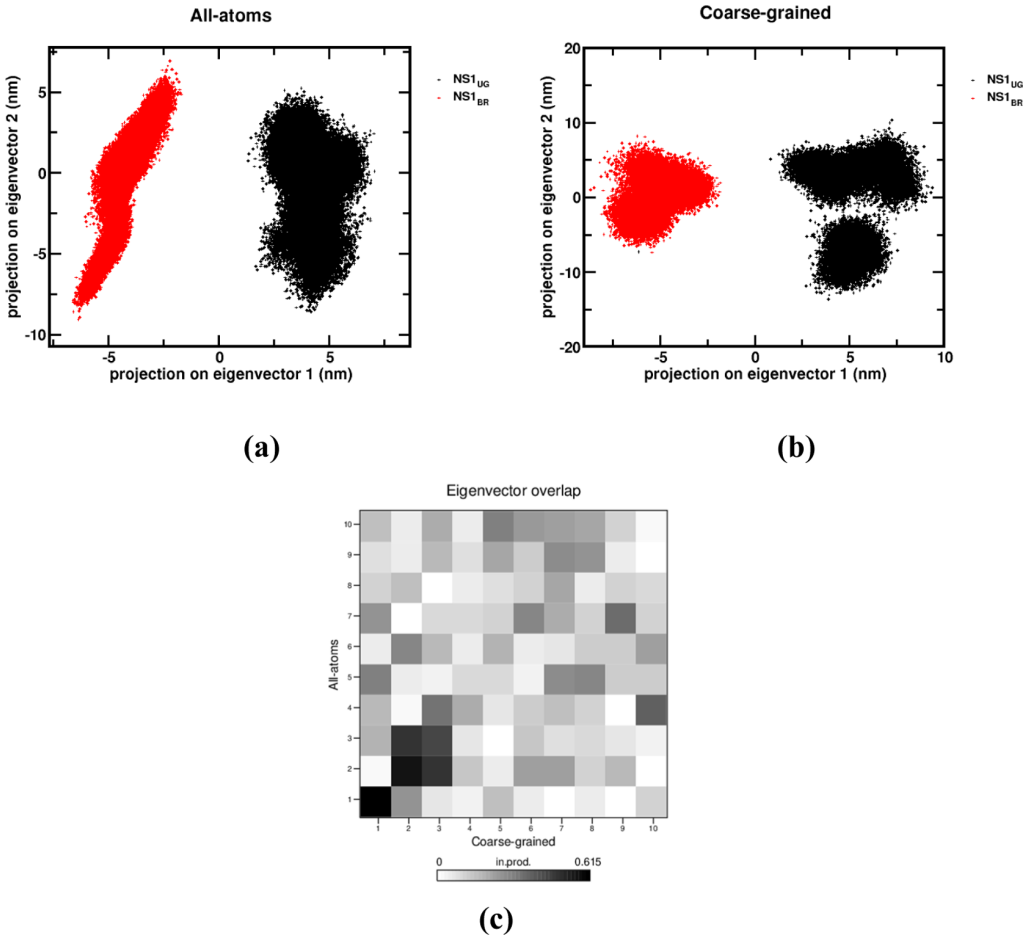



**Figure 4.** PCA for the concatenated trajectory of $NS1_{UG}$ (PDB id 5K6K) and $NS1_{BR}$ (PDB id 5GS6). (a) PCs for AA MD. (b) PCs for CG MD. (c) Eigenvector overlap between the CG and AA data.

Finally, it is worth mentioning that the $NS1_{UG}$ monomer tends to form more contacts with its partner, when compared to $NS1_{BR}$, while the distance between the two monomers is quite similar. This would indicate that the residue array of the NS1−NS1 interface is different between the two strains. Consequentially, in the following sections, we will focus on the dimer form because of its relevance in the flavivirus life cycle.

As noted above, the small changes in flexibility patterns could also be due to short timescale of simulation and, accordingly, insufficient sampling. Because of this reason, we turned out to a CG force field model (MARTINI ElNeDyn) that would allow us to better evaluate the impact of strain differences on dynamics through a larger sampling at the same cpu-time expense as AA simulations.

**Discrepancies between Two ZIKV NS1 (Dimers) Using CG Models.** As for AA simulations, we calculated time variations of different geometrical features as a function of time and averaged over the replicates for the ZIKV NS1 dimers (see Figure 3). We observed that most structural properties equilibrated around 4 $\mu$s for the two strains. RMSD values are in the same order of magnitude as for AA simulations, which confirms the adequacy of our calibration for the ElNeDyn elastic forces. The values of standard deviations associated with RMSD, SASA, average distances, the number of contacts, and the occupancy of the residues that belong to the NS1−NS1 interface (see Figure 3a,c,d,e and Table 1) prevent us from definitely concluding that the variations between the two strains are meaningful. However, the $R_g$ values of $NS1_{UG}$ and $NS1_{BR}$ appear to be significantly distinct: the $NS1_{BR}$ dimer having smaller geometrical dimensions compared to the $NS1_{UG}$ dimer. The RMSD of each monomer within the dimer, labeled $RMSD_{chain}$, is smaller compared to the dimer RMSD. Hence, the differences might originate from a change in the relative positioning of the two chains within the dimer. This is confirmed by the significant change in the average distance between the center of mass of the two chains that is larger than in the initial CG X-ray structure for $NS1_{UG}$ values. In contrast, this distance is mainly conserved for $NS1_{BR}$ (see Figure 3d). Furthermore, it should be mentioned that for most structural properties, $NS1_{UG}$ tends to exhibit higher values compared to $NS1_{BR}$. Interestingly, this tendency may be related to our previous results showing that NS1 of the Ugandan strain has a higher charge at physiological pH, favoring a strong electrostatic repulsion between the titratable residues and then producing a relative bigger structure.[36]

The analysis of CG-RMSF profiles confirms most conclusions drawn from AA simulations (see Figure 3f), with three flexible regions with higher peaks: $\beta$-roll (1−18), intertwined loop (109−127), and connector subdomain (155−165). In agreement with AA simulations, the flexible


https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

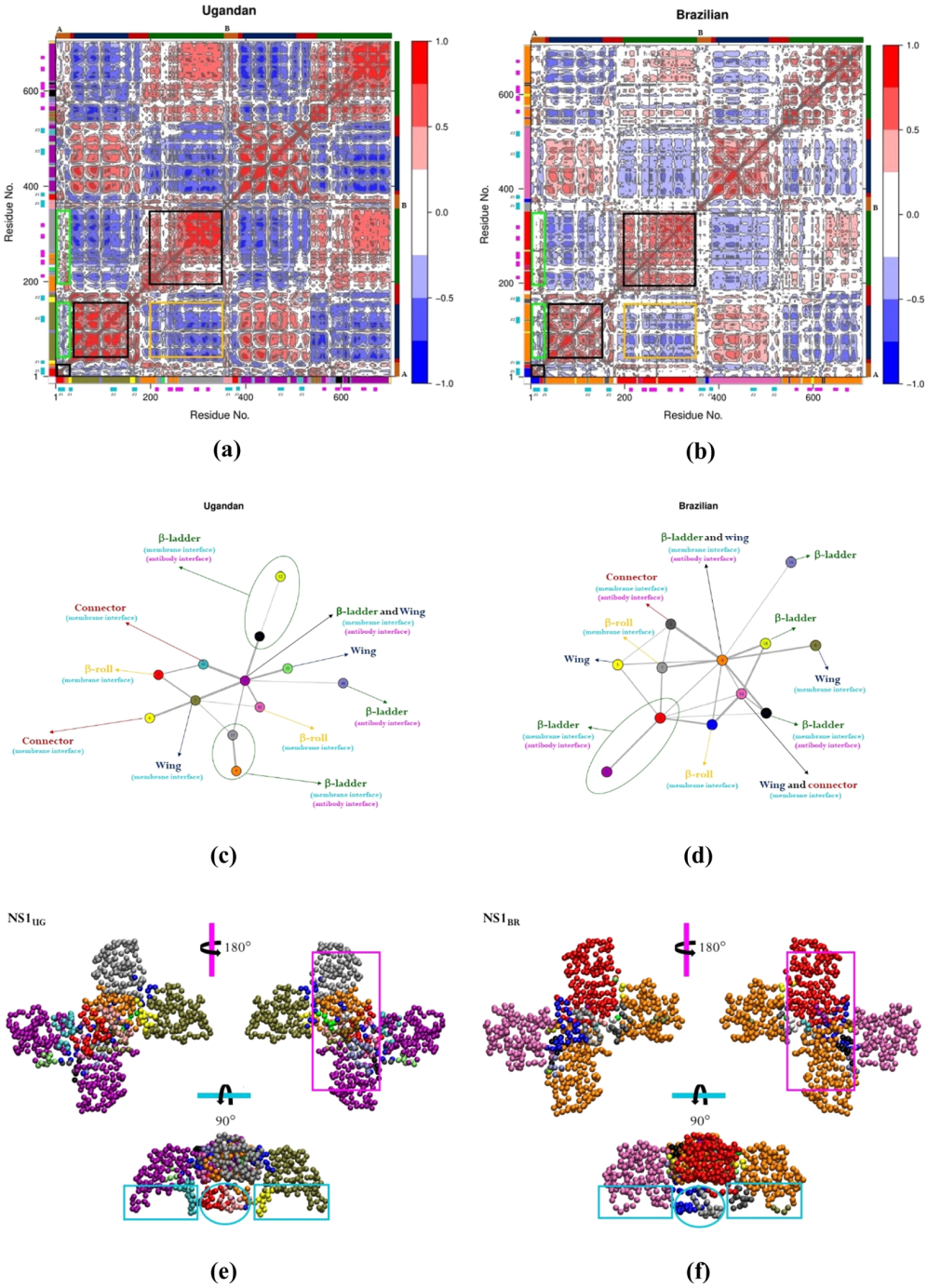



**Figure 5.** Dynamic cross-correlation matrix and 3D network representation for two $NS1_{ZIKV}$ proteins (dimers) at the CG level. (a,b) DCCM for $NS1_{UG}$ and $NS1_{BR}$. Positive, null, and negative correlations are values close to 1, 0, and −1, respectively. As shown in Figure 1a, on the top and right margins, the residues composed of $\beta$-roll, wing, and $\beta$-ladder domains are colored in orange, blue, and green respectively. The connector subdomain is colored red. At the bottom and vertical margins is shown the composition of communities estimated by the network analysis. (c,d) Network including communities with more than two members and not belonging to community "islands" for the two $NS1_{ZIKV}$. For the sake of visualization, the size of nodes was adjusted to observe the connections better (colors are the same as those displayed at the bottom and vertical margins of DCCM plots). (e,f) Distribution of the different communities estimated through network analysis at the three-dimensional level. $NS1_{UG}$ communities with one member were colored blue. Biological interfaces are also shown by rectangles in cyan (membrane) and violet (antibody), respectively. Continuous circles and rectangles refer to zones 1 and 2, respectively.

regions in the wing domain and in the connector subdomain have higher RMSFs for $NS1_{UG}$ than for $NS1_{BR}$. Putative regions of interaction with Ab do not show meaningful discrepancies of fluctuations between the two strains. Despite improvement of dynamics sampling, it is still difficult to find a correlation between RMSF profiles and sequence modifications between the two strains. Nevertheless, regarding the flexible region of the $\beta$-roll domain, fluctuations are increased


https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

compared to AA simulations for both strains, and interestingly, $NS1_{BR}$ presents significant larger values than $NS1_{UG}$ in this region (see the green curve in Figure 3f).

Concerning the number of contacts, more contacts are gained for both strains compared to their respective X-ray structures and this number tends to be higher for $NS1_{UG}$ than $NS1_{BR}$ (see Figure 3e). This result might be counterintuitive because the average separation distance is larger for $NS1_{UG}$. Yet, this apparent discrepancy might be simply due to a different arrangement of the residues at the interface for $NS1_{UG}$ compared to $NS1_{BR}$. To check this hypothesis, we analyzed in-depth the interface, which comprises residues of chain A within 0.52 nm from any residue of chain B and reciprocally. Because this list may change as a function of time, it was actualized for each frame. Therefore, we evaluated the OF that corresponds to the number of times this position is present in the list divided by the total number of frames considered. Then, we computed the difference for $OF_{i\ NS1\text{-}UG} - OF_{i\ NS1\text{-}BR}$ ($\Delta OF_i$) for each interface residue $i$ (see Figure 3g). $\Delta OF_i$ positive means that residue $i$ is more frequently observed at the interface in the Ugandan strain (e.g., 0.7 at position 31) compared to the Brazilian one. Reversely, if $\Delta OF_i$ is negative, residue $i$ is more frequently present at the interface in the Brazilian strain compared to the Uganda one. We considered only differences larger than the standard deviations calculated from the replicates. Among the 14 relevant $\Delta OF$'s positions, 10 of them were more frequently observed in the $NS1_{UG}$ interface compared to only 4 in the $NS1_{BR}$ interface, which confirms a different arrangement of residues at the dimeric interface. The smaller number of contacts in $NS1_{BR}$ might also explain the higher RMSF values observed for $NS1_{BR}$ in this region. Curiously, flexibility might represent an important structural feature when this region is associated with the membrane (see cyan rectangles in Figure 3f). Eventually, the differences in these regions with higher RMSF could be the first clue to understand the differential virulence between the two strains.

Overall, despite a very similar fold and per se a rather similar interface in the initial X-ray structures, the two strains follow a slightly different conformational dynamic landscape that leads to a significant variation in the monomer−monomer interface and in the membrane-binding region.

In order to go further and strengthen this first clue showing some difference between the two strains, we examined the ED of the systems. ED is a powerful approach to trap meaningful motions that are in general related to function.[52−55] ED is based on the principal component analysis (PCA) of covariance matrix properties. Despite some limitations, PCA has the merit to highlight differences that may exist between sets of data, in this case, the ensemble of conformations adopted during extensive dynamics sampling for each strain. Besides visualizing properties, ED helps in delineating significant motions from local and noisier movements.

**Main Motions in Two ZIKV NS1 Proteins.** ED of the two $NS1_{ZIKV}$ were calculated for the AA and CG representation (see Figure 4a,b), by considering the $C_\alpha$ and backbone beads (BB), respectively. The two first PCs for AA and CG represent 53% ($PC_1$ = 37% and $PC_2$ = 16%) and 59% ($PC_1$ = 35% and $PC_2$ = 24%) of the total variance, respectively. These values are quite large in view of the number of degrees of freedom of the systems. Thus, the discussion will be restricted to the properties of these two first components. When conformations extracted from MD replicates are projected on their respective $PC_1$−$PC_2$ plane, regardless of the force-field considered, two main clusters appear with respect to the $PC_1$ axis. Importantly, each cluster is associated with a given strain. With a significant sampling increase obtained by the CG FF, $PC_2$ reveals two distinct clusters for $NS1_{UG}$ but not for $NS1_{BR}$. Overlap calculation between the 10 first eigenvectors of the two model types shows substantial similarity, the highest values being between the two first PCs (see Figure 4c). Therefore, both FF models were able to identify similar tendencies in the main directions of motions, but more importantly, both FF models identify differences between the two strains. Altogether, these results show that the two ZIKV NS1s have discrepant ED; in other words, they highlight a distinct dynamical behavior between the two strains.

The essential motions along $PC_1$ and $PC_2$ (see Figure S2) involve regions mentioned in the previous sections, that is, the $\beta$-roll domain, the wing domain (intertwined loop), and the connector subdomain. Nevertheless, the present analyses emphasize higher amplitude of motions for $NS1_{UG}$ compared to $NS1_{BR}$ in the putative protein-membrane regions, as observed in the RMSF profiles. Interestingly, despite similar directions of movement of residues, the amplitudes are rather discrepant, the Ugandan strain being much more flexible compared to the Brazilian strain. A structural evaluation for the two $NS1_{ZIKV}$ revealed that the proportion of native contacts is higher for $NS1_{BR}$ than for $NS1_{UG}$ (data not shown), showing a correlation with the amplitudes of motions deduced by PCA.

In order to go further in the motion analysis, to provide putative explanations why distant sequence changes would impact important functional regions, we examined the dynamics cross-correlation properties. The results are detailed in the next section.

**Cross-Correlation and Protein Structure Network.** The dynamics cross-correlation maps (DCCM) were calculated for each strain from the CG trajectories (see Figure 5a,b, $NS1_{UG}$ and $NS1_{BR}$, respectively). High negative and positive correlation patterns are clearly distinguishable separated by zones with null values. Large patterns with similar values illustrate the existence of rigid bodies that move in concert (e.g., residues ∼25 to ∼87). These regions are delimitated by zones with null values (regions in white) that might correspond to hinges. Actually, at the level of one monomer (residues 1−352, for instance), three different blocks are visible. They correspond in each chain to the $\beta$-roll, the wing, and $\beta$-ladder domains, within which residues follow more or less similar directions (see black squares as an example in Figure 5a,b). On both sides of these hinges, the subdomains are moving in the opposite direction with respect to one another, for example, the wing and $\beta$-ladder domains move in opposite directions (see yellow squares as an example in Figure 5a,b). Overall, similar motions are observed for the second monomer. Nevertheless, subregions that are negatively and positively correlated with respect to wing and $\beta$-ladder domains, respectively, within the interfacial $\beta$-roll region may be pinpointed (see green squares as an example in Figure 5a,b). Apart from the interface, the motions of a given subdomain in one monomer with respect to the subdomain in the other monomer follow similar patterns. In previous studies, it was described that $NS1_{ZIKV}$ and $NS1_{DENV}$ dimers present "harmonicity" patterns in loop regions of the $\beta$-roll domain and the connector subdomain.[38] Although it was


https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

hypothesized that some few residues influence the stability of the interactions between these two loops at the intra and interatomic level, our ED results show that there is a large collectivity of residues influencing the general dynamics of $NS1_{ZIKV}$.

When comparing the two strains, a major difference appears, namely, the absolute values of correlations are much higher for $NSl_{UG}$ compared to $NS1_{BR}$. This result is in relation with the observations drawn from Figure S2 and shows that higher amplitudes of motions lead to higher correlations and vice versa (as for $NS1_{UG}$ and $NS1_{BR}$, respectively).

In order to better understand how the different regions of NS1s are dynamically interconnected, we have carried out a protein structural network (PSN) analysis. The PSN is very helpful to figure out long-range dynamics interactions between residues that are not easily visible in classical analyses. As a first step, it was necessary to determine the most suitable cutoff for each $NS1_{ZIKV}$. We took care to choose the cutoff in order to obtain networks with modularity values ranging from 0.4 to 0.7 and with minor number of communities, as recommended.[56,57] The best cutoff values were found to be higher for $NS1_{UG}$ than for $NS1_{BR,}$ which would indicate different connections between the residues of each protein. As an example, the smallest numbers of communities for $NS1_{UG}$ and $NS1_{BR}$ were 52 (modularity = 0.6) and 19 (modularity = 0.5), respectively (see Figure S3). This is not surprising because larger communities are in general associated with increased rigidity. Reciprocally, greater flexibility (as observed for $NS1_{UG}$) is commonly associated to a higher number of small communities.[58] By discarding the communities with the smallest populations,[59] we ended with 12 communities that make up the main nodes of the network displayed in Figure 5c,d for $NS1_{UG}$ and $NS1_{BR,}$ respectively. First, it is worth noting that there are more connections between the nodes for $NS1_{BR}$ compared to $NS1_{UG}$ (see Figure 5c,d for $NS1_{UG}$ and $NS1_{BR,}$ respectively). The residue composition of these communities is represented by different colors reported along the DCCM axes for each strain. Among the 12 communities represented in this figure, three large communities emerge, represented by nodes #31 (purple), #7 (tan), and #15 (gray) and nodes #2 (red), #4 (orange), and #14 (pink) for UG and BR, respectively (Figure 5c,d). However, large discrepancies exist between the two strains. The most populated community contain ~99 members for $NS1_{UG}$, while the population is much larger for $NS1_{BR}$ with ~154 members. For $NS1_{UG}$, they regroup residues located in $wing_{chainA}$ (see Figure 5c,e colored in tan), $\beta$-$ladder_{chainA}$ (see Figure 5c,e in gray), and in $wing_{chainB}$ and $\beta$-$ladder_{chainB}$ (see Figure 5c,e in purple). For $NS1_{BR,}$ the community involves residues located in similar regions as $NS1_{UG}$ (see Figure 5d,f in red and orange), but also residues in $wing_{chainB}$ (see Figure 5d,f in pink). The connections within communities and between communities can be better grasped by the representations given in Figure S5, in which red and blue lines correspond to positive and negative correlation values, respectively (see Figure 5a,b). As an example, the edges between #7, #31, and #15 for UG (Figure 5c) or #4, #2, and #14 for BR are clearly illustrated through the high density of blue lines that connect members of different communities (Figure S5a,b respectively). In addition, these views show how the dimer interface plays a pivotal role in the dynamics.

Interestingly, specific community patterns are created in regions that are supposed to interact with the membrane and were previously described as flexible regions. These communities are small and can be distributed in the same chain (yellow and gray communities in chain A of $NS1_{UG}$ and $NS1_{BR}$, respectively) or between the two monomers (red and blue communities in $\beta$-roll domains of $NS1_{UG}$ and $NS1_{BR}$, respectively). In the $\beta$-roll domain of the two strains, small community regroup residues of the $\beta$-hairpin motifs. For both strains, the so-called greasy fingers (yellow and cyan) with protruding aromatic "anchors" (specifically, F160 and F163) that might help in associating with membranes,[22] define similar but not identical communities. On summary, both proteins retain to some extent a common pattern of communities in the putative interaction with membranes. However, $NS1_{UG}$ tends to display more and smaller communities than $NS1_{BR}$.

Similarly, the putative regions of interaction with Ab (underlined in violet in Figure 5a,b) are involved in two large communities for $NS1_{BR}$, while they are distributed over a larger number of small communities for $NS1_{UG}$. This heterogeneity correlation of motion in the $NS1_{UG}$ antibody region could influence the interaction with the host's immune system, making it more difficult to recognize the Ugandan NS1 antigen. Hence, it may be assumed that a higher flexible chain will be more difficult to be recognized by produced Abs, which will tend to increase the fatality rates as seen for the UG strain.

The network betweenness measure (see Figure S4a,b, respectively) shows greater peaks of centrality for $NS1_{UG}$ compared to $NS1_{BR}$, which confirms observations detailed above. Indeed, a larger distribution of connections for $NS1_{BR}$ reduces the centrality of the node #4 (see Figure S4b). The highest centrality peaks correspond to the amino acids $I78_{chainB}$ and $N76_{chainA}$ for $NS1_{UG}$ and $NS1_{BR}$, respectively. These residues are close to the intertwined loop of the wing domain, a region assumed to interact with the membrane.[10] It is interesting to note that these two residues are close to two mutated residues: E52D and K69D (Figure 1a,b).

It is important to notice that the PSN provides a global view of the connections and long-range interactions that may establish during the dynamics. Unfortunately, it does not bring information neither about the sequence of events that lead to these connections nor on the consequence of the raising of new connections. Sequence of events would be better apprehended by establishing Markov state models, but calculations of such models require multiple replicates and much longer simulations, far beyond microseconds for such large systems. Thus, as an ersatz, movies of some trajectories are provided in the Supporting Information (Movie S1a,b) with particles colored according to their community labels.

Finally, multiple communication pathways have previously been found in protein systems. The differential connectivity between nodes represents a key topic for understanding allosteric responses in proteins.[60,61] Regarding our study, the communication routes found with the PSN for the two ZIKV NS1's bring strong support to the importance of flexibility discussed above, which should impact the way these viral proteins interact with the target host proteins.

## ■ CONCLUSIONS

In the present study, we pursue the characterization of the impact of small sequence changes between two strains of



https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

$NS1_{ZIKV}$ proteins, with the aim of understanding how these differences might be related to their virulence. Indeed, differential virulence was found between flavivirus and strains of a single virus species.[23−25]

In the continuation of the work that we have recently performed on their electrostatics properties, in the present study, we focused on another essential feature, namely, the dynamics of the strain. We sought putative differences by conducting MD analyses using AA and CG models. At first, we compared properties of monomers with dimers. At the level of monomers, we did not observe major differences between the two strains, except in the $\beta$-roll domain and the intertwined loop of the wing domain within the estimated standard deviations. At the level of dimers, the presence of the second monomer modifies the amplitude of motions, but more importantly, it has an influence on loop regions, in particular, those involved in membrane binding. Concerning the differences between strains, despite similar tendencies for domains involved in main motions, significant differences emerge with $NS1_{UG}$ exhibiting greater flexibility compared to $NS1_{BR}$. Network analyses underline the striking role that amino acids close to the intertwined loop region and two mutated residues play in the dynamic network of each strain.

It is important to notice that in the present study, we considered a constant-charge simulation protocol (fixed charges for all residues during all the MD sampling), while particles are free to move accordingly with the system forces. This is a complementary strategy to our previous studies on their electrostatic properties, where a CpH method (charges can change as a function of the solution pH and environment) for rigid structures was adopted. Interestingly, despite the intrinsic approximations of each approach, we noted that the residues with larger $pK_a$ shifts coincide with the profiles of those with more flexibility, with regions that interact with the membrane. As CpH MD methods improve their sampling efficiency and allow the simulation of larger macromolecules,[43,62] we might be able to observe an enhancement of such effects because of the proper coupling of titration with dynamical properties on future studies.

The links between virulence and dynamics are not straightforward, but we can make the following assumptions with the increase in the system dynamics: (i) it becomes more difficult for recognition by the host immune system, (ii) an enhanced penetration in the host cell membrane, and (iii) an improved facility to form larger oligomers and interact with other partners. Indeed, because $NS1_{UG}$ exhibits larger flexibility when compared to $NS1_{BR}$, the host immune system should adjust to a larger number of conformations in the recognition of this antigen. A larger conformational landscape could also improve the interaction with a specific region of the membrane (or lipid raft) and help in optimizing its activity by better interacting with other membrane proteins bearing important functions. These assumptions are not exclusive, but in combination, they should result in an even more virulent impact.

In conclusion, our present work does provide new clues for understanding putative origins of differential virulence between two major strains of NS1 and emphasize the role of their dynamical features.

## METHODS

**Modeling of NS1 Structures.** NS1 is a glycoprotein of 352 amino acids divided into three main domains: $\beta$-roll (1−30), wing (38−150), and $\beta$-ladder (181−352). The remaining residues (i.e., 31−37 and 151−180) are part of the so-called connector subdomain (see Figure 1a).[19] As was done in our previous studies,[36,37] structural analyses were carried out on two ZIKV NS1 crystallographic structures available in the RCSB Protein Data Bank: PDB id 5K6K (resolution 1.89 Å)[22] and PDB id 5GS6 (resolution 2.85 Å),[10] respectively, for Uganda strain (labeled $NS1_{UG}$) and for the Brazilian one (labeled $NS1_{BR}$). These two proteins share the same scaffold and topology and essentially differ in some loop regions (1.9 Å RMSD across all 352 pairs and 0.706 Å RMSD between 310 pruned atom pairs).

Missing regions were completed using the homology protocol as mentioned before.[36] Based on the ZIKV dimer for each strain, the chain with missing regions was modeled with the complete chain (e.g., the chain B is incomplete in $NS1_{BR}$, so the chain A was used as a template). For this process, the Modeller program version 9.18 was used.[63] Default values were chosen, that is, five models were built, and the model with the best DOPE score was selected. For simplicity, we did not include the sugar molecules that are covalently bonded to Asn residues (N130 and N207 for both $NS1_{UG}$ and $NS1_{BR}$). For the monomer exploration, we selected chain A of $NS1_{BR}$ and chain B of $NS1_{UG}$ because of the completeness of the structure.

**MD Simulation Protocol for AA Systems.** NS1 systems were built using the online CHARMM-GUI tool (http://www.charmm-gui.org).[64] Each protein was placed in a rectangular solvent box containing 0.15 M NaCl salt. The size of the box corresponded to an edge distance of 20 Å from any protein atom (30 Å for the dimeric $NS1_{UG}$). $Na^+$ counterions were added to neutralize the system. Periodic boundary conditions were adopted. Protonation states were considered assuming pH 7 and defined based on the $pK_a$ values calculated by us through CpH MC simulations.[36] The final systems contain 48,609 waters, 141 $Na^+$ ions, and 138 $Cl^-$ ions for the $NS1_{BR}$ monomer; 53,814 waters, 158 $Na^+$ ions, and 153 $Cl^-$ ions for the $NS1_{UG}$ monomer; 94,102 waters, 274 $Na^+$ ions, and 268 $Cl^-$ ions for the $NS1_{BR}$ dimer; and 137,558 waters, 402 $Na^+$ ions, and 392 $Cl^-$ ions for the $NS1_{UG}$ dimer.

Classical constant-charge MD simulations were performed using GROMACS package 5.1.5.[65] The CHARMM36m forcefield was used for protein and TIP3P model for water.[45] Long-range electrostatic interactions were treated with the particle mesh Ewald method. A minimum cutoff for nonbonded interactions of 1.2 nm was set. Covalent bonds between hydrogen and heavy atoms were restrained using the LINCS algorithm. The integration time step was 1 fs. Energy minimization was done (5000 steps) using the steepest descent method or until a maximum force value <1000 kJ $mol^{-1}$ $nm^{-1}$ was reached. Subsequently, the system was heated at 300 K and simulated in the *NVT* ensemble during 25 ns, with constraints on protein heavy atoms, with a force constant equal to 1000 kJ $nm^{-2}$. The Nose−Hoover algorithm was used for temperature coupling with a coupling constant $\tau_T$ = 1.0 ps. The resulting equilibrated systems were then used for the production phase in the *NPT* ensemble, without any constraint, maintaining the isotropic pressure at 1 bar by means of the Parrinello−Rahman barostat, with a pressure coupling constant $\tau_P$ = 5.0 ps and a compressibility of 4.5 × $10^{-5}$ bar. Three replicates differing by initial velocities were simulated during 200 ns each performing 0.6 $\mu$s of sampling.


https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

**MD Simulations Protocol for CG Systems.** In a second step, we considered CG models. The reduced number of particles allowed us to gain an order of magnitude of simulation time scale for the same amount of cpu-time ($\mu$s compared to hundreds of ns), at the benefit of a conformational sampling enhancement.

The CG models were obtained using the "martinize.py" script (version 2.5, provided by the official site: http://cgmartini.nl) applied to the AA 3D structures. The MARTINI 2.2P force field was used, combined with an elastic network for the SS's and ElNeDyn additional constraints.[46,47] Default parameters were used (i.e., the lower and upper elastic bond cutoffs set to 0.5 and 0.9 nm; the bond strengths being independent of the bond length, respectively). The elastic force constant was adjusted to 200 kJ $mol^{-1}$ $nm^{-2}$ to get slightly higher fluctuation amplitudes than with the AA force field, while maintaining the fluctuation profiles. The protein was placed in a cubic simulation box of 16.2 nm edge. Each CG system was solvated with water and salt particles (0.15 M) based on the specific beads of the MARTINI 2.2P force field. Counterions were added to neutralize the systems.

Starting structures were also protonated assuming pH equal to 7 and using the same criterion above described for the AA. Charges were kept constant during all the simulation runs. The final systems contain 33,802 water beads, 377 $Na^+$ beads, and 371 $Cl^-$ beads for the $NS1_{BR}$ dimer and 33,769 water beads, 381 $Na^+$ beads, and 371 $Cl^-$ beads for the $NS1_{UG}$ dimer. Trajectories for CG systems were performed with GROMACS 5.1.5 or later versions, with periodic boundary conditions. The time step was 20 fs as recommended with the CG force field. Reaction-field treatment was considered for electrostatic interactions with $\varepsilon_r$ equals to 2.5. A minimum cutoff of 1.1 nm was employed for the Coulomb and van der Waals interactions using the Verlet cutoff scheme. Two stages of energy minimization were carried out for 3000 steps using the steepest descent algorithm, the first one with softcore constraints and the second one without any structural constraint. Equilibration was performed during 10 ns in the *NPT* ensemble using Berendsen's algorithm for the temperature and pressure coupling, with the same pressure and temperature coupling constants as for AA systems. In the production stage, the temperature coupling $\tau_T$ was changed to 12.0 ps and the compressibility to $4.5 \times 10^{-5}$ bar, following the recommendations of the Martini-CG-MD protocol.[66] As for AA simulations, three replicates were simulated for each system during 10 $\mu$s each.

**Structural Analysis by Canonical Approaches.** Most analyses were conducted using the tools implemented in GROMACS, that is, (a) RMSD, (b) $R_g$, (c) SASA, (d) RMSF, (e) SS with DSSP, (f) average distance, (g) number of contacts, and (h) OF (for each residue in the dimeric interface) based on the trajectories of the NS1 systems. In all cases, the reference conformation was the completed model based on the corresponding X-ray structure. The variation of each property as a function of time was averaged over the three replicates. Averages over the time and their corresponding standard deviations were also assessed. In some cases, we recalculated the properties discarding the most flexible regions (labeled "no_flex" in the Results and Discussion section). In the case of dimers, some averages were calculated considering chains A and B separately for each replicate (labeled with a subscript corresponding to the chain).

**ED Analysis.** In order to highlight essential motions, PCA of the conformations adopted along the simulations was performed for the different systems examined.[52] The analysis was done considering $C_\alpha$ or BB positions for AA and CG simulations, respectively. Because the two NS1 share the same number of residues, the analysis was done on the whole set of conformations, pooling together all the replicates of the two strains. Despite the fact that $C_\alpha$ and BB positions do not coincide exactly, we calculated the similarity between the eigenvectors obtained in each case.

DCCM was calculated by employing the Bio3D tool (http://thegrantlab.org/bio3d/) with the BB conformers for each CG system. The number of conformations considered was 2700. PSN analysis was performed using the method implemented in Bio3D.[67] Communities of residues that are correlated for each protein were examined. Because network properties depend on cutoff parameters, we examined different thresholds, for example, modularities between 0.4 and 0.7 and the smallest number of communities per protein and the betweenness were considered.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01377.

Protein secondary structure, porcupine plots, correlation network analysis, centrality, and cross-correlation pattern (PDF)

Molecular dynamics for NS1UG (AVI)

Molecular dynamics for NS1BR (AVI)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Catherine Etchebest** − *Institut National de la Transfusion Sanguine, 75015 Paris, France; Institut National de la Santé et de la Recherche Médicale, UMR_S 1134, Biologie Intégrée du Globule Rouge, Equipe 2, INSERM, Dynamique des Structures et des Interactions Moléculaires, F-75015 Paris, France; Université de Paris, 75013 Paris, France; University of São Paulo and Université de Paris International Laboratory in Structural Bioinformatics, 14040-903 Ribeirão Preto, São Paulo, Brazil;* orcid.org/0000-0002-5871-2372; Email: catherine.etchebest@inserm.fr

### Authors

**Sergio A. Poveda-Cuevas** − *Programa Interunidades em Bioinformática, Universidade de São Paulo, 05508-090 São Paulo, Brazil; Departamento de Ciências Biomoleculares, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, 14040-903 Ribeirão Preto, São Paulo, Brazil; University of São Paulo and Université de Paris International Laboratory in Structural Bioinformatics, 14040-903 Ribeirão Preto, São Paulo, Brazil;* orcid.org/0000-0002-3087-6674

**Fernando Luís Barroso da Silva** − *Programa Interunidades em Bioinformática, Universidade de São Paulo, 05508-090 São Paulo, Brazil; Departamento de Ciências Biomoleculares, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, 14040-903 Ribeirão Preto, São Paulo, Brazil; Department of Chemical and Biomolecular Engineering, North Carolina State University, 27695 Raleigh, North Carolina, United States;*




https://dx.doi.org/10.1021/acs.jcim.0c01377
J. Chem. Inf. Model. 2021, 61, 1516−1530

*University of São Paulo and Université de Paris International Laboratory in Structural Bioinformatics, 14040-903 Ribeirão Preto, São Paulo, Brazil*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01377

### Author Contributions

S.A.P.-C., F.L.B.d.S., and C.E., designed the research, performed simulations, analyzed data, and wrote the paper.

### Notes

The authors declare no competing financial interest.
PDB id 5K6K. PDB id 5GS6.

## ■ ACKNOWLEDGMENTS

This work has been supported in part by the "Fundação de Amparo à Pesquisa do Estado de São Paulo" [Fapesp 2015/16116-3 and 2020/07158-2 (F.L.B.d.S.)], the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [CNPq 305393/2020-0 (F.L.B.d.S.)], the Swedish National Infrastructure for Computing [SNIC 2020/14-38 (F.L.B.d.S.)], and the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior"—CAPES (S.A.P.-C.). F.L.B.d.S. and S.A.P.-C. acknowledge the support of the University of São Paulo (USP) through the NAP-CatSinQ (Research Core in Catalysis and Chemical Synthesis) and the computing hours at HPC/USP. The authors also would like to thank CINES (Centre Informatique National de l'Enseignement Supérieur) for computational resources (project: A0060710878). F.L.B.d.S. and C.E. acknowledge the mobility support from the joint Université Sorbonne Paris Cité-USP Research program and the Fapesp-ANR-PRCI program (EMULATE project).

## ■ REFERENCES

(1) World Health Organization. Coronavirus Disease (COVID-19) Pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed 2020-10-21).
(2) Gould, E.; Solomon, T. Pathogenic Flaviviruses. *Lancet* **2008**, *371*, 500−509.
(3) Kindhauser, M. K.; Allen, T.; Frank, V.; Santhanaa, R. S.; Dyea, C. Zika: The Origin and Spread of a Mosquito-Borne Virus. *Bull. W.H.O.* **2016**, *94*, 675−686.
(4) Brady, O. J.; Gething, P. W.; Bhatt, S.; Messina, J. P.; Brownstein, J. S.; Hoen, A. G.; Moyes, C. L.; Farlow, A. W.; Scott, T. W.; Hay, S. I. Refining the Global Spatial Limits of Dengue Virus Transmission by Evidence-Based Consensus. *PLoS Neglected Trop. Dis.* **2012**, *6*, e1760.
(5) Pierson, T. C.; Diamond, M. S. The Continued Threat of Emerging Flaviviruses. *Nat. Microbiol.* **2020**, *5*, 796−812.
(6) Mackenzie, J. S.; Gubler, D. J.; Petersen, L. R. Emerging Flaviviruses: The Spread and Resurgence of Japanese Encephalitis, West Nile and Dengue Viruses. *Nat. Med.* **2004**, *10*, S98.
(7) Vasilakis, N.; Weaver, S. C. Flavivirus Transmission Focusing on Zika. *Curr. Opin. Virol.* **2017**, *22*, 30−35.
(8) Paixão, E. S.; Teixeira, M. G.; Rodrigues, L. C. Zika, Chikungunya and Dengue the Causes and Threats of New and Re-Emerging Arboviral Diseases. *BMJ Global Health* **2018**, *3*, e000530.
(9) Song, H.; Qi, J.; Haywood, J.; Shi, Y.; Gao, G. F. Zika Virus NS1 Structure Reveals Diversity of Electrostatic Surfaces among Flaviviruses. *Nat. Struct. Mol. Biol.* **2016**, *23*, 456−458.
(10) Xu, X.; Song, H.; Qi, J.; Liu, Y.; Wang, H.; Su, C.; Shi, Y.; Gao, G. F. Contribution of Intertwined Loop to Membrane Association Revealed by Zika Virus Full-Length NS1 Structure. *EMBO J.* **2016**, *35*, 2170−2178.
(11) Watterson, D.; Modhiran, N.; Young, P. R. The Many Faces of the Flavivirus NS1 Protein Offer a Multitude of Options for Inhibitor Design. *Antiviral Res.* **2016**, *130*, 7−18.
(12) Muller, D. A.; Young, P. R. The Flavivirus NS1 Protein Molecular and Structural Biology, Immunology, Role in Pathogenesis and Application as a Diagnostic Biomarker. *Antiviral Res.* **2013**, *98*, 192−208.
(13) Ye, J.; Zhu, B.; Fu, Z. F.; Chen, H.; Cao, S. Immune Evasion Strategies of Flaviviruses. *Vaccine* **2013**, *31*, 461−471.
(14) Priyamvada, L.; Quicke, K. M.; Hudson, W. H.; Onlamoon, N.; Sewatanon, J.; Edupuganti, S.; Pattanapanyasat, K.; Chokephaibulkit, K.; Mulligan, M. J.; Wilson, P. C.; Ahmed, R.; Suthar, M. S.; Wrammert, J. Human Antibody Responses after Dengue Virus Infection Are Highly Cross-Reactive to Zika Virus. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 7852−7857.
(15) Dejnirattisai, W.; Supasa, P.; Wongwiwat, W.; Rouvinski, A.; Barba-Spaeth, G.; Duangchinda, T.; Sakuntabhai, A.; Cao-Lormeau, V.-M.; Malasit, P.; Rey, F. A.; Mongkolsapaya, J.; Screaton, G. R. Dengue Virus Sero-Cross-Reactivity Drives Antibody-Dependent Enhancement of Infection with Zika Virus. *Nat. Immunol.* **2016**, *17*, 1102−1108.
(16) Gallichotte, E. N.; Young, E. F.; Baric, T. J.; Yount, B. L.; Metz, S. W.; Begley, M. C.; De Silva, A. M.; Barica, R. S. Role of Zika Virus Envelope Protein Domain III as a Target of Human Neutralizing Antibodies. *mBio* **2019**, *10*, e01485-19.
(17) Giraldo, M. I.; Xia, H.; Aguilera-Aguirre, L.; Hage, A.; van Tol, S.; Shan, C.; Xie, X.; Sturdevant, G. L.; Robertson, S. J.; McNally, K. L.; Meade-White, K.; Azar, S. R.; Rossi, S. L.; Maury, W.; Woodson, M.; Ramage, H.; Johnson, J. R.; Krogan, N. J.; Morais, M. C.; Best, S. M.; Shi, P.-Y.; Rajsbaum, R. Envelope Protein Ubiquitination Drives Entry and Pathogenesis of Zika Virus. *Nature* **2020**, *585*, 414−419.
(18) Fibriansah, G.; Tan, J. L.; Smith, S. A.; De Alwis, R.; Ng, T.-S.; Kostyuchenko, V. A.; Jadi, R. S.; Kukkaro, P.; De Silva, A. M.; Crowe, J. E.; Lok, S.-M. A Highly Potent Human Antibody Neutralizes Dengue Virus Serotype 3 by Binding across Three Surface Proteins. *Nat. Commun.* **2015**, *6*, 6341.
(19) Valente, A. P.; Moraes, A. H. Zika Virus Proteins at an Atomic Scale: How Does Structural Biology Help Us to Understand and Develop Vaccines and Drugs against Zika Virus Infection? *J. Venomous Anim. Toxins Incl. Trop. Dis.* **2019**, *25*, e20190013.
(20) Botting, C.; Kuhn, R. J. Novel Approaches to Flavivirus Drug Discovery. *Expert Opin. Drug Discovery* **2012**, *7*, 417−428.
(21) Sironi, M.; Forni, D.; Clerici, M.; Cagliani, R. Nonstructural Proteins Are Preferential Positive Selection Targets in Zika Virus and Related Flaviviruses. *PLoS Neglected Trop. Dis.* **2016**, *10*, e0004978.
(22) Brown, W. C.; Akey, D. L.; Konwerski, J. R.; Tarrasch, J. T.; Skiniotis, G.; Kuhn, R. J.; Smith, J. L. Extended Surface for Membrane Association in Zika Virus NS1 Structure. *Nat. Struct. Mol. Biol.* **2016**, *23*, 865−867.
(23) Shao, Q.; Herrlinger, S.; Zhu, Y.-N.; Yang, M.; Goodfellow, F.; Stice, S. L.; Qi, X.-P.; Brindley, M. A.; Chen, J.-F. The African Zika Virus MR-766 Is More Virulent and Causes More Severe Brain Damage than Current Asian Lineage and Dengue Virus. *Development* **2017**, *144*, 4114−4124.
(24) Simonin, Y.; Loustalot, F.; Desmetz, C.; Foulongne, V.; Constant, O.; Fournier-Wirth, C.; Leon, F.; Molès, J.-P.; Goubaud, A.; Lemaitre, J.-M.; Maquart, M.; Leparc-Goffart, I.; Briant, L.; Nagot, N.; Van de Perre, P.; Salinas, S. Zika Virus Strains Potentially Display Different Infectious Profiles in Human Neural Cells. *EBioMedicine* **2016**, *12*, 161−169.
(25) Sheridan, M. A.; Yunusov, D.; Balaraman, V.; Alexenko, A. P.; Yabe, S.; Verjovski-Almeida, S.; Schust, D. J.; Franz, A. W.; Sadovsky, Y.; Ezashi, T.; Roberts, R. M. Vulnerability of Primitive Human Placental Trophoblast to Zika Virus. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E1587−E1596.
(26) Puerta-Guardo, H.; Glasner, D. R.; Espinosa, D. A.; Biering, S. B.; Patana, M.; Ratnasiri, K.; Wang, C.; Beatty, P. R.; Harris, E. Flavivirus NS1 Triggers Tissue-Specific Vascular Endothelial

Dysfunction Reflecting Disease Tropism. *Cell Rep.* **2019**, *26*, 1598−1613.
(27) Jurado, K. A.; Simoni, M. K.; Tang, Z.; Uraki, R.; Hwang, J.; Householder, S.; Wu, M.; Lindenbach, B. D.; Abrahams, V. M.; Guller, S.; Fikrig, E. Zika Virus Productively Infects Primary Human Placenta-Specific Macrophages. *JCI Insight* **2016**, *1*, e88461.
(28) Akey, D. L.; Brown, W. C.; Dutta, S.; Konwerski, J.; Jose, J.; Jurkiw, T. J.; DelProposto, J.; Ogata, C. M.; Skiniotis, G.; Kuhn, R. J.; Smith, J. L. Flavivirus NS1 Structures Reveal Surfaces for Associations with Membranes and the Immune System. *Science* **2014**, *343*, 881−885.
(29) Wang, D.; Chen, C.; Liu, S.; Zhou, H.; Yang, K.; Zhao, Q.; Ji, X.; Chen, C.; Xie, W.; Wang, Z. A Mutation Identified in Neonatal Microcephaly Destabilizes Zika Virus NS1 Assembly in Vitro. *Sci. Rep.* **2017**, *7*, 42580.
(30) Hall, R. A.; Khromykh, A. A.; Mackenzie, J. M.; Scherret, J. H.; Khromykh, T. I.; Mackenzie, J. S. Loss of Dimerisation of the Nonstructural Protein NS1 of Kunjin Virus Delays Viral Replication and Reduces Virulence in Mice, but Still Allows Secretion of NS1. *Virology* **1999**, *264*, 66−75.
(31) Clark, D. C.; Lobigs, M.; Lee, E.; Howard, M. J.; Clark, K.; Blitvich, B. J.; Hall, R. A. In Situ Reactions of Monoclonal Antibodies with a Viable Mutant of Murray Valley Encephalitis Virus Reveal an Absence of Dimeric NS1 Protein. *J. Gen. Virol.* **2007**, *88*, 1175−1183.
(32) Oliveira, E. R. A.; de Alencastro, R. B.; Horta, B. A. C. The Mechanism by Which P250L Mutation Impairs Flavivirus-NS1 Dimerization: An Investigation Based on Molecular Dynamics Simulations. *Eur. Biophys. J.* **2016**, *45*, 573−580.
(33) Noisakran, S.; Dechtawewat, T.; Avirutnan, P.; Kinoshita, T.; Siripanyaphinyo, U.; Puttikhunt, C.; Kasinrerk, W.; Malasit, P.; Sittisombut, N. Association of Dengue Virus NS1 Protein with Lipid Rafts. *J. Gen. Virol.* **2008**, *89*, 2492−2500.
(34) Scheiffele, P.; Rietveld, A.; Wilk, T.; Simons, K. Influenza Viruses Select Ordered Lipid Domains during Budding from the Plasma Membrane. *J. Biol. Chem.* **1999**, *274*, 2038−2044.
(35) Bavari, S.; Bosio, C. M.; Wiegand, E.; Ruthel, G.; Will, A. B.; Geisbert, T. W.; Hevey, M.; Schmaljohn, C.; Schmaljohn, A.; Aman, M. J. Lipid Raft Microdomains: A Gateway for Compartmentalized Trafficking of Ebola and Marburg Viruses. *J. Exp. Med.* **2002**, *195*, 593−602.
(36) Poveda-Cuevas, S. A.; Etchebest, C.; Barroso da Silva, F. L. Insights into the ZIKV NS1 Virology from Different Strains through a Fine Analysis of Physicochemical Properties. *ACS Omega* **2018**, *3*, 16212−16229.
(37) Poveda-Cuevas, S. A.; Etchebest, C.; Barroso da Silva, F. L. Identification of Electrostatic Epitopes in Flavivirus by Computer Simulations: The PROCEEDpKa Method. *J. Chem. Inf. Model.* **2020**, *60*, 944−963.
(38) Gonçalves, R. L.; de Lima Menezes, G.; Sussuchi, L.; Moreli, M. L.; Mottin, M.; Andrade, C. H.; Pereira, M.; da Silva, R. A. Dynamic Behavior of Dengue and Zika Viruses NS1 Protein Reveals Monomer−Monomer Interaction Mechanisms and Insights to Rational Drug Design. *J. Biomol. Struct. Dyn.* **2020**, *38*, 4353−4363.
(39) Karamzadeh, R.; Karimi-Jafari, M. H.; Sharifi-Zarchi, A.; Chitsaz, H.; Salekdeh, G. H.; Moosavi-Movahedi, A. A. Machine Learning and Network Analysis of Molecular Dynamics Trajectories Reveal Two Chains of Red/Ox-Specific Residue Interactions in Human Protein Disulfide Isomerase. *Sci. Rep.* **2017**, *7*, 3666.
(40) Globisch, C.; Krishnamani, V.; Deserno, M.; Peter, C. Optimization of an Elastic Network Augmented Coarse Grained Model to Study CCMV Capsid Deformation. *PLoS One* **2013**, *8*, e60582.
(41) Machado, M. R.; González, H. C.; Pantano, S. MD Simulations of Viruslike Particles with Supra CG Solvation Affordable to Desktop Computers. *J. Chem. Theory Comput.* **2017**, *13*, 5106−5116.
(42) Barroso da Silva, F. L.; MacKernan, D. Benchmarking a Fast Proton Titration Scheme in Implicit Solvent for Biomolecular Simulations. *J. Chem. Theory Comput.* **2017**, *13*, 2915−2929.
(43) Barroso da Silva, F. L.; Dias, L. G. Development of Constant-pH Simulation Methods in Implicit Solvent and Applications in Biomolecular Systems. *Biophys. Rev.* **2017**, *9*, 699−728.
(44) Teixeira, A. A. R.; Lund, M.; Barroso da Silva, F. L. Fast Proton Titration Scheme for Multiscale Modeling of Protein Solutions. *J. Chem. Theory Comput.* **2010**, *6*, 3259−3266.
(45) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14*, 71−73.
(46) de Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schäfer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput.* **2013**, *9*, 687−697.
(47) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.
(48) Wedemeyer, W. J.; Welker, E.; Narayan, M.; Scheraga, H. A. Disulfide Bonds and Protein Folding. *Biochemistry* **2000**, *39*, 4207−4216.
(49) Wallis, T. P.; Huang, C.-Y.; Nimkar, S. B.; Young, P. R.; Gorman, J. J. Determination of the Disulfide Bond Arrangement of Dengue Virus NS1 Protein. *J. Biol. Chem.* **2004**, *279*, 20729−20741.
(50) Pryor, M. J.; Wright, P. J. The Effects of Site-Directed Mutagenesis on the Dimerization and Secretion of the NS1 Protein Specified by Dengue Virus. *Virology* **1993**, *194*, 769−780.
(51) Freire, M. C. L. C.; Pol-Fachin, L.; Coêlho, D. F.; Viana, I. F. T.; Magalhães, T.; Cordeiro, M. T.; Fischer, N.; Loeffler, F. F.; Jaenisch, T.; Franca, R. F.; Marques, E. T. A.; Lins, R. D. Mapping Putative B-Cell Zika Virus NS1 Epitopes Provides Molecular Basis for Anti-NS1 Antibody Discrimination between Zika and Dengue Viruses. *ACS Omega* **2017**, *2*, 3913−3920.
(52) David, C. C.; Jacobs, D. J. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. *Protein Dynamics*; Springer, 2014; pp 193−226.
(53) Valadié, H.; Lacapčre, J. J.; Sanejouand, Y.-H.; Etchebest, C. Dynamical Properties of the MscL of Escherichia Coli: A Normal Mode Analysis. *J. Mol. Biol.* **2003**, *332*, 657−674.
(54) Skjaerven, L.; Martinez, A.; Reuter, N. Principal Component and Normal Mode Analysis of Proteins; a Quantitative Comparison Using the GroEL Subunit. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 232−243.
(55) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. Principal Component Analysis for Protein Folding Dynamics. *J. Mol. Biol.* **2009**, *385*, 312−329.
(56) Newman, M. E. J. Modularity and Community Structure in Networks. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8577−8582.
(57) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical Networks in TRNA: Protein Complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 6620−6625.
(58) Bhattacharyya, M.; Vishveshwara, S. Elucidation of the Conformational Free Energy Landscape in *H. Pylori* LuxS and Its Implications to Catalysis. *BMC Struct. Biol.* **2010**, *10*, 27.
(59) Newman, M. E. J. The Structure and Function of Complex Networks. *SIAM Rev.* **2003**, *45*, 167−256.
(60) Scarabelli, G.; Grant, B. J. Mapping the Structural and Dynamical Features of Kinesin Motor Domains. *PLoS Comput. Biol.* **2013**, *9*, e1003329.
(61) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric Pathways in Imidazole Glycerol Phosphate Synthase. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1428−E1436.
(62) Barroso da Silva, F. L.; Sterpone, F.; Derreumaux, P. OPEP6: A New constant-pH Molecular Dynamics Simulation Scheme with OPEP Coarse-Grained Force Field. *J. Chem. Theory Comput.* **2019**, *15*, 3875−3888.
(63) Webb, B.; Sali, A. Protein Structure Modeling with MODELLER. *Protein Structure Prediction. Methods and Protocols*; Humana Press, 2014; pp 1−15.


https://dx.doi.org/10.1021/acs.jcim.0c01377
*J. Chem. Inf. Model.* 2021, 61, 1516−1530

(64) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859−1865.
(65) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1*−*2*, 19−25.
(66) Periole, X.; Cavalli, M.; Marrink, S.-J.; Ceruso, M. A. Combining an Elastic Network with a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J. Chem. Theory Comput.* **2009**, *5*, 2531−2543.
(67) Newman, M. E. J.; Girvan, M. Finding and Evaluating Community Structure in Networks. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2004**, *69*, 026113.


https://dx.doi.org/10.1021/acs.jcim.0c01377
*J. Chem. Inf. Model.* 2021, 61, 1516−1530

## Chapter 9

# NS1 interacting with a lipid bilayer

In Chapter 8, it was shown that the $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ proteins present significant dynamical differences in the IL, GF, and β-roll regions. Previous studies have hypothesized that these parts of NS1 are key for the interaction with lipid membranes (Xu et al., 2016; Brown et al., 2016). In this chapter, the NS1-membrane heterocomplex of the Ugandan and Brazilian ZIKV variants is computationally analyzed by MD CG simulations.

The main question of this part of the study is:

*Do the proteins $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ have different affinities for a lipid bilayer?*

A summary of the molecular model used, the methods to solve them, and the measures calculated appear in Table 9.1.

Table 9.1: Molecular modeling and measurements carried out in Chapter 9.

| Level of detail | Solution | Measurements |
|---|---|---|
| CG model at the 4to1 level | Classical MD | Canonical +Statistical |

Link m/s (*BioRxiv*): https://www.biorxiv.org/content/10.1101/2022.04.08.487413v1.

*Manuscript submitted to the Journal of Proteome Research (m/s ID: pr-2022-00239t) on 23/04/2022.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# NS1 from Two Zika Virus Strains Differently Interact with a Membrane: A Plausible Explanation for their Differential Virulence

Sergio A. Poveda-Cuevas [a,b,e] , Fernando L. Barroso da Silva [a,b,c,e], Catherine Etchebest[d,e*]

[a]Universidade de São Paulo, Programa Interunidades em Bioinformática, Rua do Matão, 1010, BR-05508-090 São Paulo, São Paulo, Brazil.

[b]Universidade de São Paulo, Departamento de Ciências Biomoleculares, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Av. do Café, s/no-Campus da USP, BR-14040-903 Ribeirão Preto, São Paulo, Brazil.

[c] Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27695, United States.

[d]Université Paris Cité, Biologie Intégrée du Globule Rouge, Equipe 2, INSERM, F-75015 Paris, France.

[e]University of São Paulo and Université de Paris International Laboratory in Structural Bioinformatics, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Av. do Café, s/no-Campus da USP, Bloco B, BR-14040-903 Ribeirão Preto, São Paulo, Brazil.

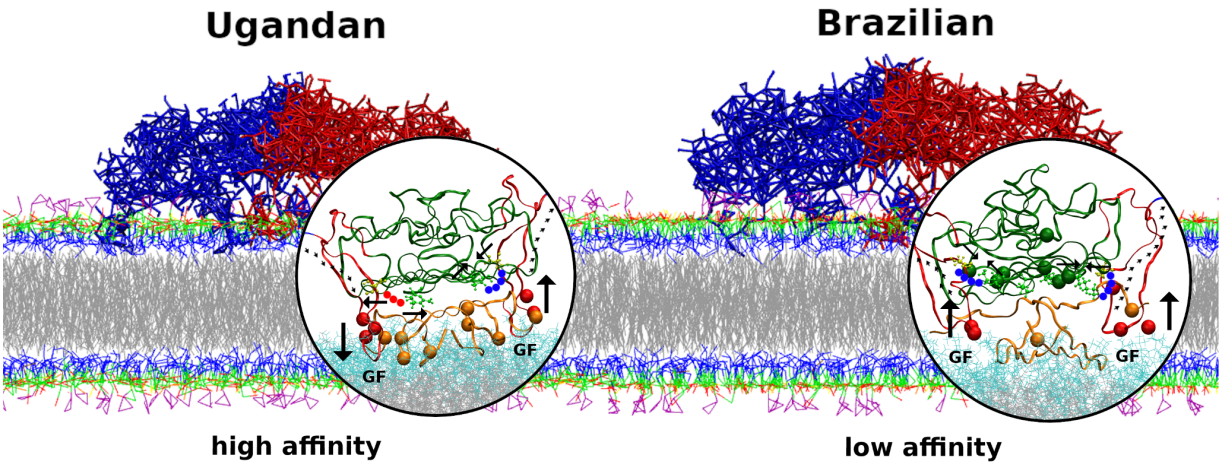
* Corresponding Author

E-mail: catherine.etchebest@inserm.fr

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


## ABSTRACT

Zika virus (ZIKV) from Uganda (UG) expresses a phenotype related to fetal loss whereas the variant from Brazil (BR) induces microcephaly in neonates. The differential virulence has a direct relation to biomolecular mechanisms that make one strain more aggressive than the other. The non-structural protein 1 (NS1) is a key viral toxin to comprehend these viral discrepancies because of its versatility in many processes of the virus life cycle. Here, we aim to examine through coarse-grained models and molecular dynamics simulations the protein-membrane interactions for both $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ dimers. A first evaluation allowed us to establish that the NS1 proteins, in the membrane presence, explore new conformational spaces when compared to systems simulated without a lipid bilayer. These events derive from both differential coupling patterns and discrepant binding affinities to the membrane. The N-terminal domain, intertwined loop, and greasy finger proposed previously as binding membrane regions were also computationally confirmed by us. The anchoring sites have aromatic and ionizable residues that manage the assembly of NS1 toward the membrane, especially for the Ugandan variant. Furthermore, the protein-membrane interaction induces perturbation in the distal part related to putative epitopes. These results open up new strategies for the treatment of flaviviruses.


**Keywords:** Protein-membrane interaction, binding affinity, non-structural protein 1, coarse-grained models, molecular dynamics.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## 1. INTRODUCTION

Zika virus (ZIKV) was isolated for the first time in 1947 in Uganda and since 2015, outbreaks related to microcephaly in newborns or fetuses have been reported in northeast Brazil [1]. To date, the most accepted hypothesis to understand this phenotypic divergence, points to discrepant neurotropism patterns for ZIKV strains from different phylogenetic origins [2–4]. For instance, both Ugandan (UG) and Brazilian (BR) ZIKV strains from African and Asiatic lineages, respectively, can be transmitted vertically and overcome the placental barriers [5,6]. However, ZIKV-UG expresses a phenotype that leads to fetal loss [7], whereas ZIKV-BR attenuates the pathogenic effects, allowing an unborn irregular development leading to birth defects [8]. All these phenomena come from differential molecular mechanisms among the ZIKV strains. This discussion is transcendental because the genomic plasticity of these biological entities can cause novel mechanisms and unleash impacts on the world's health systems. Hence, any knowledge related to these mechanisms gained at the molecular scale is valuable for the development of effective antiviral drugs or virus diagnoses.

Among the non-structural (NS) proteins that are encoded in the genome of ZIKV but do not belong to the viral envelope, the auxiliary macromolecule NS1 may independently play different roles involving host immune system evasion and pathogenesis [9–11]. Since NS1 is absent in the virion, it is a promising vaccine candidate because of the low risk of antibody-dependent enhancement (ADE) mechanism [12,13]. However, molecular mechanisms that trigger autoantibody proliferation via NS1 presence have also been recognized and are critical to understanding harmful autoimmune diseases in humans or dengue hemorrhagic fever [9,14,15]. A high concentration of NS1 monomers inside the cell allows the oligomerization to produce more liposoluble dimers [9,10]. The dimeric NS1 has a higher affinity for endoplasmic reticulum (ER) membranes (mNS1, for membrane-bound NS1), especially for lipid microdomains so-called lipid rafts [16,17]. Later, a portion of mNS1 in association with the non-structural viral protein NS4B works as a cofactor for the replication complex [18]. Concomitantly, mNS1 together with the precursor membrane protein, prM, and the Env envelope protein are

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

responsible for the infectious viral particle production [19]. A second stage shows that mNS1 traffics from the ER through the Golgi apparatus, where dimers associate to form soluble homo-hexamers (sNS1, for soluble hexamer) further secreted through cellular secretion pathways [20,21]. The extracellular barrel-like sNS1 accommodates in a central channel, a lipid cargo that stabilizes the structure [20]. While sNS1 is the oligomer most secreted into the extracellular environment, a small portion of mNS1 has also been detected upon the surface of infected cells [16]. Moreover, fluorescence assays have revealed that mNS1 is more intracellularly frequent, contributing to the structural remodeling of the ER membrane for the replication organelle [22].

The versatility of NS1 in the different infection/virulence stages of flaviviruses makes it an important target. Some comparative studies of Dengue virus (DENV), West Nile Virus (WNV), and ZIKV show that distinct electrostatic profiles could have implications on interactions of NS1 with different partners [*e.g.*, membrane, antibody (Ab), etc] [23–25]. Interestingly, we have found that electrostatic properties are also affected among ZIKV strains sharing more than 96% sequence identity [26]. Moreover, by examining the electrostatic properties of the antigen-Ab interaction through constant pH (CpH) Monte Carlo (MC) simulations and the PROCEEDpK$_a$ method, we found that antigenic regions of NS1$_{\text{ZIKV-UG}}$ and NS1$_{\text{ZIKV-BR}}$ exhibit discrepant "electrostatic epitopes" [27] (defined as titratable residues whose the p$K_a$'s are shifted by the presence of other charged body in the vicinity, *e.g.*, the occupancy of an antibody). On the other hand, besides electrostatic properties, our previous results of classical molecular dynamics (MD) simulations, have revealed that the dynamical profile of the two strains differs, particularly, in the putative regions of membrane interaction localized in the N-terminal (NTD) and wing domains, in addition, to the connector subdomain [28]. Furthermore, we observed that NS1 from African origin is more flexible than Asian ones. MD simulations have also brought noteworthy outcomes to understanding the monomer–monomer interaction mechanisms for NS1$_{\text{ZIKV}}$ and NS1$_{\text{DENV}}$ providing keys for the rational development of flaviviral-specific drugs [29]. Overall, these studies have provided important insights, supporting the idea of a close relationship between the

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

molecular mechanisms adopted by NS1 and the differential virulence expressed by several ZIKV strains.

In the present study, we go further and fully characterize the protein-membrane (PM) interaction of ZIKV NS1 for the same two strains as in our previous study [28], namely one originating from Uganda and the other one from Brazil. Endoplasmic reticulum being the first organelle encountered along the infectivity pathway of the NS1 protein, we have chosen a membrane composition close to the ER membrane. For the first time, very long (~20 μs) and multiple MD simulations using coarse-grained (CG) force-field models were performed to provide relevant statistical measures of the behavior of the two $NS1_{ZIKV}$ strains relative to the membrane. Moreover, we have compared several structural properties between $NS1_{ZIKV}$ in the absence (*apo*) and the presence (*holo*) of the lipid bilayer. The conformational spaces explored by the systems were examined using essential dynamics (ED) and dynamical cross-correlation analysis (DCCM). The impact of membrane on protein-protein (PP) and putative protein-Ab (PA) interfaces was also assessed. In support of our previous conclusions, the present results reveal that $NS1_{ZIKV\text{-}UG}$ exhibits more membrane-interacting residues compared with $NS1_{ZIKV\text{-}BR}$ and is more deeply anchored in the membrane. This striking difference brings into light the role of the membrane itself in the infectivity process, a topic that recently has started to receive more attention [30].

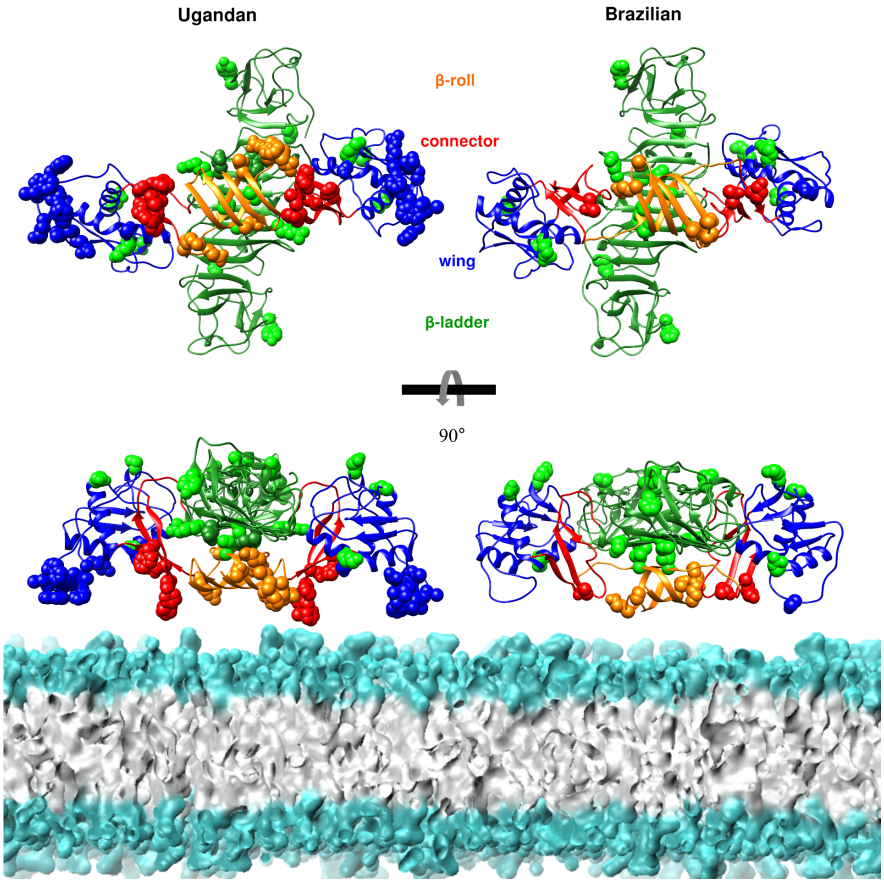
## 2. METHODOLOGY

### 2.1. Modeling *apo* systems

The crystallographic structures of the dimers $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ were extracted from the RCSB Protein Data Bank [PDB id 5K6K (1.89 Å) and 5GS6 (2.85 Å), respectively]. These two proteins share a similar scaffold and topology. They differ only in some loop regions (1.9 Å RMSD across all 352 pairs and 0.706 Å RMSD between 310 pruned atom pairs). Missing regions were completed following the same protocol as previously described and detailed by us (see Figure 1) [26,28].

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AA proteins were converted to CG with the online CHARMM-GUI tool [31]. The Martini 2.2P force field was used, combined with an elastic network for the secondary structures and harmonic additional constraints [32] to preserve most of the native contacts as observed in our previous AA-MD simulations of the NS1 proteins [28]. Parametrization was by default (*i.e.*, the lower and upper elastic bond cutoffs were set to 0 and 0.9 nm; and the bond strengths were independent of the bond length). The elastic bond force constant was defined as 500 kJ mol $^{-1}$ nm$^{-2}$. Counterions were added to neutralize the systems and additional CG ions were placed to reach the salt concentration of 0.15 M. Proteins were protonated assuming pH equal to 7 and p$K_a$ values calculated by us through CpH MC simulations [26]. The final systems contain 31,082 water beads, 349 $Na^+$ beads, and 343 $Cl^-$ beads for the $NS1_{ZIKV\text{-}BR}$ dimer and 31,105 water beads, 353 $Na^+$ beads, and 343 $Cl^-$ beads for the $NS1_{ZIKV\text{-}UG}$ dimer.



bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Figure 1.** Domain organization of NS1 using ribbon's representation of the X-ray crystal structure of the $NS1_{ZIKV\text{-}UG}$ (PDB id 5K6K) and $NS1_{ZIKV\text{-}BR}$ dimer (PDB id 5GS6). β-roll, wing, and β-ladder domains are colored in orange, blue, and green, respectively. The connector subdomain is colored in red. The light green spheres represent the point mutations between strains. It is to be noted that the 90° rotation about the *x*-axis for the dimers represents the estimated position of interaction with the membrane. A membrane model in surface representation is shown. The headgroups and aliphatic chains of the membrane appear as cyan and white, respectively. Orange, red, blue, and dark green spheres represent the residues with more frequent contact based on the occupied fraction shift for the residues involved in the protein-membrane interface ($\Delta OF_{PM}$) for $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ (for more details, see the Results and Discussion section).

### 2.2. Modeling *holo* systems

Again, employing CHARMM-GUI, the CG double-layer membrane (based on Martini 2.2P force field) was placed in a cubic box simulation of 50 nm edge. The membrane model was built to roughly reproduce the lipid composition of an ER membrane [33]. Thus, the lipid composition had the ratios 18:7:2:2:1, for respectively, POPC (PC), POPE (PE), POPI (PI), POPS (PS), and POSM (SM) per leaflet (area total estimated = 47158 $Å^2$). Each protein was aligned along the *z*-axis of the CG membrane based on OPM server (https://opm.phar.umich.edu/ppm_server2) predictions [34]. Figure 1 shows a scheme to exemplify the initial arrangement of the NS1 systems. Note that the systems have been moved up relative to the membrane for better visualization of the regions involved in the anchoring. The final systems contain 52,884 water beads, 834 $Na^+$ beads, and 600 $Cl^-$ beads for the $NS1_{ZIKV\text{-}UG}$; and 52,299 water beads, 823 $Na^+$ beads, and 593 $Cl^-$ beads for the $NS1_{ZIKV\text{-}BR}$. Both systems contain 424 PC, 160 PE, 56 PI, 56 PS, and 24 SM per layer.

### 2.3. MD simulations protocol

Simulations were performed in a NPT ensemble with GROMACS 5.1.5 or later versions, applying periodic boundary conditions [35]. During all simulations, titratable groups were kept

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

constant at their initial ionizable states. The electrostatic interactions were treated with the reaction-field algorithm with a cutoff of 1.1 nm and $\varepsilon_r$ equal to 2.5. The Verlet cutoff scheme and the potential-shift modifier were used for van der Waals forces with a cutoff of 1.1 nm. The time step was 20 fs as recommended for the used CG force field [36].

For the *apo* systems, two minimizations were carried out for 3,000 steps using the steepest descent algorithm, the first one with softcore constraints and the second one without any structural constraint. Later, an equilibration was performed during 10 ns in the NPT ensemble using Berendsen's algorithm for the temperature and pressure coupling, with $T$=300 K ($\tau_T$ = 1 ps) and $P$=1 bar ($\tau_P$ = 5 ps). In the production phase, the temperature coupling $\tau_T$ was changed to 12.0 ps and the compressibility to $4.5 \times 10^{-5}$ bar using an NPT ensemble. This protocol was based on recommendations for the Martini CG force field [36].

Regarding *holo* systems, we employed three different steps as it is suggested for the used force field [32,37]: *i*) a steepest descent minimization using softcore constraints for 3,000 steps followed by *ii*) a similar run without any structural restraints, then *iii*) an equilibration phase in NPT ensemble (P=1 bar, $\tau_P$ = 5 ps) and (T=300 K, $\tau_T$ = 1 ps) using Berendsen algorithms, and comprising 5 successive steps (for 1 ns each), where the integration time step was gradually increased (2, 5, 10, 15 and 20 fs), while restrictions imposed on the protein and lipid head beads were gradually removed. Subsequently, the production phase in the NPT ensemble was done with the Parrinello-Rahman method with $\tau_P$ = 12 ps. For both equilibration and production runs the compressibility was set as a $3.0 \times 10^{-5}$ bar.

For both *apo* and *holo* systems, three replicates differing by initial velocities were simulated during 10 and 20 μs, respectively.

### 2.4. Canonical measurements

Standard calculations like root mean square deviation (RMSD), the radius of gyration ($R_g$), solvent accessible surface area (SASA), and root mean square fluctuation (RMSF) were

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

performed with the GROMACS tools from the trajectories of each system [28]. For all cases the reference was their respective modeled crystallographic structures. The average and standard deviation for each measure were computed. An additional averaging was performed from the RMSD of chains A and B separating them (this was labeled with the subscript "chain").

### 2.5. Evaluation of the biological interfaces

To evaluate the persistence time of monomer-monomer contacts (noted PP interface), calculations of the occupied fraction ($OF_{PP}$) provided by GROMACS tools were used (see Figure S1, step *a*). A residue-residue contact is defined when the distance between any CG bead of the two residues is smaller or equal to the distance threshold of 0.52 nm, which corresponds to the sum of the van der Waals radii of the standard Martini CG beads (0.26 nm) [38]. Hence, the $OF_{PP}$ by definition is the number of contacts of a residue from chain A with another from chain B (or the contrary) along the trajectory, divided by the total number of frames considered (*i.e.*, the frequency). Following a similar protocol given in our previous work [28], the difference between $OF_{i\ PP_NS1_ZIKV\text{-}UG}$ - $OF_{i\ PP_NS1_ZIKV\text{-}BR}$ for each interface residue *i* is calculated for both *apo* ($\Delta OF_{PP_(-)_membrane}$) and *holo* ($\Delta OF_{PP_(+)_membrane}$) systems. $\Delta OF_{PP} > 0$ indicates that residue *i* is having more contacts in $NS1_{ZIKV\text{-}UG}$ compared to $NS1_{ZIKV\text{-}BR}$, whereas $\Delta OF_{PP} < 0$ means the reverse (residue *i* presents more frequent contacts for $NS1_{ZIKV\text{-}BR}$ than $NS1_{ZIKV\text{-}UG}$) (see Figure S1, step *d*). To highlight the most significant and relevant residues among the strains, we considered 3 conditions: *i*) the differences in the measures between the strains must be larger than the standard deviations (see Figure S1, step *c*), *ii*) the $|\Delta OF_{PP}|$ should be greater than 0.1 (see Figure S1, step *e*), and *iii*) the $OF_{PP}$ for any system must be greater than to a specific threshold (see Figure S1, step *b*). Regarding item *iii*, we examined $OF_{PP}$ values between 0 and 1 every 0.1 to reduce interpretation bias. Whether one amino acid meets these *three* conditions (see Figure S1, step *f*), it will be considered a "relevant residue" for *one* of the strains. Figure S1 shows the different steps to define a relevant residue.

The interactions between the protein and the membrane have been evaluated similarly, namely through the occupied fraction ($OF_{PM}$). In this case, the $OF_{PM}$ is defined as the number of contacts that a residue of the dimer has with any CG bead of lipid molecules. We computed the difference

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$OF_{i\ PM_NS1_ZIKV\text{-}UG}$ - $OF_{i\ PM_NS1_ZIKV\text{-}BR}$ ($\Delta OF_{PM}$) for each PM interface residue *i* to evaluate the possible discrepancies between the strains. All interpretations or conventions for $\Delta OF_{PM}$ followed the same rules as $\Delta OF_{PP}$ and the relevant and significant residues involved in the membrane interaction were similarly defined.

The preferential location of the PM relevant residues defined above was also determined through the calculation of density profiles (employing the density GROMACS tool [35]) along the *z*-axis of the simulation box. A comparison with the density profiles of a specific lipid moiety (INO, NC3, NH3, CNO, PO4, GL, and AM) was achieved. The names given in parentheses are the "nicknames" given in the Martini force field for the different functional chemical groups of the lipid headgroups [39]. The impact of the protein on the membrane itself was evaluated by calculating lipid density calculations in the XY plane (using the densmap in GROMACS tools [35]), which allows highlighting the phospholipids close to the protein. The averaging direction and bin size were set as *z* and 0.1 nm, respectively.

**2.6. Statistical measurements**

Essential dynamics of all systems were established using principal component analysis (PCA). We considered the backbone bead (BB) positions of the CG models. Analyses took into account the whole set of conformations saved every 1,000 ps along the simulations, pooling together all the replicates of the two $NS1_{ZIKV}$ in the absence and presence of the membrane. Besides, additional PCAs were carried out but took into account just one of the strains. In other words, only the *apo* and *holo* form trajectories of $NS1_{ZIKV\text{-}UG}$ were used without $NS1_{ZIKV\text{-}BR}$ or vice versa.

The dynamic cross-correlation matrix (DCCM) of the backbone beads (BB) was computed employing the Bio3D tool (http://thegrantlab.org/bio3d/) for each CG system [40]. The number of conformations considered was 2,700 (*apo*) and 5,700 (*holo*). The difference matrix between the two conditions, $|DCCM_{C\ NS1_strain_a}|$ - $|DCCM_{C\ NS1_strain_h}|$ was computed (noted $\Delta DCCM_{strain}$) for each strain. Note that the absolute value was applied for each DCCM matrix. For a *specific strain*, a positive $\Delta DCCM_{strain}$ means that the coupling of residues is favored in the *apo* form and

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the reverse for a $\Delta DCCM_{strain}$ negative. This protocol was also applied to compare the behavior *between* the strains in each *apo* or *holo* condition. Consequently, for a *specific condition*, *apo* or *holo*, | $DCCM_{C\ NS1_ZIKV\text{-}UG_form}$| - |$DCCM_{C\ NS1_ZIKV\text{-}BR_form}$| was calculated and referred to as $\Delta DCCM_h$ or $\Delta DCCM_a$, respectively. In the *holo* condition, a negative $\Delta DCCM_h$ means that the coupling of residues is favored for $NS1_{ZIKV\text{-}UG_h}$ over $NS1_{ZIKV\text{-}BR_h}$ whereas a positive value means the reverse. A similar interpretation for the *apo* systems is given with $\Delta DCCM_a$ .

### 2.7. Selection of putative epitope regions

To date, there is no structure for $NS1_{ZIKV}$ in association with Ab in the RCSB PDB. However, X-ray crystallographic structures of the NS1-Ab complex have been determined for some flaviviruses, *e.g.*, $NS1_{WNV}$ [PDB id 4OII (3 Å)] [41] interacting with the Fab fragment 22NS1 ($Ab_{WNV}$) and $NS1_{DENV}$ [PDB ids 6WEQ (3.2 Å), 6WER (4 Å), and 7K93 (3 Å)] [42] in complexation with the Ab 2B7 ($Ab_{DENV}$). Interestingly, the cross-reactivity of a given antibody between different flaviviruses has been demonstrated [42–45]. This may provide some clues on how interactions between $NS1_{ZIKV}$ and putative Ab's might occur. Consequently, based on the homology and the high structural similarity between the three flaviviruses (*i.e.*, $NS1_{ZIKV}$, $NS1_{WNV,}$ and $NS1_{DENV}$), it was assumed that the putative antigenic regions for $NS1_{ZIKV}$ are similar to those observed in the complexes $NS1_{WNV}$-$Ab_{WNV}$ and $NS1_{DENV}$-$Ab_{DENV}$ [26,27].

## 3. RESULTS AND DISCUSSION

Previously, we have shown how both electrostatic and dynamical properties can vary as a function of point mutations on $NS1_{ZIKV}$ proteins from two different phylogenetic origins [26–28]. In these proceeding calculations, the protein was considered in a pure soluble environment (electrolyte solution). Interestingly, regarding the dynamical properties, we found significant discrepancies between the two strains, particularly in regions corresponding to putative membrane binding sites. Hence, in the present study, we go further in the understanding of the molecular mechanisms that govern the infectivity of the two strains, by examining the presence of an *explicit* lipid bilayer on the dynamics of NS1. In this aim, as we did in our previous study

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[28], a CG representation of the interacting molecules was employed. CG models are efficient to describe this type of macromolecular complexes, in addition to agreeing with AA models. The gain of several temporal orders of magnitude when compared to AA allows the detection of molecular phenomena that demand higher time scales.

This study focuses on answering the following four main questions: *i*) what is the influence of the membrane on the NS1 proteins for each strain?, *ii*) what are the differences of the behavior of the NS1 proteins between the two strains in the presence of the membrane?, *iii*) what are the detailed molecular mechanisms that govern the protein-membrane interaction for each one?, and *iv*) how are the biological interfaces, *i.e.*, monomer-monomer interface and putative epitope regions impacted by the presence of the membrane? In this aim, we compared the structural/dynamic properties of the $NS1_{ZIKV\text{-}UG}$ system without (*apo*, labeled "*a"* in the following) and with the membrane (*holo*, labeled "*h"* in the following). The same comparison was done for $NS1_{ZIKV\text{-}BR}$. Subsequently, the two $NS1_{ZIKV}$s were compared to each other in their two different form versions.

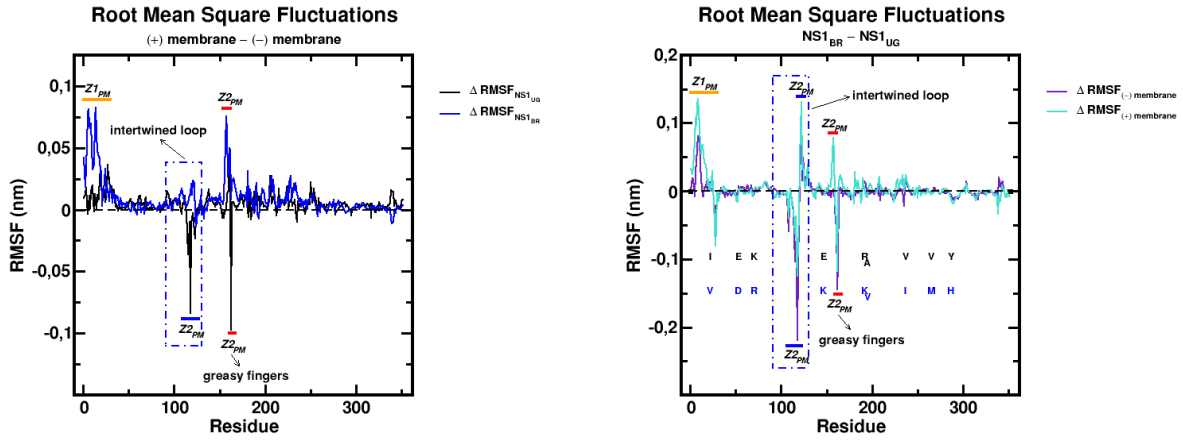
### 3.1. NS1 proteins explore new structural configurations in the membrane presence

As expected, RMSF calculations (see Figure 2) revealed that the structural parts being the most affected by the presence of the membrane for both $NS1_{ZIKV_h}$ match with the PM putative locations. They correspond to those regions previously suggested [28], namely zone 1 [$Z1_{PM}$, 1-29 (β-roll)] and zone 2 [$Z2_{PM}$, 91-130 (wing), 135-145 (wing), and 155-165 (connector)]. $Z2_{PM}$ could be subdivided into two key segments, the so-called intertwined loop (IL, 91-130) and greasy fingers (GF, 159-162), which are shown to be involved in biological functions [24,25,46].

To trap the main dynamical features of the viral proteins, the essential dynamics (ED) of the systems as examined by calculating the principal components (PCs) provided by PCA.[43–45] PCA is actually a valuable tool to visualize clusters of representative conformations that may be difficult to identify. It is also helpful to assess whether the two systems visit similar conformational states or not. In this aim, all the replicates for *holo* and *apo* systems were merged altogether. The PCA of the BB positions was calculated for a set of frames extracted from this

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ensemble. This approach was successfully applied in our previous study and permitted us to show that CG models follow similar EDs when compared to AA $NS1_{ZIKV}$ systems [28].



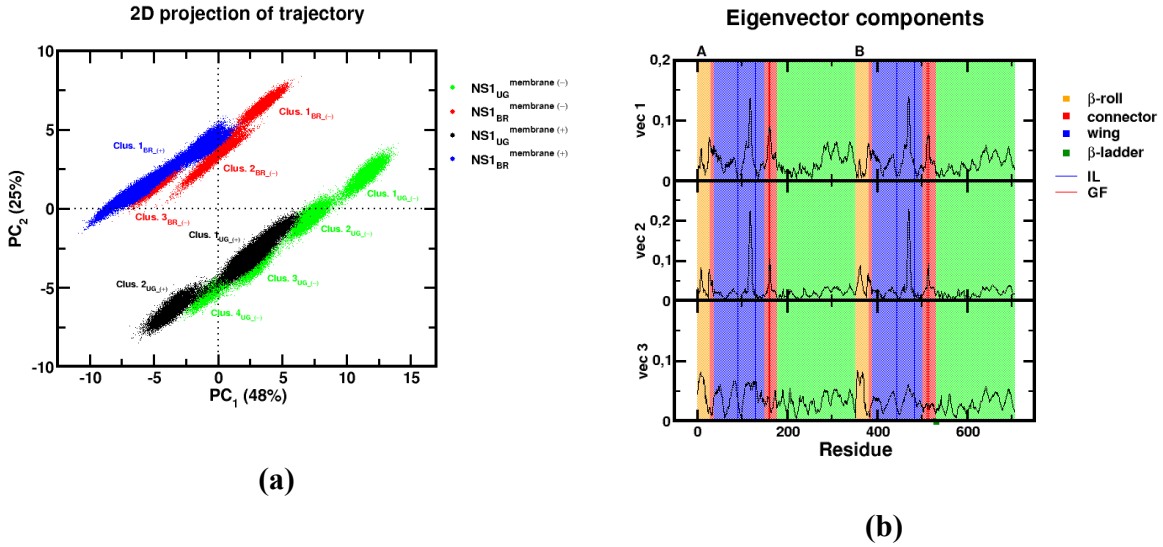

**(a)** **(b)**

**Figure 2.** Averaged structural properties for the dimeric $NS1_{ZIKV\text{-}UG}$ (PDB id 5K6K) and $NS1_{ZIKV\text{-}BR}$ (PDB id 5GS6) at the presence of the membrane. (a) Difference $RMSF_{(+)\ membrane} - RMSF_{(-)\ membrane}$. (b) Difference $RMSF_{NS1_ZIKV\text{-}BR} - RMSF_{NS1_ZIKV\text{-}UG}$. Sequence changes between the $NS1_{ZIKV\text{-}UG}$ (black letters) and $NS1_{ZIKV\text{-}BR}$ (blue letters) are shown below the curves in plot (b). For further explanation see the supplementary material.

In the present case, the first two PCs capture 73% ($PC_1$=48% and $PC_2$=25%) of the total variance (see Figure S2a), which is rather large considering the size of the systems. Consequently, even by restricting our detailed analyses to these two PCs, we can be confident in the relevance of the large-scale protein motions observed in the $NS1_{ZIKV}$ systems. The projection of the ensemble of conformations in the $PC_1$-$PC_2$ plane is given in Figure 3a. As a first result, it is clear that the two strains visit distinguishable conformational spaces. The effect of the membrane is also visible but to a lesser extent (black versus green points and red versus blue points in Figure 3 for $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$, respectively). The greatest contribution to the variance for these systems is due to flexible regions identified through eigenvector component calculations for $PC_1$ and $PC_2$ (see

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3b). They are related to the same $Z1_{PM}$ and $Z2_{PM}$ zones found in the RMSF profiles (see Figure 2b).




**(a)** **(b)**

**Figure 3.** (a) Principal component analysis for two $NS1_{ZIKV}$ in the presence [membrane (+)] and absence of membrane [membrane (-)]. (b) Eigenvector components calculations for the first three PCs that represent 79% of the total variance ($PC_1$ = 48%, $PC_2$ = 25%, and $PC_3$ = 6%).

Despite some overlap between the *holo* and *apo* forms, the conformational space visited in the presence of the membrane looks slightly reduced or at least less diverse, particularly for the Brazilian strain, compared to the situation in the absence of the membrane. Indeed, in the case of $NS1_{ZIKV\text{-}UG}$, four clusters (labeled Clus.x$_{_UG(-)}$) can be distinguished for the *apo* compared to two clusters for the *holo* forms (labeled Clus.x$_{_UG(+)}$). In the same way, three clusters (labeled Clus.x$_{_BR(-)}$) can be discerned for the *apo* structure and only one cluster for the *holo* one (labeled Clus.x$_{_BR(+)}$), in the case of $NS1_{ZIKV\text{-}BR}$. This observation is confirmed by examining the PCA of *apo and holo* simulations for each strain separately: there are unshared conformations between the *apo* and *holo* forms of $NS1_{ZIKV\text{-}UG}$. [see $PC_1$ (65% of total variance) and $PC_2$ (9%) in Figure S2b], as well for the two Brazilian forms [$PC_1$ (56%) and $PC_2$ (10%) in Figure S2c].

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 3.2. Comparison of the nature of motions for two NS1 in the presence of the membrane

The DCCMs were calculated for each strain in the two situations. We focused on the differences between maps (see Figures S3 and Figure 4), which are *apo* versus *holo* for each strain, and $NS1_{ZIKV\text{-}BR}$ versus $NS1_{ZIKV\text{-}UG}$ to highlight the characteristics of correlated motions in each situation. Each block of residues indicated in the figures corresponds to β-roll, wing, connector, and β-ladder regions. A similar comparative strategy was followed for the *apo* forms in our previous study [28]. First, the results (see Figures S3a-b) show that the membrane mainly influences the correlated (or anticorrelated) motions between the β-ladder (in green) and the wing regions (in blue) within and between chains. This is true for both strains, but more striking for the $NS1_{ZIKV\text{-}UG}$, in particular in the putative epitope regions (PA). These PA regions were deduced by similarity with those found in the antibody-NS1 complex from two other members of the Flavivirus, namely DENV and WNV (noted $Ab_{DENV}$ and $Ab_{WNV}$ in Figure 4). The $Ab_{WNV}$ region seems particularly affected by the presence of the membrane for UG as shown by the significant modifications of the (anti)-correlated motions between this region located in the β-ladder and the wings. Interestingly, Biering and collaborators have proposed that for the case $NS1_{DENV}$-$Ab_{DENV}$ complex, the occlusion of the PM interaction sites on the wing domain would trigger steric impediments in the β-ladder one, reducing the infectivity of DENV [42]. Therefore, our calculations supply additional clues for explaining the reduction of infectivity. Importantly, they could also contribute to providing a plausible explanation for the discrepancy of virulence between the two strains.

Second, in the presence of the membrane, at the strain level, we observe significant discrepancies between the patterns of correlation (see Figure 4). Similar behavior was found in the *apo* form (see Figure S3c), as also detected previously [28]. Furthermore, the greatest impact involves *functional* regions, namely the PP, and PM (NTD, IL, and GF) interfaces, and PA. Consequently, $\Delta DCCM_h$ matrix shows that still, the two strains behave slightly differently in these regions in the presence of the membrane.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

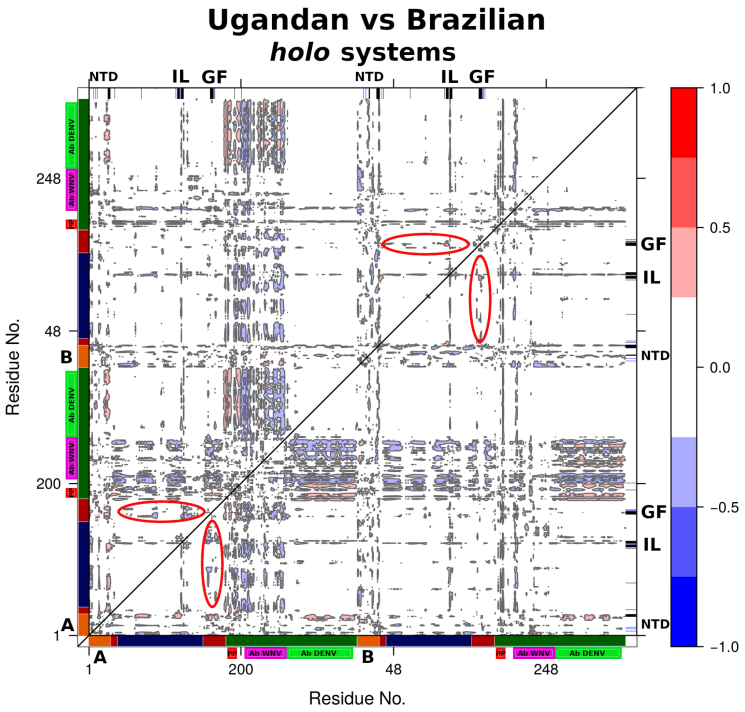



**Figure 4.** Dynamic cross-correlation matrix (DCCM) for two NS1 ZIKV proteins at the CG level. $\Delta DCCM_h$ based on the difference $|DCCM_{C\ NS1_ZIKV\text{-}UG_h}|$ - $|DCCM_{C\ NS1_ZIKV\text{-}BR_h}|$. The bottom and left margins are showing the β-roll (orange), connector (red), wing (blue), and β-ladder (green) regions. Relevant residues derived from $\Delta OF_{PM}$ analysis ($OF_{PM}$ = 0.5) appear on the top and right margins (the regions are distinguished as NTD, IL, and GF as well). The zone $Z2_{PP}$ of the β-ladder is shown as red rectangles. The putative epitopes based on $NS1_{WNV}$-$Ab_{WNV}$ and $NS1_{DENV}$-$Ab_{DENV}$ are roughly represented as violet and light green rectangles, respectively. More details are given in the text.

### 3.3. The membrane impacts the protein-protein interface

The PP interface is composed of two main sites, a first one encompassing the connector/β-ladder regions (noted $Z2_{PP}$) and a second one comprising the β-roll/connector domain (noted $Z1_{PP}$, similar to $Z1_{PM}$) (see Figure 5a). These two regions are intimately intricated, through contacts involving as well charged/polar as hydrophobic residues [D1, V19, I21, N23, D24, V25 ($Z1_{PP}$ of the β-roll), and G185, A187, K189, E192, A194V, S196, L198, W201 ($Z2_{PP}$ of the β-ladder), see

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these residues colored in green in Figure S4b]. All these contacts obviously contribute to the dimer stability [24,25].

Along with the simulations, the protein-protein interface of the *holo* form is once again modified by conformational adjustments *between* the monomers, as for the *apo* form. Indeed, since the conformational change quantified by the RMSD is smaller for a given chain than for the whole complex, the RMSD increase should be attributed to a change in the relative positioning of the two chains (see Table S1). Therefore, the dynamical behavior of the NS1 PP interface was assessed by calculating the occupied fraction of the residues at the monomer-monomer interface ($OF_{PP}$). We tested several thresholds (ranging from 0 to 1 by steps of 0.1) for defining what is considered as "relevant" residues, namely those that mainly participate in the interface for one of the $NS1_{ZIKV}$ all along with the simulations. The results (see Figure S4a) show that *i*) for both strains, only about ten residues are conserved at the interface for more than 80% of the time, *ii*) the very few residues that are maintained *all along the whole set* of simulations (OF = 1) are slightly larger for $NS1_{ZIKV\text{-}UG}$ compared to $NS1_{ZIKV\text{-}BR,}$ *iii*) in both cases, the presence of the membrane tends to reinforce the interface by increasing the number of contacts, and *iv*) except when the threshold > 0.8 (namely for the very few persistent contacts, see Figure S4a), for the two conditions, *apo* and *holo*, $NS1_{ZIKV\text{-}UG}$ has less relevant residues than $NS1_{ZIKV\text{-}BR}$ (compare green and red curves in Figure S4a, respectively). This discrepancy is likely due to the nature of the residues and their different arrangement within the PP interface between the two strains [28]. It would also illustrate larger stability of the "core" of the interface defined by persistent residues, *i.e.*, the roughly ten residues of the PP interface mentioned previously in *i)*, for $NS1_{ZIKV\text{-}UG}$ compared to $NS1_{ZIKV\text{-}BR}$, in accordance with the RMSF values.

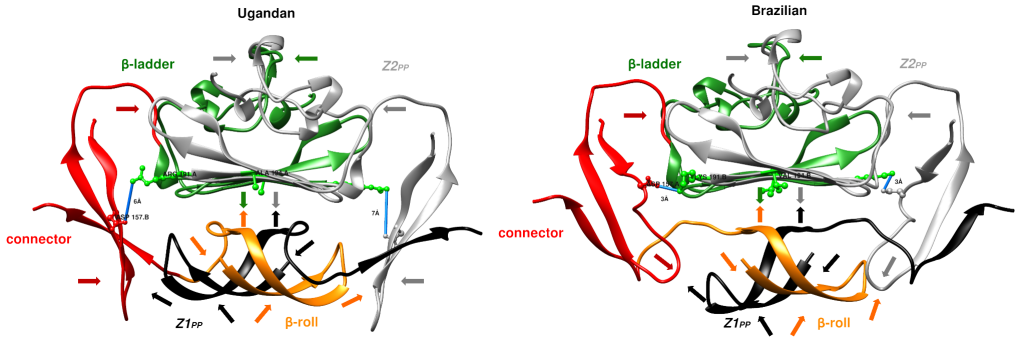
Next, we detailed the differences of the occupied fractions between the two strains in the two situations (*i.e.*, absence and presence of the membrane), which permits better characterization of the interface features. The results are depicted in Figure 5 (b-e), where a positive difference means that the residue is maintained for a longer time at the interface for the Uganda strain, whereas a negative difference means a longer time at the interface for the Brazilian strain. Since

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

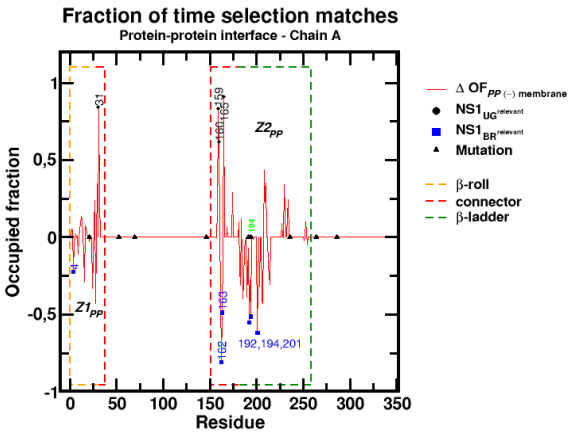
the systems are no more symmetrical due to the stochasticity of the simulations, each chain was considered separately. The "relevant" residues, *i.e.,* those that are in a specific OF range (for this case it was delimit considering OF ≥ 0.5, that is when the interaction happens at least half the time), their differences in OF are larger than the standard deviations, and |ΔOF| > 0.1 are indicated by colored points.

In the absence of the membrane, 11 "relevant" residues were found in the Ugandan strain compared to 13 for the Brazilian one. The distribution of these residues for chains A and B are shown in Figures 5b and 5c, respectively. Interestingly, there is a significant difference between the two contact regions $Z1_{PP}$ and $Z2_{PP}$: the $NS1_{ZIKV\text{-}UG_a}$ $Z1_{PP}$ region contains a larger number of relevant residues. This might contribute to a more stable interface region for $NS1_{ZIKV\text{-}UG_a}$ along the simulations, reducing the fluctuation of NTD when compared to $NS1_{ZIKV\text{-}BR_a}$ (see the violet curve in Figure 2b). Reversely, a larger number of residues is found in the $Z2_{PP}$ for the Brazilian strain. It is noteworthy that among the most frequently contacted residues for $NS1_{ZIKV\text{-}BR_a}$, there are two amino acid changes between the two variants, R191⇔K and A194⇔V. In the $NS1_{ZIKV\text{-}BR}$ X-ray structure, K191 and D157 are engaged in a salt bridge (Figure 5a), which is maintained in the AA simulations (data not shown). In the CG simulations, the existence of a salt bridge is more difficult to precisely assess due to the nature and size of the bead. Nonetheless, based on the X-ray CG model, a distance threshold smaller than 0.6 nm looks appropriate. Thus, in this case, the distance between K191 and D157 remains smaller than this threshold for $NS1_{ZIKV\text{-}BR_a}$ when compared to $NS1_{ZIKV\text{-}UG_a}$ (see Figure S4c) whatever the chain is considered, which means that the system $NS1_{ZIKV\text{-}BR_a}$ is stabilized by two salt bridges in the $Z2_{PP}$. Noteworthy, the role of salt bridges in the dimer stabilization previously underlined for $NS1_{ZIKV\text{-}UG_a}$ and $NS1_{DENV_a}$ [29] is highlighted for two $NS1_{ZIKV_a}$ in the present study, which confirms the importance of this kind of electrostatic interactions.

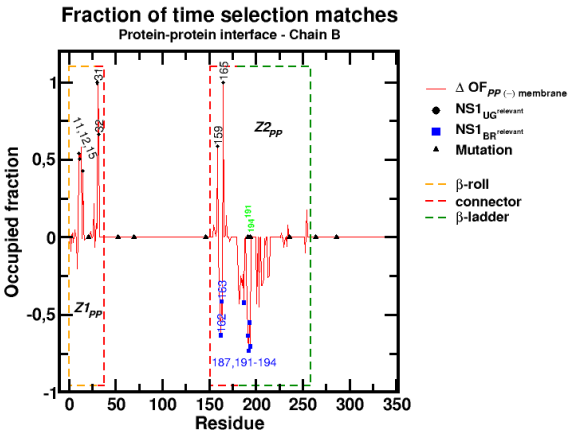
bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



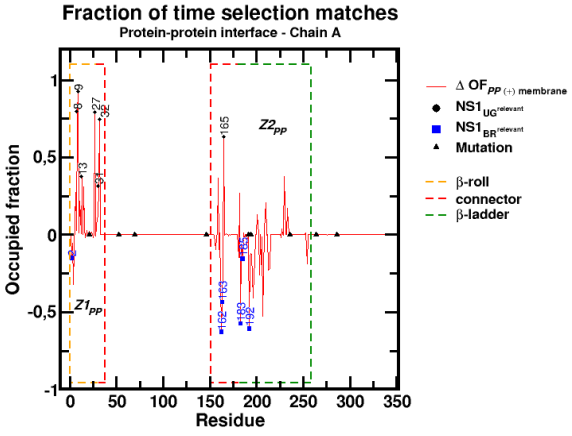

**(a)**



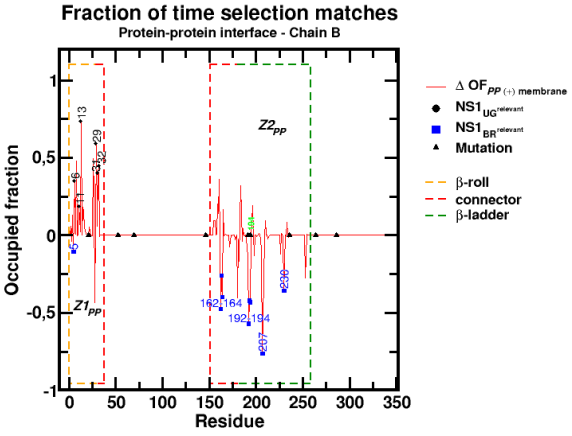

**(b)**




**(c)**




**(d)**




**(e)**

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Figure 5.** Protein-protein interfaces for two $NS1_{ZIKV}$ in the absence and presence of the membrane. (a) Ribbon's representation for the PP interface zones based on X-ray crystal structures of two ZIKV NS1 proteins (Uganda and Brazil with PDB id's 5K6K and 5GS6, respectively). One of the chains shows the structural domains and another the $Z1_{PP}$ and $Z2_{PP}$ zones. Point mutations are represented as balls and sticks in light green. The distance between the pair D157 and R191K is shown in blue. (b,c) $\Delta OF_{PP_(-)_membrane}$ for chains A and B, respectively. The differences were calculated as $OF_{i\ PP_NS1_ZIKV\text{-}UG}$ - $OF_{i\ PP_NS1_ZIKV\text{-}BR}$. Relevant residues for $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ are shown in black and blue symbols, respectively. Point mutations appear like black triangles (when a relevant residue matches with a mutation, the position residue is highlighted with light green). $Z1_{PP}$ and $Z2_{PP}$ zones are marked with dashed rectangles. (d,e) $\Delta OF_{PP_(+)_membrane}$ for chains A and B, respectively. Plots follow the same convention as (b,c).

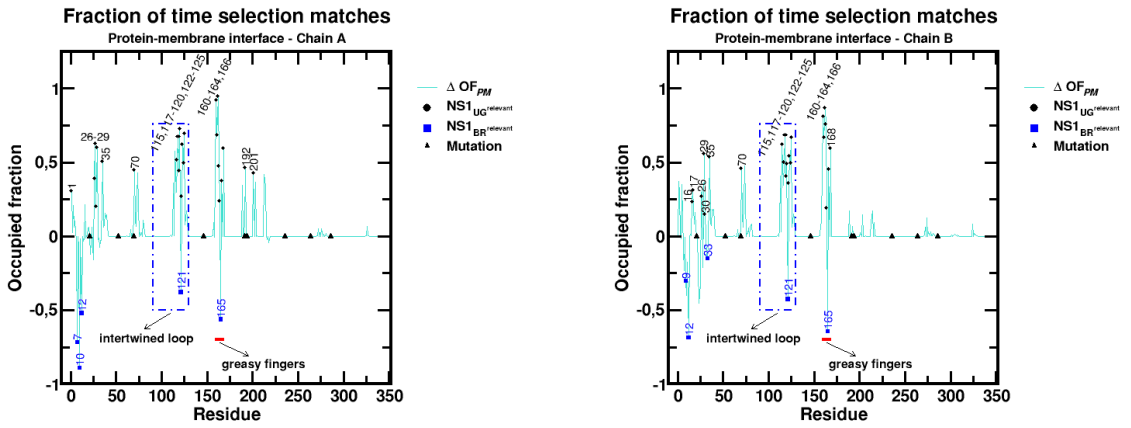
In the presence of the membrane, 13 relevant residues were found for $NS1_{ZIKV\text{-}UG_h}$ compared to 15 for $NS1_{ZIKV\text{-}BR_h}$. The distribution of these residues in chains A and B are shown in Figures 5d and 5e, respectively. Apart from a slight increase in the number of relevant residues for both strains compared to the *apo* case, most of the observations detailed above still stand: namely the $Z1_{PP}$ zone presents a higher number of relevant residues for $NS1_{ZIKV\text{-}UG_h}$ than $NS1_{ZIKV\text{-}BR_h}$, while $Z2_{PP}$ is dominated by the relevant residues of $NS1_{ZIKV\text{-}BR_h}$. Nevertheless, the presence of the membrane tends to affect the stability of the salt bridges [*i.e.* those formed between R191 and D157 ($NS1_{ZIKV\text{-}UG_h}$), and between K191 and D157 ($NS1_{ZIKV\text{-}BR_h}$)] and, in this particular case, the discrepancy between the two cases tends to vanish (see Figure S4d).

### 3.4. Differential PM interfaces for two $NS1_{ZIKV}$ proteins

Firstly, the total number of residues that remain close to any atom of the membrane is considerably higher for the Uganda strain compared to the Brazilian one, whatever the OF threshold for PM interactions. For example, the number of residues with OF > 0.8 is twice for $NS1_{ZIKV\text{-}UG_h}$ compared to $NS1_{ZIKV\text{-}BR_h}$ (Figure S5). Hence, whereas $NS1_{ZIKV\text{-}UG_h}$ is strongly anchored to the membrane through NTD, IL and GF regions, $NS1_{ZIKV\text{-}BR_h}$ has a tendency to have

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

more contacts through the NTD domain. Contacts exist in the GF region for $NS1_{ZIKV\text{-}BR_h}$ but they are much more transient. Interestingly, residue F8 which was previously proposed as responsible for the membrane anchoring of the protein [20] is located close to the membrane for most of the time. ($OF_{PM_strain} > 0.9$).




**(a)** **(b)**

**Figure 6.** Protein-membrane interfaces for two $NS1_{ZIKV}$. (a,b) $\Delta OF_{PM}$ for chains A and B, respectively. The differences were calculated as $OF_{i\ PM_NS1_ZIKV\text{-}UG}$ - $OF_{i\ PM_NS1_ZIKV\text{-}BR}$. Relevant residues for $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ are shown in black and blue symbols, respectively. Point mutations appear like black triangles. The intertwined loop and the greasy fingers are indicated in the plots.

As for the PP interface, focusing on the *relevant* residues enables highlighting the differences between the two strains. Regardless of the OF threshold for PM interactions, the Ugandan strain has a considerably higher number of relevant residues that bind to the membrane when compared to the Brazilian one (*e.g.*, 20 compared to 5 for OF equals 0.8, Figure S5). Surprisingly, assuming an OF of roughly 0.8 the residues Y122, F123, and V124, which were proposed by experimental studies to contribute to membrane associations of Brazilian $NS1_{ZIKV\text{-}BR_h}$, were only found for $NS1_{ZIKV\text{-}UG_h}$.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The differences in the occupied fractions between the two strains were evaluated, considering a reasonable threshold for $\Delta OF_{PM}$ equal to 0.5. The results for $\Delta OF_{PM} > 0.5$ for chains A and B are given in Figures 6a and 6b, respectively. Interestingly, besides the membrane interacting regions that were previously suggested [*i.e.*, 1-29 ($Z1_{PM}$), 91-130 ($Z2_{PM}$, IL), 135-145 ($Z2_{PM}$), and 155-165 ($Z2_{PM}$)] [28], the present analysis brings into light additional sites able to bind to the membrane, namely residues 30-37, 70-75, 166, 168, 192, 201, 213-214.

Importantly, in the present study, we show that the PM interaction for all these binding sites is especially favored in the African strain compared to the Brazilian variant, for which the contact regions are limited to a part of the NTD region, one residue in IL (S121) and near to GF (T165). These relevant residues partly coincide with the regions of high fluctuation for the two $NS1_{ZIKV}$ (see Figures 2b and 3b). This tends to show that flexibility helps in optimizing PM interaction, as was suggested experimentally.

### 3.5. Specificity of lipid interactions with proteins residues in each strain:

In this section, we focus on the interplay between the relevant protein residues and the lipid molecules and seek possible specific interactions with some *types* of lipids. The membrane model was composed of a ratio of 18 (PC): 7 (PE): 2 (PI): 2 (PS): 1 (SM) in order to reproduce the endoplasmic reticulum membrane composition (host membrane) [33]. Accordingly, the density profile of the lipid chemical headgroups represented by the different CG beads was studied: PO4 and GL, common to most of the phospholipids; NC3 for PC and SM; NH3 for PE, INO for PI; CNO for PS. In SM, GL is replaced by AM (see Figure S6). The partial density of the lipid moieties (LM) along the *z*-axis (perpendicular to the membrane plane) shows the following order, INO, NC3, NH3, CNO, PO4, GL, and AM, from the outer leaflet to the membrane center. The distance between the two PO4 peaks is ~4.0 nm, which is compatible with the membrane thickness of a pure PC [50–52]. This indicates that the presence of the protein has no major effect on this bilayer property. Nevertheless, the impact of the protein may be attested as the peaks of PO4, NC3, and GL for instance, are not purely symmetrical (see Figure S6).

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

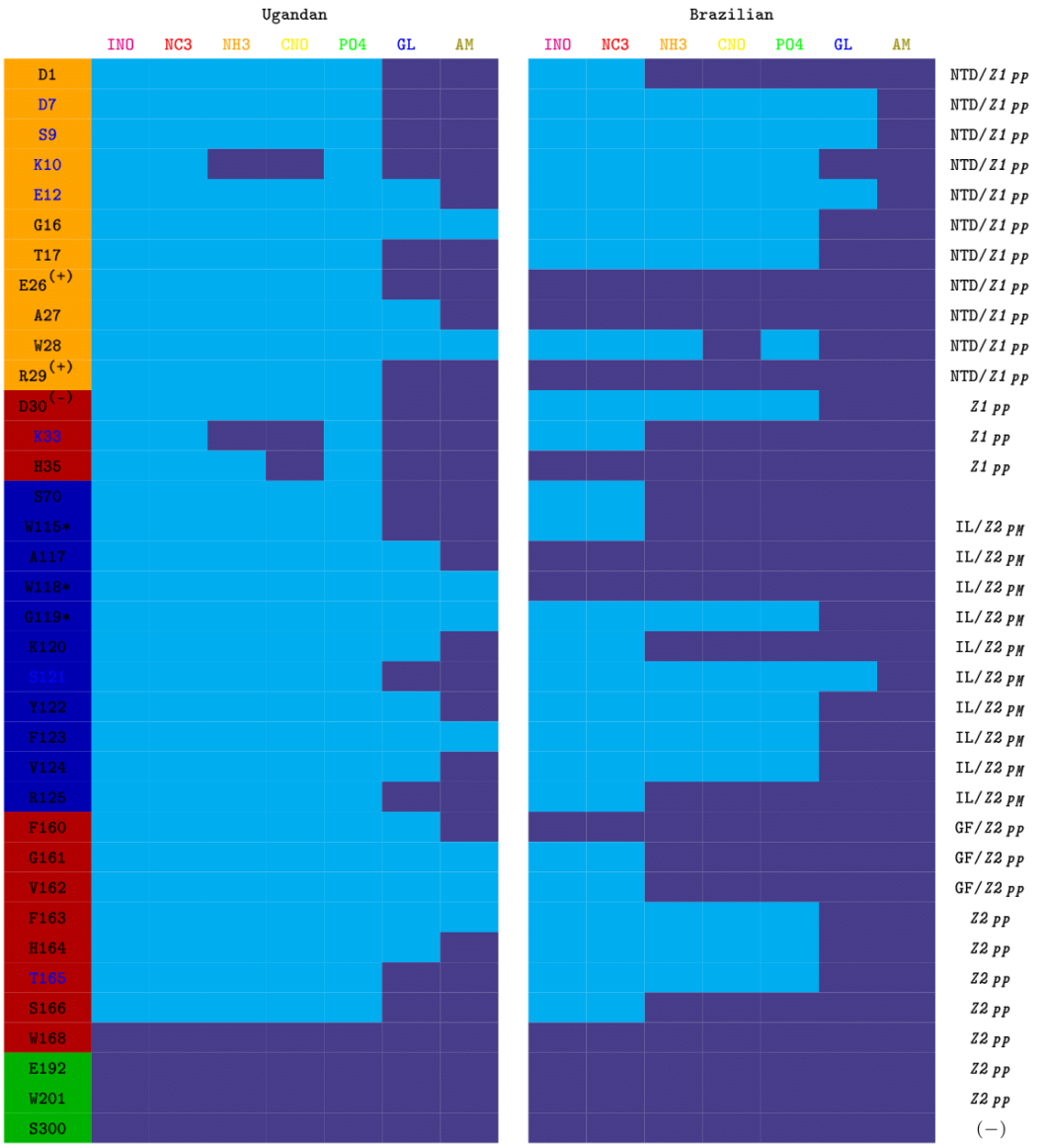



**Figure 7.** Summary of the position of each relevant residue estimated through $\Delta OF_{PM}$ analysis ($OF_{PM}$ = 0.5) and density calculations with respect to a particular lipid moiety. Note that amino acids are distinguished themselves as black or blue when they have relevance to $NS1_{ZIKV\text{-}UG_h}$ or $NS1_{ZIKV\text{-}BR_h}$ respectively. Titratable groups with symbols in parentheses indicate a positive (+) or negative (-) $\Delta pK_a$ [26]. Asterisks mark the WXXWG motif. The orange, red, blue, and green background colors in the first column are the regions β-roll, connector, wing, and β-ladder, respectively. When a residue matches the level of a specific lipid moiety is colored in background cyan. The last column shows important zone identifiers.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Moreover, we calculated the density profiles for the relevant residues defined by the $OF_{PM}$ for both strains (for details, see Figure S7). The preferred position of a residue relative to a specific LM was determined when its highest density peak coincides with the density of the LM. The results summarized in Figure 7 show that $NS1_{ZIKV\text{-}UG_h}$ has a higher number of residues matching with the lipid headgroups when compared to $NS1_{ZIKV\text{-}BR_h}$. These residues belong to the different regions detailed above, namely, the β-roll, the wing, and the connector. Furthermore, the present analysis clearly shows that residues of the African variant are more deeply embedded in the membrane. As an example, 16 and 8 residues are located at the level of GL and AM moieties (deepest functional groups) respectively for $NS1_{ZIKV\text{-}UG_h}$ compared to 4 and 0 for $NS1_{ZIKV\text{-}BR_h}$.

### 3.6. Role of hydrophobic and ionizable residues on the PM interface:

As expected, the depth of residues within the membrane is directly related to their physicochemical properties, that is, the more hydrophobic or aliphatic are more frequently located at the GL and AM levels (Figures 7 and 8). This is particularly true for $NS1_{ZIKV\text{-}UG_h}$, for which our results support the importance of the hydrophobic regions in NTD, IL, and GF in tune with experimental studies [22,24,25,46].

The interfacial zone, which covers the region from INO to PO4, is dominated by titratable groups (mostly positives) in the NTD/$ZI_{PP}$ and $Z1_{PP}$ regions (see the last column in Figure 7). Some of these residues coincide with high electrostatic coupling groups [26]. For instance, E26, and R29, which were identified with a positive $\Delta pK_a$ in our previous studies, have an impairment for membrane association in $NS1_{ZIKV\text{-}BR_h}$. In contrast, D30 with a negative $\Delta pK_a$ matches for both strains at the PO4 level. This means that the pH sensitivity would play a significant role in membrane association, even though further investigations are required to clearly decipher its impact.

Finally, we also examined the motif W115-X-X-W118-G119 (WXXWG) of the IL [24,25], which is considered a critical site for membrane interaction. In our previous work, the relevance of WXXWG in part to the groups with high electrostatic coupling like K116 in $NS1_{ZIKV}$ was

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

attributed [26]. Although $OF_{PM_strain}$ of K116 was below the threshold we choose ($OF_{PM_strain}$ < 0.5), the three conserved residues of WXXWG were highlighted for the African strain, while only the residue G119 was detected for the *holo* form of $NS1_{ZIKV\text{-}BR}$.

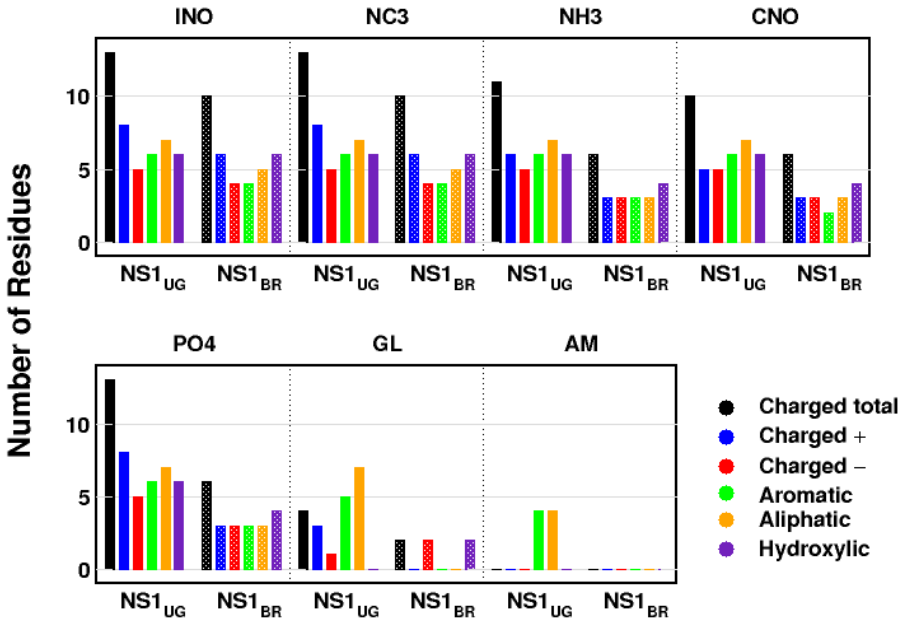


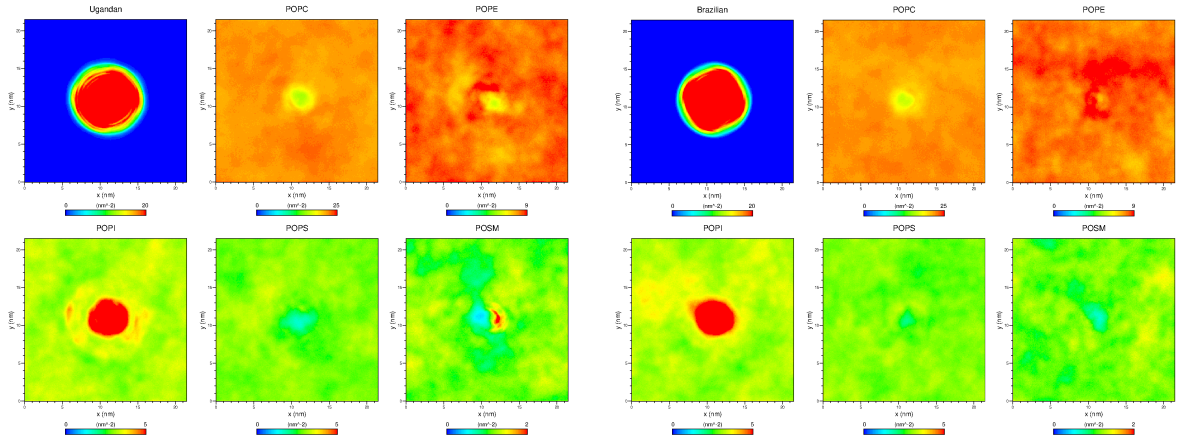

**Figure 8.** Barchart of the number of relevant residues per lipid headgroup for two $NS1_{ZIKV}$. Positively charged (R, H, K, Y), negatively charged (D, E), aromatic (F, W), aliphatic (A, G, V), and hydroxylic (S, T) residues are shown.

### 3.7. $NS1_{ZIKV}$ proteins have a special affinity for PI

To go further in deciphering the putative affinity of the two $NS1_{ZIKV}$ for a particular phospholipid, we examined the density of the CG beads for each protein and lipid (*i.e.*, PC, PE, PI, PS, and SM) in the XY plane (see Figure 9). Firstly, this measure indicates a slightly different shape of the two proteins in the plane. The distribution of the most abundant phospholipids (PC and PE) is relatively homogeneous in the simulation box for the two strains. In contrast, SM lipids seem to have closer proximity with the Ugandan strain compared to $NS1_{ZIKV\text{-}BR_h}$. The phospholipids PS and PI have the same proportions in the membrane (*i.e.*, 56 lipids per leaflet) but PI shows a more striking colocation compared to PS for the two *holo* $NS1_{ZIKV_h}$, in particular

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for $NS1_{ZIKV\text{-}UG_h}$. This feature may be related to the positively charged ionizable groups of the two proteins [53], which should contribute to a selective interaction of the anionic PI due to electrostatic affinity.




**(a)** **(b)**

**Figure 9.** Density maps along the XY plane for two NS1ZIKVs and their respective phospholipids. (a) Data for $NS1_{ZIKV\text{-}UG_h}$. (b) Data for $NS1_{ZIKV\text{-}BR_h}$. As colors become warmer, they indicate greater density. Regarding phospholipids, note that the scales vary according to concentration.

### 3.8. Biological discussion

The outcomes of the present study confirm that electrostatic interactions by their long-range nature contribute to stabilizing the dimer toward the membrane. Nevertheless, as frequently observed for peripheral membrane proteins, hydrophobic regions also participate in the anchoring of mNS1. Importantly, a recent study has shown that the cooperation of hydrophobic and positively charged residues at the PM interface are crucial to inducing curvature upon negatively charged ER membranes for the subsequent formation of the replication organelle [53]. Our results show that these mechanisms may subtly differ between ZIKV strains, due to a varied hydrophobic/charged PM interaction profile, and also specific lipid affinities. These results shed light upon the key regions with particular physical chemistry and dynamical nature (such as NTD, IL, and GF).

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ultimately, the hydrophobic face of the β-roll and wing domains assist the interaction between dimers to form hexameric sNS1 [46]. Both regions cooperate to attach lipid cargo giving greater structural stability in a hydrophobic core [20]. The lipid cargo is composed of an internal region that carries mostly triglycerides and cholesteryl esters, while externally there is a greater presence of phosphatidylcholine, phosphatidylethanolamine, and sphingomyelin. Our results show that aromatic residues have a high affinity for sites close to the aliphatic chains of the membrane. Furthermore, aromatic groups in conjunction with aliphatic residues also play an important role in the dimerization of monomers [29]. This means that the hydrophobic core has dual functionality contributing to the dimer-lipid and dimer-dimer association of sNS1.

## 4. CONCLUSIONS

We have examined how the structural and dynamical properties of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ are impacted by the presence of a lipid bilayer. A first analysis showed that the NS1 *holo* systems explore different configurational spaces when compared to *apo* forms. The discrepancies derive from particular coupling patterns that have direct consequences on the NS1 biological interfaces. Membrane binding sites such as NTD, IL, and GF are unlike for the two strains, impacting also the monomer-monomer interface. Despite the simplicity of the model that neglects different partners that might also contribute to NS1 functioning (*e.g.*, in association with sugars or GPI anchor, presence of triglycerides or other proteins in the membrane), the results detailed here allow capturing anchoring processes by aromatic residues (an essential step for PM assembly) to a model of ER lipid bilayer in conjunction with ionizable groups that also participate the PM binding. Both $NS1_{ZIKV}$ have a predilection for anionic phospholipids like PI, probably as a consequence of electrostatic coupling due to positively charged amino acids abundantly observed on PM interfaces. Furthermore, a particular cross-correlation dynamic pattern is exhibited for each strain as a result of a discrepant interaction with the membrane. Finally, our study definitively establishes that the Ugandan strain has a higher affinity for the membrane compared to the Brazilian variant and that the interaction with the membrane could trigger allosteric effects

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that impact the putative epitope regions. Overall, these results bring a new view on the differential virulence between the two strains that can originate from the different abilities of NS1 to interact with the membrane and their consequence on the recognition capacity of the immune system. As a perspective, the dynamical-electrostatic features of NS1 can be exploited for the efficient diagnosis of flaviviruses or drug development due to a low risk of ADE.

## 5. ACKNOWLEDGMENTS

This work has been supported in part by the "Fundação de Amparo à Pesquisa do Estado de São Paulo" [Fapesp 2015/16116-3 and 2020/07158-2 (F.L.B.d.S.)], the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [CNPq 305393/2020-0 (F.L.B.d.S.)], the Swedish National Infrastructure for Computing [SNIC 2020/14-38 and SNIC 2021/3-6 (F.L.B.d.S.)] and the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior"—CAPES (S.A.P.C.). F.L.B.d.S. and S.A.P.C. acknowledge the support of the University of São Paulo (USP) through the NAP-CatSinQ (Research Core in Catalysis and Chemical Synthesis) and the computing hours at HPC/USP. The authors also would like to thank CINES (Centre Informatique National de l'Enseignement Supérieur) for computational resources (project: A0060710878). F.L.B.d.S. and C.E. acknowledge the mobility support from the joint Université Sorbonne Paris Cité-USP Research program.

**Author Contributions**: S.A.P.C., F.L.B.d.S., and C.E., designed the research, performed simulations, analyzed data, and wrote the paper.

## 6. BIBLIOGRAPHY

[1] N.R. Faria, R. do S. da Silva Azevedo, M.U. Kraemer, R. Souza, M.S. Cunha, S.C. Hill, J. Thézé, M.B. Bonsall, T.A. Bowden, I. Rissanen, Zika virus in the Americas: early epidemiological and genetic findings, Science. (2016) aaf5036.

[2] Y. Simonin, D. van Riel, P. Van de Perre, B. Rockx, S. Salinas, Differential virulence between Asian and African lineages of Zika virus, PLoS Negl. Trop. Dis. 11 (2017) e0005821.

[3] M.A. Sheridan, D. Yunusov, V. Balaraman, A.P. Alexenko, S. Yabe, S. Verjovski-Almeida,

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

D.J. Schust, A.W. Franz, Y. Sadovsky, T. Ezashi, R. Roberts, Vulnerability of primitive human placental trophoblast to Zika virus, Proc. Natl. Acad. Sci. 114 (2017) E1587–E1596. https://doi.org/10.1073/pnas.1616097114.

[4] F. Aubry, S. Jacobs, M. Darmuzey, S. Lequime, L. Delang, A. Fontaine, N. Jupatanakul, E.F. Miot, S. Dabo, C. Manet, X. Montagutelli, A. Baidaliuk, F. Gámbaro, E. Simon-Lorière, M. Gilsoul, C.M. Romero-Vivas, V.-M. Cao-Lormeau, R.G. Jarman, C.T. Diagne, O. Faye, O. Faye, A.A. Sall, J. Neyts, L. Nguyen, S.J.F. Kaptein, L. Lambrechts, Recent African strains of Zika virus display higher transmissibility and fetal pathogenicity than Asian strains, Nat. Commun. 12 (2021) 916. https://doi.org/10.1038/s41467-021-21199-z.

[5] L. de Noronha, C. Zanluca, M.L.V. Azevedo, K.G. Luz, C.N.D. dos Santos, Zika virus damages the human placental barrier and presents marked fetal neurotropism, Mem. Inst. Oswaldo Cruz. 111 (2016) 287–293. https://doi.org/10.1590/0074-02760160085.

[6] K.A. Jurado, M.K. Simoni, Z. Tang, R. Uraki, J. Hwang, S. Householder, M. Wu, B.D. Lindenbach, V.M. Abrahams, S. Guller, others, Zika virus productively infects primary human placenta-specific macrophages, JCI Insight. 1 (2016).

[7] Q. Shao, S. Herrlinger, Y.-N. Zhu, M. Yang, F. Goodfellow, S.L. Stice, X.-P. Qi, M.A. Brindley, J.-F. Chen, The African Zika virus MR-766 is more virulent and causes more severe brain damage than current Asian lineage and dengue virus, Development. 144 (2017) 4114–4124.

[8] F.R. Cugola, I.R. Fernandes, F.B. Russo, B.C. Freitas, J.L.M. Dias, K.P. Guimarães, C. Benazzato, N. Almeida, G.C. Pignatari, S. Romero, C.M. Polonio, I. Cunha, C.L. Freitas, W.N. Brandão, C. Rossato, D.G. Andrade, D. de P. Faria, A.T. Garcez, C.A. Buchpigel, C.T. Braconi, E. Mendes, A.A. Sall, P.M. de A. Zanotto, J.P.S. Peron, A.R. Muotri, P.C.B. Beltrão-Braga, The Brazilian Zika virus strain causes birth defects in experimental models, Nature. 534 (2016) 267–271. https://doi.org/10.1038/nature18296.

[9] D.A. Muller, P.R. Young, The Flavivirus NS1 protein: molecular and structural biology, immunology, role in pathogenesis and application as a diagnostic biomarker, Antiviral Res. 98 (2013) 192–208. https://doi.org/10.1016/j.antiviral.2013.03.008.

[10] D. Watterson, N. Modhiran, P.R. Young, The many faces of the flavivirus NS1 protein offer a multitude of options for inhibitor design, Antiviral Res. 130 (2016) 7–18.

[11] J. Ye, B. Zhu, Z.F. Fu, H. Chen, S. Cao, Immune evasion strategies of flaviviruses, Vaccine. 31 (2013) 461–471.

[12] A. Li, J. Yu, M. Lu, Y. Ma, Z. Attia, C. Shan, M. Xue, X. Liang, K. Craig, N. Makadiya, J.J. He, R. Jennings, P.-Y. Shi, M.E. Peeples, S.-L. Liu, P.N. Boyaka, J. Li, A Zika virus vaccine expressing premembrane-envelope-NS1 polyprotein, Nat. Commun. 9 (2018) 3067. https://doi.org/10.1038/s41467-018-05276-4.

[13] S. Sridhar, A. Luedtke, E. Langevin, M. Zhu, M. Bonaparte, T. Machabert, S. Savarino, B. Zambrano, A. Moureau, A. Khromava, Z. Moodie, T. Westling, C. Mascareñas, C. Frago, M. Cortés, D. Chansinghakul, F. Noriega, A. Bouckenooghe, J. Chen, S.-P. Ng, P.B. Gilbert, S. Gurunathan, C.A. DiazGranados, Effect of Dengue Serostatus on Dengue Vaccine Safety and Efficacy, N. Engl. J. Med. 379 (2018) 327–340. https://doi.org/10.1056/NEJMoa1800820.

[14] A.L. St. John, A.P.S. Rathore, Adaptive immune responses to primary and secondary dengue virus infections, Nat. Rev. Immunol. 19 (2019) 218–230.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

https://doi.org/10.1038/s41577-019-0123-x.

[15] D.H. Libraty, P.R. Young, D. Pickering, T.P. Endy, S. Kalayanarooj, S. Green, D.W. Vaughn, A. Nisalak, F.A. Ennis, A.L. Rothman, High Circulating Levels of the Dengue Virus Nonstructural Protein NS1 Early in Dengue Illness Correlate with the Development of Dengue Hemorrhagic Fever, J. Infect. Dis. 186 (2002) 1165–1168. https://doi.org/10.1086/343813.

[16] G. Winkler, S.E. Maxwell, C. Ruemmler, V. Stollar, Newly synthesized Dengue–2 virus nonstructural protein NS1 is a soluble protein but becomes partially hydrophobic and membrane–associated after dimerization, Virology. 171 (1989) 302–305. https://doi.org/10.1016/0042-6822(89)90544-8.

[17] S. Noisakran, T. Dechtawewat, P. Avirutnan, T. Kinoshita, U. Siripanyaphinyo, C. Puttikhunt, W. Kasinrerk, P. Malasit, N. Sittisombut, Association of dengue virus NS1 protein with lipid rafts, J. Gen. Virol. 89 (2008) 2492–2500.

[18] S. Youn, T. Li, B.T. McCune, M.A. Edeling, D.H. Fremont, I.M. Cristea, M.S. Diamond, Evidence for a genetic and physical interaction between nonstructural proteins NS1 and NS4B that modulates replication of West Nile virus, J. Virol. 86 (2012) 7360–7371. https://doi.org/10.1128/JVI.00157-12.

[19] P. Scaturro, M. Cortese, L. Chatel-Chaix, W. Fischl, R. Bartenschlager, Dengue Virus Non-structural Protein 1 Modulates Infectious Particle Production via Interaction with the Structural Proteins, PLOS Pathog. 11 (2015) e1005277. https://doi.org/10.1371/journal.ppat.1005277.

[20] I. Gutsche, F. Coulibaly, J.E. Voss, J. Salmon, J. d'Alayer, M. Ermonval, E. Larquet, P. Charneau, T. Krey, F. Mégret, Secreted dengue virus nonstructural protein NS1 is an atypical barrel-shaped high-density lipoprotein, Proc. Natl. Acad. Sci. 108 (2011) 8003–8008.

[21] D.A. Muller, M.J. Landsberg, C. Bletchly, R. Rothnagel, L. Waddington, B. Hankamer, P.R. Young, Structure of the dengue virus glycoprotein non-structural protein 1 by electron microscopy and single-particle analysis, J. Gen. Virol. 93 (2012) 771–779. https://doi.org/10.1099/vir.0.039321-0.

[22] Y. Ci, Z.-Y. Liu, N.-N. Zhang, Y. Niu, Y. Yang, C. Xu, W. Yang, C.-F. Qin, L. Shi, Zika NS1–induced ER remodeling is essential for viral replication, J. Cell Biol. 219 (2020) e201903062. https://doi.org/10.1083/jcb.201903062.

[23] H. Song, J. Qi, J. Haywood, Y. Shi, G.F. Gao, Zika virus NS1 structure reveals diversity of electrostatic surfaces among flaviviruses, Nat. Struct. Mol. Biol. 23 (2016) 456–458. https://doi.org/10.1038/nsmb.3213.

[24] W.C. Brown, D.L. Akey, J.R. Konwerski, J.T. Tarrasch, G. Skiniotis, R.J. Kuhn, J.L. Smith, Extended surface for membrane association in Zika virus NS1 structure, Nat. Struct. Mol. Biol. 23 (2016) 865–867. https://doi.org/10.1038/nsmb.3268.

[25] X. Xu, H. Song, J. Qi, Y. Liu, H. Wang, C. Su, Y. Shi, G.F. Gao, Contribution of intertwined loop to membrane association revealed by Zika virus full–length NS1 structure, EMBO J. 35 (2016) 2170–2178. https://doi.org/10.15252/embj.201695290.

[26] S.A. Poveda-Cuevas, C. Etchebest, F.L. Barroso da Silva, Insights into the ZIKV NS1 Virology from Different Strains through a Fine Analysis of Physicochemical Properties, ACS Omega. 3 (2018) 16212–16229.

[27] S.A. Poveda-Cuevas, C. Etchebest, F.L. Barroso da Silva, Identification of Electrostatic

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Epitopes in Flavivirus by Computer Simulations: The PROCEEDpKa Method, J. Chem. Inf. Model. 60 (2020) 944–963.

[28] S.A. Poveda-Cuevas, F.L. Barroso da Silva, C. Etchebest, How the Strain Origin of Zika Virus NS1 Protein Impacts Its Dynamics and Implications to Their Differential Virulence, J. Chem. Inf. Model. 61 (2021) 1516–1530. https://doi.org/10.1021/acs.jcim.0c01377.

[29] R.L. Gonçalves, G. de Lima Menezes, L. Sussuchi, M.L. Moreli, M. Mottin, C.H. Andrade, M. Pereira, R.A. da Silva, Dynamic behavior of Dengue and Zika viruses NS1 protein reveals monomer–monomer interaction mechanisms and insights to rational drug design, J. Biomol. Struct. Dyn. (2019) 1–11.

[30] V. Monje-Galvan, G.A. Voth, Molecular interactions of the M and E integral membrane proteins of SARS-CoV-2, Faraday Discuss. 232 (2021) 49–67. https://doi.org/10.1039/D1FD00031D.

[31] S. Jo, T. Kim, V.G. Iyer, W. Im, CHARMM-GUI: a web-based graphical user interface for CHARMM, J. Comput. Chem. 29 (2008) 1859–1865.

[32] C. Globisch, V. Krishnamani, M. Deserno, C. Peter, Optimization of an Elastic Network Augmented Coarse Grained Model to Study CCMV Capsid Deformation, PLoS ONE. 8 (2013) e60582. https://doi.org/10.1371/journal.pone.0060582.

[33] G. van Meer, D.R. Voelker, G.W. Feigenson, Membrane lipids: where they are and how they behave, Nat. Rev. Mol. Cell Biol. 9 (2008) 112–124. https://doi.org/10.1038/nrm2330.

[34] M.A. Lomize, I.D. Pogozheva, H. Joo, H.I. Mosberg, A.L. Lomize, OPM database and PPM web server: resources for positioning of proteins in membranes, Nucleic Acids Res. 40 (2012) D370–D376. https://doi.org/10.1093/nar/gkr703.

[35] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX. 1–2 (2015) 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

[36] X. Periole, M. Cavalli, S.-J. Marrink, M.A. Ceruso, Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition, J. Chem. Theory Comput. 5 (2009) 2531–2543.

[37] C. Arnarez, J.J. Uusitalo, M.F. Masman, H.I. Ingólfsson, D.H. de Jong, M.N. Melo, X. Periole, A.H. de Vries, S.J. Marrink, Dry Martini, a Coarse-Grained Force Field for Lipid Membrane Simulations with Implicit Solvent, J. Chem. Theory Comput. 11 (2015) 260–275. https://doi.org/10.1021/ct500477k.

[38] D.H. de Jong, G. Singh, W.F.D. Bennett, C. Arnarez, T.A. Wassenaar, L.V. Schäfer, X. Periole, D.P. Tieleman, S.J. Marrink, Improved Parameters for the Martini Coarse-Grained Protein Force Field, J. Chem. Theory Comput. 9 (2013) 687–697. https://doi.org/10.1021/ct300646g.

[39] M. Aghaaminiha, A.M. Farnoud, S. Sharma, Quantitative relationship between cholesterol distribution and ordering of lipids in asymmetric lipid bilayers, Soft Matter. 17 (2021) 2742–2752. https://doi.org/10.1039/D0SM01709D.

[40] B.J. Grant, A.P. Rodrigues, K.M. ElSawy, J.A. McCammon, L.S. Caves, Bio3d: an R package for the comparative analysis of protein structures, Bioinformatics. 22 (2006) 2695–2696.

[41] M.A. Edeling, M.S. Diamond, D.H. Fremont, Structural basis of Flavivirus NS1 assembly and antibody recognition, Proc. Natl. Acad. Sci. 111 (2014) 4285–4290. https://doi.org/10.1073/pnas.1322036111.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487413; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[42] S.B. Biering, D.L. Akey, M.P. Wong, W.C. Brown, N.T.N. Lo, H. Puerta-Guardo, F. Tramontini Gomes de Sousa, C. Wang, J.R. Konwerski, D.A. Espinosa, N.J. Bockhaus, D.R. Glasner, J. Li, S.F. Blanc, E.Y. Juan, S.J. Elledge, M.J. Mina, P.R. Beatty, J.L. Smith, E. Harris, Structural basis for antibody inhibition of flavivirus NS1–triggered endothelial dysfunction, Science. 371 (2021) 194–200. https://doi.org/10.1126/science.abc0476.

[43] L. Priyamvada, K.M. Quicke, W.H. Hudson, N. Onlamoon, J. Sewatanon, S. Edupuganti, K. Pattanapanyasat, K. Chokephaibulkit, M.J. Mulligan, P.C. Wilson, others, Human antibody responses after dengue virus infection are highly cross-reactive to Zika virus, Proc. Natl. Acad. Sci. 113 (2016) 7852–7857.

[44] W. Dejnirattisai, P. Supasa, W. Wongwiwat, A. Rouvinski, G. Barba-Spaeth, T. Duangchinda, A. Sakuntabhai, V.-M. Cao-Lormeau, P. Malasit, F.A. Rey, others, Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus, Nat. Immunol. 17 (2016) 1102.

[45] K. Stettler, M. Beltramello, D.A. Espinosa, V. Graham, A. Cassotta, S. Bianchi, F. Vanzetta, A. Minola, S. Jaconi, F. Mele, others, Specificity, cross-reactivity, and function of antibodies elicited by Zika virus infection, Science. 353 (2016) 823–826.

[46] D.L. Akey, W.C. Brown, S. Dutta, J. Konwerski, J. Jose, T.J. Jurkiw, J. DelProposto, C.M. Ogata, G. Skiniotis, R.J. Kuhn, J.L. Smith, Flavivirus NS1 structures reveal surfaces for associations with membranes and the immune system, Science. 343 (2014) 881–885.

[47] C.C. David, D.J. Jacobs, Principal component analysis: a method for determining the essential dynamics of proteins, in: Protein Dyn., Springer, 2014: pp. 193–226.

[48] L. Skjaerven, A. Martinez, N. Reuter, Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit: Comparison of PCA and NMA for Proteins, Proteins Struct. Funct. Bioinforma. 79 (2011) 232–243. https://doi.org/10.1002/prot.22875.

[49] G.G. Maisuradze, A. Liwo, H.A. Scheraga, Principal component analysis for protein folding dynamics, J. Mol. Biol. 385 (2009) 312–329.

[50] N. Kučerka, M.-P. Nieh, J. Katsaras, Fluid phase lipid areas and bilayer thicknesses of commonly used phosphatidylcholines as a function of temperature, Biochim. Biophys. Acta BBA - Biomembr. 1808 (2011) 2761–2771. https://doi.org/10.1016/j.bbamem.2011.07.022.

[51] R.W. Benz, F. Castro-Román, D.J. Tobias, S.H. White, Experimental Validation of Molecular Dynamics Simulations of Lipid Bilayers: A New Approach, Biophys. J. 88 (2005) 805–817. https://doi.org/10.1529/biophysj.104.046821.

[52] G. Shahane, W. Ding, M. Palaiokostas, M. Orsi, Physical properties of model biological lipid bilayers: insights from all-atom molecular dynamics simulations, J. Mol. Model. 25 (2019) 76. https://doi.org/10.1007/s00894-019-3964-0.

[53] Y. Ci, Y. Yang, C. Xu, C.-F. Qin, L. Shi, Electrostatic Interaction Between NS1 and Negatively Charged Lipids Contributes to Flavivirus Replication Organelles Formation, Front. Microbiol. 12 (2021) 641059. https://doi.org/10.3389/fmicb.2021.641059.

## Chapter 10

# General Conclusions

The flavivirus NS1 protein is an ideal target because of a broad range of functions essential for the pathogenicity, virulence, and survival of viruses such as ZIKV. $NS1_{ZIKV}$ from Uganda and Brazil exhibit a high degree of structural conservation (in terms of the scaffold) and sequence identity (97%). Through a computational approach, several physicochemical and dynamic-structural properties were evaluated. The main biological interfaces of NS1 examined are related to the *i)* monomer-monomer or dimer-dimer interactions, *ii)* Ab recognition parts, *iii)* and lipid membrane binding.

Point mutations of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ impact the electrostatic features, especially the protein net charge ($Z_P$) and, therefore, the p$K_a$'s. When applying the $\Delta$p$K_a$ analysis, the protein regions that are highlighted are located in β-roll, connector (specifically GF), and wing (specifically IL). Interestingly, these regions also have a high degree of flexibility which may be related to a greater coupling between charged residues and the aqueous environment. However, an analysis of the more global movements shows that there is coupling between specific parts of these proteins that influence their whole dynamics and the PSN.

The β-ladder domain (C-terminal) presents low fluctuation compared to the remaining parts of the 3D structure (*i.e.* N-terminal). This structural feature may be conditioning these systems for better recognition by Ab in the C-terminal part of NS1. This is because less flexibility could reduce the range of structural arrangements that NS1 can have. Nevertheless, the specificity for an Ab can be due to the electrostatic properties that a specific antigen like NS1 has. A detailed analysis of $\Delta$p$K_a$ from the β-ladder domain by means of the PROCEEDp$K_a$ method shows that the titratable residues (*i.e.* the so-called EE's) of $NS1_{ZIKV\text{-}UG}$ have greater interaction with the $Ab_{WNV}$ when compared to $NS1_{ZIKV\text{-}BR}$.

In terms of the dimerizations and tetramerizations of NS1 for various flaviviruses, we found that attractive free energies prevail under physiological conditions of pH and salt. For this case, the contributions that drive the association are mostly van der Waals forces. In acidic pH regimes, the systems are repulsive because Coulombic interactions prevail. This is true for both taxonomic conditions (*i.e.* at the interspecific and intraspecific levels). On the other hand, the $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ proteins show discrepant protein-

protein interaction profiles (especially for dimer-dimer interactions). These differences are presumably given by the IL regions (zones where there are coincidentally greater differences in terms of RMSF), which produce more steric hindrances in the African NS1 when compared to the Asian one.

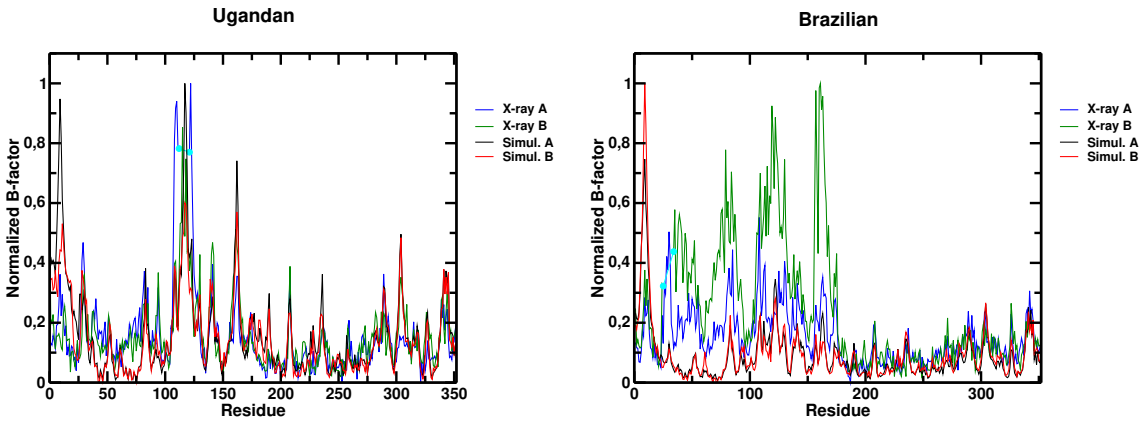
The $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ proteins were evaluated in the presence of a membrane. We found that the regions with high flexibility are involved in the anchoring to the lipid membrane. The $NS1_{ZIKV\text{-}UG}$ has a higher affinity for the membrane compared to $NS1_{ZIKV\text{-}BR}$ since the residues (mostly aromatic) embed more towards the aliphatic chains of the membrane model.

In general, our results prove that $NS1_{ZIKV\text{-}UG}$ interacts differently when compared to $NS1_{ZIKV\text{-}BR}$, despite a high identity protein sequence. This also shows an important insights into the differential virulence expressed by African and Asian strains, giving new perspectives at the molecular level.

## Appendix A

# Comparison between experimental and simulation NS1 data

From structures of $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ solved by X-ray assays, we can extract the thermal B-factor. This thermodynamic variable allows us to measure the motion of a certain part of a particular protein. Generally, a high $B-factor$ indicates high thermal fluctuation. The $B-factor$ is displayed as a function of the residue position for each $NS1_{ZIKV}$ crystal in Figure A.1.




(a) (b)

Figure A.1: $B-factor$ as a function of the residue position for two $NS1_{ZIKV}$ from experimental and simulated data. The $B-factor$ for each chain is visualized. Note that the missing regions in the crystallographic structures have been highlighted in cyan.

It is important to mention that each structure was solved at different resolutions (1.89 and 2.85 Å for $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$, respectively), so the B-factor had to be normalized. When it is observed the curves of the crystal $NS1_{ZIKV\text{-}UG}$, we identify that the chains present different fluctuation profiles among them. Even chain A has missing residues in loop regions (wing domain), while chain B is complete. Missing regions are a consequence of high thermal fluctuations that prevent them from being detected in this type of X-ray diffraction experiment. Moreover, it is worth mentioning that the same happens for $NS1_{ZIKV\text{-}BR}$ (for this case chain A is complete, while chain B has some missing

amino acids in the $\beta$-roll and connector regions). The differences found between the chains of each $NS1_{ZIKV}$ are because of the intrinsic structural stochasticity of the systems.

The RMSF profiles for the CG models of two $NS1_{ZIKV}$ obtained in Chapter 8 were converted to $B - factor$ (see black and red curves in Figure A.1). Our CG model reproduces several peaks that are found in the crystallographic structures. However, the N-terminal of $NS1_{ZIKV\text{-}BR}$ part was less reproducible (especially for chain B). We suppose that it is a consequence of the low resolution for the $NS1_{ZIKV\text{-}BR}$ crystal (*i.e.* 2.85 Å). As it was argued in Chapters 8 and 9 the regions with the greatest flexibility are directly related to key interface regions for lipid membrane binding.

## Appendix B

# Comparison between glycosylated and non-glycosylated proteins

The glycosylation is essential for the structural stability of NS1 (Somnuke et al., 2011; Watterson et al., 2016; Yap et al., 2017). Understanding the effects that sugar molecules produced in these systems is an important subject of study. Initially, we started from protein models assuming a dimeric NS1 with complex and high mannose oligosaccharides presented in N130 and N207, respectively (Figures 2.2a and B.1). This should correspond to a mature NS1 after undergoing a sugar editing process in the Golgi apparatus (see Subsection 2.3.2). The glycosylated proteins were compared with structures without sugar derived from Chapter 8. The AA models were simulated until reaching 200 ns and 3 replicates were done per system. We calculated the ED through PCA, in order to assess the impact produced by the glycans in $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$. The trajectories for all replicates and strains were pooling together. The final trajectory consisted of 23000 frames and the coordinates of the $C_\alpha$'s were employed for the analysis.

The first three PC's capture 55.9% ($PC_1 = 34.8\%$, $PC_2 = 12.2\%$, and $PC_3 = 8.9\%$) of the total variance (Figure B.2). These values are enough to characterize the large-scale protein motions for the $NS1_{ZIKV}$ systems and, therefore, we will restrict our discussion to these PCs. Our results show that $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ in the absence of sugar molecules differ significantly from each other. This is consistent with what was previously found by us (see Chapter 8). When comparing the two $NS1_{ZIKV}$ in the presence of glycans, it is reached the same conclusions. Surprisingly, we clearly observe that two clusters are formed (representing the essential modes for $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$) regardless of whether or not the proteins are glycosylated. An analysis of the eigenvector components for the first three PC's (data not shown) shows that the structural discrepancies between the two strains coincide with the regions of greatest fluctuation reported by the B-factor profiles (Figure A.1).

Due to both $NS1_{ZIKV\text{-}UG}$ or $NS1_{ZIKV\text{-}BR}$ have equal ED's at the presence or absence of oligosaccharides, these outcomes give great support to the previously performed analyzes (see Chapters 8 and 9). The presence of oligosaccharides should be a more relevant topic when studying complex phenomena such as NS1-Ab interactions or stability of NS1

hexamers (Somnuke et al., 2011).

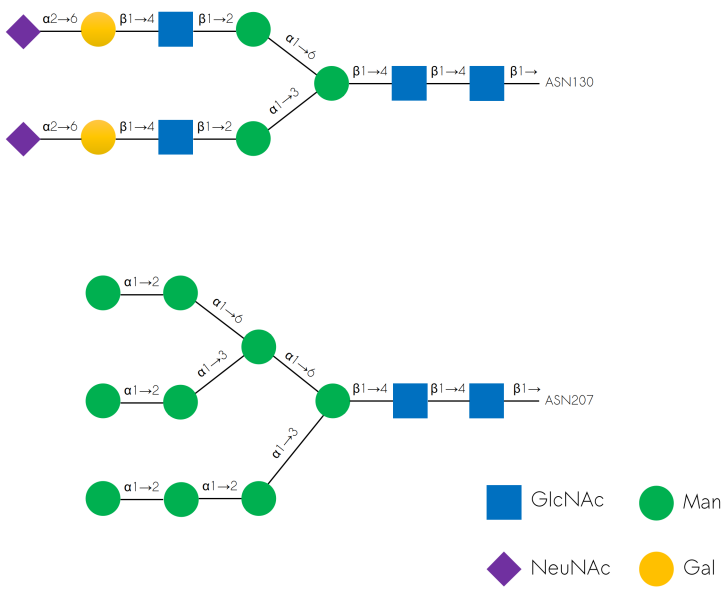



Figure B.1: Schematic representation of the complex- (top) and high-mannose-type (bellow) oligosaccharides used to construct NS1 glycosylated forms. N-acetyl-D-glucosamine, D-mannose, N-acetyl-D-neuraminic acid, and D-galactose are represented as blue squares, green circles, violet diamonds, and yellow circles, respectively.

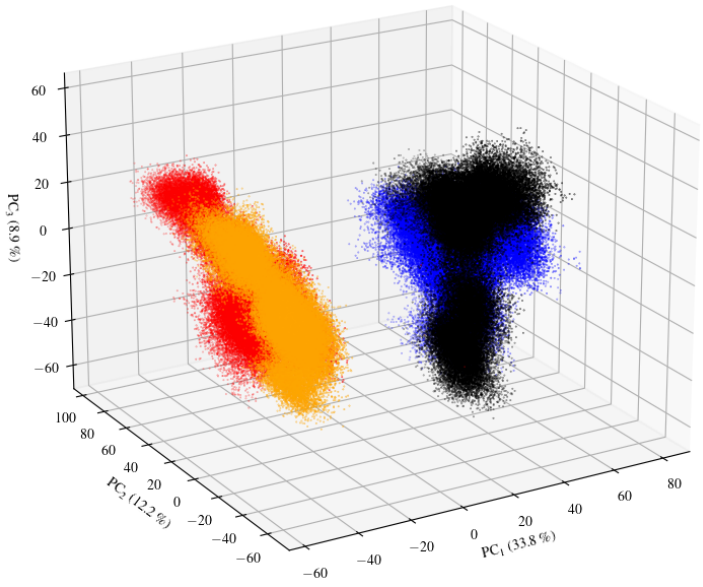



Figure B.2: Principal component analysis for two $NS1_{ZIKV}$ at the presence and absence of sugar molecules. $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ without sugar are represented with black and red scatter points, respectively. $NS1_{ZIKV\text{-}UG}$ and $NS1_{ZIKV\text{-}BR}$ glycosylated are represented with blue and orange scatter points, respectively.

## Appendix C

# CPRISMA (Coloring PRoteins by Inputs and Multiple Alignments)

The present appendix is showing the user guide of CPRISMA (Coloring PRoteins by Inputs and Multiple Alignments) 1.0 software, published in 2021 and with free access in the GitHub repository (https://github.com/seapovedac/cprisma). This program was implemented to build some protein sequence alignments for Chapters 5, 6, 7, and 8.



# CPRISMA

## Coloring Proteins By Inputs and Sets of Multiple Alignments

# CPrisma User's Guide

Version 1.0

2021

Contributions from

Sergio Alejandro Poveda Cuevas

seapovedac@gmail.com

Laboratory of Computational Biophysical Chemistry

https://barrosolab.fcfrp.usp.br/

University of São Paulo

Ribeirão Preto - SP

Brazil

## DESCRIPTION

CPRISMA is a bioinformatics program that gives color to multiple sequence alignment based on an input of numerical data.

## LICENSE

CPRISMA is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License, version 3 (GPLv3).



# Contents


**1 Introduction** **1**

**2 Setup** **3**

2.1 Installation/unistallation . . . 3
2.2 Inputs . . . 3
2.3 Outputs . . . 7

**3 Basic commands** **8**

3.1 Target-residues (`-tr`) . . . 11
3.2 Accuracy (`-a`) . . . 12
3.3 Alignment visualization (`-va`) . . . 13
3.4 Join (`-j`) . . . 15
3.5 Numbering (`-n`) . . . 15
3.6 Number of residues per line (`-l`) . . . 15
3.7 Name sequence (`-ns`) . . . 16
3.8 Hide conservation line (`-hc`) . . . 17
3.9 Check reference method (`-ck`) . . . 17
3.10 Reference method (`-rf`) . . . 19
3.10.1 Default . . . 20
3.10.2 Pair . . . 20
3.10.3 Multiple . . . 22
3.11 Total information (`-t`) . . . 25

**4 Operation (`descriptor_ope`)** **26**

4.1 None (`n`) . . . 26
4.2 Delta (`d`) . . . 27
4.3 Absolute delta (`da`) . . . 27
4.4 Multiply (`m`) . . . 28
4.5 Operation descriptors dictionary (`-dop`) . . . 28




**5 Color (descriptor_col) 31**

5.1 No color (nc) . . . 32
5.2 Same sequence color (ssc) . . . 32
5.3 Free sequence color (fsc) . . . 33
5.4 Free amino acid color (fac) . . . 34
5.5 Free amino acid and mutation color (fmac) . . . 35
5.6 Color by position index (pic) . . . 36
5.7 Color by position index and mutation (pimc) . . . 38
5.8 Color by threshold (tc) . . . 39
5.9 Color by threshold and mutation (tmc) . . . 40
5.10 p$K_a$ color (pkac) . . . 41
5.11 Color descriptors dictionary (-dco) . . . 42
5.12 Color sequence (-sco) . . . 44
5.13 Color mutation (-mco) . . . 45
5.14 Color intensity (-ico) . . . 45
5.15 Color on three-dimensional structure (-tco) . . . 45
5.16 List of colors (-lco) . . . 46

**6 Visualization (descriptor_vis) 48**

6.1 Visualization descriptors dictionary (-dvi) . . . 51

**7 Maximum restriction (descriptor_mxr) 54**

7.1 No restriction (nr) . . . 55
7.2 Restriction by mutation (rm) . . . 56
7.3 Restriction by amino acid (ra) . . . 57
7.4 Restriction by amino acid and mutation (ram) . . . 58
7.5 Restriction by index position (rpi) . . . 59
7.6 Restriction by threshold (rt) . . . 59
7.7 Restriction by sequence (rs) . . . 60
7.8 Restriction by sequence and mutation (rsm) . . . 61
7.9 Restriction by sequence and amino acid (rsa) . . . 62
7.10 Restriction by sequence, amino acid, and mutation (rsam) . . . 62
7.11 Restriction by sequence and index position (rspi) . . . 63
7.12 Restriction by sequence, index position, and mutation (rspim) . . . 64
7.13 Restrictions by target-residues (Y) . . . 64



7.14 Maximum restriction descriptors dictionary (`-dmx`) . . . . . 67

**8 Practical Examples** **69**

8.1 Example 1: $\Delta pK_a$ for several $NS1_{ZIKV}$ . . . . . 69

8.2 Example 2: B-cell epitope predictions on $NS1_{WNV(176-352)}$ . . . . . 71

8.3 Example 3: Protein structural domains for two $NS1_{ZIKV}$ and its biological interfaces . . . . . 74

**Bibliography** **77**



# 1 Introduction

CPRISMA (**C**oloring **PR**oteins by **I**nputs and **S**ets of **M**ultiple **A**lignments) is a command-line program designed to give color to protein multiple sequence alignments. CPRISMA was written in Python language and runs on Unix systems. Since we often want to compare sets of numerical protein variables at the primary structural level, CPRISMA is an ideal program to visualize and/or detect possible differences in specific amino acids or protein regions through a color palette. CPRISMA is a versatile tool that allows the user to incorporate the features that are more convenient to highlight.

The development of this program was inspired by previous studies where we sought to highlight regions for the non-structural (NS) protein 1 from several *Flavivirus* with particular physical-chemistry or structural behavior [1, 2, 3]. We thought it would be interesting to have a capable tool to make relationships between numerical information and sequence alignments, but for any external data set.

In this documentation, we have divided the Chapters following a logical order and using practical examples to understand each CPRISMA parameter. In Chapter 2, the installation/uninstallation and the type of input and output (I/O) files needed to run a simulation in CPRISMA will be treated. Chapter 3 will describe the basic functions, and Chapters 4, 5, 6, and 7 more complex variables. These six Chapters, we used as examples the NS1 proteins from Zika virus (ZIKV), Dengue variant 2 (DENV2), and West Nile viruses (WNV). Regarding ZIKV, we employ two strains from Uganda (UG) and Brazil (BR). Finally, in Chapter 8, we show three practical examples that correlate directly with three publications where we explored *i)* p$K_a$ shifts for several NS1$_{\text{ZIKV}}$ [1], *ii)* B-cell epitope prediction for NS1$_{\text{WNV}(176-352)}$ [2], and *iii)* the structural domains and biological interfaces for two ZIKV NS1 proteins [3].

Command lines, alert messages, or variables will be distinguished with the `Typewriter` font. Also, when a particular word/letter is preceded by a hyphen (-) in any title of this documentation it should be interpreted as a line command, otherwise, it should be interpreted



as a variable.

The current developer of CPRISMA is Sergio Alejandro Poveda Cuevas at the Laboratory of Computational Biophysical Chemistry (University of São Paulo). If you have any suggestions, bug reports, or general comments about CPRISMA, please contact us at seapovedac@gmail.com.



# 2 Setup

## 2.1 Installation/unistallation

To install CPRISMA you must directly download the code from the GitHub repositories https://github.com/seapovedac/cprisma/ or by executing the following command:

```
git clone https://github.com/seapovedac/cprisma.git
```

Then, you can invoke in that directory:

```
pip install .
```

Now your program should be in the OS environment variables. Typing in the terminal:

`cprisma`[1]

or

```
python -m cprisma
```

should return an error, requesting one of the input files for execution. If you want to be sure that there are no problems with the program, you can try to generate outputs from the "examples" directory.

To uninstall the program you can simply do it through the following command:

```
pip uninstall cprisma
```

## 2.2 Inputs

To run CPRISMA, a basic knowledge of Python is required, so a good understanding of the data types that this language uses (*e.g.* sets, tuples, lists, etc.) is essential to achieve the

[1] We will use this command throughout the documentation.



desired results[2].

Hereafter, we will use as an example a data set of p$K_a$'s[3] for four NS1 from ZIKV (two different strains), DENV2, and WNV. Protein sequences were extracted from the NCBI (National Center for Biotechnology Information)[4], and have the next accession codes: AY632535 (ZIKV-UG), KU365777 (ZIKV-BR), PR159-S1 (DENV2), and Q9Q6P4 (WNV). We will only take into account incomplete sequences (from residues 1 to 120) and the focus will be given to the titratable residues: Asp, Glu, Arg, His, Lys, and Tyr. Cys residues are not included, due to the fact that they are part of disulfide bridges.

CPRISMA is a program that requires at least three file kinds in the run directory[5]:

1. The alignment to put the color called "alignment.dat".
2. A CSV file called "data_input.csv".
3. A Python script called "array_get.py".

```
ZIKV-UG         DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSV
ZIKV-BR         DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSV
DENV2           DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSV
WNV             DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSV
                * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **

ZIKV-UG         SRMENIMWKSVEGELNAILEENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
ZIKV-BR         SRMENIMWRSVEGELNAILEENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
DENV2           TRLENLMWKQITPELNHILSENEVKLTIMTGDIKGIMQAGKRSLRPQPTELKYSWKTWGK
WNV             SRLEHQMWEAVKDELNTLLKENGVDLSVVVEKQEGMYKSAPKRLTATTEKLEIGWKAWGK
                :*:*: **  :  *** :*.** *.*:::. . :.  . . . *     :*  .**:***
```

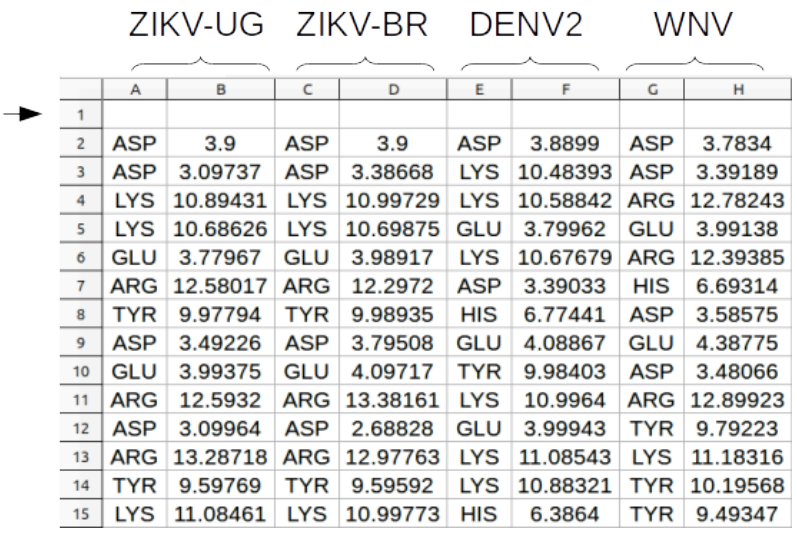
Figure 2.2.1: MUSCLE output in ClustalW format.

In addition to these files, others directories will appear (*e.g.*, "cprisma", "examples", so on), which contain all the scripts and other stuff necessary for the program. For now, let us focus on the three files mentioned above. To get the alignment data, you can directly go to the website program called MUSCLE (MUltiple Sequence Comparison by Log-Expectation)[6] and run it with the FASTA format protein sequences of your interest. You can modify the input parameters as wanted, nevertheless, it is *essential* that you keep ClustalW as the output format. You can also generate these multiple sequence alignments by invoking MUSCLE

[2]See also the official Python documentation: https://docs.python.org/3.7/.
[3]p$K_a$'s were obtained through charges of ionizable residues calculated by Monte Carlo simulations [4, 5].
[4]See https://www.ncbi.nlm.nih.gov/.
[5]These files can be found in the directory "examples/Example_0".
[6]See https://www.ebi.ac.uk/Tools/msa/muscle/.

through Biopython[7]. A typical MUSCLE result appears in Figure 2.2.1. All the details about this type of output is previously described [6]. An improper alignment format can produce the following output:

```
The program stop!  Improper alignment format.  Check ClustalW format at
https://www.ebi.ac.uk/Tools/msa/muscle.
```




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | ASP | 3.9 | ASP | 3.9 | ASP | 3.8899 | ASP | 3.7834 |
| 3 | ASP | 3.09737 | ASP | 3.38668 | LYS | 10.48393 | ASP | 3.39189 |
| 4 | LYS | 10.89431 | LYS | 10.99729 | LYS | 10.58842 | ARG | 12.78243 |
| 5 | LYS | 10.68626 | LYS | 10.69875 | GLU | 3.79962 | GLU | 3.99138 |
| 6 | GLU | 3.77967 | GLU | 3.98917 | LYS | 10.67679 | ARG | 12.39385 |
| 7 | ARG | 12.58017 | ARG | 12.2972 | ASP | 3.39033 | HIS | 6.69314 |
| 8 | TYR | 9.97794 | TYR | 9.98935 | HIS | 6.77441 | ASP | 3.58575 |
| 9 | ASP | 3.49226 | ASP | 3.79508 | GLU | 4.08867 | GLU | 4.38775 |
| 10 | GLU | 3.99375 | GLU | 4.09717 | TYR | 9.98403 | ASP | 3.48066 |
| 11 | ARG | 12.5932 | ARG | 13.38161 | LYS | 10.9964 | ARG | 12.89923 |
| 12 | ASP | 3.09964 | ASP | 2.68828 | GLU | 3.99943 | TYR | 9.79223 |
| 13 | ARG | 13.28718 | ARG | 12.97763 | LYS | 11.08543 | LYS | 11.18316 |
| 14 | TYR | 9.59769 | TYR | 9.59592 | LYS | 10.88321 | TYR | 10.19568 |
| 15 | LYS | 11.08461 | LYS | 10.99773 | HIS | 6.3864 | TYR | 9.49347 |

Figure 2.2.2: Data input with p$K_a$ values for four proteins. All the information has not been shown.

Our second type of file with CSV extension is the numerical data that will be used as a basis for defining the color. Figure 2.2.2 shows an example of how this file should be. We are working with residues up to position 120, however, Figure 2.2.2 does not show all the charged amino acids up to that position. Each system is divided into 2 columns where the information on the residue kind (with code of three letters in capital) and the numerical value (in our case p$K_a$) will be contained.

*Note: If, for example, you eliminated the p$K_a$ value of* 2.68828 *from row* 12 *and column D, CPRISMA will consider the absence of it as equal to* 0.

Notice that there are 4 data sets for the 4 systems. In our example, the absence of any 8

[7]For more details: https://biopython.org/docs/1.75/api/Bio.Align.Applications.html.



columns in the CSV file can return the following error message:

```
The program stop!  Number of columns are not complete in data_input.csv.
```

```
def array_get():

    # Tuple of residues
    target_residues = ()

    # Tuple of sequences name
    name_sequence = ()

    # Operations
    descriptor_ope = { 'n' }

    # Color
    descriptor_col = { 'nc' }

    # Visualization
    descriptor_vis = { 'ReY','DeN','LeY' }

    # Maximum restriction
    descriptor_mxr = { 'nr' }

    # Array for comparisons
    dict_ref = {}

    # Array for operations
    dict_ope = {}

    # Array for color
    dict_col = {}

    # Array for visualization
    dict_vis = {}

    # Array for maximum restriction
    dict_mxr = {}

    return target_residues, name_sequence, descriptor_ope, descriptor_col, descriptor_vis, descriptor_mxr, dict_ref, dict_ope, dict_col, dict_vis, dict_mxr
```

Figure 2.2.3: Example of how looks like the Python script "array_get.py".

On the other hand, the order as each protein appears in the alignment should match with the numerical data of the CSV file. For instance, the first dataset (highlighted by brackets) is related to the first protein (labeled as "ZIKV-UG") in the alignment sequence and so on.

*Note: The program does not have the ability to distinguish which data set each sequence belongs to, so it is your responsibility to put this information in the correct order.*

Finally, it is *crucial* that the first row is empty, as indicated by the arrow in Figure 2.2.2.

*Note: Sometimes the numerical data length will not be the same between the systems, however, you can ignore this as long as there is a match regarding the order of appearance between the alignment and numerical data files.*

The third type of file is a Python script with the variables that will be used to give different instructions to the program and obtain the desired output (Figure 2.2.3)[8].

[8]These variables and how they work will be better described in next Chapters.



## 2.3 Outputs

The two types of output files that CPRISMA generates are a log file called "cprisma.log" which shows relevant data from the comparisons that are made or data directly related to the input parameters, and a HTML file called "alignment.html" where the colored alignment will appear. As we will see later, due to the various features that CPRISMA has, it is better to observe with some examples how these two files work. Besides, it is worth mentioning that all the warning messages or error reports will appear in the log file. Finally, you can generate other types of files with specific commands (see Section 5.15).



# 3 Basic commands

Our starting point will be the files located in "examples/Example_0". To run CPRISMA, you can simply type `cprisma` in the terminal (at the working directory). Based on the files previously created (*i.e.*, "alignment.dat", "data_input.csv", and "array_get.py"), something you will notice is that a new log file called "cprisma.log" has been generated in your directory. In this file will appear a count of all the amino acids per system registered from "alignment.dat" (Figure 3.0.1). On the other hand, a count of total charged or hydrophobic residues will also show right away (Figure 3.0.2).

```
                    --- Number of residues in alignment ---
          A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V   Z
ZIKV-UG   7   8   6   5   3   3   10  11  2   5   6   8   3   2   7   8   3   6   3   14  0
ZIKV-BR   7   9   6   6   3   3   9   11  2   4   6   7   3   2   7   8   3   6   3   15  0
DENV2     4   4   5   3   3   6   10  9   3   10  9   13  3   2   5   10  9   5   2   5   0
WNV       8   6   3   6   3   5   12  10  3   6   10  12  3   1   3   6   6   4   4   9   0
```

Figure 3.0.1: Amino acid count in the log file. Note that the last column is showing the residue called "Z", this is only a way to identify those residues that do not belong to the 20 essential amino acids.

```
                    --- Kind of residues in alignment ---
ZIKV-UG charged=39 ['D', 'E', 'R', 'H', 'K', 'Y', 'C'], hydrophobic=50 ['A', 'I', 'L', 'M', 'F', 'P', 'W', 'V']
ZIKV-BR charged=39 ['D', 'E', 'R', 'H', 'K', 'Y', 'C'], hydrophobic=50 ['A', 'I', 'L', 'M', 'F', 'P', 'W', 'V']
DENV2   charged=38 ['D', 'E', 'R', 'H', 'K', 'Y', 'C'], hydrophobic=43 ['A', 'I', 'L', 'M', 'F', 'P', 'W', 'V']
WNV     charged=46 ['D', 'E', 'R', 'H', 'K', 'Y', 'C'], hydrophobic=44 ['A', 'I', 'L', 'M', 'F', 'P', 'W', 'V']

Number of columns in data_input.csv is ok!
```

Figure 3.0.2: Kind of residues count in the log file.

*Note: For our example, in the output shown in Figure 3.0.2 it was considered the Cys residues as charged, and this should not be. Nonetheless, CPRISMA by default counts all charged amino acids without discriminating whether or not Cys' belong to disulfide bridges.*

You can see that the log file will report whether the number of columns in the CSV file is equivalent to the number of systems in the alignment file (see the last text in Figure 3.0.2).



Other additional information that appears in the log file will be the numerical data of the CSV file (Figure 3.0.3). Although it may be redundant information because you know about that in advance, what is sought with this is to propose the user a double-check of the numerical data to avoid future problems.

--- Information in data_input.csv ---

| ZIKV-UG | ZIKV-UG_dat | ZIKV-BR | ZIKV-BR_dat | DENV2 | DENV2_dat | WNV | WNV_dat |
|---|---|---|---|---|---|---|---|
| ASP | 3.90000 | ASP | 3.90000 | ASP | 3.88990 | ASP | 3.78340 |
| ASP | 3.09737 | ASP | 3.38668 | LYS | 10.48393 | ASP | 3.39189 |
| LYS | 10.89431 | LYS | 10.99729 | LYS | 10.58842 | ARG | 12.78243 |
| LYS | 10.68626 | LYS | 10.69875 | GLU | 3.79962 | GLU | 3.99138 |
| GLU | 3.77967 | GLU | 3.98917 | LYS | 10.67679 | ARG | 12.39385 |
| ARG | 12.58017 | ARG | 12.29720 | ASP | 3.39033 | HIS | 6.69314 |
| TYR | 9.97794 | TYR | 9.98935 | HIS | 6.77441 | ASP | 3.58575 |
| ASP | 3.49226 | ASP | 3.79508 | GLU | 4.08867 | GLU | 4.38775 |
| GLU | 3.99375 | GLU | 4.09717 | TYR | 9.98403 | ASP | 3.48066 |
| ARG | 12.59320 | ARG | 13.38161 | LYS | 10.99640 | ARG | 12.89923 |
| ASP | 3.09964 | ASP | 2.68828 | GLU | 3.99943 | TYR | 9.79223 |
| ARG | 13.28718 | ARG | 12.97763 | LYS | 11.08543 | LYS | 11.18316 |
| TYR | 9.59769 | TYR | 9.59592 | LYS | 10.88321 | TYR | 10.19568 |
| LYS | 11.08461 | LYS | 10.99773 | HIS | 6.38640 | TYR | 9.49347 |
| TYR | 9.99884 | TYR | 9.68700 | GLU | 3.78878 | GLU | 3.99996 |
| HIS | 6.27657 | HIS | 6.19907 | GLU | 3.48118 | LYS | 10.69714 |

Figure 3.0.3: Numerical data in the log file. All the information has not been shown.

Assuming both the alignment and the numerical data match, the directory would be expected to have another new file called "alignment.html", which contains the multiple sequence alignment with color. For our example, we will not get this file, instead, it displays the following message:

```
The program stop!  Number of target-residues [ZIKV-UG] is not the same in the
input files.  Residue A: 0 (data_input.csv) != 7 (alignment.dat).
```

This error message is telling us that the number of Ala/Alanine residues (or A) for ZIKV-UG is not the same in both the alignment and the numerical data files. This error can be corrected by applying a specific command line, but, we must know the different possible parameters that can be used through CPRISMA at first. Executing:

```
cprisma -h
```



... the following optional arguments will be displayed:

| | |
|---|---|
| `-h --help` | show this help message and exit |
| `-v --version` | show program's version number and exit |
| `-ns [bool]` | rename the sequences based on a tuple |
| `-va [int]` | method to visualize the alignment |
| `-j [bool]` | join sequences (only available for `va = 2`) |
| `-hc [bool]` | hide the conservation line |
| `-n [int]` | change the first and subsequent numbers in the alignment |
| `-l [int]` | number of residues per line |
| `-tr [bool]` | get a tuple of target-residues |
| `-a [int]` | define the accuracy of the input data |
| `-t [bool]` | get all information for the different comparisons in log file |
| `-ck [bool]` | check reference method |
| `-rf [str]` | method to compare sequences (`'default'`, `'pair'`, `'multiple'`) |
| `-lco [bool]` | display a list of the available colors |
| `-ico [int]` | multiply the intensity of the color |
| `-sco [int]` | sequence color (when `'ssc'` descriptor is applied) |
| `-mco [int]` | mutation color (when `'fmac'` or `'pimc'` descriptors are applied) |
| `-tco [bool]` | display color at the level of 3D structure through pymol script |
| `-dop [bool]` | get a operation descriptors dictionary (only available for `rf = 'multiple'`) |
| `-dco [bool]` | get a color descriptors dictionary (only available for `rf = 'multiple'`) |
| `-dvi [bool]` | get a visualization descriptors dictionary (only available for `rf = 'multiple'`) |
| `-dmx [bool]` | get a maximum restriction descriptors dictionary (only available for `rf = 'multiple'`) |

Above you can see the possible flags implemented in CPRISMA. Note that the Python variable type is appearing inside of square brackets, and a short description is displayed as well.

Next, a description of them will be given in the next Sections and Chapters, bringing practical examples to better understand how they work. Some flags should be parsed in parallel with the "array_get.py" script to avoid future misinterpretations and others may be interdependent.

The positional arguments `-h` (`--help`) and `-v (--version)` are the default commands that will display the possible flags that can be written in the terminal and the version of the



program, respectively. As these two commands are by default and self-explicit, in the following sections we will just focus on the CPRISMA parameters.

## 3.1 Target-residues (`-tr`)

The problem with the residue A of ZIKV-UG shown above is caused by the absence of numerical data for that amino acid. A possible solution could be to include in the CSV file, new rows named "ALA" followed by a column with p$K_a$ values equal to 0 and respecting the order in which they appear for each protein.

*Note: Here we assume $pK_a = 0$ as equivalent to no color, but this may vary depending on the type of data.*

But this process can become time-consuming because the p$K_a$ values for residues of Val, Ser, Trp, and so on, would also have to be included in the CSV file. A more efficient way to exclude residues that are not needed is to run the `-tr` command. However, when running it you may find the following problem:

```
The program stop!  The variable to check the target-residues is True, but the
tuple is equal to 0.
```

The flag `-tr` is a Boolean command, which means that by default it is "False", but when you invoke it becomes "True". `-tr` calls the `target_residues` variable in the script "array_get.py" (Figure 2.2.3). This one is a *tuple*[1] that must contain residues in one-letter format. When `-tr` is executed, the correlation is made between the CSV file and the residues reported in the tuple. So you should try to use the same types of amino acids that appear in the numerical data file to build the tuple. Common errors that can occur in the construction of the tuple are listed below:

1. An empty tuple will return an error like the one we see above.
2. Do not put all the kinds of residues that appear in the CSV file in the tuple.
3. To include a residue in the tuple that is not in the CSV file.

For our example, the tuple should be as follows:

```
target_residues = ('D', 'E', 'R', 'H', 'K', 'Y')
```

[1] See https://docs.python.org/3.7/tutorial/datastructures.html#tuples-and-sequences.



If we again run the program with the next command line:

```
cprisma -tr
```

... our log file will successfully return the following message:

```
Target-residues ('D', 'E', 'R', 'H', 'K', 'Y') match in alignment.dat and
data_input.csv.
```

*Note: It is important to mention that if you have the numerical values of all residues that appear in a specific multiple sequence alignment, it is not necessary to explicitly define a tuple with the 20 essential amino acids in the script "array_get.py" because the program already defines those 20 residues by default (see the practical example of* Section 8.3*).*

## 3.2 Accuracy (`-a`)

Sometimes the color may not have enough intensity depending on the numerical data employed. This happens due to the limits that the HTML color code itself may have or because the limits of the human vision. To improve the color intensity to highlight the regions that matter most to us in the sequence alignment is necessary to make additional adjustments.

In this section, let us focus on the accuracy of the data, which is changeable by means of `-a` command[2]. In Figure 3.0.3, it is observed that the precision of the data goes up to the fifth decimal place. As shown in Figure 3.2.1, after program processing one decimal per value will be displayed (which corresponds to the default condition). When two or more decimals are considered, it is necessary imperative to invoke the command `-a` followed by a positive integer number which refers to the decimal places, as follows:

```
cprisma -tr -a 3
```

In our example, we are ordering the program to only consider taking into account three decimal places. Remember that this command will only affect the possible values obtained from operations, and then, the color applied when a gradient is executed (see Chapters 4, 5, and 6, respectively). Nevertheless, some information in the log file will not be affected by this parameter, like the statistical data reported there (Figure 3.3.1).

[2]This is one of the possible methods to improve the color intensity. Other methods can also be applied (see Section 5.14 and Chapter 7).



--- Data processed ---

| | ZIKV-UG | ZIKV-UG_dat | ZIKV-BR | ZIKV-BR_dat | DENV2 | DENV2_dat | WNV | WNV_dat |
|---|---|---|---|---|---|---|---|---|
| 0 | ASP | 3.9 | ASP | 3.9 | ASP | 3.9 | ASP | 3.8 |
| 1 | ASP | 3.1 | ASP | 3.4 | XXX | 0.0 | ASP | 3.4 |
| 2 | XXX | 0.0 | XXX | 0.0 | LYS | 10.5 | XXX | 0.0 |
| 3 | LYS | 10.9 | LYS | 11.0 | XXX | 0.0 | ARG | 12.8 |
| 4 | LYS | 10.7 | LYS | 10.7 | LYS | 10.6 | XXX | 0.0 |
| 5 | GLU | 3.8 | GLU | 4.0 | GLU | 3.8 | GLU | 4.0 |
| 6 | ARG | 12.6 | ARG | 12.3 | LYS | 10.7 | ARG | 12.4 |
| 7 | TYR | 10.0 | TYR | 10.0 | XXX | 0.0 | HIS | 6.7 |
| 8 | XXX | 0.0 | XXX | 0.0 | ASP | 3.4 | XXX | 0.0 |
| 9 | ASP | 3.5 | ASP | 3.8 | XXX | 0.0 | ASP | 3.6 |
| 10 | GLU | 4.0 | GLU | 4.1 | HIS | 6.8 | GLU | 4.4 |
| 11 | ARG | 12.6 | ARG | 13.4 | XXX | 0.0 | XXX | 0.0 |
| 12 | ASP | 3.1 | ASP | 2.7 | GLU | 4.1 | ASP | 3.5 |
| 13 | ARG | 13.3 | ARG | 13.0 | XXX | 0.0 | ARG | 12.9 |
| 14 | TYR | 9.6 | TYR | 9.6 | TYR | 10.0 | TYR | 9.8 |
| 15 | LYS | 11.1 | LYS | 11.0 | LYS | 11.0 | LYS | 11.2 |
| 16 | TYR | 10.0 | TYR | 9.7 | XXX | 0.0 | TYR | 10.2 |
| 17 | HIS | 6.3 | HIS | 6.2 | XXX | 0.0 | TYR | 9.5 |

Figure 3.2.1: Numerical data processed in the log file. Note that some rows are named "XXX", which indicates that a gap has been inserted. These gaps are directly related to the alignment and are necessary for the reading and coloring of the sequences later. Notice that there is an index starting from 0 that number each row. All the information has not been shown.

## 3.3 Alignment visualization (`-va`)

At this point, you will have noticed that the file "alignment.html" appears in your directory (Figure 3.3.2). Note that the file does not have any color still[3]. By default, CPRISMA will separate each protein sequence keeping the conservation line for each of them [*i.e.*, the line that contains dots (“.” , “:”), asterisks (“*”) or spaces (“ ”)]. Now it seems not to have much relevance to observe our alignment without comparisons among sequences. Nevertheless, as described later this kind of visualization can be very useful to detect independent color behaviors from sets of numerical data.

Another way to visualize our alignment can be calling method 2 through the `-va` command, as follows:

```
cprisma -tr -va 2
```

[3]This aspect will be covered in Chapter 5.



```
                    --- Statistics ---

              ZIKV-UG    ZIKV-BR      DENV2        WNV
count   36.000000  36.000000  35.000000  43.000000
mean     8.027069   8.065460   7.998239   7.820189
std      3.957723   4.020073   3.578109   3.732822
min      3.097370   2.688280   3.390330   3.384640
25%      3.964313   3.990678   3.892795   3.993175
50%      9.787815   9.641460   9.984030   9.493470
75%     11.078850  11.322598  10.883975  11.034720
max     13.392970  13.485050  13.184100  13.491570
```

Figure 3.3.1: Statistical data for each protein in the log file.

```
        1                 20                  40                  60
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

        81                100                 120
ZIKV-UG ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
        ** *.*:::. . :.  . . . *     :*  .**:***

        1                 20                  40                  60
ZIKV-BR DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

        81                100                 120
ZIKV-BR ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
        ** *.*:::. . :.  . . . *     :*  .**:***

        1                 20                  40                  60
DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

        81                100                 120
DENV2   ENEVKLTIMTGDIKGIMQAGKRSLRPQPTELKYSWKTWGK
        ** *.*:::. . :.  . . . *     :*  .**:***

        1                 20                  40                  60
WNV     DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSVSRLEHQMWEAVKDELNTLLK
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

        81                100                 120
WNV     ENGVDLSVVVEKQEGMYKSAPKRLTATTEKLEIGWKAWGK
        ** *.*:::. . :.  . . . *     :*  .**:***
```

Figure 3.3.2: Alignment processed through visualization method 1 without sequence comparison.



An output using this method appears in Figure 3.6.1. Notice that each sequence is being labeled at the end with the word "comparison" followed by a number (see brackets in Figure 3.6.1). This means that CPRISMA is not comparing the sequences but only analyzing each protein separately, like in the previous method.

*Note: We can apply these same visualization methods but comparing our sequences. To see this aspect in more detail it is recommended to check Sections 3.9 and 3.10.*

## 3.4 Join (`-j`)

The `-j` command is a complement for the visualization method 2. With this command the space between the comparisons is reduced (compare Figures 3.6.1 and Figure 3.7.1). To get an alignment without spaces, you can invoke the following command:

```
cprisma -tr -va 2 -j
```

Practical examples using this flag appear in Chapter 8.

## 3.5 Numbering (`-n`)

To change the first and subsequent numbers that appear at the top of the alignment, the command `-n` can be executed followed by a positive integer number corresponding to the desired position. For instance, we can assume that our first number is 101, so:

```
cprisma -tr -va 2 -j -n 101
```

See the practical example of Section 8.2.

## 3.6 Number of residues per line (`-l`)

You can change the number of amino acids per line using the command `-l` followed by a positive integer number.

*Note: Remember that 80 residues per line will appear by default, so we recommend only use values greater than and multiples of 80 to have an optimal display of the numbering. You can use whatever amount of residues per line you prefer, but there may be times when the numbers will be cut off.*




Figure 3.6.1: Alignment processed through visualization method 2 without sequence comparison.

Next, an example of a command-line for this feature appears:

```
cprisma -tr -va 2 -j -l 160
```

## 3.7 Name sequence (-ns)

Something that is useful when working with different sequences is renaming them with the command `-ns`. This command calls the variable `name_sequence` of the script "array_get.py" (Figure 2.2.3) and makes a relationship between this variable (which is a tuple) and the number of proteins analyzed. Let us imagine that we build the following tuple:

```
name_sequence = (''Sequence 1'', ''Sequence 2'', ''Sequence 3'')
```

Then, the next command-line is executed:

```
cprisma -tr -va 2 -j -ns
```

Since the number of "name sequences" is only 3 and the total of proteins is 4, CPRISMA will return the following error:



`Variable check_name_seq = True but the sequences' number does not match with tuple input ('Sequence 1', 'Sequence 2', 'Sequence 3')`

Adding one more name to the tuple `('Sequence 1', 'Sequence 2', 'Sequence 3', 'Sequence 4')`, the program will finally return the following message:

`Variable check_name_seq = True, so the sequences' names have been changed.`

Now you can see that both your alignment (Figure 3.7.1) and all the information in the log file have changed the sequences' names.

```
                1                 20                 40                 60
Sequence 1      DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE  comparison 1
Sequence 2      DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE  comparison 2
Sequence 3      DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS  comparison 3
Sequence 4      DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSVSRLEHQMWEAVKDELNTLLK  comparison 4
                * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

                81                100                120
Sequence 1      ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK  comparison 1
Sequence 2      ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK  comparison 2
Sequence 3      ENEVKLTIMTGDIKGIMQAGKRSLRPQPTELKYSWKTWGK  comparison 3
Sequence 4      ENGVDLSVVVEKQEGMYKSAPKRLTATTEKLEIGWKAWGK  comparison 4
                ** *.*:::. . :.  . . . *     :*  .**:***
```

Figure 3.7.1: Alignment with renamed sequences.

## 3.8 Hide conservation line (`-hc`)

The `-hc` command hides the conservation line of the alignment (check the line with "*", ".", ":" and spaces " " in Figure 3.7.1, for instance). `-hc` can be invoked following the command-line:

```
cprisma -tr -hc
```

Hiding the conservation line can be useful when we are using transcendental comparison methods (see Section 3.10).

See the practical example of Section 8.2 where `-hc` is employed.

## 3.9 Check reference method (`-ck`)

So far you have surely noticed the following message in your log file:



```
No comparison among the data.

-- Set of comparisons:  main_0 --

     ZIKV-UG X no comparison

-- Set of comparisons:  main_1 --

     ZIKV-BR X no comparison

-- Set of comparisons:  main_2 --

      DENV2 X no comparison

-- Set of comparisons:  main_3 --

       WNV X no comparison
```

For our example, the data of the protein systems are not being compared. But if you want to apply mathematical operations between your numerical data (see Chapter 4), it will be mandatory to compare the sequences. For instance, in our previous study, we wanted to compare the p$K_a$'s of ionizable residues among NS1's proteins from several ZIKV strains [1]. For that, it was calculated the difference of p$K_a$'s ($\Delta$p$K_a$) among proteins and they were visualized with specific colors (see Section 5.10 and the practical example described in Section 8.1 as well).

Before doing any calculation, you must inform the program that it is desired to compare the data by means of the `-ck` flag. The command line to execute comparisons between the protein sequences appears below:

```
cprisma -tr -ck
```

Now, the following information will be found in the log file:

```
-- Set of comparisons:  main_0 --

   ZIKV-UG X ZIKV-BR DENV2 WNV
```

Note that now the sequences are being compared, so ZIKV-UG is the reference sequence, and ZIKV-BR, DENV2, and WNV are the target-sequences. Also, this comparison is being coupled to a group of comparisons called "`main_0`". This is not the same as what was seen previously where different "`main`'s" were generated. At the same time, you will have perceived that your



HTML file (Figure 3.9.1) is different when compared to the previous outcomes (Figures 3.3.2 and 3.6.1). Technically, the alignment is the same as the input shown in Figure 2.2.1 but with a slightly different format.

*Note: In Section 3.10, you will find more information on how to apply more transcendental comparisons.*

```
        1                 20                 40                 60
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE
ZIKV-BR DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE
DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS
WNV     DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSVSRLEHQMWEAVKDELNTLLK
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

        81                100                120
ZIKV-UG ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
ZIKV-BR ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
DENV2   ENEVKLTIMTGDIKGIMQAGKRSLRPQPTELKYSWKTWGK
WNV     ENGVDLSVVVEKQEGMYKSAPKRLTATTEKLEIGWKAWGK
        ** *.*:::. . :.  . . . *     :*  .**:***
```

Figure 3.9.1: Alignment processed through visualization method 1 with sequence comparison.

## 3.10 Reference method (`-rf`)

Whether `-ck` was executed, other additional information will appear in the log file:

```
Method used 'default' for the array:  | main_0 [{'0':  [1, 2, 3]}] |
```

To understand what this means, a basic understanding of Python's data structure is necessary. Let us go a little back, just to correctly interpret these outcomes. Based on the previous message, note that you have one comparison that belongs to `main_0`. Figure 3.10.1 is showing an example of the CPRISMA default variable when `-rf` is invoked (pay attention to the Roman numerals). In Python, whenever one string type (II) and any type variables (VI) appear separated by a colon (V) inside brackets (I), it will be interpreted by the language as a *dictionary*[4] . The numerals II and VI are also called *keys* and *values*, respectively. In our case, the key of our dictionary is the string '0' and the value is a *list* of numbers from 1 to 3. Note that the lists have square brackets (III), which differentiates them from dictionaries[5].

[4]See https://docs.python.org/3.7/tutorial/datastructures.html#dictionaries.
[5]See https://docs.python.org/3.7/tutorial/datastructures.html#more-on-lists.

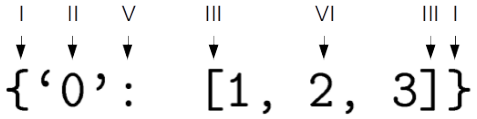



Figure 3.10.1: Dictionary created by CPRISMA. the meaning of Roman numerals is explained in the text.

*Note: Based on the message cited at the beginning of this Section, see that your dictionary is also inside square brackets:* `[{'0': [1, 2, 3]}]`[6]. *In Subsection 3.10.3, we will explain in more detail the reason for that.*

Please keep all these features in mind as they will be helpful in understanding the following sections. Furthermore, in some moments we will call the sets of comparisons as *arrays*.

To interpret the numbers in our dictionary (Figure 3.10.1), we simply have to relate the order of appearance for each sequence based on the multiple alignments. Thus, '0' will be ZIKV-UG, '1' will be ZIKV-BR, and so on.

There are three comparison methods used by CPRISMA: `default`, `pair`, and `multiple`, which are described below.

### 3.10.1 Default

The `default` method of CPRISMA compares the first sequence with the rest. You can explicitly invoke this method via the `-rf` flag followed by the word "default" like this:

```
cprisma -tr -ck -rf default
```

*Note: Remember, due to this comparison method is the optional argument by default (as its name implies), CPRISMA will always return a dictionary with this structure when you run* `-ck` *(i.e., you do not need to type* `-rf default`*, explicitly)*

### 3.10.2 Pair

Another comparison method is called `pair`. To invoke it, we can type in terminal:

```
cprisma -tr -ck -rf pair
```

[6]We have changed the color to purple for some square brackets to distinguish easily each list.



Now, our log file will contain the following information:

```
Method used 'pair' for the array:  | main_0 [{'0':  [1]}] | | main_2 [{'2':
[3]}] |

-- Set of comparisons:  main_0 --

        ZIKV-UG X ZIKV-BR

-- Set of comparisons:  main_2 --

           DENV2 X WNV
```

**Comparison: 1**

```
        1                 20                 40                 60
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE
ZIKV-BR DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

        81                100                120
ZIKV-UG ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
ZIKV-BR ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
        ** *.*:::. . :.  . . . *     :*  .**:***
```

**Comparison: 2**

```
        1                 20                 40                 60
DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS
WNV     DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSVSRLEHQMWEAVKDELNTLLK
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

        81                100                120
DENV2   ENEVKLTIMTGDIKGIMQAGKRSLRPQPTELKYSWKTWGK
WNV     ENGVDLSVVVEKQEGMYKSAPKRLTATTEKLEIGWKAWGK
        ** *.*:::. . :.  . . . *     :*  .**:***
```

Figure 3.10.2: Alignment processed with `pair` comparison method.

Note that two arrays have now been created, which can be interpreted as two independent sets of comparisons and they belong to the `main`'s: `main_0` and `main_2`. One where 0 (ZIKV-UG) and 1 (ZIKV-BR) sequences are the $\text{reference}_{\text{comparison1}}$ and $\text{target}_{\text{comparison1}}$, respectively; and another where 2 (DENV2) and 3 (WNV) sequences are the $\text{reference}_{\text{comparison2}}$ and



target$_{comparison2}$, respectively. This will also be reflected in our HTML file (Figure 3.10.2)[7].

It is important to mention that for CPRISMA to work correctly with the `pair` comparison method, the total number of sequences is needed to be *even*. Otherwise, it will return to the method for `default` with the following message:

`Attention!!! Method selected by user is 'pair' but number of sequences (=3`[8]`) is not a even number. The method return to 'default'.`

### 3.10.3 Multiple

Among all the comparison methods, `multiple` can be the most complex to understand. With this method, many types of comparisons can be made between a set of sequences (*e.g.* see all the practical examples in Chapter 8). However, before invoking it you must have built a dictionary compatible with the CPRISMA format array. In the script "array_get.py" a variable called `dict_ref` will appear (Figure 2.2.3). From this, you can build an array like this[9]:

```
dict_ref = { 'main_0' :  [ {'0':[1, 2, 3]}, {'0':[]} ] ,
             'main_1' :  [ {'1':[0, 2]}, {'1':[0, 2, 3]} ] ,
             'main_2' :  [ {'2':[0]} ] }
```

At first glance, this type of arrangement may seem tricky, but give us a chance to explain it to make it easier.

The CPRISMA convention is going to call the "external layer" of the array with the names: "*first-key*" (or "*main key*") and "*first-value*" (or "*main value*"). Then, the first-keys will be all `main`'s typed in the array, *i.e.*, `main_0`, `main_1`, and `main_2`; and the first-values will be the lists: `[{'0': [1, 2, 3]}, {'0': []}]`, `[{'1': [0, 2]}, {'1': [0, 2, 3]}]`, and `[{'2': [0]}]`[10]. Each element of the list/first-value is equivalent to a comparison and they have a "*second-key*" (or "*subkey*") and a "*second-value*" (or "*subvalue*"). The subkeys are the string numbers that represent the sequence references and the subvalues are the lists with the target proteins. Notice that each cluster of first-key/first-value pairs is separated by a

[7] We suggest the user to also explore visualization method 2 (see Section 3.3).
[8] A data set of 3 proteins was used to compare them (only as an example).
[9] Note that we have differentiated each comparison with a different color to aid the reader.
[10] We have changed the color to purple for some square brackets to distinguish easily each list.

comma and they are so-called like "set of comparisons"[11]. Besides, each comparison in the list/first-value is separated by a comma as well. Please, avoid confusing this! For the sake of convenience, in Figure 3.10.3, we have made a sketch for `main_0`, where you can follow each characteristic of the CPRISMA array.

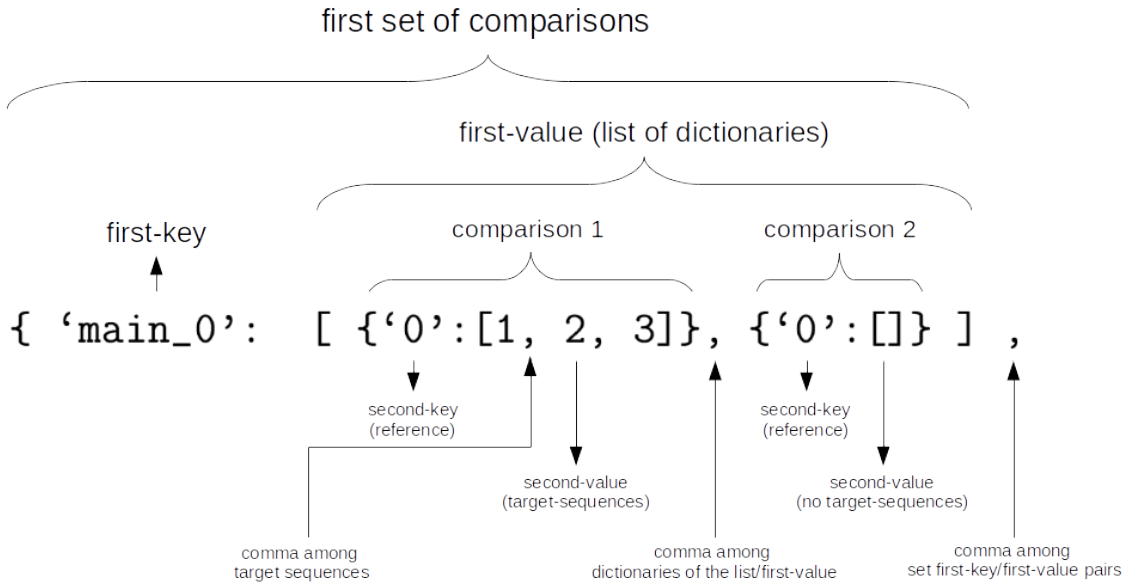



Figure 3.10.3: Example of the structure of the CPRISMA array when is used `multiple` method.

Maybe, you can also perceive that each `main` has the same reference number of a particular set of comparisons. For example, `main_1` has comparisons 3 and 4 which are `{'1': [0, 2]}` and `{'1': [0, 2, 3]}`, respectively, so the reference sequences `'1'` match with the number of the `main` (*i.e.*, `main_1`). The same goes for `main_0` and `main_2`. However, it should be noted that you can use whatever name you prefer for each "`main`", to better identify your comparisons. Just make sure they are different from each other. For practical purposes, we will use the standard name as already seen in the previous and following examples.

*Note: Check that comparison 2 has a list with no elements or empty:* `{'0': []}`. *CPRISMA will interpret this dictionary as "without comparison" (or without target-sequences).*

[11] Identify this same structure when using the `default` and `pair` methods.



```
        1                 20                 40                 60
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE  comparison 1
ZIKV-BR DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE
DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS
WNV     DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSVSRLEHQMWEAVKDELNTLLK

ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE  comparison 2

ZIKV-BR DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE  comparison 3
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE
DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS

ZIKV-BR DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE  comparison 4
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE
DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS
WNV     DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSVSRLEHQMWEAVKDELNTLLK

DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS  comparison 5
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE

        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.
```

Figure 3.10.4: Alignment processed with `multiple` comparison method. Not all alignment is being shown.

If the array is well constructed, we can invoke the following command:

```
cprisma -tr -ck -rf multiple
```

Then, in our log file the comparisons will appear as follows:

```
-- Set of comparisons:  main_0 --

   ZIKV-UG X ZIKV-BR DENV2 WNV

     ZIKV-UG X no comparison

-- Set of comparisons:  main_1 --

     ZIKV-BR X ZIKV-UG DENV2

   ZIKV-BR X ZIKV-UG DENV2 WNV

-- Set of comparisons:  main_2 --

          DENV2 X ZIKV-UG
```



Our HTML file will show an output like the one shown in Figure 3.10.4 (to facilitate the visualization, method 2 was used, it is recommended to see Section 3.3).

Finally, if the array is not well built, CPRISMA will always return to the `default` value. In addition, it will inform you in detail where the errors are. Based on this, you can correct them yourself (the program will guide you). Warning/error messages for `multiple` command will not appear in this documentation, as we prefer that the user be trained in understanding the construction of CPRISMA arrays.

See in Chapter 8 the possibilities that can be declared in `dict_ref` using `multiple` comparison method .

## 3.11 Total information (`-t`)

To have the verbose version of the log file, you can invoke the following command:

```
cprisma -tr -t
```

Now, the log file will show in detail the operations that you are applying to each set of comparisons (see Chapter 4). In turn, a relationship between each comparison and the color assigned to it will be shown (see Chapter 5). Also, if a maximum restriction is applied (see Chapter 7), you can visualize the lists of values used to normalize the color displayed on the HTML file (whether or not math operations are performed).



# 4 Operation (`descriptor_ope`)

CPRISMA has a feature called "operation" that allows the user to compute basic mathematical operations among the data. To execute this feature, it is *mandatory* to have at least one comparison between a particular reference and target proteins. For the following examples, we will always assume the `default` comparison method (see Subsection 3.10.1), with the exception of Section 4.5.

Generally, the program will always read the variable called `descriptor_ope` from the Python script "array_get.py" (Figure 2.2.3). Then, CPRISMA will perform a certain operation between the vector $\mathbf{r} = [r_j]$ (reference) and the matrix $\mathbf{T} = [t_{ij}]$ [target(s)]. Based on the data processed and displayed in the log file, for our example using p$K_a$ values, the vector $\mathbf{r}$, and matrix $\mathbf{T}$ have dimensions $56 \times 1$ and $56 \times 3$, respectively. Something that should be noted is that CPRISMA takes advantage of the `numpy` features to apply broadcasting and thus improve the speed of matrix operations. For this version of CPRISMA, you can perform 3 types of calculations: delta/`d` ($\Delta$), absolute delta/`da` ($|\Delta|$), and multiplication/`m`. Although, it is possible to not calculate any operation (none/`n`) as well. For CPRISMA, these 4 possibilities are called *descriptors*.

*Note: We highlight that this nomenclature (i.e., descriptor) also applies to other features in addition to operations, such as color, visualization, and maximum restriction (see* Chapters 5, 6, *and* 7, *respectively).*

Finally, the numerical data generated through the operations will serve as the basis to put the color in the multiple sequence alignments.

## 4.1 None (`n`)

When it is desired to keep the same numerical values of the input (*i.e.*, the data of the CSV file), you can do this by declaring the descriptor `n` in the "array_get.py" file, as follows:

```
descriptor_ope = { 'n' }
```



Perceive that the `n` descriptor is a string variable inside a pair of brackets. This data structure corresponds to a *set* variable type and should not be confused with a dictionary[1].

After having run CPRISMA, your log file will have various information about the operation applied. For now, we should only focus on the next[2]:

| main | comparison | descriptor |
|---|---|---|
| main_0 | 1 | n |

This statement is simply saying that `n` descriptor was considered for comparison 1, in the `main_0` comparison set.

*Note: Sometimes, it could be declared this descriptor or another wrongly. To avoid that the program stops, it will always return to* `n` *descriptor (default parameter), with some additional error messages in your log file, which will guide you to correct the problem.*

## 4.2 Delta (`d`)

When the `d` descriptor is invoked like this:

```
descriptor_ope = { 'd' }
```

... the program will do a conventional $\Delta$ calculation between $\mathbf{r}$ and $\mathbf{T}$ (Equation 4.2.1).

$$\mathbf{W} = \mathbf{T} - \mathbf{r}, \tag{4.2.1}$$

where $\mathbf{W}$ is the new data to put color (see Chapter 5).

## 4.3 Absolute delta (`da`)

To declare the `da` descriptor, it is possible as follow:

```
descriptor_ope = { 'da' }
```

[1] See https://docs.python.org/3.7/tutorial/datastructures.html#sets.
[2] When using a different operation descriptor, respectively, its letter will appear in the `descriptor` column.



CPRISMA will apply the $\Delta$ calculation as before (Section 4.2), but using the absolute value:

$$\mathbf{W} = |\mathbf{T} - \mathbf{r}| \tag{4.3.1}$$

## 4.4 Multiply (`m`)

To declare the `m` descriptor, you can invoke:

```
descriptor_ope = { 'm' }
```

The program will multiply the data as shown in Equation 4.4.1.

$$\mathbf{W} = \mathbf{T} \times \mathbf{r} \tag{4.4.1}$$

## 4.5 Operation descriptors dictionary (`-dop`)

One can run the following command using `multiple` comparison method:

```
cprisma -tr -ck -rf multiple
```

We have started from the next CPRISMA array (see Subsection 3.10.3)[3]:

```
dict_ref = { 'main_0':  [ {'0':[1, 2, 3]}, {'0':[]} ] ,
             'main_1':  [ {'1':[0, 2]}, {'1':[0, 2, 3]} ] ,
             'main_2':  [ {'2':[0]} ] }
```

Suppose we are doing a $\Delta$ operation (*i.e.*, `descriptor_ope = { 'd' }`). Then, the following information will appear in the log file:

```
Comparing dict_ref with dict_ope:
                                {'0':  [1, 2, 3]} ........  d
                                {'0':  []} ........  d
                                {'1':  [0, 2]} ........  d
```

[3] Note that we have differentiated each comparison with a different color to aid the reader.



```
                                {'1':  [0, 2, 3]} ........  d

                                {'2':  [0]} ........  d
```

```
The array of the feature 'operation' dict_ope is compatible with the array of
comparison sequences dict_ref!
```

*Note: Comparison* `2` *does not have a target-sequence (i.e.,* `{'0':[]}`*), so operations like* `d`*,* `da`*, or* `m` *are not applied.*

This information is showing us a relationship between each comparison and the type of operation that it is applied for each one. For all comparisons, a $\Delta$ is computed among its specific reference and target-sequences.

Nonetheless, we are not always going to calculate a single operation for all comparisons. To perform a different math calculation for each reference/target(s) relationship, an array of operation descriptors called `dict_ope` should be built in our "array_get.py" script. We shall follow very similar rules to those we saw earlier for the `multiple` method (see Subsection 3.10.3). Our dictionary of operations must have an equivalent number of `main`'s as well as in `dict_ref` (*i.e.*, `main_0`, `main_1`, and `main_2`). In turn, for each `main` (first-key), a list of operation descriptors (first-values) should be invoked[4]. Based on this, suppose we construct the following operation-array[5]:

```
dict_ope = { 'main_0' :  [ {'d'} , {'n'} ] ,

             'main_1' :  [ {'m'} , {'d'} ] ,

             'main_2' :  [ {'da'} ] }
```

Invoking the `-dop` command you will order to CPRISMA to take into account the `dict_ope` dictionary of the "array_get.py" script:

```
cprisma -tr -ck -rf multiple -dop
```

Now, our log file will show the operation applied for each comparison:

```
Comparing dict_ref with dict_ope:
```

[4]For the other features (*i.e.*, color, visualization, and maximum restriction), these first-values will always be a list of descriptors with the same number of elements equivalent to the number of comparisons.

[5]Note that we have differentiated each operation descriptor for each comparison with a different color to aid the reader.



```
{'0':  [1, 2, 3]} ........  d

{'0':  []} ........  n

{'1':  [0, 2]} ........  m

{'1':  [0, 2, 3]} ........  d

{'2':  [0]} ........  da
```

```
The array of the feature 'operation' dict_ope is compatible with the array of
comparison sequences dict_ref!
```

... and the next table:

| main | comparison | descriptor |
|---|---|---|
| main_0 | 1 | d |
| main_0 | 2 | n |
| main_1 | 3 | m |
| main_1 | 4 | d |
| main_2 | 5 | da |

It is important to mention that `-dop` will only work with the `multiple` comparison method.

*Note: To check how the numerical calculations are applied between each row of the processed input file, the* `-t` *flag can be employed (see* Section 3.11*).* Figure 7.0.1 *shows an example of how that part of the log file looks like.*

*Note: If the array of* `dict_ope` *is wrongly built, all descriptors will be transformed to* `n` *(default parameter) for each comparison and additional information about where the errors are will be displayed.*



# 5 Color (`descriptor_col`)

So far, we have intentionally seen situations where our multiple sequence alignment is colorless. This is because we did not want to work with many features simultaneously. This Chapter will describe how to use the CPRISMA feature called "color".

Remember that the color applied is working with the HTML language. The file employed by CPRISMA to apply the color is located at "cprisma/colors.csv". Figure 5.0.1 shows some rows from this file. The cells are divided into four columns with the heads "General", "Name", "Hex", and "RGB". CPRISMA will only invoke the column labeled RGB to give color, the other information is just to help the user to identify each color easily. It is important to mention that all this stuff is based on the official HTML site about color code[1]. For quick access to this list of colors, you can invoke the `-lco` command described in Section 5.16.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | General | Name | Hex | RGB |
| 2 | Red HTML | IndianRed | #CD5C5C | rgb(205, 92, 92 |
| 3 | Red HTML | LightCoral | #F08080 | rgb(240, 128, 128 |
| 4 | Red HTML | Salmon | #FA8072 | rgb(250, 128, 114 |
| 5 | Red HTML | DarkSalmon | #E9967A | rgb(233, 150, 122 |
| 6 | Red HTML | LightSalmon | #FFA07A | rgb(255, 160, 122 |
| 7 | Red HTML | Crimson | #DC143C | rgb(220, 20, 60 |
| 8 | Red HTML | Red | #FF0000 | rgb(255, 0, 0 |

Figure 5.0.1: CSV file with color information. All the rows have not been shown.

For the color feature, we also have a set variable called `descriptor_col` in the script "array_get.py" (Figure 2.2.3). The color feature has 10 different descriptors (`nc`, `ssc`, `fsc`, `fac`, `fmac`, `pic`, `pimc`, `tc`, `tmc`, and `pkac`) that will be described below.

For most next examples the following command is being executed:

[1] See https://htmlcolorcodes.com/color-names/.



```
cprisma -tr -ck
```

Furthermore, the variable `descriptor_ope` in the Python script "array_get.py" was typed as follows:

```
descriptor_ope = { 'd' }
```

## 5.1 No color (nc)

To use the descriptor `nc`, you can declare it like this:

```
descriptor_col = { 'nc' }
```[2]

As the name implies, you are not adding any color to the alignment. Now, your log file will have various information about the color feature applied. We should only focus on the next:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | nc | NaN | NaN | NaN | NaN | NaN |

This table is simply saying that the `nc` descriptor was considered for comparison 1, in the `main_0` comparison set (for now ignore the rest of columns in that log message).

*Note: Sometimes, some descriptors can be declared wrongly. To avoid that the program stops, it will always return to the* `nc` *descriptor (default parameter), with some additional error messages in your log file, which will guide you to correct the problem.*

## 5.2 Same sequence color (ssc)

Suppose that you want to color a set of sequences with the same color. For this we declare:

```
descriptor_col = { 'ssc' }
```

The log file will report as `[same]` in the `sequence` column:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | ssc | [same] | NaN | NaN | NaN | NaN |

Furthermore, the HTML file now has a colored alignment (Figure 5.2.1).

[2] For all alignment examples of Chapter 3, this was the variable executed.

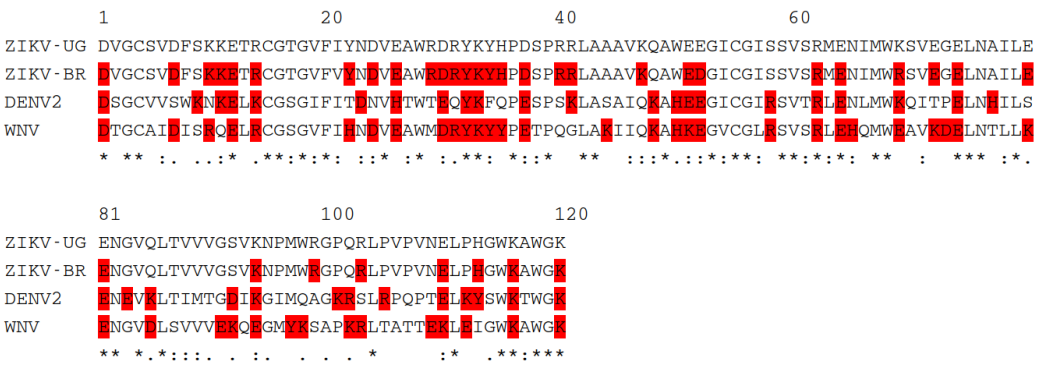



Figure 5.2.1: Alignment colored through `ssc` descriptor.

Note that only the target-residues in all the proteins are being colored with Red (see Section 3.1) and the reference (*i.e.*, ZIKV-UG) is not being examined. The default color for the `ssc` descriptor is Red, which corresponds to the 6th row when executed `-lco` (see Section 5.16). You can change that color by invoking the `-sco` flag described in Section 5.12.

## 5.3 Free sequence color (`fsc`)

To give a different color to each *target-sequence*, the descriptor `fsc` can be declared as follows:

```
descriptor_col = { 'fsc' :  [6, 54, 94] }
```

Notice that now is not being invoked a set variable type but a dictionary. Our key will be the descriptor `fsc` (string type) and our value is a list of number colors (see Section 5.16) that corresponds to the total number of target-sequences and their order of appearance (*i.e.*, colors `6`, `54`, and `94` for ZIKV-BR, DENV2, and WNV, respectively). After running the program, the log file will show the following information:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | fsc | [6, 54, 94] | NaN | NaN | NaN | NaN |

... and the HTML file will now display a color for each target-sequence (Figure 5.3.1). Assuming you will do quick tests and do not care about the color kind assigned to each target-sequence, it is possible to declare your variable like this:

```
descriptor_col = { 'fsc' :  ['random'] }
```

When the word ‘random’ is typed into the list, CPRISMA will give each target-sequence a random color.




Figure 5.3.1: Alignment colored through `fsc` descriptor.

## 5.4 Free amino acid color (`fac`)

To color each *target-residue* for the target-sequence(s), the `descriptor_col` can be invoked like this:

```
descriptor_col = { ‘fac’ :  [6, 20, 22, 54, 94, 46] }
```

Look at the list has 6 different color numbers. That number of elements is correlated to the number of target-residues declared in the tuple `target_residues` of the script “array_get.py” (see Section 3.1).

*Note: If you need to assess the* 20 *essential amino acids, you will need to declare* 20 *color numbers in that list or use* `random` *option.*

Later, the log file will report this list of color numbers in the `residue` column like this:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | fac | NaN | [6, 20...] | NaN | NaN | NaN |

... and the HTML file will now display a color for each target-residue (Figure 5.4.1). As with the `fsc` descriptor, random colors for your target-residues can be declared as follows:

```
descriptor_col = { ‘fac’ :  [‘random’] }
```

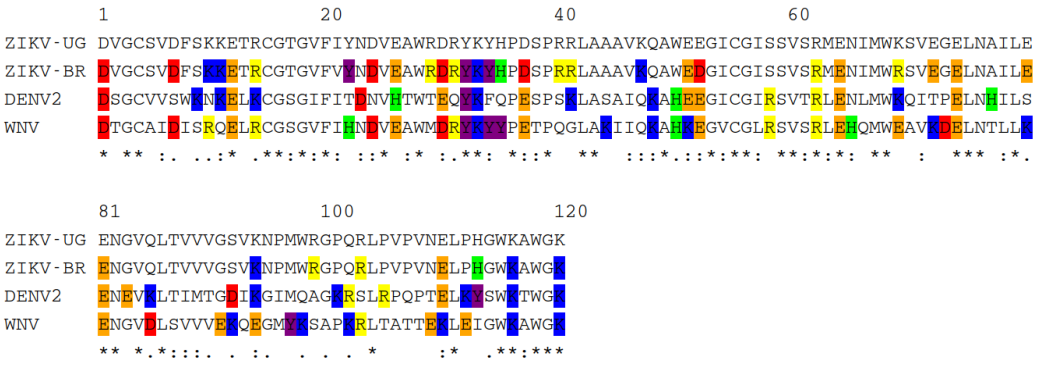



Figure 5.4.1: Alignment colored through `fac` descriptor.

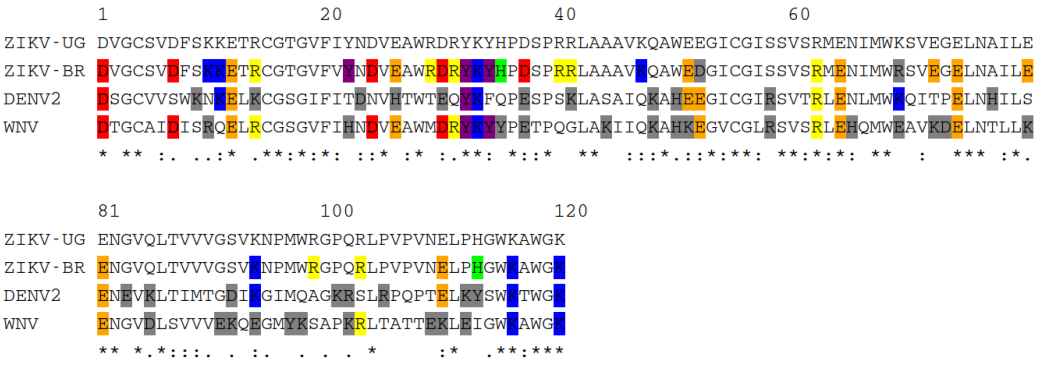



Figure 5.4.2: Alignment colored through `fmac` descriptor.

## 5.5 Free amino acid and mutation color (`fmac`)

The `fmac` descriptor is a copy of the `fac` one, but with the capacity to distinguish regions where there are mutations between the reference and target-sequence(s). To execute it, you can declare your variable `descriptor_col` like this:

```
descriptor_col = { 'fmac' :  [6, 20, 22, 54, 94, 46][3] }
```

Now the log file will also inform you that mutations are being differentiated by means of the `True` Boolean in the `mutation` column:

[3]Remember that the `random` method could be employed here.



| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | fmac | NaN | [6, 20...] | True | NaN | NaN |

A typical HTML outcome using the descriptor `fmac` appears in Figure 5.4.2. See that mutations are colored in gray. Whether you want to change the Gray color of the mutations (default), it is possible through the `-mco` command (see Section 5.13).

## 5.6 Color by position index (`pic`)

When it is desired to highlight only specific regions of the alignment, the `pic` descriptor can be invoked[4]:

```
descriptor_col = { 'pic' :  [ { '1' :  [6, 'n', '1-2', '3', '4-8'] } ,
                              { '1' :  [94, 'b', '11-27'] } ,
                              { '2' :  [75, 'd', '3', '30-55'] } ,
                              { '3' :  [54, 'u', '20-35'] } ] }
```

This can be one of the most complex color descriptors to implement. The descriptor `pic` is a dictionary type where its key will be the string `'pic'`, and the value is a list of dictionaries with several parameters (the orange square brackets delimit the list and purple brackets the dictionaries).

We have seen this arrangement structure before (see Subsection 3.10.3). However, when `pic` is called, we have some differences. Our second-key will be a number (string type), from one of the target-sequence(s) (for our example, `'1'`, `'2'`, and `'3'`, or ZIKV-BR, DENV2, and WNV, respectively).

*Note: It is imperative that you respect the order in which the target-sequences appear in a specific comparison. However, note that you can call the same sequence more than once as* `'1'` *in our example.*

The second-value is a list of parameters where the first (in red letters) and second (in green letters) refer, respectively, to the positive integer color number (see Section 5.16) and the font style [`'n'` ("none"), `'b'` ("**bold**"), `'i'` ("*italic*"), `'d'` ("~~deleted~~"), and `'u'` ("underlined")].

[4]Some brackets, letters, and numbers are colored to aid the reader.



From the third (in blue), we find the intervals (string type) that we want to color with the format: `'number1-number2'`. The intervals must be from a lower to a higher number and sometimes only one can be considered (for our example see `'3'` in blue). You should respect this consistency! The numbers of the intervals that you must implement are the same when the input data is processed (see the index of each row in Figure 3.2.1 or the "Data processed" section of the log file). Remember, your intervals should be related to that index and not the alignment position!

*Note: Be careful about overlaps that can occur between the intervals.*

When CPRISMA is run, the program will display the following information in the log file:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | pic | [1] | [6...] | Nan | n | NaN |
| main_0 | 1 | pic | [1] | [94...] | Nan | b | NaN |
| main_0 | 1 | pic | [2] | [75...] | Nan | d | NaN |
| main_0 | 1 | pic | [3] | [54...] | Nan | u | NaN |

```
           1                 20                  40                  60
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE
ZIKV-BR DVGCSVDFSKKETRCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE
DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS
WNV     DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSVSKLEHQMWHAVKDELNTLLK
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

           81                100                 120
ZIKV-UG ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
ZIKV-BR ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
DENV2   ENEVKLTIMTGDIKGIMQAGKRSLRPQPTELKYSWKTWGK
WNV     ENGVDLSVVVEKQEGMYKSAPKRLTATTEKLEIGWKAWGK
        ** *.*:::. . :.  . . . *     :*  .**:***
```

Figure 5.6.1: Alignment colored through `pic` descriptor.

Notice that for the `pic` descriptor, the `sequence` column is not showing the color as we saw with the `fsc` descriptor, instead, it is displaying the number of the target protein that is being colored. To find the color, you must see the first element in the list that appears in the `residue` column [for our example the color values will be `6` (Red), `94` (Blue), `75` (Cyan), and `54` (Lime)].



The next elements on the same list will be the highlighted residues[5]. Moreover, the font letter that is being applied for each region will appear in the `style` column. The HTML file is going to display specific regions based on the position index of the processed data (Figure 5.6.1). Make correlations with both that graph and the table above. Note that if the interval matches with a gap, this region will not be colored. For instance, the interval `'1-2'` (sequence$_1$/ZIKV-BR) presents a gap at position 2 (Figure 3.2.1), as this position coincides with a Lys residue for DENV2 and one Ser residue for ZIKV-BR, no color should be assigned to that region of ZIKV-BR, due to the Ser is not a target-residue.

```
           1                 20                  40                  60
ZIKV-UG DVGCSVDFSKKETRCGTGVFIYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEEGICGISSVSRMENIMWKSVEGELNAILE
ZIKV-BR DVGCSVLFSKKETHCGTGVFVYNDVEAWRDRYKYHPDSPRRLAAAVKQAWEDGICGISSVSRMENIMWRSVEGELNAILE
DENV2   DSGCVVSWKNKELKCGSGIFITDNVHTWTEQYKFQPESPSKLASAIQKAHEEGICGIRSVTRLENLMWKQITPELNHILS
WNV     DTGCAIDISRQELRCGSGVFIHNDVEAWMDRYKYYPETPQGLAKIIQKAHKEGVCGLRSVSKLEHQMWEAVKDELNTLLK
        * ** :. ..:* .**:*:*: ::* :* :.**: *::*  **  :::*.::*:**: **:*:*: **  :  *** :*.

           81                100                 120
ZIKV-UG ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
ZIKV-BR ENGVQLTVVVGSVKNPMWRGPQRLPVPVNELPHGWKAWGK
DENV2   ENEVKLTIMTGDIKGIMQAGKRSLRPQPTELKYSWKTWGK
WNV     ENGVDLSVVVEKQEGMYKSAPKRLTATTEKLEIGWKAWGK
        ** *.*:::. . :.  . . . *     :*  .**:***
```

Figure 5.6.2: Alignment colored through `pimc` descriptor.

## 5.7 Color by position index and mutation (`pimc`)

The `pimc` descriptor can be declared when it is desired to put color based on an index and distinguish point mutations[6]. `pimc` is a copy of `pic`, so we can simply change `pic` by `pimc` from the previous array to get the following information:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | pimc | [1] | [6...] | True | n | NaN |
| main_0 | 1 | pimc | [1] | [94...] | True | b | NaN |
| main_0 | 1 | pimc | [2] | [75...] | True | d | NaN |
| main_0 | 1 | pimc | [3] | [54...] | True | u | NaN |

[5] In the table they are being replaced by points.

[6] Imagine that this descriptor is equivalent to `fmac` or `tmc` (see Sections 5.5 or 5.9, respectively).

Check that the `mutation` column appears with the Boolean value equal to `True`. If you prefer another color for the mutations, remember that it is also possible to implement the `-mco` flag (see Section 5.13).

The HTML file will be like shown in Figure 5.6.2.

A practical example using this descriptor appears in Section 8.3. Also, the implementation of the `pimc` descriptor appears in a previously published study [3].

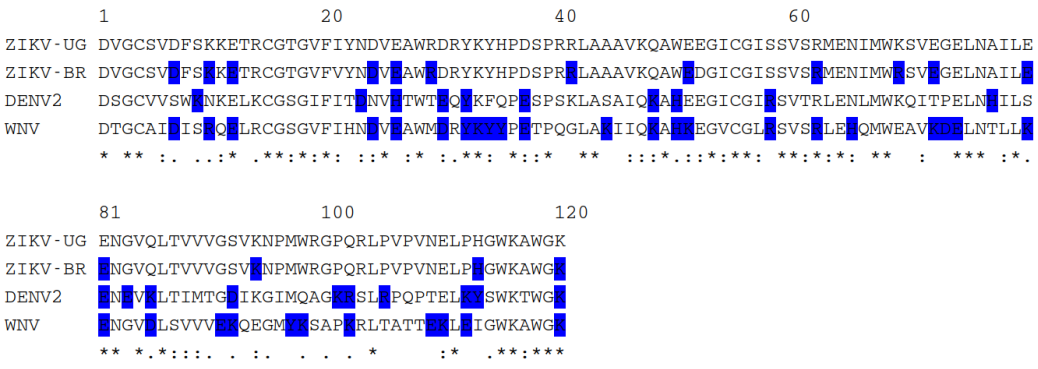



Figure 5.7.1: Alignment colored through `tc` descriptor.

## 5.8 Color by threshold (`tc`)

To discriminate the numeric values of the input data by means of color and using a threshold, it is possible through the following declaration:

```
descriptor_col = { 'tc' :  [94, 0.1, '>='] }
```

The descriptor `tc` is a dictionary where its value is a list with three elements. In the first position of that list, you must put a positive integer number that is related to the type of color desired (see Section 5.16). In the second position, there is an `int`/`float` number that defines the threshold, and the third element, must be a string that refers to a symbol of order relation (*i.e.*, `'>='`, `'<='`, `'>'`, `'<'`, `'=='`, `'!='`). For our example, we are ordering the program to color in Blue only those $\Delta pK_a$ of the target-residues that are greater equal than 0.1.

Subsequently, the log file will report the color number in the `sequence` column, and the two last parameters of the `tc` descriptor list in the `threshold` column, like this:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | tc | [94] | NaN | NaN | NaN | 0.1, >= |

The HTML file will now display a color for each target-residue following the threshold condition (Figure 5.7.1).

A practical example using this descriptor appears in Section 8.2. The implementation of the `tc` descriptor appears in a previously published study [2].

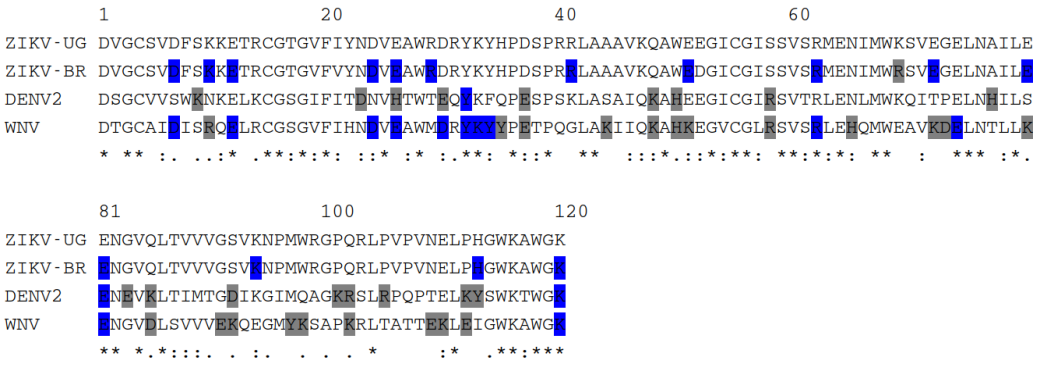



Figure 5.8.1: Alignment colored through `tmc` descriptor.

## 5.9 Color by threshold and mutation (`tmc`)

The descriptor `tmc` is the version of `tc` that will simultaneously discriminate mutations. This follows the same `tc` rules and is invoked as follows:

```
descriptor_col = { 'tmc' :  [94, 0.1, '>='] }
```

The log file will report the next information:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | tc | [94] | NaN | True | NaN | 0.1, >= |

Look that a Boolean value equal to `True` appears in the `mutation` column. If you prefer another color for the mutations remember that it is also possible to implement the `-mco` flag (see Section 5.13).

The HTML file will be like shown in Figure 5.8.1. Note that CPRISMA only taking into account mutations that are matching the declared threshold condition.

## 5.10 p$K_a$ color (pkac)

To use the `pkac` descriptor you must *always* do a $\Delta$ operation among *ionizable residues* (see Section 4.2). To execute this descriptor, your "array_get.py" script should contain the following instructions:

```
descriptor_col = { 'pkac' :  True/False[7] }
```

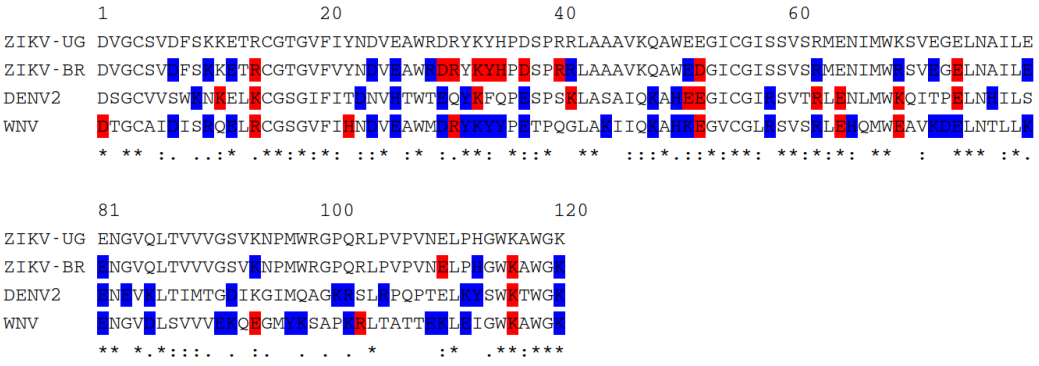



Figure 5.10.1: Alignment colored through `pkac` descriptor and invoking `False` as value.

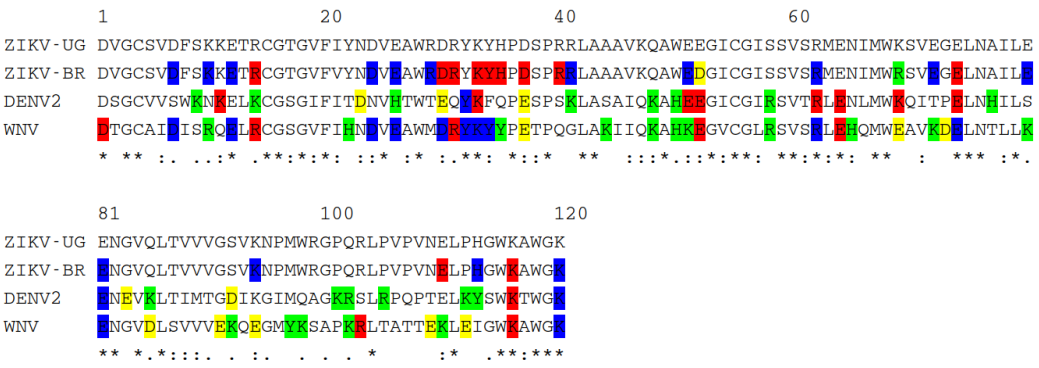



Figure 5.10.2: Alignment colored through `pkac` descriptor and invoking `True` as value.

[7] Select only one Boolean option.



The descriptor `pkac` is able to distinguish between those negative and positive $\Delta pK_a$ of the ionizable groups and will assign Red and Blue colors, respectively. Observe that the `descriptor_col` dictionary has a Boolean variable as a value. You can interpret `False` or `True` for cases where you do not require (as `fac`/`pic`/`tc` descriptors) or require (as `fmac`/`pimc`/`tmc` descriptors) to differentiate point mutations, respectively. If `True`, mutations will be highlighted in two ways:

1. When there is a shift from acid residue to a basic one (Lime color will be assigned).
2. Vice versa (Yellow color will be assigned).

A typical output in the log file when the `pkac` descriptor is invoked with `False` Boolean appears below:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | pkac | NaN | [ionizable] | NaN | NaN | NaN |

*Note: Remember, if you invoke the Boolean* `True` *as a value in the dictionary of* `pkac` *descriptor, then this parameter will be displayed on the* `mutation` *column as* `True`.

Whenever the `pkac` descriptor is run, the `residue` column will show the information: `[ionizable]`. Examples of colored alignments by means of the `pkac` descriptor using `False` and `True` appear in Figures 5.10.1 and 5.10.2, respectively. See that this descriptor can be quite useful for data with negative or positive numbers even if they are not $pK_a$ values!

A practical example appears in Section 8.1. Nevertheless, note that for these cases other features are being incorporated, such as color gradient (see Chapter 6). The implementation of the `pkac` descriptor appears in a previously published study [1].

## 5.11 Color descriptors dictionary (`-dco`)

When the `multiple` comparison method is executed, maybe you need to distinguish each reference/target(s) relationship with a specific color descriptor. For that, we must build an array of color descriptors through the `dict_col` variable of the script "array_get.py". To build it, you should follow the same rules as for `dict_ope` (see Section 4.5). For this example, we used the same dictionary for the `multiple` method as a base (see Subsection 3.10.3). Then,



our CPRISMA color-array will be as follows[8]:

```
dict_col = { 'main_0' :  [ { 'pimc' :  [ { '1' :  [63, 'b', '0-20'] } ,
                              { '3' :  [96, 'u', '20-35'] } ] } , { 'ssc' } ] ,
           'main_1' :  [ { 'fsc' :  [2,80] } , { 'pkac' :  True } ] ,
           'main_2' :  [ { 'fac' :  ['6, 20, 22, 54, 94, 46'] } ] }
```

Now we can execute the `-dco` flag to call the array `dict_col`:[9]

```
cprisma -tr -ck -rf multiple -va 2 -hc -sco 20 -mco 75 -dco
```

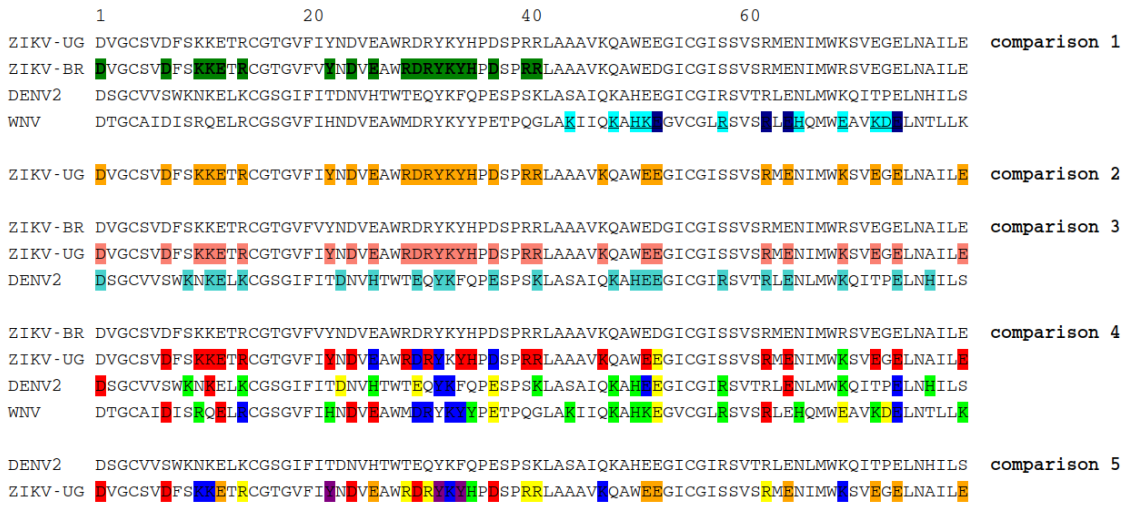



Figure 5.11.1: Alignment for `multiple` comparison method and colored through an array of color descriptors `dict_col`. Not all alignment is being shown.

The log file will show the relationships between each comparison and color descriptor:

```
Comparing dict_ref with dict_col:

                              {'0':  [1, 2, 3]} ........  { 'pimc' :  ...  }
```

[8]Note that we have differentiated each color descriptor for each comparison with a different color to aid the reader.

[9]Notice that additional parameters are being implemented.



```
{'0':  []} ........  { 'ssc' }

{'1':  [0, 2]} ........  { 'fsc' :  [2,80] }

{'1':  [0, 2, 3]} ........  { 'pkac' :  True }

{'2':  [0]} ........  { 'fac' :  [6, 20, 22...]  }
```

```
The array of the feature 'color' dict_col is compatible with the array of
comparison sequences dict_ref!
```

... and a table with a summary for all color descriptor information per comparison:

| main | comparison | descriptor | sequence | residue | mutation | style | threshold |
|---|---|---|---|---|---|---|---|
| main_0 | 1 | pimc | [1] | [63...] | True | b | NaN |
| main_0 | 1 | pimc | [3] | [96...] | True | u | NaN |
| main_0 | 2 | ssc | [same] | NaN | NaN | NaN | NaN |
| main_1 | 3 | fsc | [2, 80] | NaN | NaN | NaN | NaN |
| main_1 | 4 | pkac | NaN | [ionizable] | True | NaN | NaN |
| main_2 | 5 | fac | NaN | [6...] | NaN | NaN | NaN |

Figure 5.11.1 shows the alignment for the typed commands and the built `dict_col` array.

*Note: If the array of* `dict_col` *is wrongly built, all descriptors will be transformed to* `nc` *(default parameter) for each comparison and additional information about where the errors are will be displayed.*

*Note: Remember that these complex color-arrays can be simultaneously combined (or not) with other transcendental CPRISMA arrangements of the other features as well (i.e., operation, visualization, and maximum restriction described in* Sections 4.5, 6.1, *and* 7.14*). See also the practical examples in* Chapter 8.

## 5.12 Color sequence (`-sco`)

To change the default color of the `ssc` descriptor, you should leave your variable `descriptor_col` as declared above in Section 5.2 and execute the `-sco` flag followed by a positive integer number in the terminal, as follows:



```
cprisma -tr -ck -sco 51
```

Remember that this number should be equivalent to a positive integer number less equal than 142. For the command above we have changed the alignment color from Red to GreenYellow, which is at row-position 51 when `-lco` is executed (see Section 5.16).

## 5.13 Color mutation (`-mco`)

It is possible to change the color of the mutations by executing a command like the one below:

```
cprisma -tr -ck -mco 75
```

Remember to use the same rules as for the `-sco` flag (see Section 5.12). For the command above, the color of point mutations was changed from Gray to Cyan.

*Note: The flag* `-mco` *only works for* `fmac`, `pimc`, *and* `tmc` *color descriptors (see* Sections 5.5, 5.7, *and* 5.9, *respectively).*

## 5.14 Color intensity (`-ico`)

Sometimes when we apply a gradient color (see Chapter 6) the intensity of this can not be the best due to several problems (as mentioned earlier in Section 3.2). An alternative to fix these dilemmas is to multiply the intensity of the color by a positive integer number invoking the `-ico` flag. As an example, you can execute it as appears below:

```
cprisma -tr -ck -ico 4
```

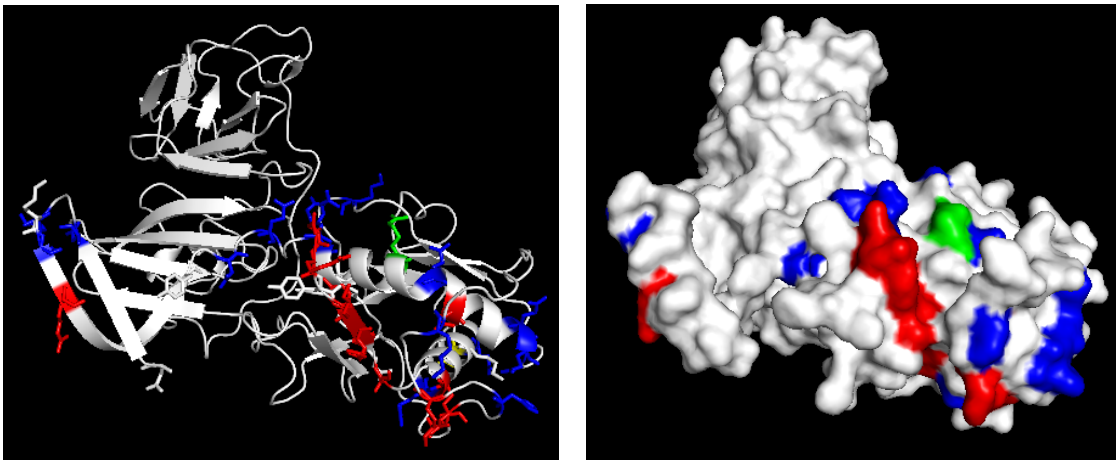
For the previous command line, we are multiplying by 4 the intensity of the color. We recommend the user to test with gradient colors and implement `-ico`.

## 5.15 Color on three-dimensional structure (`-tco`)

To observe the regions colored by CPRISMA at the three-dimensional structure level, the `-tco` flag can be useful. Using the case of Section 5.10 as an example[10], we can execute:

```
cprisma -tr -ck -tco
```

[10] Differences between mutations are being considered.



(a) (b)

Figure 5.15.1: Three-dimensional structure for the monomeric $NS1_{ZIKV\text{-}BR}$ (PDB id 5GS6) using the `pkac` descriptor. (a) Ribbons representation (target-residues appear as sticks). (b) Surface representation. Visualizations were made from Pymol scripts generated through the `-tco` flag.

This command will generate a directory called "3D_representation" with various Pymol scripts for all proteins and comparisons. To visualize the protein, type the next command line:

```
pymol your-protein.pdb 3D_representation/main_0/main_0_comparison1_your-protein.pml
```

An example for ZIKV-BR appears in Figure 5.15.1. Notice that only the target-residues are taken into account. Furthermore, only the first 120 amino acids are being evaluated.

## 5.16 List of colors (`-lco`)

To have quick access to the list of colors you can execute the command:

```
cprisma -lco
```

Figure 5.16.1 shows the first 10 rows displayed in the terminal as an example. See that an index appears numbering each color. This index will be useful to access the colors that CPRISMA applies. Remember, these colors are based on the website https://htmlcolorcodes.com/color-names/.

*Note: The possible number of colors is from 0 to 142.*

| | General | Name | Hex | RGB |
|---|---|---|---|---|
| 0 | Red HTML | IndianRed | #CD5C5C | rgb(205, 92, 92 |
| 1 | Red HTML | LightCoral | #F08080 | rgb(240, 128, 128 |
| 2 | Red HTML | Salmon | #FA8072 | rgb(250, 128, 114 |
| 3 | Red HTML | DarkSalmon | #E9967A | rgb(233, 150, 122 |
| 4 | Red HTML | LightSalmon | #FFA07A | rgb(255, 160, 122 |
| 5 | Red HTML | Crimson | #DC143C | rgb(220, 20, 60 |
| 6 | Red HTML | Red | #FF0000 | rgb(255, 0, 0 |
| 7 | Red HTML | FireBrick | #B22222 | rgb(178, 34, 34 |
| 8 | Red HTML | DarkRed | #8B0000 | rgb(139, 0, 0 |
| 9 | Pink HTML | Pink | #FFC0CB | rgb(255, 192, 203 |
| 10 | Pink HTML | LightPink | #FFB6C1 | rgb(255, 182, 193 |

Figure 5.16.1: First 10 rows when invoked `-lco` command.



# 6 Visualization (`descriptor_vis`)

Previously, in Chapters 3 and 5, we already saw some commands and variables that alter the display of alignments. In this Chapter, we will give an expansion focusing on three possible functionalities available on CPRISMA through the so-called "visualization" feature:

1. Reference display (`Re`).
2. Target-sequence(s) degrade (`De`).
3. Visualization of the letters (`Le`).

These three variables will be the descriptors of the "visualization" feature, and they are always declared simultaneously through the `descriptor_vis` of the "array_get.py" script (Figure 2.2.3), as follows:

```
descriptor_vis = { 'ReY', 'DeN', 'LeY' }
```

The `descriptor_vis` is a set variable with three elements. Besides, each descriptor is accompanied by a letter `Y` (Yes) or `N` (No), which indicates whether or not this visualization feature is enabled, respectively.

Suppose we respectively declare the operation (see Chapter 4) and color (see Chapter 5) variables like this,

```
descriptor_ope = { 'd' }
```

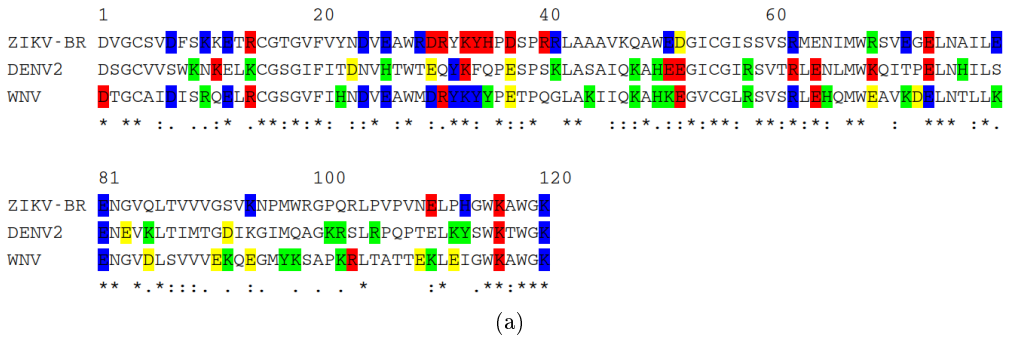
```
descriptor_col = { 'pkac' :  True }
```

... and we run the next command[1]:

```
cprisma -tr -ck
```

You may have noticed that your log file shows the following information:

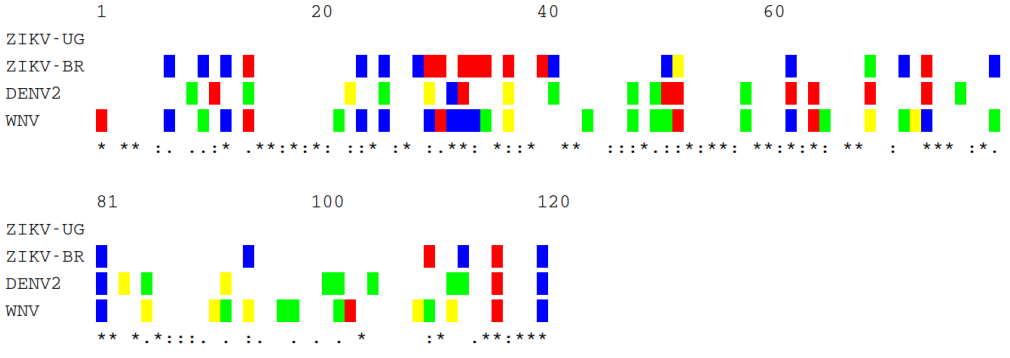
[1] It is important to mention that for most of the examples in this Chapter we have used this command and these variables as declared.




(a)




(b)




(c)

Figure 6.0.1: Several alignments using different visualization descriptors. (a) `Re` descriptor with `False` Boolean (`N`). (b) `De` descriptor with `True` Boolean (`Y`). (c) `Le` descriptor with `False` Boolean (`N`).



| | main | comparison | reference | degraded | letters |
|---|---|---|---|---|---|
| main_0 | | 1 | True | False | True |

Now, for each visualization descriptor, a Boolean variable is shown based on the relations: `Y` (`True`) and `N` (`False`).

*Note: Sometimes, some descriptors can be declared wrongly. To avoid that the program stops, it will always return to* `'ReY', 'DeN', 'LeY'` *descriptors (default parameters), with some additional error messages in your log file, which will guide you to correct the problem.*

Early, the output alignment for these initial parameters was displayed (Figure 5.10.2). But now, let us change reference descriptor from `ReY` to `ReN`, as follows:

```
descriptor_vis = { 'ReN', 'DeN', 'LeY' }
```

As shown in Figure 6.0.1a, the "reference line" has disappeared. Since this sequence may not be relevant because it does not have any color, the `Re` descriptor allows us to remove it or not as appropriate.

Enabling the degrade, like this:

```
descriptor_vis = { 'ReY', 'DeY', 'LeY' }
```

... we can see the effect that our operations have when it is invoked the `d` descriptor on the `descriptor_ope` (Figure 6.0.1b). Check that the color is on a scale from $-13.3$ to 13.3. This scale is normalized based on the maximum value derived from $\Delta$ operation[2]. We observe that the point mutations have a greater impact on the rest of the $\Delta \mathrm{p}K_a$ values. Nevertheless, it is possible to apply maximum restriction methods to improve the color alignment display (see Chapter 7).

Finally, if we are not interested in visualizing the letters of the protein sequences, you could eliminate them by means of the following declaration:

```
descriptor_vis = { 'ReY', 'DeN', 'LeN' }
```

Figure 6.0.1c shows an alignment without the letters.

We recommend checking the practical examples in Sections 8.1 and 8.2 for a better understanding of the visualization descriptors.

[2]See explicitly the $\Delta$ calculations in your log file with the `-t` flag (see Section 3.11).



## 6.1 Visualization descriptors dictionary (-dvi)

When it is applied complex comparisons with the `multiple` method (see Subsection 3.10.3), perhaps you will require to make a specific visualization based on `Re`, `De`, and `Le`, for each reference/target(s) relationship. For this, we can invoke the `-dvi` flag to get a visualization descriptors dictionary `dict_vis` from the "array_get.py" script. It is important to mention that the same rules as already described for `dict_ope` were followed (see Section 4.5). For our example, the `dict_col` arrangement mentioned in Section 5.11 was employed. On the other hand, it will be assumed to be `d` for all comparisons (*i.e.*, `descriptor_ope = { 'd' }`) and the same array of comparisons described in Subsection 3.10.3. Our `dict_vis` variable will be declared as follows[3]:

```
dict_vis = { 'main_0' :  [ { 'ReN', 'DeN', 'LeY' } , { 'ReY', 'DeY', 'LeY' } ] ,

             'main_1' :  [ { 'ReY', 'DeN', 'LeN' } , { 'ReY', 'DeY', 'LeY' } ] ,

             'main_2' :  [ { 'ReN', 'DeY', 'LeY' } ] }
```

Now we can execute the `-dvi` flag to call the array `dict_vis`[4]:

```
cprisma -tr -ck -rf multiple -va 2 -hc -sco 20 -mco 75 -dco -dvi
```

The log file will show the relationships between each comparison and visualization descriptors:

```
Comparing dict_ref with dict_vis:

                                {'0':  [1, 2, 3]} ........  ['ReN', 'DeN', 'LeY']

                                {'0':  []} ........  ['ReY', 'DeY', 'LeY']

                                {'1':  [0, 2]} ........  ['ReY', 'DeN', 'LeN']

                                {'1':  [0, 2, 3]} ........  ['ReY', 'DeY', 'LeY']

                                {'2':  [0]} ........  ['ReN', 'DeY', 'LeY']

The array of the feature 'visualization' dict_vis is compatible with the array
of comparison sequences dict_ref!
```

[3] Note that we have differentiated each set of visualization descriptors for each comparison with a different color to aid the reader.

[4] Notice that additional parameters are being implemented.

... and a table with a summary for all visualization descriptor information per comparison:

| main | comparison | reference | degraded | letters |
|---|---|---|---|---|
| main_0 | 1 | False | False | True |
| main_0 | 2 | True | True | True |
| main_1 | 3 | True | False | False |
| main_1 | 4 | True | True | True |
| main_2 | 5 | False | True | True |

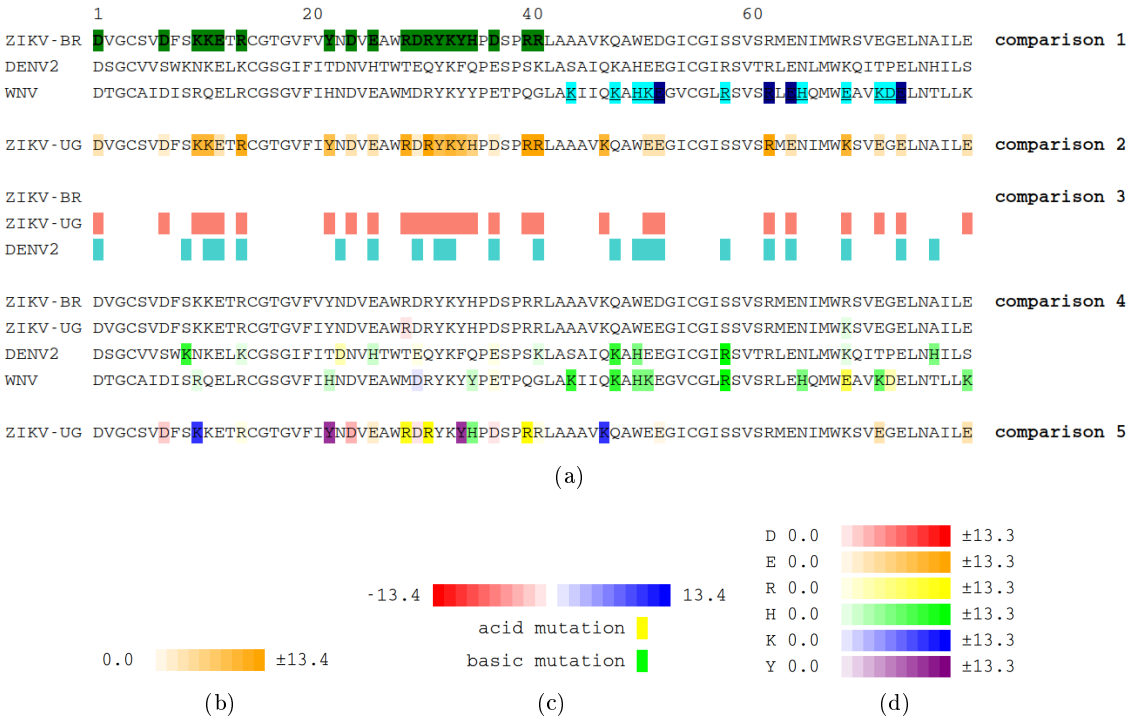



Figure 6.1.1: (a) Alignment for `multiple` comparison method and colored through the arrays of color (`dict_col`) and visualization (`dict_vis`) descriptors. Not all alignment is being shown. (b) Scale for comparison 2. (c) Scale for comparison 4. (d) Scale for comparison 5.

The HTML output for the previous conditions invoked is displayed in Figure 6.1.1a. Make the respective relationships between the previous table and that graph. Since we are invoking `DeY` for comparisons `2`, `4`, and `5`, we are obtaining, respectively, their scales in the log file (Figures 6.1.1b, 6.1.1c, and 6.1.1d). Although comparison `2` does not have any target-sequence (*i.e.*,



`{'0':[]}`), the gradient color that you are seeing for this case is related to the reference protein. The case of comparison `4` is similar to what we saw earlier (Figure 6.0.1b), but now the reference sequence is ZIKV-BR. Finally, see that comparison `5` has separated a different color scale for each target-residue. This is because the color descriptor `fac` is being declared for that comparison (see Section 5.4). Nevertheless, the scale is normalized based on the greater value of the result of the operation `d` that is being applied for that comparison (*i.e.*, ±13.3). You can see the independent behavior for each amino acid, but for that, you need to apply special maximum restriction methods (see Section 7.13).

*Note: Whether the array of* `dict_vis` *is wrongly built, all descriptors will be transformed to* `ReY, DeN, LeY` *(default parameters) for each comparison and additional information about where the errors are will be displayed.*



# 7 Maximum restriction (`descriptor_mxr`)

To understand how the CPRISMA feature called "maximum restriction" works, let us suppose the next declared variables in the "array_get.py" script[1]:

```
descriptor_ope = { 'd' }

descriptor_col = { 'pkac' :  True }

descriptor_vis = { 'ReY', 'DeY', 'LeY' }
```

Then the following command is run in terminal[2]:

```
cprisma -tr -ck -t
```

The HTML output for this example was already cited earlier (Figure 6.0.1b). Although this output is correct, we are not clearly observing the p$K_a$ shifts for many target-residues. Since we have invoked the `-t` flag (see Section 3.11), our log file will show in the operation feature part, how the $\Delta$ data is being obtained (see Figure 7.0.1 as an example). The value of $\Delta$p$K_a$ highlighted in row 13 (*i.e.*, $-13.3$) is based on the subtraction $\text{p}K_{a-\text{DENV2}} - \text{p}K_{a-\text{ZIKV-UG}}$. This value matches with a mutation, due to a genetic indel (see row 13 in Figure 3.2.1). Thus, the scale of Figure 6.0.1b is being normalized as a function of this value, which corresponds to the maximum among the entire data set. Nonetheless, the regions that should be more Red or Blue cannot display that color because their $\Delta\text{p}K_a$ are very low with respect to the greater value. For instance, when is calculated $\left|0.1^{\Delta\text{p}K_a} / -13.3^{\Delta pK_a(\max)}\right|$, it is observed that 0.1 is $7.5 \times 10^{-3}$ smaller than the maximum value. From the point of view of the HTML code, this is negligible and the color should be equal to White.

To improve the color displayed when is invoked a gradient by `DeY` (see Chapter 6), we can use a maximum restriction method like: `rm`, `ra`, `ram`, `rpi`, `rt`, `rs`, `rsm`, `rsa`, `rsam`, `rspi`, and `rspim`, through the variable `descriptor_mxr` of the script "array_get.py" (Figure 2.2.3).

[1] For a further explanation of these variables see Chapters 4, 5, and 6, respectively.
[2] Most of the examples below are using this command line.



| | ZIKV-UG | X | ZIKV-BR | DENV2 | WNV | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | | 1 | 2 | 3 | d | | |
| 0 | 3.9 | | 3.9 | 3.9 | 3.8 | -> | 0.0 | 0.0 | -0.1 |
| 1 | 3.1 | | 3.4 | 0.0 | 3.4 | -> | 0.3 | -3.1 | 0.3 |
| 2 | 0.0 | | 0.0 | 10.5 | 0.0 | -> | 0.0 | 10.5 | 0.0 |
| 3 | 10.9 | | 11.0 | 0.0 | 12.8 | -> | 0.1 | -10.9 | 1.9 |
| 4 | 10.7 | | 10.7 | 10.6 | 0.0 | -> | 0.0 | -0.1 | -10.7 |
| 5 | 3.8 | | 4.0 | 3.8 | 4.0 | -> | 0.2 | 0.0 | 0.2 |
| 6 | 12.6 | | 12.3 | 10.7 | 12.4 | -> | -0.3 | -1.9 | -0.2 |
| 7 | 10.0 | | 10.0 | 0.0 | 6.7 | -> | 0.0 | -10.0 | -3.3 |
| 8 | 0.0 | | 0.0 | 3.4 | 0.0 | -> | 0.0 | 3.4 | 0.0 |
| 9 | 3.5 | | 3.8 | 0.0 | 3.6 | -> | 0.3 | -3.5 | 0.1 |
| 10 | 4.0 | | 4.1 | 6.8 | 4.4 | -> | 0.1 | 2.8 | 0.4 |
| 11 | 12.6 | | 13.4 | 0.0 | 0.0 | -> | 0.8 | -12.6 | -12.6 |
| 12 | 3.1 | | 2.7 | 4.1 | 3.5 | -> | -0.4 | 1.0 | 0.4 |
| 13 | 13.3 | | 13.0 | 0.0 | 12.9 | -> | -0.3 | -13.3 | -0.4 |
| 14 | 9.6 | | 9.6 | 10.0 | 9.8 | -> | 0.0 | 0.4 | 0.2 |
| 15 | 11.1 | | 11.0 | 11.0 | 11.2 | -> | -0.1 | -0.1 | 0.1 |
| 16 | 10.0 | | 9.7 | 0.0 | 10.2 | -> | -0.3 | -10.0 | 0.2 |
| 17 | 6.3 | | 6.2 | 0.0 | 9.5 | -> | -0.1 | -6.3 | 3.2 |

Figure 7.0.1: Example of the p$K_a$ data for each target protein and its $\Delta$p$K_a$ obtained after running `d` as an operation. The number highlighted in blue refers to the maximum value for this data set. All the information has not been shown.

## 7.1 No restriction (`nr`)

When it is not considered maximum restrictions we declare the `descriptor_mxr` as follows:

```
descriptor_mxr = { 'nr' }
```
[3]

Note that `nr` descriptor is a string variable inside a pair of brackets.

When we run CPRISMA, our log file will show the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | nr | NaN | NaN | NaN | NaN | NaN | NaN |

This statement is simply saying that the `nr` descriptor was considered for comparison 1, in the `main_0` comparison set (for now ignore the rest of the columns in that log message). Besides this table, in the maximum restriction part of the log file, you will find a list with all values for the operations invoked, like this[4]:

[3] For all alignment examples of the preceding Chapters, this was the variable executed.
[4] These lists only appear when `-t` flag is executed (see Section 3.11).



```
list_max = [0.0, 0.3, 0.0, 0.1, 0.0, 0.2, -0.3, 0.0, 0.0, 0.3, 0.1, 0.8, -0.4, -0.3, 0.0,
-0.1, -0.3, -0.1, -0.5, -0.2, 0.1, 0.0, 0.0, 0.0, 0.0, 0.1, -0.6, 0.0, 0.1, 0.0, 0.0,
1.7, 0.2, 0.0, -0.1, 0.0, 0.1, 0.3, 0.0, 0.0, 0.0, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, -0.2, 0.0, 0.1, -0.5, 0.3, 0.0, -3.1, 10.5, -10.9, -0.1, 0.0, -1.9, -10.0,
3.4, -3.5, 2.8, -12.6, 1.0, -13.3, 0.4, -0.1, -10.0, -6.3, 0.7, -13.0, -1.8, 0.0, -11.0,
10.9, 6.4, -0.3, -0.8, 12.9, -0.2, -0.1, 0.0, -0.2, -3.8, 0.0, -0.2, 6.8, -4.2, 0.1, 4.2,
10.8, 0.0, 3.5, 0.0, 0.0, 0.0, -12.7, 10.2, 12.4, -12.9, 12.6, 0.0, 0.0, 11.2, 3.3, -
0.2, 0.1, -0.1, 0.3, 0.0, 1.9, -10.7, 0.2, -0.2, -3.3, 0.0, 0.1, 0.4, -12.6, 0.4, -0.4,
0.2, 0.1, 0.2, 3.2, 0.7, -13.0, -12.9, 10.7, -11.0, 10.6, 6.1, 6.5, -0.9, 13.2, 0.1, -
0.2, 6.8, -7.3, 7.9, 3.4, 0.1, 0.0, 7.2, 0.2, 0.0, 3.7, 4.0, 11.5, -6.7, 9.9, 10.9, -
12.7, 0.0, 10.2, -0.5, 0.0, 3.6, 7.6, 4.2, -6.4, -0.1, 0.2]
```

A summary of this information should appear immediately, as follows:

| main | minimum (abs) | maximum (abs) | lenght |
|---|---|---|---|
| main_0 | 0.0 | 13.3 | 168 |

Notice that this table is showing the maximum and minimum numbers for the last list after applying the absolute value. This is because the HTML code will only work with positive numbers. Furthermore, the column `length` is showing us the total number of elements for that list ($56^{\text{rows}} \times 3^{\text{target-proteins}} = 168$). As no restriction applies, all data is being considered.

*Note: Sometimes, some descriptors can be declared wrongly. To avoid that the program stops, it will always return to the* `nr` *descriptor (default parameter), with some additional error messages in your log file, which will guide you to correct the problem.*

## 7.2 Restriction by mutation (`rm`)

For our example, notice that most of the high $\Delta \text{p}K_a$ values correspond to regions where mutations occurred (compare Figures 3.2.1 and 7.0.1). We can declare the `rm` descriptor to avoid taking into account these "high $\Delta \text{p}K_a$ values" through:

```
descriptor_mxr = { 'rm' }
```

After executing the program, our table for this CPRISMA feature will appear in the log file like this:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rm | NaN | NaN | NaN | True | NaN | NaN |

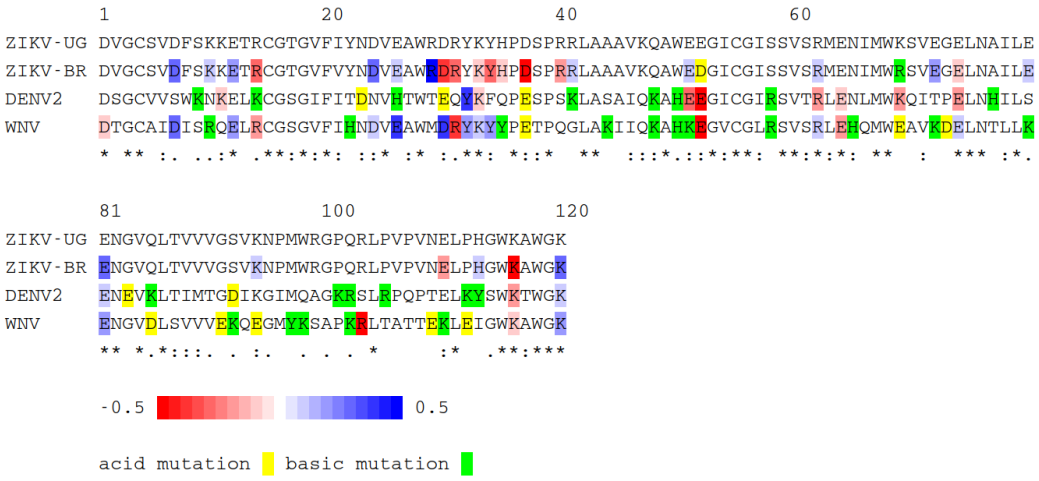



Figure 7.2.1: Alignment colored applying `rm` descriptor.

Notice that the `mutation` column has a Boolean equal to `True`. The list of $\Delta pK_a$ values that are being considered to normalize the color now is reduced:

```
list_max = [0.0, 0.2, 0.0, -0.1, 0.1, 0.0, -0.1, 0.3, -0.5, 0.3, 0.0, 0.0, 0.4, -0.1, -0.2,
-0.1, -0.2, 0.1, -0.2, 0.1, -0.1, 0.2, 0.2, 0.1, 0.1, -0.2, 0.1, 0.2, -0.1, 0.2]
```

| main | minimum (abs) | maximum (abs) | lenght |
|---|---|---|---|
| main_0 | 0.0 | 0.5 | 30 |

*Note: For the following sections, this information will not be displayed or discussed anymore. When a maximum restriction is invoked, it is understood that CPRISMA is removing or transforming the data to normalize and improve the color visual quality of the alignment.*

Changing the scale of these data allows us to observe with greater definition the color for each $pK_a$ shifts (Figure 7.2.1). At this point, we can finally see the impact of the numerical data of the CSV input file after the $\Delta$ operation is applied.

## 7.3 Restriction by amino acid (`ra`)

Suppose you do not want to consider the numerical values of some target-residues (see Section 3.1). To delete them you can invoke the `ra` descriptor, as mentioned bellow:

```
descriptor_mxr = { 'ra' :  ['R', 'K'] }
```

The dictionary structure of the `ra` descriptor is composed by a key as a string (*i.e.*, `'ra'`) and



a value as a list of target-residues (*i.e.*, `['R', 'K']`). For this example, CPRISMA is being instructed to do not take into account the numerical values of the row where Arg and Lys residues match in the processed data (for instance, row 2, 3, and 4 in Figure 3.2.1).

*Note: When the* `ra` *descriptor is run, CPRISMA will looks for all the rows where the residue declared appears and the program will "remove" the entire row if it matches. Later, CPRISMA will calculate the maximum from these transformations. However, it is possible to use other specific descriptors like* `rsa` *(see* Section 7.9*) to have a more refined definition of maximum restriction.*

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | ra | NaN | ['R', 'K'] | NaN | NaN | NaN | NaN |

See that the `residue` column is showing the list declared. The color scale is now normalized based on this condition.

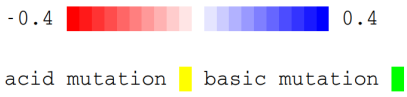



Figure 7.3.1: Scale example invoking `ram` descriptor.

## 7.4 Restriction by amino acid and mutation (`ram`)

The numerical values of target-residues and mutations can simultaneously constrained by means of:

```
descriptor_mxr = { 'ram' :  ['R', 'K'] }
```

The `ram` descriptor is a combination of `rm` (see Section 7.2) and `ra` (see Section 7.3), therefore it will follow the same rules as for those descriptors.

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | ram | NaN | ['R', 'K'] | NaN | True | NaN | NaN |



Notice that the `mutation` column has a Boolean equal to `True`. Removing rows with Arg and Lys residues has changed the color scale. Now, it is displays a maximum of ±0.4 instead of ±0.5 (Figures 7.3.1 and 7.2.1, respectively).

## 7.5 Restriction by index position (`rpi`)

To have a maximum restriction for the exact position of specific operation values, the variable `descriptor_mxr` can be declared like this:

```
descriptor_mxr = { 'rpi' :  [13, 45, 48] }
```

The `rpi` descriptor has a dictionary structure similar to `ra` (see Section 7.3). Although the list has numbers that match with the row index of the processed data (Figures 3.2.1 and 7.0.1), instead to target-residues. Remember, the numbers of the list should be related to that index and not the alignment position!

*Note: When the* `rpi` *descriptor is typed, CPRISMA will looks for all the rows based on the index positions declared and the program will "remove" the entire row if matches. Later, CPRISMA will calculate the maximum from these transformations. Nevertheless, it is possible to use other specific descriptors like* `rspi` *(see* Section 7.11*) to have a more refined definition of maximum restriction.*

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rpi | NaN | NaN | [13, 45, 48] | NaN | NaN | NaN |

See that in the `position` column, the list declared for the `rpi` descriptor appears. Based on this, the color scale will be normalized.

## 7.6 Restriction by threshold (`rt`)

A maximum restriction based on a threshold can be invoked through:

```
descriptor_mxr = { 'rt' :  [0.0, '>='] }
```

Look that `rt` descriptor has a dictionary structure. The value of the dictionary is a list with two elements where the first shows a numerical value (float type), meanwhile the second is a



string that refers to a symbol of order relation (*i.e.*, `'>='`, `'<='`, `'>'`, `'<'`, `'=='`, `'!='`). Based on our example, we are telling the program to only take into account those rows with values greater equal than 0.0. Whether the row presents any number in any column that does not meet that condition, then the value will be converted to 0.0 immediately.

*Note: Occasionally, too strict threshold can be invoked and this could cause all values to be ignored. To avoid the program stops, by default, CPRISMA will normalize everything based on* 1.0.

The table when invoked `rt` descriptor looks like this:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rt | NaN | NaN | NaN | NaN | [0.0, >=] | NaN |

When `-t` flag is invoked (see Section 3.11), the log file will display the following information:

| main | minimum (abs) | maximum (abs) | lenght |
|---|---|---|---|
| main_0 | 0.0 | 13.2 | 168 |

See that the `length` column reports the same amount of data as when we have no restriction (*i.e.*, when `nr` is applied [see Section 7.1)]. This is because all values that do not satisfy the threshold condition are being converted to 0.0 and are not being "eliminated".

Finally, the new maximum for the color scale will be 13.2. Our maximum constraint has not changed much despite the conversion of the highest negative value (*i.e.*, from $-13.3$ to 0.0). As the threshold condition cited above does not take into account values less than 0.0, the problem persists because, even with this restriction, the data still have "high" positive values like 13.2.

*Note: Based on the last statement, it is important to mention that* `rt` *descriptor is more suitable for data with only one type of mathematical sign. For that, you could use* `da` *as an operation descriptor (see* Section 4.3*).*

## 7.7 Restriction by sequence (`rs`)

In CPRISMA you can not consider the data of a whole sequence by invoking `descriptor_mxr` like this:



```
descriptor_mxr = { 'rs' :  [2] }
```

The dictionary structure of the `rs` descriptor is similar to `ra` (see Section 7.3) or `rpi` (see Section 7.5). But now, the number in the list refers to a target-sequence.

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rs | [2] | NaN | NaN | NaN | NaN | NaN |

For our example, we are not taking into account the data of the DENV2 sequence. Based on this, the color scale will be normalized.

*Note: Regarding the descriptor* `rs`*, the* `pair` *comparison method, array of comparisons like* `{# :  []}`*, or cases where* `-ck` *is not invoked, CPRISMA will always return to the default value of the maximum restriction feature (i.e.,* `nr`*).*

If you want to restrict the maximum values of target-residues or row position index from a specific sequence, we recommend Sections 7.9 and 7.11, respectively.

## 7.8 Restriction by sequence and mutation (`rsm`)

The data from a sequence and the places where mutations appear can simultaneously "removed" by means of:

```
descriptor_mxr = { 'rsm' :  [2] }
```

The descriptor `rsm` is a combination of `rm` (see Section 7.2) and `rs` (see Section 7.7), so the same rules will apply for it.

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rsm | [2] | NaN | NaN | True | NaN | NaN |

Notice that the `mutation` column has a Boolean equal to `True`. The scale of your HTML output will be normalized based on these parameters.



## 7.9 Restriction by sequence and amino acid (rsa)

Sometimes, it could be very radical to apply `ra` (see Section 7.3) due to many rows with useful values may be being removed. Nevertheless, to focus only on target-residues of a specific sequence, the `rsa` descriptor can be invoked. To declare this maximum restriction descriptor you can do it as follows:

```
descriptor_mxr = { 'rsa' : [{ '1' : ['R', 'K'] }, { '2' : ['D', 'Y'] }] }
```

Notice that we saw this array structure before (see Subsection 3.10.3). Now, our value is a list of dictionaries. Each second-key represents the target-sequences of interest (*i.e.*, `'1'` and `'2'`) and the second-values are the lists of target-residues to be discard (*i.e.*, `['R', 'K']` and `['D', 'Y']`, respectively). For our example, CPRISMA is being instructed to just consider the sequences of ZIKV-BR (`'1'`) and DENV2 (`'2'`) and "to delete" the Arg and Lys residues; and the Asp and Tyr residues of that proteins, respectively. The descriptor `rsa` follows similar rules like `ra` but just examining the target-residues of the target-sequence of interest.

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rsa | [1] | ['R', 'K'] | NaN | NaN | NaN | NaN |
| main_0 | 1 | rsa | [2] | ['D', 'Y'] | NaN | NaN | NaN | NaN |

The color scale will be normalized based on these conditions.

*Note: Regarding the descriptor* `rsa`*, the* `pair` *comparison method, array of comparisons like* `{# : []}`*, or cases where* `-ck` *is not invoked, CPRISMA will always return to the default value of the maximum restriction feature (i.e.,* `nr`*).*

## 7.10 Restriction by sequence, amino acid, and mutation (rsam)

We can use a version of `rsa` that also restricts mutations simultaneously. For that, we invoke the maximum restriction variable by means of:

```
descriptor_mxr = { 'rsam' : [{ '1' : ['R', 'K'] }, { '2' : ['D', 'Y'] }] }
```

The `rsam` is a combination of `rm` and `rsa` descriptors (see Sections 7.2 and 7.9, respectively), so it will follow similar rules as them. For this case, `rsam` will not take into account the mutations



that occur between the target-sequence of the second-keys [for our example, `'1'` (ZIKV-BR) and `'2'` (DENV2)] and the reference sequence [for our example, `'0'` (ZIKV-UG)].

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rsa | [1] | ['R', 'K'] | NaN | True | NaN | NaN |
| main_0 | 1 | rsa | [2] | ['D', 'Y'] | NaN | True | NaN | NaN |

Notice that the `mutation` column has a Boolean equal to `True`. The color scale will be normalized based on these conditions.

## 7.11 Restriction by sequence and index position (`rspi`)

The `rspi` descriptor is similar to `rsa`, however, the second-value must have the index positions of the processed data (see the index in Figure 7.0.1). To declare `rspi`, we can do it by means of:

```
descriptor_mxr = { 'rspi' :  [ { '1' :  [0, 29] } , { '2' :  [48] } ] }
```

The `rspi` descriptor shares the same rules as `rpi` (see Section 7.5) and `rsa` (see Section 7.9). For our example, CPRISMA is being instructed to just "delete" the positions `0` and `29` of ZIKV-BR (`'1'`) and `48` of DENV2 (`'2'`).

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rspi | [1] | NaN | [0, 29] | NaN | NaN | NaN |
| main_0 | 1 | rspi | [2] | NaN | [48] | NaN | NaN | NaN |

The color scale will be normalized based on these conditions.

*Note: Regarding the descriptor* `rspi`*, the* `pair` *comparison method, array of comparisons like* `{# :  []}`*, or cases where* `-ck` *is not invoked, CPRISMA will always return to the default value of the maximum restriction feature (i.e.,* `nr`*).*



## 7.12 Restriction by sequence, index position, and mutation (rspim)

The version of `rspi` that restricts places where point mutations occur is called `rspim` and can be declared as follows:

```
descriptor_mxr = { 'rspim' : [ { '1' : [0, 29] }, { '2' : [48] } ] }
```

The `rspim` is a combination of `rm` and `rspi` descriptors (see Sections 7.4 and 7.11, respectively), so it will follow similar rules as them[5].

The log file will display the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rspim | [1] | NaN | [0, 29] | True | NaN | NaN |
| main_0 | 1 | rspim | [2] | NaN | [48] | True | NaN | NaN |

Notice that the `mutation` column has a Boolean equal to `True`. The scale of your HTML output will be normalized taking into account these parameters.

## 7.13 Restrictions by target-residues (Y)

Sometimes some target-residues have more influence than others when we apply a maximum restriction method. We can evaluate the independent maximum behavior for each target-residue by adding the letter `'Y'` to our descriptor in the following way:

```
descriptor_mxr = { 'nrY' }
```

When we run CPRISMA, our log file will show the following information:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | nr | NaN | NaN | NaN | NaN | NaN | True |

Notice that the `separate` column is showing a Boolean variable as `True`. Whether `-t` flag is invoked (see Section 3.11), the log file will display the following information:

[5] See other considerations on how numerical data is removed at the places where mutations happen in Section 7.10 (although `rsam` and `rspim` are different maximum restriction methods they follow the same pattern of execution)

| main | minimum (abs) | maximum (abs) | lenght |
|---|---|---|---|
| main_0 | 0.0 | [3.7, 7.3, 13.2, 6.8, 11.5, 9.9] | NaN |

Now, the `maximum (abs)` column is displaying the maximum for each target-residue. In other words, the values for each amino acid will be normalized according to their respective maxima. This list and its order will match according to the target-residues that were declared in the tuple `target_residues` (see Section 3.1). The alignment of the HTML file for this new condition appears in Figure 7.13.1. Note that the scale for each target-residue can be observed. A first impression may seem like there is no difference between what we got earlier (Figure 6.0.1b). Although some regions are fainter in color (see some amino acids between positions 20 and 40 on the alignment). The intensity of the color is still not enough because of the maximum values are still very high due to the presence of mutations.

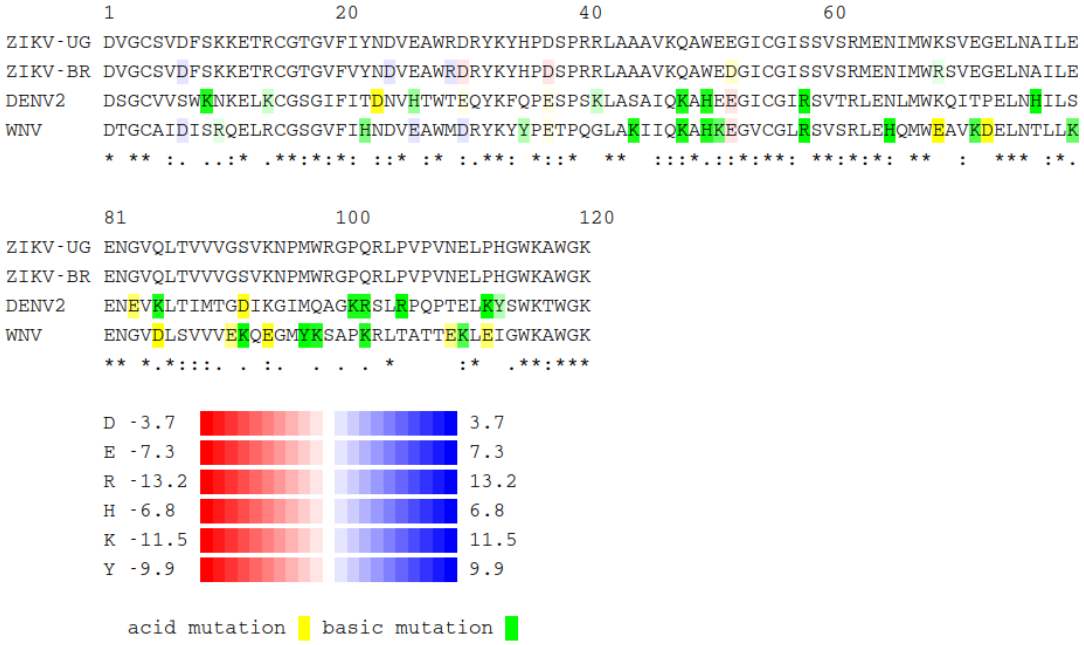



Figure 7.13.1: Alignment colored applying `nr` descriptor and the condition `Y` to separate the maximum per target-residue.

We can confirm this by invoking the color descriptor like this:

```
descriptor_col = { 'fac' :  [6, 20, 22, 54, 94, 46] }
```

Based on that, we can clearly see the influence of the maximum for each target-residue (Figure 7.13.2). Furthermore, check how the most intense colors match with mutations.

Finally, you can test different descriptors and observe the behavior separately based on maximums for each target-residue. The base structure for invoking these conditions is shown as follows:

```
descriptor_mxr = { 'maximum_descriptorY' }
```

```
descriptor_mxr = { 'maximum_descriptorY' : additional_parameters }
```

Note that we are considering situations where descriptors can be sets (*e.g.* `nr`, `rm`, and so on) or dictionaries (*e.g.*, `rt`, `rsa`, `rspi`, and so on), respectively.

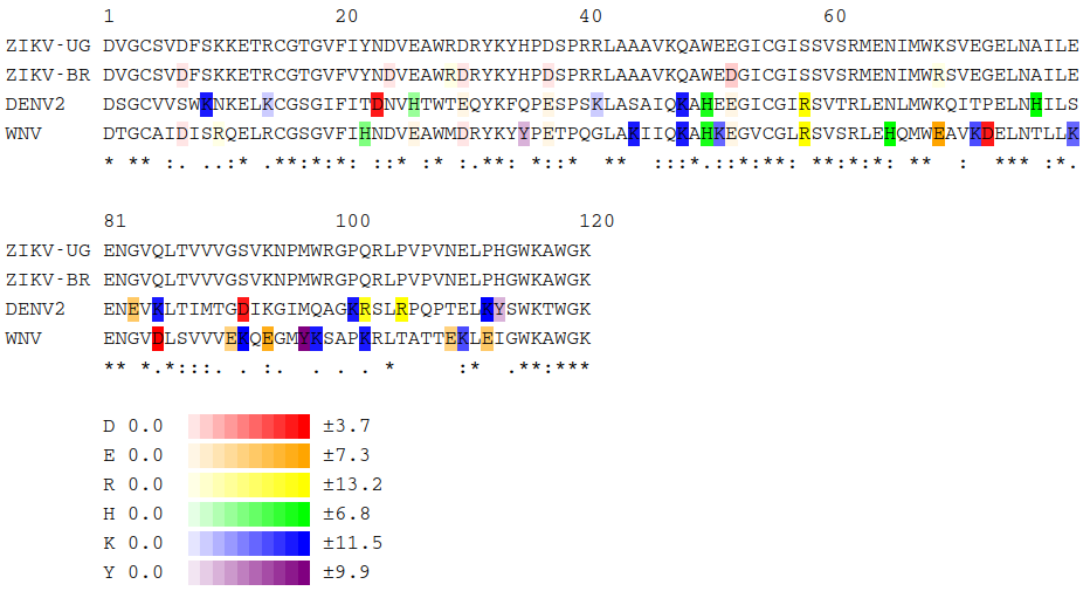



Figure 7.13.2: Alignment colored applying `nr` descriptor and the condition `Y` to separate the maximum per target-residue but using `fac` descriptor.

*Note: We recommend the user to test different combinations of maximum restriction and color (see* Chapter 5*) descriptors, to examine the impact of each case. Keep in mind the same rules for each maximum restriction descriptor that we saw previously.*



## 7.14 Maximum restriction descriptors dictionary (-dmx)

When the `multiple` comparison method is declared, it is possible that you need to distinguish each reference/target(s) relationship with a specific maximum restriction descriptor. For that, we must build an array of maximum descriptors through the `dict_mxr` variable of the script "array_get.py". To build it, you should follow the same rules as for `dict_ope` (see Section 4.5). For this example, we used the same dictionary for the `multiple` method as a base (see Subsection 3.10.3). The CPRISMA maximum-restriction-array will be declared as follows[6]:

```
dict_mxr = { 'main_0' :  [ { 'rmY' } , { 'nr' } ] ,
             'main_1' :  [ { 'rspim' :  [ {'2' :  [47, 48, 50] }] } ,
                           { 'rsaY' : [ { '0' :  ['R'] } , { '3':  ['D'] } ]
,
             'main_2' :  [ { 'rt' :  [0, '>='] } ] }
```

Now we can execute the `-dmx` flag to call the array `dict_mxr`[7]:

```
cprisma -tr -ck -rf multiple -dmx
```

The log file will show the relationships between each comparison and maximum restriction descriptors:

```
Comparing dict_ref with dict_mxr:
                                {'0':  [1, 2, 3]} ........  { 'rmY' }
                                {'0':  []} ........  { 'nr' }
                                {'1':  [0, 2]} ........  { 'rspim' :  ...  }
                                {'1':  [0, 2, 3]} ........  { 'rsaY' : ...  }
                                {'2':  [0]} ........  { 'rt' :  [0, '>='] }
The array of the feature 'maximum' dict_mxr is compatible with the array of
comparison sequences dict_ref!
```

[6] Note that we have differentiated each maximum restriction descriptor for each comparison with a different color to aid the reader.

[7] Notice that additional parameters are being implemented.

... and a table with a summary for all maximum descriptor information per comparison:

| main | comparison | descriptor | sequence | residue | position | mutation | threshold | separate |
|---|---|---|---|---|---|---|---|---|
| main_0 | 1 | rmY | NaN | NaN | NaN | True | NaN | True |
| main_0 | 2 | nr | NaN | NaN | NaN | NaN | NaN | NaN |
| main_1 | 3 | rspim | 2 | NaN | [47...] | True | NaN | NaN |
| main_1 | 4 | rsaY | 0 | [R] | NaN | NaN | NaN | True |
| main_1 | 4 | rsaY | 3 | [D] | NaN | NaN | NaN | True |
| main_2 | 5 | rt | NaN | NaN | NaN | NaN | [0, >=] | NaN |

*Note: If the array of* `dict_mxr` *is wrongly built, all descriptors will be transformed to* `nr` *(default parameter) for each comparison and additional information about where the errors are will be displayed.*



# 8 Practical Examples

The following examples bring together some of the features that we described earlier. Each is related to a previously published study [1, 2]. Furthermore, in the "examples" directory, you will find the scripts, alignments and other files that help as the basis for achieving the desired visualizations. Many details will not be given as we wish to motivate the reader to test and explore the characteristics of each example on their own. Each case follows the next order:

1. Command-line invoked.
2. Declared variables (as shown in Figure 2.2.3).
3. An image of the alignment.

Each example tries to approximate a certain image from a specific publication. In example 1 (see Section 8.1) we try to reproduce Figure 8 [1], in example 2 (see Section 8.2) Figure 5 (only chain A) [2], and in example 3 (see Section 8.3) Figure 1a [3].

*Note: For more details on the methodological aspects, data collection, purpose, and others, you can find them in the respective publication.*

## 8.1 Example 1: $\Delta pK_a$ for several $NS1_{ZIKV}$

Command-line:

```
cprisma -tr -ns -ck -rf multiple -va 2 -j -sco 142 -dop -dco -dvi -dmx
```

Variables:

```
target_residues = (‘D’, ‘E’, ‘R’, ‘H’, ‘K’, ‘Y’)
```

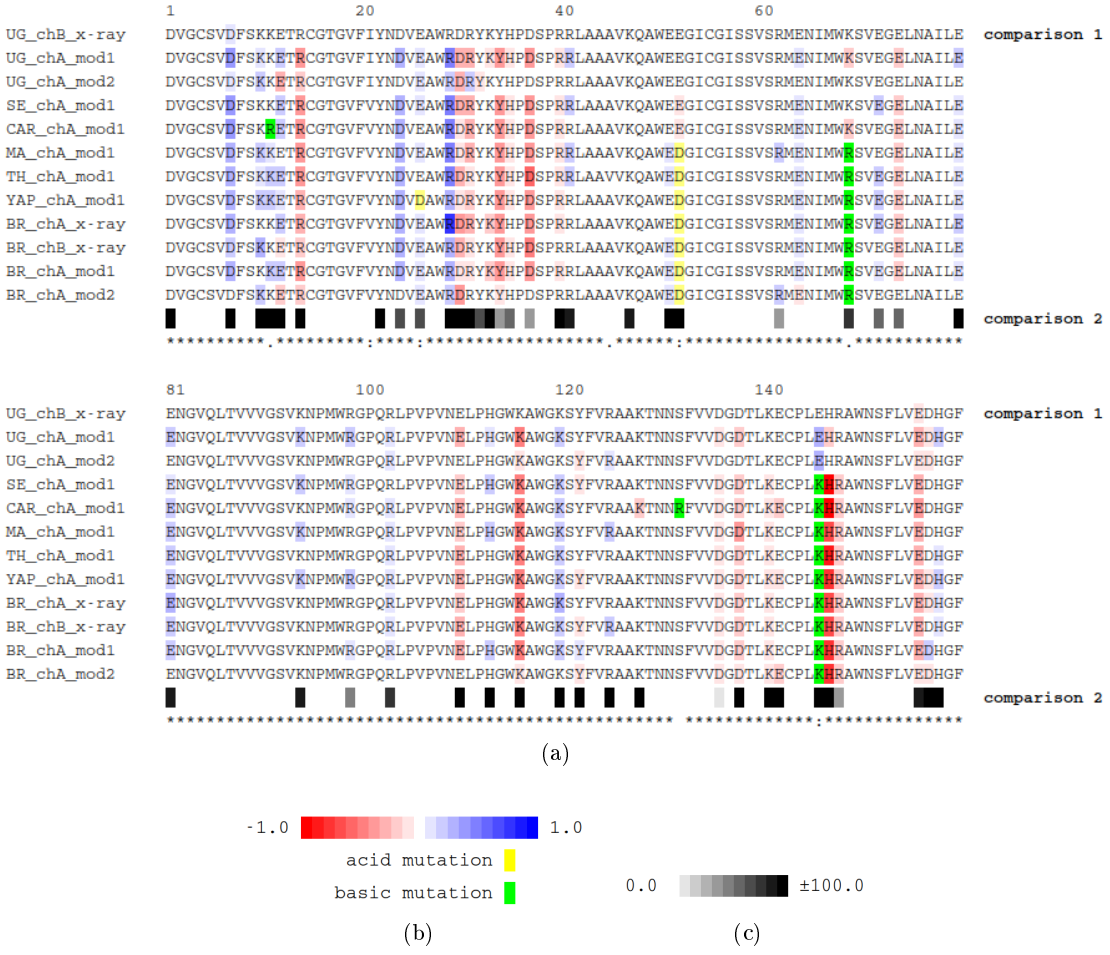
```
name_sequence = (‘UG_chA_x-ray’, ‘UG_chB_x-ray’, ‘UG_chA_mod1’, ‘UG_chA_mod2’,
‘SE_chA_mod1’, ‘CAR_chA_mod1’, ‘MA_chA_mod1’, ‘TH_chA_mod1’, ‘YAP_chA_mod1’,
‘BR_chA_x-ray’, ‘BR_chB_x-ray’, ‘BR_chA_mod1’, ‘BR_chA_mod2’, ‘’)
```

```
descriptor_ope = {‘n’}
```

```
descriptor_col = { 'nc' }

descriptor_vis = { 'ReY','DeN','LeY' }

descriptor_mxr = { 'nr' }
```




(a)




(b)




(c)

Figure 8.1.1: (a) Alignment for Example 1. Not all alignment is being shown. (b) Scale for comparison 1. (c) Scale for comparison 2.

```
dict_ref = { 'main_0' :  [ {'0' :  [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] } ] ,

             'main_13' :  [ {'13' :  [] } ] }
```



```
dict_ope = { 'main_0' :  [ { 'd' } ] ,
             'main_13' :  [ { 'n' } ] }
dict_col = { 'main_0' :  [ { 'pkac' :  True } ] ,
             'main_13' :  [ { 'ssc' } ] }
dict_vis = { 'main_0' :  [ { 'ReN','DeY','LeY' } ] ,
             'main_13' :  [ { 'ReY','DeY','LeN' } ] }
dict_mxr = { 'main_0' :  [ { 'rm' } ] ,
             'main_13' :  [ { 'nr' } ] }
```

From these input parameters, we obtain the alignment that appears in Figure 8.1.1. Note that the last element of the tuple `name_sequence` does not have a name. Due to that, the alignment will not display any name for that sequence.

## 8.2 Example 2: B-cell epitope predictions on $NS1_{WNV(176-352)}$

Command-line:

```
cprisma -tr -ck -rf multiple -ns -va 2 -j -hc -n 176 -sco 142 -a 3 -dop -dco
-dvi
```

Variables

```
target_residues = ('D', 'E', 'R', 'H', 'K', 'Y')
name_sequence = ('SASA', 'pKa_NS1 (alone)','PROCEEDpKa', 'PISA', 'DiscoTope',
'ElliPro', 'SEPPA2', 'SEPPA3', 'Consensus', '')
descriptor_ope = { 'n' }
descriptor_col = { 'nc' }
descriptor_vis = { 'ReY','DeN','LeY' }
descriptor_mxr = { 'nr' }
dict_ref = { 'main_0' :  [ {'0' :  [] } ] ,
             'main_1' :  [ {'1' :  [2] } ] ,
```



```
              'main_3' :  [ {'3' :  [] } ] ,
              'main_4' :  [ {'4' :  [] } ] ,
              'main_5' :  [ {'5' :  [] } ] ,
              'main_6' :  [ {'6' :  [] } ] ,
              'main_7' :  [ {'7' :  [] } ] ,
              'main_8' :  [ {'8' :  [] } ] ,
              'main_9' :  [ {'9' :  [] } ] }
dict_ope = { 'main_0' :  [ { 'n' } ] ,
              'main_1' :  [ { 'da' } ] ,
              'main_3' :  [ { 'n' } ] ,
              'main_4' :  [ { 'n' } ] ,
              'main_5' :  [ { 'n' } ] ,
              'main_6' :  [ { 'n' } ] ,
              'main_7' :  [ { 'n' } ] ,
              'main_8' :  [ { 'n' } ] ,
              'main_9' :  [ { 'n' } ] }
dict_col = { 'main_0' :  [ { 'tc' :  [54, 0, '>'] } ] ,
              'main_1' :  [ { 'tc' :  [6, 0.007, '>'] } ] ,
              'main_3' :  [ { 'tc' :  [75, 0, '>'] } ] ,
              'main_4' :  [ { 'tc' :  [20, -6.85, '>='] } ] ,
              'main_5' :  [ { 'tc' :  [64, 0, '>'] } ] ,
              'main_6' :  [ { 'tc' :  [22, 0.049, '>='] } ] ,
              'main_7' :  [ { 'tc' :  [96, 0.093, '>='] } ] ,
              'main_8' :  [ { 'tc' :  [92, 0, '>'] } ] ,
              'main_9' :  [ { 'ssc' } ] }
```

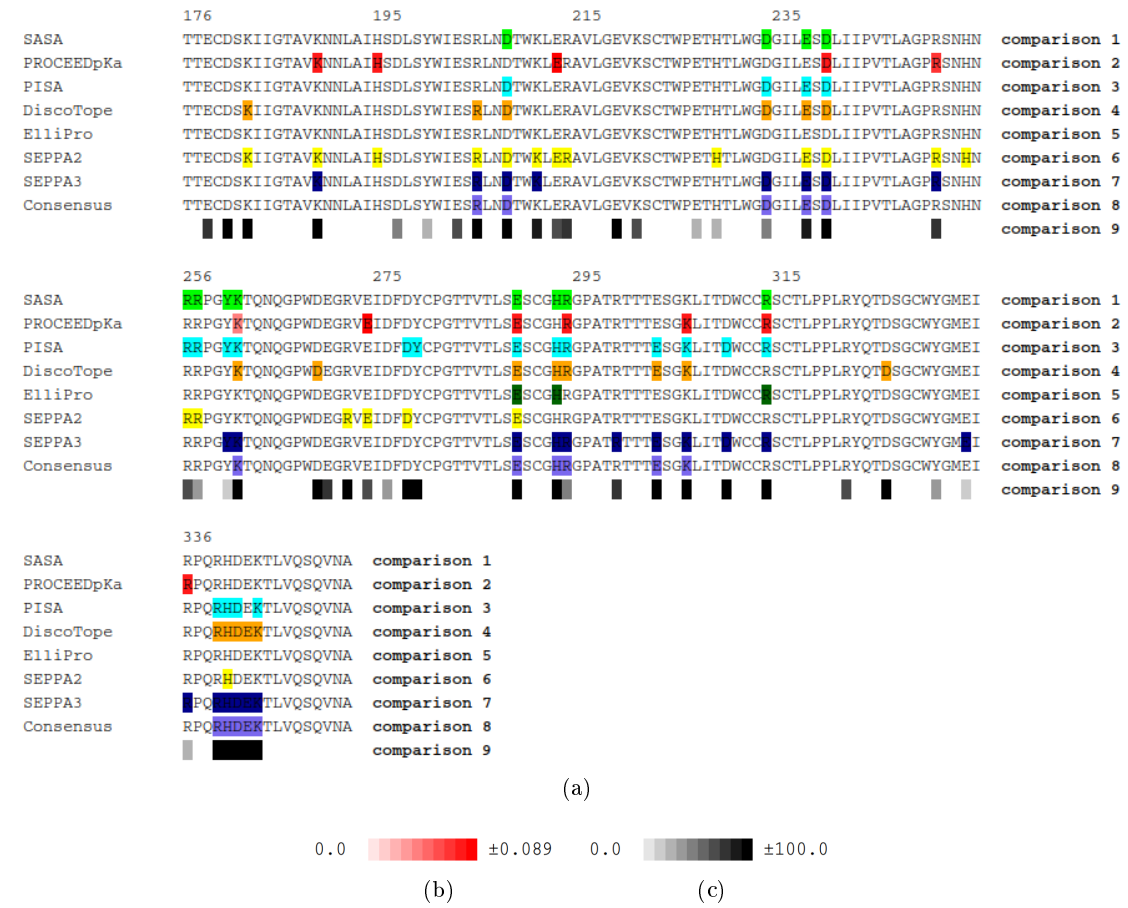



Figure 8.2.1: (a) Alignment for Example 2. (b) Scale for comparison 2. (c) Scale for comparison 9.

```
dict_vis = { 'main_0' :  [ { 'ReY','DeN','LeY' } ] ,

             'main_1' :  [ { 'ReN','DeY','LeY' } ] ,

             'main_3' :  [ { 'ReY','DeN','LeY' } ] ,

             'main_4' :  [ { 'ReY','DeN','LeY' } ] ,

             'main_5' :  [ { 'ReY','DeN','LeY' } ] ,

             'main_6' :  [ { 'ReY','DeN','LeY' } ] ,

             'main_7' :  [ { 'ReY','DeN','LeY' } ] ,
```



```
        'main_8' :  [ { 'ReY','DeN','LeY' } ] ,

        'main_9' :  [ { 'ReY','DeY','LeN' } ] }
dict_mxr={}
```

As we are implementing the `tc` color descriptor (see Section 5.8) and the numerical data is dependent on the threshold, we recommend taking care with the accuracy. The output for example 2 appears in Figure 8.2.1. To obtain this outcome, we have considered an accuracy of 3 decimal places with the `-a` flag (see Section 3.2).

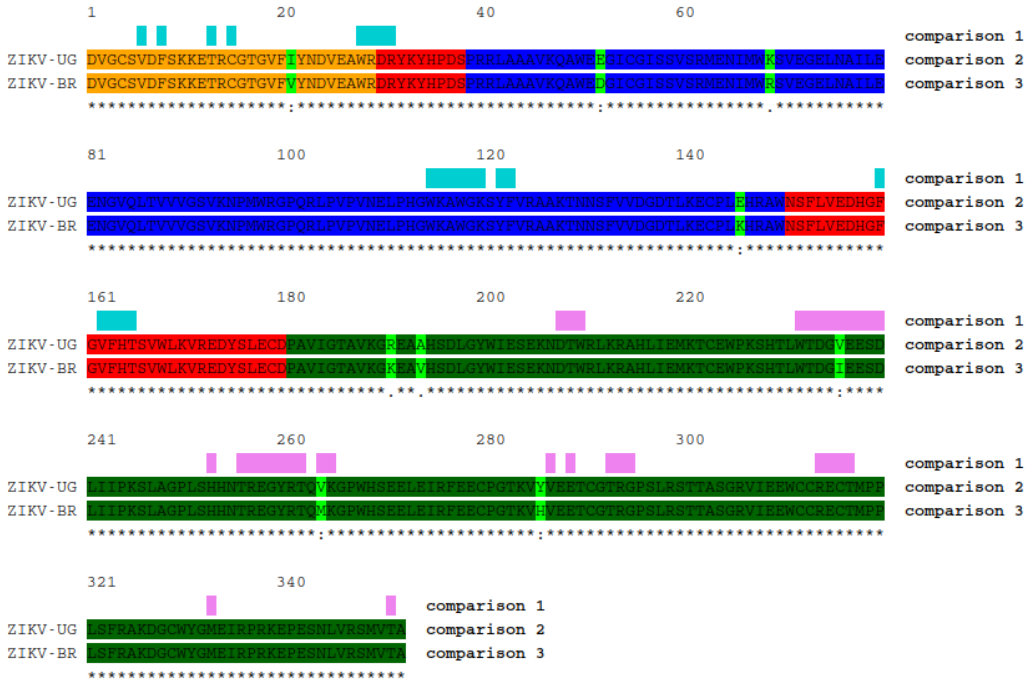



Figure 8.2.2: Alignment for Example 3.

## 8.3 Example 3: Protein structural domains for two $NS1_{ZIKV}$ and its biological interfaces

Command-line:

```
cprisma -va 2 -j -ns -ck -rf multiple -mco 54 -dco -dvi
```



Variables:

```
target_residues = ()

name_sequence = (‘’, ‘ZIKV-UG’, ‘ZIKV-BR’)

descriptor_ope = { ‘n’ }

descriptor_col = { ‘nc’ }

descriptor_vis = { ‘ReY’,‘DeN’,‘LeY’ }

descriptor_mxr = { ‘nr’ }

dict_ref = { ‘main_0’ :  [ {‘0’ :  [] } ] ,

             ‘main_2’ :  [ {‘2’ :  [1] } ] ,

             ‘main_1’ :  [ {‘1’ :  [2] } ] }

dict_ope = {}

dict_col = { ‘main_0’ :  [ { ‘pic’ :  [ { ‘0’ :  [81, ‘n’, ‘5‘ ,‘7’, ‘12’, ‘14’,
                                                  ‘27-30’, ‘114-119’, ‘121-122’,
                                                  ‘159’, ‘161-164’] } ,
                                       { ‘0’ :  [35, ‘n’, ‘207-209’, ‘231-239’,
                                                  ‘252’, ‘255-261’, ‘263-264’,
                                                  ‘286’, ‘288’, ‘292-294’,
                                                  ‘313-316’, ‘332’, ‘350’] } ] } ] ,
             ‘main_2’ :  [ { ‘pimc’ :  [ { ‘1’ :  [20, ‘n’, ‘0-28’] } ,
                                         { ‘1’ :  [6, ‘n’, ‘29-37’] } ,
                                         { ‘1’ :  [94, ‘n’, ‘38-149’] } ,
                                         { ‘1’ :  [6, ‘n’, ‘150-179’] } ,
                                         { ‘1’ :  [64, ‘n’, ‘180-351’] } ] } ] ,
             ‘main_1’ :  [ { ‘pimc’ :  [ { ‘2’ :  [20, ‘n’, ‘0-28’] } ,
                                         { ‘2’ :  [6, ‘n’, ‘29-37’] } ,
```



```
                                        { '2' :  [94, 'n', '38-149'] } ,
                                        { '2' :  [6, 'n', '150-179'] } ,
                                        { '2' :  [64, 'n', '180-351'] } ] } ] }
dict_vis = { 'main_0' :  [ { 'ReY','DeN','LeN' } ] ,
             'main_2' :  [ { 'ReN','DeN','LeY' } ] ,
             'main_1' :  [ { 'ReN','DeN','LeY' } ] }
dict_mxr = {}
```

The output for example 3 appears in Figure 8.2.2. See that for `pimc` descriptor to distinguish the mutations (see Section 5.7), the `main`'s `1` and `2` in `dict_ref` have the sequences exchanged and we have hidden the references with the `ReN` visualization method (see Chapter 6).



# Bibliography


[1] Sergio Alejandro Poveda-Cuevas, Catherine Etchebest, and Fernando Luís Barroso da Silva. Insights into the ZIKV NS1 virology from different strains through a fine analysis of physicochemical properties. *ACS omega*, 3(11):16212–16229, 2018.

[2] Sergio Alejandro Poveda-Cuevas, Catherine Etchebest, and Fernando Luís Barroso da Silva. Identification of electrostatic epitopes in flavivirus by computer simulations: The PROCEEDpKa method. *Journal of Chemical Information and Modeling*, 60(2):944–963, 2020.

[3] Sergio Alejandro Poveda-Cuevas, Fernando Luís Barroso da Silva, and Catherine Etchebest. How the strain origin of ZIKV NS1 protein impacts its dynamics and implications to their differential virulence. *Journal of Chemical Information and Modeling*, 61(3):1516–1530, 2021.

[4] Fernando Luís Barroso da Silva and Donal MacKernan. Benchmarking a fast proton titration scheme in implicit solvent for biomolecular simulations. *Journal of Chemical Theory and Computation*, 13:2915–2929, 2017.

[5] Fernando Luís Barroso da Silva and Luis Gustavo Dias. Development of constant-pH simulation methods in implicit solvent and applications in biomolecular systems. *Biophysical Reviews*, 9(5):699–728, 2017.

[6] Robert Edgar. MUSCLE a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113, 2004.

*"One should never try to prove anything that is not almost obvious"*

Alexander Grothendieck



# Bibliography


Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25.

Agrelli, A., de Moura, R. R., Crovella, S., and Brandão, L. A. C. (2019). Zika virus entry mechanisms in human cells. *Infection, Genetics and Evolution*, 69:22–29.

Akbayrak, I. Y., Caglayan, S. I., Durdagi, S., Kurgan, L., Uversky, V. N., Ulver, B., Dervisoğlu, H., Haklidir, M., Hasekioglu, O., and Coskuner-Weber, O. (2021). Structures of MERS-CoV macro domain in aqueous solution with dynamics: Impacts of parallel tempering simulation techniques and CHARMM36m and AMBER99SB force field parameters. *Proteins: Structure, Function, and Bioinformatics*.

Akey, D. L., Brown, W. C., Dutta, S., Konwerski, J., Jose, J., Jurkiw, T. J., DelProposto, J., Ogata, C. M., Skiniotis, G., Kuhn, R. J., and Smith, J. L. (2014). *Flavivirus* NS1 structures reveal surfaces for associations with membranes and the immune system. *Science*, 343(6173):881–885.

Alhossary, A., Awuni, Y., Kwoh, C. K., and Mu, Y. (2018). Proposing drug fragments for dengue virus NS5 protein. *Journal of bioinformatics and computational biology*, 16(03):1840017.

Allen, M. P. and Tildesley, D. J. (2017). *Computer simulation of liquids*. Oxford university press.

Alves, A. M. B., Costa, S. M., and Pinto, P. B. A. (2021). Dengue virus and vaccines: How can dna immunization contribute to this challenge? *Frontiers in Medical Technology*, 3:13.

Anasir, M. I., Ramanathan, B., and Poh, C. L. (2020). Structure-based design of antivirals against envelope glycoprotein of dengue virus. *Viruses*, 12(4):367.

Andrew, R. L. (2001). Molecular modeling principles and applications. *Prentice Hall, London*.

Arthur, E. J. and Brooks III, C. L. (2016). Efficient implementation of constant pH molecular dynamics on modern graphics processors. *Journal of computational chemistry*, 37(24):2171–2180.

Aubry, F., Jacobs, S., Darmuzey, M., Lequime, S., Delang, L., Fontaine, A., Jupatanakul, N., Miot, E. F., Dabo, S., Manet, C., et al. (2021). Recent african strains of zika virus display higher transmissibility and fetal pathogenicity than asian strains. *Nature communications*, 12(1):1–14.

Avirutnan, P., Zhang, L., Punyadee, N., Manuyakorn, A., Puttikhunt, C., Kasinrerk, W., Malasit, P., Atkinson, J. P., and Diamond, M. S. (2007). Secreted ns1 of dengue virus attaches to the surface of cells via interactions with heparan sulfate and chondroitin sulfate e. *PLoS pathogens*, 3(11):e183.

Bajar, B. T., Wang, E. S., Zhang, S., Lin, M. Z., and Chu, J. (2016). A guide to fluorescent protein FRET pairs. *Sensors*, 16(9):1488.

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological Reviews*, 35(3):235.

Barrett, A. D. (2001). *Flaviviruses*. John Wiley & Sons, Ltd.

Barroso da Silva, F. L. (2000). *Statistical mechanical studies of aqueous solutions and biomolecular systems*. PhD thesis, University of Lund.

Barroso da Silva, F. L., Carloni, P., Cheung, D., Cottone, G., Donnini, S., Foegeding, E. A., Gulzar, M., Jacquier, J. C., Lobaskin, V., MacKernan, D., et al. (2020). Understanding and controlling food protein structure and function in foods: perspectives from experiments and computer simulations. *Annual review of food science and technology*, 11:365–387.

Barroso da Silva, F. L., Derreumaux, P., and Pasquali, S. (2017). Fast coarse–grained model for RNA titration. *The Journal of Chemical Physics*, 146(3):035101.

Barroso da Silva, F. L., Derreumaux, P., and Pasquali, S. (2018). Protein-RNA complexation driven by the charge regulation mechanism. *Biochemical and Biophysical Research Communications*, 498(7):264–273.

Barroso da Silva, F. L. and Dias, L. G. (2017). Development of constant-pH simulation methods in implicit solvent and applications in biomolecular systems. *Biophysical Reviews*, 9(5):699–728.

Barroso da Silva, F. L. and Jönsson, B. (2009). Polyelectrolyte–protein complexation driven by charge regulation. *Soft Matter*, 5(15):2862–2868.

Barroso da Silva, F. L. and MacKernan, D. (2017). Benchmarking a fast proton titration scheme in implicit solvent for biomolecular simulations. *Journal of Chemical Theory and Computation*, 13:2915–2929.

Barroso da Silva, F. L., Pasquali, S., Derreumaux, P., and Dias, L. G. (2016). Electrostatics analysis of the mutational and pH effects of the N-terminal domain self–association of the major ampullate spidroin. *Soft Matter*, 12(25):5600–5612.

Beck, C., Jimenez-Clavero, M. A., Leblond, A., Durand, B., Nowotny, N., Leparc-Goffart, I., Zientara, S., Jourdain, E., and Lecollinet, S. (2013). Flaviviruses in europe: complex circulation patterns and their consequences for the diagnosis and control of west nile disease. *International journal of environmental research and public health*, 10(11):6049–6083.

Beckham, J. D., Pastula, D. M., Massey, A., and Tyler, K. L. (2016). Zika virus as an emerging global pathogen: Neurological complications of Zika virus. *JAMA Neurology*, 73(7):875–879.

Bennett, W. D., Chen, A. W., Donnini, S., Groenhof, G., and Tieleman, D. P. (2013). Constant pH simulations with the coarse-grained MARTINI model – Application to oleic acid aggregates. *Canadian Journal of Chemistry*, 91(9):839–846.

Beresford-Smith, B. and Chan, D. Y. (1983). Electrical double-layer interactions in concentrated colloidal systems. *Faraday Discussions of the Chemical Society*, 76:65–75.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Biering, S. B., Akey, D. L., Wong, M. P., Brown, W. C., Lo, N. T., Puerta-Guardo, H., de Sousa, F. T. G., Wang, C., Konwerski, J. R., Espinosa, D. A., et al. (2021). Structural basis for antibody inhibition of flavivirus NS1–triggered endothelial dysfunction. *Science*, 371(6525):194–200.

Blanco, A. and Blanco, G. (2017). Chapter 8 – Enzymes. *Medical Biochemistry; Blanco, A., Blanco, G., Eds*, pages 153–175.

Blitvich, B. J., Scanlon, D., Shiell, B. J., Mackenzie, J. S., Pham, K., and Hall, R. A. (2001). Determination of the intramolecular disulfide bond arrangement and biochemical identification of the glycosylation sites of the nonstructural protein ns1 of murray valley encephalitis virus. *Journal of General Virology*, 82(9):2251–2256.

Bollati, M., Alvarez, K., Assenberg, R., Baronti, C., Canard, B., Cook, S., Coutard, B., Decroly, E., de Lamballerie, X., Gould, E. A., et al. (2010). Structure and functionality

in flavivirus NS-proteins: perspectives for drug design. *Antiviral research*, 87(2):125–148.

Bonheur, A. N., Thomas, S., Soshnick, S. H., McGibbon, E., Dupuis 2nd, A. P., Hull, R., Slavinski, S., Del Rosso, P. E., Weiss, D., Hunt, D. T., McCabe, M. E., Dean, A. B., Folkerth, R., Laib, A. M., and Wong, S. J. (2021). A fatal case report of antibody-dependent enhancement of dengue virus type 1 following remote Zika virus infection. *BMC Infectious Diseases*, 21:1–8.

Bou-Rabee, N. (2014). Time integrators for molecular dynamics. *Entropy*, 16(1):138–162.

Brady, O. J., Gething, P. W., Bhatt, S., Messina, J. P., Brownstein, J. S., Hoen, A. G., Moyes, C. L., Farlow, A. W., Scott, T. W., and Hay, S. I. (2012). Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Neglected Tropical Diseases*, 6(8):e1760.

Branden, C. I. and Tooze, J. (1999). *Introduction to protein structure*. Garland Science.

Breitbart, M. and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology*, 13(6):278–284.

Bressanelli, S., Stiasny, K., Allison, S. L., Stura, E. A., Duquerroy, S., Lescar, J., Heinz, F. X., and Rey, F. A. (2004). Structure of a flavivirus envelope glycoprotein in its low-ph-induced membrane fusion conformation. *The EMBO journal*, 23(4):728–738.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. a., and Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187–217.

Brown, W. C., Akey, D. L., Konwerski, J. R., Tarrasch, J. T., Skiniotis, G., Kuhn, R. J., and Smith, J. L. (2016). Extended surface for membrane association in Zika virus NS1 structure. *Nature Structural & Molecular Biology*, 23(9):865–867.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., et al. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451.

Cao-Lormeau, V.-M., Blake, A., Mons, S., Lastère, S., Roche, C., Vanhomwegen, J., Dub, T., Baudouin, L., Teissier, A., Larre, P., et al. (2016). Guillain-Barré syndrome outbreak associated with Zika virus infection in French Polynesia: a case-control study. *The Lancet*, 387(10027):1531–1539.

Cao-Lormeau, V.-M., Roche, C., Teissier, A., Robin, E., Berry, A.-L., Mallet, H.-P., Sall, A. A., and Musso, D. (2014). Zika virus, french polynesia, south pacific, 2013. *Emerging Infectious Diseases*, 20(6):1085.

Cecchetto, J., Fernandes, F. C., Lopes, R., and Bueno, P. R. (2017). The capacitive sensing of NS1 *Flavivirus* biomarker. *Biosensors and Bioelectronics*, 87:949–956.

Chandra, N., Anand, P., and Yeturu, K. (2010). Structural bioinformatics: deriving biological insights from protein structures. *Interdisciplinary Sciences: Computational Life Sciences*, 2(4):347–366.

Chaudhury, S., Ripoll, D. R., and Wallqvist, A. (2015). Structure-based p$K_a$ prediction provides a thermodynamic basis for the role of histidines in pH-induced conformational transitions in dengue virus. *Biochemistry and Biophysics Reports*, 4:375–385.

Chou, K.-C. (2004). Structural bioinformatics and its impact to biomedical science. *Current medicinal chemistry*, 11(16):2105–2134.

Chuang, C.-H., Chiou, S.-j., Cheng, T.-L., and Wang, Y.-T. (2018). A molecular dynamics simulation study decodes the zika virus NS5 methyltransferase bound to SAH and RNA analogue. *Scientific reports*, 8(1):1–9.

Ci, Y., Liu, Z.-Y., Zhang, N.-N., Niu, Y., Yang, Y., Xu, C., Yang, W., Qin, C.-F., and Shi, L. (2020). Zika NS1–induced ER remodeling is essential for viral replication. *Journal of Cell Biology*, 219(2).

Ci, Y., Yang, Y., Xu, C., Qin, C.-F., and Shi, L. (2021). Electrostatic interaction between NS1 and negatively charged lipids contributes to Flavivirus replication organelles formation. *Frontiers in Microbiology*, 12:1002.

Clark, A. C. (2008). Protein folding: are we there yet? *Archives of biochemistry and biophysics*, 469(1):1.

Coêlho, D. F., Ferraz, M. V., Marques, E. T., Lins, R. D., and Viana, I. F. (2019). 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 and Modelling*, 93:107442.

Collins, M. H. and Metz, S. W. (2017). Progress and works in progress: update on flavivirus vaccine development. *Clinical therapeutics*, 39(8):1519–1536.

Cox, B. D., Stanton, R. A., and Schinazi, R. F. (2015). Predicting Zika virus structural biology: challenges and opportunities for intervention. *Antiviral Chemistry and Chemotherapy*, 24(3-4):118–126.

Cox, M. M. and Nelson, D. L. (2000). *Lehninger principles of biochemistry.* New York, Worth.

Cramer, C. J. (2013). *Essentials of computational chemistry: theories and models.* John Wiley & Sons.

Cruz-Oliveira, C., Freire, J. M., Conceição, T. M., Higa, L. M., Castanho, M. A., and Da Poian, A. T. (2015). Receptors and routes of dengue virus entry into the host cells. *FEMS microbiology reviews*, 39(2):155–170.

Curry, S. (2015). Structural biology: a century-long journey into an unseen world. *Interdisciplinary Science Reviews*, 40(3):308–328.

da Penha Valente, R. P., de Souza, R. C., de Medeiros Muniz, G., Ferreira, J. E. V., de Miranda, R. M., e Lima, A. H. L., Vianez, J. L. d. S. G., et al. (2020). Using accelerated molecular dynamics simulation to elucidate the effects of the T198F mutation on the molecular flexibility of the West Nile virus envelope protein. *Scientific reports*, 10(1):1–6.

Dahirel, V. and Jardat, M. (2010). Effective interactions between charged nanoparticles in water: What is left from the dlvo theory? *Current Opinion in Colloid & Interface Science*, 15(1-2):2–7.

David, C. C. and Jacobs, D. J. (2014). Principal component analysis: a method for determining the essential dynamics of proteins. In *Protein dynamics*, pages 193–226. Springer.

de Lima, T. W., Caliri, A., Barroso da Silva, F. L., Tinós, R., Travieso, G., da Silva, I. N., Lopes, P. S., de Souza, E. M., Delbem, A. C. B., and Bonatto, V. (2009). Some modeling issues for protein structure prediction using evolutionary algorithms. *Evolutionary Computation*, page 153.

Dejnirattisai, W., Supasa, P., Wongwiwat, W., Rouvinski, A., Barba-Spaeth, G., Duangchinda, T., Sakuntabhai, A., Cao-Lormeau, V.-M., Malasit, P., Rey, F. A., et al. (2016). Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. *Nature immunology*, 17(9):1102.

Delboni, L. A. and Barroso da Silva, F. L. (2016). On the complexation of whey proteins. *Food Hydrocolloids*, 55:89–99.

Deller, M. C., Kong, L., and Rupp, B. (2016). Protein stability: a crystallographer's perspective. *Acta Crystallographica Section F: Structural Biology Communications*, 72(2):72–95.

Dey, S., Nandy, A., Basak, S. C., Nandy, P., and Das, S. (2017). A Bioinformatics approach to designing a Zika virus vaccine. *Computational Biology and Chemistry*, 68:143–152.

Dick, G., Kitchen, S., and Haddow, A. (1952). Zika virus (I). Isolations and serological specificity. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 46(5):509–520.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, 148(4):1667–1686.

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., and Scharf, M. (1995). The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry*, 16(3):273–284.

Elliott, J. A. (2011). Novel approaches to multiscale modelling in materials science. *International Materials Reviews*, 56(4):207–225.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of chemical physics*, 103(19):8577–8593.

Falconar, A. K. and Young, P. R. (1990). Immunoaffinity purification of native dimer forms of the *Flavivirus* non-structural glycoprotein, NS1. *Journal of Virological Methods*, 30(3):323–332.

Faria, N. R., da Silva Azevedo, R. d. S., Kraemer, M. U., Souza, R., Cunha, M. S., Hill, S. C., Thézé, J., Bonsall, M. B., Bowden, T. A., and Rissanen, I. (2016). Zika virus in the americas: early epidemiological and genetic findings. *Science*, page aaf5036.

Faye, O., Freire, C. C., Iamarino, A., Faye, O., de Oliveira, J. V. C., Diallo, M., and Zanotto, P. M. (2014). Molecular evolution of Zika virus during its emergence in the $20^{th}$ century. *PLoS Neglected Tropical Diseases*, 8(1):e2636.

Fields, B. N. (2014). *Fields virology*. Lippincott Williams and Wilkins, $6^{th}$ edition.

Figueiredo, A., Vieira, N. C., Dos Santos, J. F., Janegitz, B. C., Aoki, S. M., Junior, P. P., Lovato, R. L., Nogueira, M. L., Zucolotto, V., and Guimãraes, F. E. (2015). Electrical detection of dengue biomarker using egg yolk immunoglobulin as the biological recognition element. *Scientific Reports*, 5:7865.

Flamand, M., Megret, F., Mathieu, M., Lepault, J., Rey, F. A., and Deubel, V. (1999). Dengue virus type 1 nonstructural glycoprotein ns1 is secreted from mammalian cells as

a soluble hexamer in a glycosylation-dependent fashion. *Journal of virology*, 73(7):6104–6110.

Freire, M. C., Pol-Fachin, L., Coêlho, D. F., Viana, I. F., Magalhães, T., Cordeiro, M. T., Fischer, N., Loeffler, F. F., Jaenisch, T., Franca, R. F., Marques, E. T. A., and Lins, R. D. (2017). Mapping putative B–cell Zika virus NS1 epitopes provides molecular basis for anti-NS1 antibody discrimination between Zika and Dengue viruses. *ACS Omega*, 2(7):3913–3920.

Fuzo, C. A. and Degrève, L. (2014). The pH dependence of flavivirus envelope protein structure: insights from molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics*, 32(10):1563–1574.

Geoghegan, J. L. and Holmes, E. C. (2018). The phylogenomics of evolving virus virulence. *Nature Reviews Genetics*, 19(12):756–769.

Giron, C. C., Laaksonen, A., and Barroso da Silva, F. L. (2020). On the interactions of the receptor-binding domain of SARS-CoV-1 and SARS-CoV-2 spike proteins with monoclonal antibodies and the receptor ACE2. *Virus Research*, (285).

Giron, C. C., Laaksonen, A., and Barroso da Silva, F. L. (2021). Up state of the SARS-CoV-2 spike homotrimer favors an increased virulence for new variants. *Frontiers in Medical Technology*, 3:1–20.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

Gonçalves, R. L., de Lima Menezes, G., Sussuchi, L., Moreli, M. L., Mottin, M., Andrade, C. H., Pereira, M., and da Silva, R. A. (2019). Dynamic behavior of Dengue and Zika viruses NS1 protein reveals monomer–monomer interaction mechanisms and insights to rational drug design. *Journal of Biomolecular Structure and Dynamics*, pages 1–11.

Gong, Z., Gao, Y., and Han, G.-Z. (2016). Zika virus: two or three lineages? *Trends in Microbiology*, 24(7):521–522.

Govindjee, J. K., Messinger, J., and Whitmarsh, J. (2010). Encyclopedia of life sciences (els).

Gräbner, C. (2018). How to relate models to reality? an epistemological framework for the validation and verification of computational models. *Journal of Artificial Societies and Social Simulation*, 21(3).

Gregory, R. (1995). *Protein-solvent interactions*. CRC Press.

Grudinin, S., Laine, E., and Hoffmann, A. (2020). Predicting protein functional motions: an old recipe with a new twist. *Biophysical journal*, 118(10):2513–2525.

Grünewald, F., Souza, P. C., Abdizadeh, H., Barnoud, J., de Vries, A. H., and Marrink, S. J. (2020). Titratable Martini model for constant pH simulations. *The Journal of chemical physics*, 153(2):024118.

Gubler, D. J. (1998). Dengue and dengue hemorrhagic fever. *Clinical Microbiology Reviews*, 11(3):480–496.

Gutsche, I., Coulibaly, F., Voss, J. E., Salmon, J., d'Alayer, J., Ermonval, M., Larquet, E., Charneau, P., Krey, T., and Mégret, F. (2011). Secreted dengue virus nonstructural protein NS1 is an atypical barrel-shaped high-density lipoprotein. *Proceedings of the National Academy of Sciences*, 108(19):8003–8008.

Guzmán, M. G. and Kouri, G. (2002). Dengue: an update. *The Lancet Infectious Diseases*, 2(1):33–42.

Hadinegoro, S. R., Arredondo-García, J. L., Capeding, M. R., Deseda, C., Chotpitayasunondh, T., Dietze, R., Hj Muhammad Ismail, H., Reynales, H., Limkittikul, K., and Rivera-Medina, D. M. (2015). Efficacy and long-term safety of a dengue vaccine in regions of endemic disease. *New England Journal of Medicine*, 373(13):1195–1206.

Hamel, R., Ferraris, P., Wichit, S., Diop, F., Talignani, L., Pompon, J., Garcia, D., Liégeois, F., Sall, A. A., Yssel, H., and Missé, D. (2017). African and Asian Zika virus strains differentially induce early antiviral responses in primary human astrocytes. *Infection, Genetics and Evolution*, 49:134–137.

Harrison, J. A., Schall, J. D., Maskey, S., Mikulski, P. T., Knippenberg, M. T., and Morrow, B. H. (2018). Review of force fields and intermolecular potentials used in atomistic computational materials research. *Applied Physics Reviews*, 5(3):031104.

Harte, W., Swaminathan, S., Mansuri, M., Martin, J., Rosenberg, I., and Beveridge, D. (1990). Domain communication in the dynamical structure of human immunodeficiency virus 1 protease. *Proceedings of the National Academy of Sciences*, 87(22):8864–8868.

Hermans, J., Berendsen, H. J., Van Gunsteren, W. F., and Postma, J. P. (1984). A consistent empirical potential for water–protein interactions. *Biopolymers: Original Research on Biomolecules*, 23(8):1513–1518.

Hollingsworth, S. A. and Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143.

Howard, M. J. (1998). Protein NMR spectroscopy. *Current biology*, 8(10):R331–R333.

Hu, T., Wu, Z., Wu, S., Chen, S., and Cheng, A. (2021). The key amino acids of E protein involved in early flavivirus infection: viral entry. *Virology Journal*, 18(1):1–12.

Huang, J. and MacKerell Jr, A. D. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of computational chemistry*, 34(25):2135–2145.

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., and MacKerell, A. D. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods*, 14(1):71–73.

Huber, R. G., Marzinek, J. K., Boon, P. L., Yue, W., and Bond, P. J. (2021). Computational modelling of flavivirus dynamics: The ins and outs. *Methods*, 185:28–38.

Ichiye, T. and Karplus, M. (1991). Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics*, 11(3):205–217.

ICTV (2021). Taxonomy – *International Committee on Taxonomy of Viruses*. `https://talk.ictvonline.org/taxonomy/`. Accessed: 2021-12-05.

Israelachvili, J. (2011). *Intermolecular and surface forces: revised third edition.* Elsevier Science.

Izmirly, A. M., Alturki, S. O., Alturki, S. O., Connors, J., and Haddad, E. K. (2020). Challenges in dengue vaccines development: pre-existing infections and cross-reactivity. *Frontiers in Immunology*, 11:1055.

Jacobs, M. G., Robinson, P. J., Bletchly, C., Mackenzie, J. M., and Young, P. R. (2000). Dengue virus nonstructural protein 1 is expressed in a glycosyl-phosphatidylinositol-linked form that is capable of signal transduction. *The FASEB Journal*, 14(11):1603–1610.

John, A. L. S. and Rathore, A. P. (2019). Adaptive immune responses to primary and secondary dengue virus infections. *Nature Reviews Immunology*, 19(4):218–230.

Jones, M. L., Legge, F. S., Lebani, K., Mahler, S. M., Young, P. R., Watterson, D., Treutlein, H. R., and Zeng, J. (2017). Computational identification of antibody epitopes on the dengue virus NS1 protein. *Molecules*, 22(4):607.

Jones, S. and Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. *Progress in biophysics and molecular biology*, 63(1):31–65.

Jönsson, B., Lund, M., and Barroso da Silva, F. L. (2007). Electrostatics in macromolecular solutions. In *Conference on Food Colloids 2006*, volume 302, pages 129–154. Royal Society of Chemistry.

Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236.

Jurado, K. A., Simoni, M. K., Tang, Z., Uraki, R., Hwang, J., Householder, S., Wu, M., Lindenbach, B. D., Abrahams, V. M., Guller, S., et al. (2016). Zika virus productively infects primary human placenta-specific macrophages. *JCI insight*, 1(13).

Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652.

Kasahara, M. (2010). Genome duplication and T cell immunity. *Progress in molecular biology and translational science*, 92:7–36.

Kasprzykowski, J. I., Fukutani, K. F., Fabio, H., Fukutani, E. R., Costa, L. C., Andrade, B. B., and Queiroz, A. T. (2020). A recursive sub-typing screening surveillance system detects the appearance of the zikv african lineage in brazil: Is there a risk of a new epidemic? *International Journal of Infectious Diseases*, 96:579–581.

Kim, D.-G., Choi, Y., and Kim, H.-S. (2021). Epitopes of protein binders are related to the structural flexibility of a target protein surface. *Journal of Chemical Information and Modeling*, 61(4):2099–2107.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chemical reviews*, 116(14):7898–7936.

Kochakarn, T., Kotanan, N., Kümpornsin, K., Loesbanluechai, D., Thammasatta, M., Auewarakul, P., Wilairat, P., and Chookajorn, T. (2016). Comparative genome analysis between southeast Asian and south American Zika viruses. *Asian Pacific Journal of Tropical Medicine*, 9(11):1048–1054.

Kuhn, R. J., Dowd, K. A., Post, C. B., and Pierson, T. C. (2015). Shake, rattle, and roll: impact of the dynamics of flavivirus particles on their interactions with the host. *Virology*, 479:508–517.

Lanciotti, R. S., Kosoy, O. L., Laven, J. J., Velez, J. O., Lambert, A. J., Johnson, A. J., Stanfield, S. M., and Duffy, M. R. (2008). Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. *Emerging Infectious Diseases*, 14(8):1232.

Le Breton, N., Martinho, M., Mileo, E., Etienne, E., Gerbaud, G., Guigliarelli, B., and Belle, V. (2015). Exploring intrinsically disordered proteins using site-directed spin labeling electron paramagnetic resonance spectroscopy. *Frontiers in molecular biosciences*, 2:21.

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. (2016). 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*, 12(1):405–413.

Lemkul, J. (2018). From proteins to perturbed hamiltonians: A suite of tutorials for the GROMACS-2018 molecular simulation package [article v1. 0]. *Living Journal of Computational Molecular Science*, 1(1):5068.

Li, A., Yu, J., Lu, M., Ma, Y., Attia, Z., Shan, C., Xue, M., Liang, X., Craig, K., Makadiya, N., et al. (2018). A Zika virus vaccine expressing premembrane-envelope-NS1 polyprotein. *Nature Communications*, 9(3067):1–17.

Lindenbach, B. D. and Rice, C. M. (2003). Molecular biology of Flaviviruses. *Advances in Virus Research*, 59:23–61.

Liu, L., Dong, H., Chen, H., Zhang, J., Ling, H., Li, Z., Shi, P.-Y., and Li, H. (2010). Flavivirus RNA cap methyltransferase: structure, function, and inhibition. *Frontiers in Biology*, 5(4):286–303.

Lund, M. and Jönsson, B. (2003). A mesoscopic model for protein–protein interactions in solution. *Biophysical Journal*, 85(5):2940–2947.

Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, 40(4):346–358.

Machado, M. R., González, H. C., and Pantano, S. (2017). MD simulations of virus–like particles with supra CG solvation affordable to desktop computers. *Journal of chemical theory and computation*, 13(10):5106–5116.

Mackenzie, J. M., Jones, M. K., and Young, P. R. (1996). Immunolocalization of the dengue virus nonstructural glycoprotein NS1 suggests a role in viral RNA replication. *Virology*, 220(1):232–240.

Mackenzie, J. S., Gubler, D. J., and Petersen, L. R. (2004). Emerging flaviviruses: the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses. *Nature Medicine*, 10(12s):S98.

Maisuradze, G. G., Liwo, A., and Scheraga, H. A. (2009). Principal component analysis for protein folding dynamics. *Journal of molecular biology*, 385(1):312–329.

Malavige, G., Fernando, S., Fernando, D., and Seneviratne, S. (2004). Dengue viral infections. *Postgraduate Medical Journal*, 80(948):588–601.

Manrubia, S. C. and Lázaro, E. (2006). Viral evolution. *Physics of Life Reviews*, 3(2):65–92.

Mansfield, K., Johnson, N., Phipps, L., Stephenson, J., Fooks, A., and Solomon, T. (2009). Tick-borne encephalitis virus–a review of an emerging zoonosis. *Journal of General Virology*, 90(8):1781–1794.

Marrink, S. J., De Vries, A. H., and Mark, A. E. (2004). Coarse grained model for semiquantitative lipid simulations. *The Journal of Physical Chemistry B*, 108(2):750–760.

Marrink, S. J. and Tieleman, D. P. (2013). Perspective on the Martini model. *Chemical Society Reviews*, 42(16):6801–6822.

McGrath, E. L., Rossi, S. L., Gao, J., Widen, S. G., Grant, A. C., Dunn, T. J., Azar, S. R., Roundy, C. M., Xiong, Y., Prusak, D. J., Loucas, B. D., Wood, T. G., Yu, Y., FernÃ¡ndez-Salas, I., Weaver, S. C., Vasilakis, N., and Wu, P. (2017). Differential responses of human fetal brain neural stem cells to Zika virus infection. *Stem Cell Reports*, 8(3):715–727.

McQuarrie, D. (1976). *Statistical mechanics*. Harper and Row.

McQuarrie, D. A. and Simon, J. D. (1997). *Physical chemistry: a molecular approach*, volume 1. Sterling Publishing Company.

Mendonça, D. C., Macedo, J. N., Guimaraes, S. L., Barroso da Silva, F. L., Cassago, A., Garratt, R. C., Portugal, R. V., and Araujo, A. P. (2019). A revised order of subunits in mammalian septin complexes. *Cytoskeleton*, 76(9-10):457–466.

Menezes, G. d. L., Gonçalves, R. L., Lima, R. M., Pereira, M., Moreli, M. L., and da Silva, R. A. (2020). Mutation of critical residues reveals insights of yellow fever virus nonstructural protein 1 (ns1) stability and its formation. *Journal of Biomolecular Structure and Dynamics*, 38(17):5277–5286.

Mirza, M. U., Rafique, S., Ali, A., Munir, M., Ikram, N., Manan, A., Salo-Ahen, O. M., and Idrees, M. (2016). Towards peptide vaccines against Zika virus: Immunoinformatics combined with molecular dynamics simulations to predict antigenic epitopes of Zika viral proteins. *Scientific Reports*, 6:37313.

Mlakar, J., Korva, M., Tul, N., Popović, M., Poljšak-Prijatelj, M., Mraz, J., Kolenc, M., Resman Rus, K., Vesnaver Vipotnik, T., and Fabjan Vodušek, V. (2016). Zika virus associated with microcephaly. *The New England Journal of Medicine*, 2016(374):951–958.

Modhiran, N., Watterson, D., Muller, D. A., Panetta, A. K., Sester, D. P., Liu, L., Hume, D. A., Stacey, K. J., and Young, P. R. (2015). Dengue virus NS1 protein activates cells via Toll-like receptor 4 and disrupts endothelial cell monolayer integrity. *Science Translational Medicine*, 7(304):304ra142–304ra142.

Modis, Y., Ogata, S., Clements, D., and Harrison, S. C. (2004). Structure of the dengue virus envelope protein after membrane fusion. *Nature*, 427(6972):313.

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. (2008). The martini coarse-grained force field: extension to proteins. *Journal of chemical theory and computation*, 4(5):819–834.

Mottin, M., Braga, R. C., da Silva, R. A., da Silva, J. H. M., Perryman, A. L., Ekins, S., and Andrade, C. H. (2017). Molecular dynamics simulations of Zika virus NS3 helicase: Insights into RNA binding site activity. *Biochemical and biophysical research communications*, 492(4):643–651.

Muller, D. A. and Young, P. R. (2013). The *Flavivirus* NS1 protein: molecular and structural biology, immunology, role in pathogenesis and application as a diagnostic biomarker. *Antiviral Research*, 98(2):192–208.

Murata, K. and Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1862(2):324–334.

Mushegian, A. (2020). Are there 1031 virus particles on earth, or more, or fewer? *Journal of bacteriology*, 202(9):e00052–20.

Musso, D. and Desprès, P. (2020). Serological diagnosis of flavivirus-associated human infections. *Diagnostics*, 10(5):302.

Mustafa, M., Rasotgi, V., Jain, S., and Gupta, V. (2015). Discovery of fifth serotype of dengue virus (denv-5): A new public health dilemma in dengue control. *Medical journal armed forces India*, 71(1):67–70.

Nassar, R., Dignon, G. L., Razban, R. M., and Dill, K. A. (2021). The protein folding problem: The role of theory. *Journal of Molecular Biology*, page 167126.

Neamtu, A., Mocci, F., Laaksonen, A., and Barroso da Silva, F. L. (2022). Towards an optimal monoclonal antibody with higher binding affinity to the receptor-binding domain of SARS-CoV-2 spike proteins from different variants. *bioRxiv*.

Nesmelov, Y. E. and Thomas, D. D. (2010). Protein structural dynamics revealed by site-directed spin labeling and multifrequency EPR. *Biophysical reviews*, 2(2):91–99.

Neufeldt, C. J., Cortese, M., Acosta, E. G., and Bartenschlager, R. (2018). Rewiring cellular networks by members of the Flaviviridae family. *Nature Reviews Microbiology*, 16(3):125–142.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Noisakran, S., Dechtawewat, T., Avirutnan, P., Kinoshita, T., Siripanyaphinyo, U., Puttikhunt, C., Kasinrerk, W., Malasit, P., and Sittisombut, N. (2008). Association of dengue virus NS1 protein with lipid rafts. *Journal of General Virology*, 89(10):2492–2500.

Nomaguchi, M., Fujita, M., Miyazaki, Y., and Adachi, A. (2012). Viral tropism. *Frontiers in Microbiology*, 3:1.

Nozaki, Y. and Tanford, C. (1967). [84] Examination of titration behavior. *Methods in Enzymology*, 11:715–734.

Oliveira, E. R., de Alencastro, R. B., and Horta, B. A. (2016). The mechanism by which P250L mutation impairs flavivirus-NS1 dimerization: an investigation based on molecular dynamics simulations. *European Biophysics Journal*, 45(6):573–580.

Olson, M. A. (2018). Conformational selection of a polyproline peptide by Ebola virus VP30. *Proteomics*, 18(21-22):1800081.

Paixão, E. S., Teixeira, M. G., and Rodrigues, L. C. (2018). Zika, Chikungunya and Dengue the causes and threats of new and re–emerging arboviral diseases. *BMJ Global Health*, 3(Suppl 1):e000530.

Periole, X., Cavalli, M., Marrink, S.-J., and Ceruso, M. A. (2009). Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition. *Journal of chemical theory and computation*, 5(9):2531–2543.

Periole, X. and Marrink, S.-J. (2013). The Martini coarse-grained force field. In *Biomolecular Simulations: Methods and Protocols, Methods in Molecular Biology*, pages 533–565. Springer.

Persson, B. A., Lund, M., Forsman, J., Chatterton, D. E., and Åkesson, T. (2010). Molecular evidence of stereo–specific Lactoferrin dimers in solution. *Biophysical Chemistry*, 151(3):187–189.

Perutz, M. (1978). Electrostatic effects in proteins. *Science*, 201(4362):1187–1191.

Petersen, E., Wilson, M. E., Touch, S., McCloskey, B., Mwaba, P., Bates, M., Dar, O., Mattes, F., Kidd, M., Ippolito, G., Azhar, E., and Zumla, A. (2016). Rapid spread of Zika virus in the Americas–implications for public health preparedness for mass gatherings at the 2016 Brazil Olympic Games. *International Journal of Infectious Diseases*, 44:11–15.

Phillies, G. D. (1974). Excess chemical potential of dilute solutions of spherical polyelectrolytes. *The Journal of Chemical Physics*, 60(7):2721–2731.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26(16):1781–1802.

Pierson, T. C. and Diamond, M. S. (2020). The continued threat of emerging flaviviruses. *Nature microbiology*, 5(6):796–812.

Poveda-Cuevas, S. A. (2017). Coarse–grained modeling with constant pH of the protein complexation phenomena. Master's thesis, University of São Paulo.

Prakash, M. K., Barducci, A., and Parrinello, M. (2010). Probing the mechanism of pH-induced large-scale conformational changes in dengue virus envelope protein using atomistic simulations. *Biophysical Journal*, 99(2):588–594.

Priyamvada, L., Quicke, K. M., Hudson, W. H., Onlamoon, N., Sewatanon, J., Edupuganti, S., Pattanapanyasat, K., Chokephaibulkit, K., Mulligan, M. J., Wilson, P. C., et al. (2016). Human antibody responses after dengue virus infection are highly cross-reactive to zika virus. *Proceedings of the National Academy of Sciences*, 113(28):7852–7857.

Puerta-Guardo, H., Glasner, D. R., Espinosa, D. A., Biering, S. B., Patana, M., Ratnasiri, K., Wang, C., Beatty, P. R., and Harris, E. (2019). Flavivirus NS1 triggers tissue-specific vascular endothelial dysfunction reflecting disease tropism. *Cell reports*, 26(6):1598–1613.

Puerta-Guardo, H., Glasner, D. R., and Harris, E. (2016). Dengue virus NS1 disrupts the endothelial glycocalyx, leading to hyperpermeability. *PLoS pathogens*, 12(7):e1005738.

Queen, J. P., Quinn, G. P., and Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

Rapaport, D. C. (2004). *The art of molecular dynamics simulation*. Cambridge university press.

Sahu, I. D. and Lorigan, G. A. (2020). Electron paramagnetic resonance as a tool for studying membrane proteins. *Biomolecules*, 10(5):763.

Samish, I., Bourne, P. E., and Najmanovich, R. J. (2015). Achievements and challenges in structural bioinformatics and computational biophysics. *Bioinformatics*, 31(1):146–150.

Satoh, A. (2010). *Introduction to practice of molecular simulation: molecular dynamics, Monte Carlo, Brownian dynamics, Lattice Boltzmann and dissipative particle dynamics*. Elsevier.

Scarabelli, G. and Grant, B. J. (2013). Mapping the structural and dynamical features of kinesin motor domains. *PLoS computational biology*, 9(11).

Scaturro, P., Cortese, M., Chatel-Chaix, L., Fischl, W., and Bartenschlager, R. (2015). Dengue virus non-structural protein 1 modulates infectious particle production via interaction with the structural proteins. *PLoS pathogens*, 11(11):e1005277.

Scaturro, P., Stukalov, A., Haas, D. A., Cortese, M., Draganova, K., Płaszczyca, A., Bartenschlager, R., Götz, M., and Pichlmair, A. (2018). An orthogonal proteomic survey uncovers novel zika virus host factors. *Nature*, 561(7722):253–257.

Schmid, M., Speiseder, T., Dobner, T., and Gonzalez, R. A. (2013). DNA virus replication compartments. *Journal of Virology*, 88(3):1404–1420.

Schmitz, K., editor (1994). *Macro-ion Characterization: From Dilute Solutions to Complex Fluids*. American Chemistry Society, Washington.

Scott, W. R., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Krüger, P., and van Gunsteren, W. F. (1999). The GROMOS biomolecular simulation program package. *The Journal of Physical Chemistry A*, 103(19):3596–3607.

Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA: protein complexes. *Proceedings of the National Academy of Sciences*, 106(16):6620–6625.

Shao, Q., Herrlinger, S., Zhu, Y.-N., Yang, M., Goodfellow, F., Stice, S. L., Qi, X.-P., Brindley, M. A., and Chen, J.-F. (2017). The african zika virus mr-766 is more virulent and causes more severe brain damage than current asian lineage and dengue virus. *Development*, 144(22):4114–4124.

Sheridan, M. A., Yunusov, D., Balaraman, V., Alexenko, A. P., Yabe, S., Verjovski-Almeida, S., Schust, D. J., Franz, A. W., Sadovsky, Y., Ezashi, T., and Roberts, R. (2017). Vulnerability of primitive human placental trophoblast to Zika virus. *Proceedings of the National Academy of Sciences*, 114(9):E1587–E1596.

Shi, Y. (2014). A glimpse of structural biology through X-ray crystallography. *Cell*, 159(5):995–1014.

Simmonds, P., Becher, P., Bukh, J., Gould, E. A., Meyers, G., Monath, T., Muerhoff, S., Pletnev, A., Rico-Hesse, R., and Smith, D. B. (2017). ICTV virus taxonomy profile: *Flaviviridae*. *Journal of General Virology*, 98(1):2–3.

Simonin, Y., Loustalot, F., Desmetz, C., Foulongne, V., Constant, O., Fournier-Wirth, C., Leon, F., Molès, J.-P., Goubaud, A., Lemaitre, J.-M., Maquart, M., Leparc-Goffart, I., Briant, L., Nagot, N., Van de Perre, P., and S, S. (2016). Zika virus strains potentially display different infectious profiles in human neural cells. *EBioMedicine*, 12:161–169.

Singh, N. and Li, W. (2019). Recent advances in coarse-grained models for biomolecules and their applications. *International journal of molecular sciences*, 20(15):3774.

Sironi, M., Forni, D., Clerici, M., and Cagliani, R. (2016). Nonstructural proteins are preferential positive selection targets in Zika virus and related flaviviruses. *PLoS neglected tropical diseases*, 10(9).

Sniegowski, P. D., Gerrish, P. J., Johnson, T., and Shaver, A. (2000). The evolution of mutation rates: separating causes from consequences. *Bioessays*, 22(12):1057–1066.

Somnuke, P., Hauhart, R. E., Atkinson, J. P., Diamond, M. S., and Avirutnan, P. (2011). N–linked glycosylation of dengue virus NS1 protein modulates secretion, cell–surface expression, hexamer stability, and interactions with human complement. *Virology*, 413(2):253–264.

Song, H., Qi, J., Haywood, J., Shi, Y., and Gao, G. F. (2016). Zika virus NS1 structure reveals diversity of electrostatic surfaces among flaviviruses. *Nature Structural & Molecular Biology*, 23(5):456–458.

Songprakhon, P., Thaingtamtanha, T., Limjindaporn, T., Puttikhunt, C., Srisawat, C., Luangaram, P., Dechtawewat, T., Uthaipibull, C., Thongsima, S., Yenchitsomanus, P.-

t., et al. (2020). Peptides targeting dengue viral nonstructural protein 1 inhibit dengue virus production. *Scientific reports*, 10(1):1–16.

Souza, P. C., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., Patmanidis, I., Abdizadeh, H., Bruininks, B. M., Wassenaar, T. A., et al. (2021). Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature methods*, 18(4):382–388.

Sridhar, S., Luedtke, A., Langevin, E., Zhu, M., Bonaparte, M., Machabert, T., Savarino, S., Zambrano, B., Moureau, A., Khromava, A., et al. (2018). Effect of dengue serostatus on dengue vaccine safety and efficacy. *New England Journal of Medicine*, 379(4):327–340.

Srivastava, D., Santiso, E., Gubbins, K., and Barroso da Silva, F. L. (2017). Computationally mapping pKa shifts due to the presence of a polyelectrolyte chain around whey proteins. *Langmuir*, 33(42):11417–11428.

Stettler, K., Beltramello, M., Espinosa, D. A., Graham, V., Cassotta, A., Bianchi, S., Vanzetta, F., Minola, A., Jaconi, S., Mele, F., et al. (2016). Specificity, cross-reactivity, and function of antibodies elicited by zika virus infection. *Science*, 353(6301):823–826.

Stiasny, K., Allison, S. L., Schalich, J., and Heinz, F. X. (2002). Membrane interactions of the tick-borne encephalitis virus fusion protein E at low pH. *Journal of virology*, 76(8):3784–3790.

Sugiki, T., Kobayashi, N., and Fujiwara, T. (2017). Modern technologies of solution nuclear magnetic resonance spectroscopy for three-dimensional structure determination of proteins open avenues for life scientists. *Computational and structural biotechnology journal*, 15:328–339.

Talero-Gutiérrez, C., Rivera-Molina, A., Pérez-Pavajeau, C., Ossa-Ospina, I., Santos-García, C., Rojas-Anaya, M., and de-la Torre, A. (2018). Zika virus epidemiology from: Uganda to world pandemic, an update. *Epidemiology & Infection*, 146(6):673–679.

Teixeira, A. A. R., Lund, M., and Da Silva, F. L. B. (2010). Fast proton titration scheme for multiscale modeling of protein solutions. *Journal of Chemical Theory and Computation*, 6(10):3259–3266.

Teunissen, A. J., Pérez-Medina, C., Meijerink, A., and Mulder, W. J. (2018). Investigating supramolecular systems using Förster resonance energy transfer. *Chemical Society Reviews*, 47(18):7027–7044.

Thiemmeca, S., Tamdet, C., Punyadee, N., Prommool, T., Songjaeng, A., Noisakran, S., Puttikhunt, C., Atkinson, J. P., Diamond, M. S., and Ponlawat, A. (2016). Secreted

NS1 protects Dengue virus from mannose-binding lectin–mediated neutralization. *The Journal of Immunology*, page 1600323.

Tozzini, V. (2009). Multiscale modeling of proteins. *Accounts of Chemical Research*, 43(2):220–230.

Tribello, G. A. and Gasparotto, P. (2019). Using dimensionality reduction to analyze protein trajectories. *Frontiers in molecular biosciences*, 6:46.

Tsuchida, E. and Abe, K. (1982). Interactions between macromolecules in solution and intermacromolecular complexes. In *Interactions between macromolecules in solution and intermacromolecular complexes*, pages 1–119. Springer.

van der Kamp, M. W., Shaw, K. E., Woods, C. J., and Mulholland, A. J. (2008). Biomolecular simulation and modelling: status, progress and prospects. *Journal of the Royal Society Interface*, 5(suppl_3):173–190.

Vasilakis, N. and Weaver, S. C. (2017). *Flavivirus* transmission focusing on Zika. *Current Opinion in Virology*, 22:30–35.

Voth, G. A. (2008). *Coarse-graining of condensed phase and biomolecular systems*. CRC press, Boca Raton, 1st edition.

Wallis, T. P., Huang, C.-Y., Nimkar, S. B., Young, P. R., and Gorman, J. J. (2004). Determination of the disulfide bond arrangement of dengue virus ns1 protein. *Journal of Biological Chemistry*, 279(20):20729–20741.

Waman, V. P., Sen, N., Varadi, M., Daina, A., Wodak, S. J., Zoete, V., Velankar, S., and Orengo, C. (2021). The impact of structural bioinformatics tools and resources on sars-cov-2 research and therapeutic strategies. *Briefings in Bioinformatics*, 22(2):742–768.

Watterson, D., Modhiran, N., and Young, P. R. (2016). The many faces of the flavivirus ns1 protein offer a multitude of options for inhibitor design. *Antiviral research*, 130:7–18.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784.

Wessel, A. W., Doyle, M. P., Engdahl, T. B., Rodriguez, J., Crowe Jr, J. E., and Diamond, M. S. (2021). Human monoclonal antibodies against NS1 protein protect against lethal West Nile virus infection. *Mbio*, 12(5):e02440–21.

White, K. A., Enjuanes, L., and Berkhout, B. (2011). RNA virus replication, transcription and recombination. *RNA Biology*, 8(2):182–183.

WHO (2021a). *Dengue – World Health Organization*. `http://www.who.int/denguecontrol/epidemiology/en/`. Accessed: 2021-12-05.

WHO (2021b). *Zika – World Health Organization*. `http://www.who.int/emergencies/zika-virus/classification-tables/en/`. Accessed: 2021-12-05.

Winkler, G., Maxwell, S. E., Ruemmler, C., and Stollar, V. (1989). Newly synthesized Dengue–2 virus nonstructural protein NS1 is a soluble protein but becomes partially hydrophobic and membrane–associated after dimerization. *Virology*, 171(1):302–305.

Winkler, G., Randolph, V. B., Cleaves, G. R., Ryan, T. E., and Stollar, V. (1988). Evidence that the mature form of the *Flavivirus* nonstructural protein NS1 is a dimer. *Virology*, 162(1):187–196.

Xu, X., Song, H., Qi, J., Liu, Y., Wang, H., Su, C., Shi, Y., and Gao, G. F. (2016). Contribution of intertwined loop to membrane association revealed by Zika virus full–length NS1 structure. *The EMBO Journal*, 35(20):2170–2178.

Yang, H. and Rao, Z. (2021). Structural biology of sars-cov-2 and implications for therapeutic development. *Nature Reviews Microbiology*, pages 1–16.

Yap, S. S., Nguyen-Khuong, T., Rudd, P. M., and Alonso, S. (2017). Dengue virus glycosylation: what do we know? *Frontiers in microbiology*, 8:1415.

Ye, J., Zhu, B., Fu, Z. F., Chen, H., and Cao, S. (2013). Immune evasion strategies of flaviviruses. *Vaccine*, 31(3):461–471.

Yu, A., Pak, A. J., He, P., Monje-Galvan, V., Casalino, L., Gaieb, Z., Dommer, A. C., Amaro, R. E., and Voth, G. A. (2021). A multiscale coarse-grained model of the SARS-CoV-2 virion. *Biophysical journal*, 120(6):1097–1104.

Zhang, Z., Witham, S., and Alexov, E. (2011). On the role of electrostatics in protein–protein interactions. *Physical biology*, 8(3):035001.