

**UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ**

**Colegio de Ciencias e Ingenierías**

**Identificación de sitios potenciales de unión de moléculas  
farmacológicas en la proteasa del Virus del Dengue  
Proyecto de Investigación**

**Linda Estefanía Robayo Riofrío**

**Química**

Trabajo de titulación presentado como requisito  
para la obtención del título de  
Licenciada en Química

Quito, 18 de diciembre de 2015

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ  
COLEGIO DE CIENCIAS E INGENIERÍAS

**HOJA DE CALIFICACIÓN  
DE TRABAJO DE TITULACIÓN**

**Identificación de sitios potenciales de unión de moléculas farmacológicas en  
la proteasa del Virus del Dengue**

**Linda Estefanía Robayo Riofrío**

Calificación:

Nombre del profesor, Título académico

Miguel Ángel Méndez , Ph.D.

Firma del profesor

---

Quito, 18 de diciembre de 2015

## Derechos de Autor

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante: \_\_\_\_\_

Nombres y apellidos: Linda Estefanía Robayo Riofrío

Código: 00108449

Cédula de Identidad: 1720981685

Lugar y fecha: Quito, 18 de diciembre de 2015

## RESUMEN

La proteasa NS3/NS2B es esencial en el proceso de replicación del Virus del Dengue y constituye un blanco para el desarrollo de antivirales. Sin embargo, no ha sido posible desarrollar un fármaco clínicamente efectivo. El objetivo de este artículo es desarrollar un método para identificar "sitios de unión" alternativos para moléculas farmacológicas, utilizando un algoritmo de multilayer perceptron entrenado para clasificar los residuos que pueden ser significante para la actividad de la enzima de los residuos no esenciales. Varios factores obtenidos mediante mutagénesis de alanina, secuencia y conservación de la estructura y otros factores basados en la geometría fueron utilizados para entrenar al modelo. Tres sitios alternativos fueron identificados: NS3-Leu58 y residuos cercanos, el cluster NS3-His72, Phe116, Thr156; y los residuos NS2B-Thr77, NS2B-Thr83 y NS2B-Met84.

Palabras clave: Virus del Dengue, Escaneo de Mutagénesis computacional de Alanina, DV NS3

## ABSTRACT

Proteasa, DV NS2B cofactor, aprendizaje de máquina, ingeniería molecular, sitios de unión, inhibidor de proteasa The NS3/NS2B protease is essential on the Dengue Virus replication process and constitutes a desirable target for antiviral development. Nevertheless, it has not being possible to develop a clinically effective drug. The aim of this article is to develop an approach to identify alternative "bindable sites" for druglike molecules using an multilayer perceptron algorithm trained to classify residues that may be significant for the enzyme activity from the non essential. Several features derived from computational alanine scanning mutagenesis, sequence and structure conservation and other geometry-based features were used to train the model. Three alternative sites were identified: NS3-Leu58 and nearby residues, NS3-His72,Phe116, Thr156 cluster, and NS2B-Thr77, NS2B-Thr83, and NS2B Met84.

Keywords: Dengue Virus, Computational Alanine Scanning Mutagenesis, DV NS3 Protease, DV NS2B cofactor, Machine Learning, Molecular engineering, bindability sites, protease inhibitor

## TABLA DE CONTENIDO

Índice de tablas .....	7
Índice de figuras.....	8
Introducción .....	10
Metodología .....	11
Resultados.....	15
Discusiones.....	21
Referencias.....	23

## ÍNDICE DE TABLAS

**Tabla #1. Resumen del set de entrenamiento y validación cruzada**

**28**

## ÍNDICE DE FIGURAS

<b>Figura 1: Diagrama de flujo de métodos</b>	<b>29</b>
<b>Figura 2: Estructura de la proteasa NS3 y el cofactor NS2B</b>	<b>30</b>
<b>Figura 3: Conservación de los residuos identificados como sitio activo</b>	<b>31</b>
<b>Figura 4: Residuos clase A, sitio diana identificado</b>	<b>32</b>
<b>Figura 5: Residuos clase A, sitio diana identificado</b>	<b>33</b>
<b>Figura 6: Residuos clase A, sitio diana identificado</b>	<b>34</b>
<b>Figura 7: Residuos clase A, cluster identificado</b>	<b>35</b>
<b>Figura 8: Representación de interacciones</b>	<b>36</b>



---

## Identification of bindability sites for drug-like molecules at Dengue Virus protease

M. A. Méndez · D. Aguilera-Pesantes ·  
P. E. Méndez · D. Moyocana · L. E.  
Robayo · F. J. Torres

Received: date / Accepted: date

---

**M. A. Méndez**

Universidad San Francisco de Quito, Instituto de Simulación Computacional (ISC-USFQ),  
Diego de Robles sn y Vía Interoceánica, 17-1200-841, Quito, Ecuador  
Universidad San Francisco de Quito, Escuela de Medicina, Diego de Robles sn y Vía Interoceánica, 17-1200-841, Quito, Ecuador  
Tel. + 593-2-297-1700  
E-mail: mmendez@usfq.edu.ec

**D. Aguilera**

Universidad San Francisco de Quito, Instituto de Simulación Computacional (ISC-USFQ)  
Diego de Robles sn y Vía Interoceánica, 17-1200-841, Quito, Ecuador

**P. E. Méndez**

Universidad de Las Americas, Departamento de Estadística, Granados E12-41 y Colimes,  
Quito, Ecuador

**D. Moyocana**

Universidad San Francisco de Quito, Grupo de Química Computacional y Teórica, Departamento de Ingeniería Química, Diego de Robles sn y Vía Interoceánica, 17-1200-841, Quito, Ecuador

**L. E. Robayo**

Universidad San Francisco de Quito, Instituto de Simulación Computacional (ISC-USFQ)  
Diego de Robles sn y Vía Interoceánica, 17-1200-841, Quito, Ecuador

**F. J. Torres**

Universidad San Francisco de Quito, Grupo de Química Computacional y Teórica (QCT-USFQ), Departamento de Ingeniería Química, Diego de Robles sn y Vía Interoceánica, 17-1200-841, Quito, Ecuador  
Universidad San Francisco de Quito, Instituto de Simulación Computacional (ISC-USFQ), Diego de Robles sn y Vía Interoceánica, 17-1200-841, Quito, Ecuador  
Université de Bordeaux, ISM, UMR 5255, 351, Cours de la Libération, Talence F-33405, France

**Abstract** The NS3/NS2B protease is essential on the Dengue Virus replication process and constitutes a desirable target for antiviral development. Nevertheless, it has not being possible to develop a clinically effective drug. The aim of this article is to develop an approach to identify alternative "bindable sites" for drug-like molecules using an multilayer perceptron algorithm trained to classify residues that may be significant for the enzyme activity from the non essential. Several features derived from computational alanine scanning mutagenesis, sequence and structure conservation and other geometry-based features were used to train the model. Three alternative sites were identified: NS3-Leu58 and nearby residues, NS3-His72, Phe116, Thr156 cluster, and NS2B-Thr77, NS2B-Thr83, and NS2B-Met84.

**Keywords** Dengue Virus · Computational Alanine Scanning Mutagenesis · DV NS3 Protease · DV NS2B cofactor · Machine Learning · Molecular engineering · bindability sites · protease inhibitor

## 1 Introduction

The NS3 protease domain bound to its cofactor NS2B (NS2B-NS3Pro) is essential on the replication process of the Dengue Virus (DV) [1–4] and therefore for the infection process that can result on dengue fever[5], hemorrhagic dengue [6] and dengue shock syndrome[4]. The worldwide incidence has being steadily increasing these last years threatening the health of millions of people.[4, 6, 7]. For these reasons it has being considered a global health priority to develop antiviral drugs against DV.[4, 8] The development of inhibitors has being an aim for industry and academia alike for several years nevertheless a clinical drug has not being obtained.[4, 9–11] Indeed, mainly supportive fluid therapy constitutes the only option available for patient treatment.[11] Vaccines are now on clinical trials[12, 13] though there is not any on the market yet. If a vaccine is released in the near future it will provide a much needed tool for fighting this virus but their effectiveness on the global health landscape will need to be tested in the coming years. A combined approach for the fight against DV should include antiviral drugs as the best option against this public health threat.

Several strategies are currently being used on the search of effective drug-like candidates for DV antivirals. Inhibitors for viral proteases are proven targets since exist at least ten clinical available inhibitors for HIV protease and two for HCV.[9]. Nevertheless, there is not yet a clinical DV protease inhibitor available.[14] The availability of the DV protease structures for almost all the serotypes has allowed that methods as docking[15, 16] to be used with the aim of identifying lead compounds. Here we suggest a straightforward and complementary strategy on the search of these compounds. In particular, we aim to predict and to find bindability sites for drug-like molecules. We built a classifier model using machine learning methods where for the training and validation of the model we generated quantitative and qualitative features. These were features related to the function and structure of the protease in its functional folding. Finally, we used site directed mutagenesis results previously reported in the literature for establishing biologically relevant classes for the classifier model. The broad contribution from this article

is that residues and sites found by our methodology as important for the function of the protease should be targets for further characterization on the development process of drug-like molecules for DV NS2B-NS3Pro.

## 2 Methodology

First, we obtained several quantitative and qualitative features by analyzing the functional structure and sequence of a DV NS2B-NS3 protease. The methodology followed here makes use of a crystallographic structure previously reported (PDB 3U1I)[17] as input for computational alanine scanning mutagenesis and the calculation of several properties of the enzyme. Second, the amino acid sequence was used as input for an analysis using several tools provided by the online suite Phyre2[18] in order to find conservation not only at the level of sequence but at the level of protein structure. Finally, we compiled from several sources site directed mutagenesis results to create classes (categories) regarding the importance of a residue for the DV protease activity. Lastly, all this information was used to train a machine learning classifier that would predict for the residues where not experimental site directed mutagenesis results were provided, if the residue will be potentially important for the activity of the enzyme. The methodology is summarized on Figure 1.

### 2.1 The DV protease

The structure chosen for all the analysis corresponds to the one reported by Nitsche et al corresponding to the serotype 3 DV protease [17]. -An important part on the replication process of the virus is the processing of the DV poliprotein that contains structural and non structural viral proteins. Host proteases as well as the DV protease are needed. This poliprotein contains the non structural protein NS3 whose N terminus has the protease activity (NS3Pro, 180 residues) and it also contains the NS3Pro's peptide cofactor NS2B (aprox. 40 residues).[7] Figure 2 shows this structure where the three most important residues for the catalytic activity (His51, Asp75 and Ser135) of the enzyme are highlighted. This structure was selected because the NS3Pro protein is folded in its functional conformation being able to interact with the ligand and forms the active catalytic site. Other structures are available that may give other kind of insights but many do not include the protein bound to the ligand in its functional folding.[4]

The full structure of DV serotype 4 NS3-NS2B has being solved,[19] and contrary to the Hepatitis C Virus the position of the protease domain relative to the helicase domain does not include a interface of the protease domain's catalytic site in direct contact with the helicase domain. This justifies our choice of working only with the protease domain/NS2B in this article. We have used the 3U1I structure but this crystal structure contains a dimer that contains two protease/cofactor assemblies that are not equal. For this study only the A and B chains were used. The chains C, D and the ligands (E, F) were deleted. The sequence of the chosen serine protease subunit NS2B [PDB: 3U1I, chain A] follows: GPLGSDLTVE KAADVWEEEE AEQTVGVSHNL MITVDDDGTM RIKDDETENI L

The Sequence of the serine protease NS3Pro domain [PDB: 3U1I, chain B] follows: GGGGSGGGGS GVLWDVPSPP ETQKAELEEG VYRIKQQGIF GK-TQVGVG VQ KEGVFHTMWH VTRGAVLTHN GKRLEPNWAS VKKDLISYGG GWRLSAQWQK GEEVQVIAVE PGKNPKNFQT MPGTFQTTTG EIGAIALDFK PGTSGSPIIN REGKVVGLYG NGVVTKNGGY VSGIAQTNAE PDGPTPELEE E

## 2.2 Data collection for class definition

A machine learning classifier needs for the training and validation sets a known class or category for each group of features of a certain object, an amino acid residue on the NS3Pro or the NS2B in this study. The algorithm is going to be trained to recognize to which specific class each residue belongs to. In the current case, the three most important classes of interest are the cases where a mutation in a given residue eliminates, modifies, or it does not have an effect on the enzymatic activity of the DV protease. We proceeded to collect data reported in the existing literature from site directed mutagenesis studies where the effect of mutating some residues *in vitro* for Alanine, and/or other residues was evaluated. We assigned our findings for the literature collected experimental mutagenesis results within categories: A, total loss of activity; B, major loss of activity; C, moderate-activity lost; D slight activity lost and WT, same catalytic activity as wild type. These data are presented on Table 1.

## 2.3 Protease structural and function related feature generation

### *Computational alanine scanning mutagenesis derived features*

Although there are programs for predicting by homology the structure of a protein such as the online tool Phyre2 that we used for several analysis, these type of software do not have the functionality to predict the wider structural effect of a point mutation. Molecular dynamics simulations (MD) in principle could predict such effects. The MD results will depend on the force field used and the calculation could be very time consuming since it will need to re-run all the simulations for each residue or mutated residues. An alternative and useful approach is to use Prime/Molecular Mechanics Generalized Born Surface Area (Prime/MM-GBSA) method [20, 21] that performs a series of minimizations of the receptor and ligand and point energy calculations. Prime MM-GBSA uses an implicit (continuum) solvation model. This allows to efficiently characterize an interaction with a ligand and allows in fact to mutate every residue on a certain ligand to study the effect on the interaction. This allowed us to perform an alanine scanning mutagenesis analysis where each residue is mutated for an alanine residue.[22] The underlining principle behind this analysis is that if a certain residue is important for the interaction ligand-receptor when mutated for alanine, the total  $\Delta G$  binding will decrease.[23] The software then calculates the  $\Delta\Delta G$  binding (referred here simply as  $\Delta$  affinity). Additionally, the change in stability of a protein caused by the mutation can also be calculated,  $\Delta\Delta G$  stability (referred here simply as  $\Delta$  stability). These parameters give us insight on the importance of a residue for the whole protein structure.(BioLuminate 1.1 user manual)

Our aim with this computational alanine mutagenesis scanning (CAMS) is to understand the driving forces underlying protein-protein interactions by analyzing the changes in protein binding affinity and/or protein stability of the complex NS2B-NS3Pro. For CAMS a known starting structure is required. Experimental molecular structures can be acquired as pdb files from databases as the Protein Data Bank or the European Bioinformatics Institute. The structures on the repositories were obtained by X Ray Crystallography PDB ID: 3U1I as discussed above.[17] The Crystallographic structure lacks hydrogen atoms and it does not resemble necessarily its aqueous state. Hence, protein preparation must be performed prior calculations.

In this work we report  $\Delta$  protein binding affinity,  $\Delta$  protein stability predictions from computational alanine scanning mutagenesis of the NS2B-NS3 Dengue Virus (DV) Protease using MM-GBSA approach, OPLS2005 force field, VSGB solvent model, Prime (version 3.1, Schrodinger, LLC, New York, NY, 2012), rotamer search algorithms included in the program BioLuminate (version 1.0, Schrodinger, LLC, New York, NY, 2012). For each mutagenesis, energy minimizations were carried out according to the standard MM-GBSA protocol but not a full molecular dynamics of the starting structure since we are using a crystal structure in a known functional conformation. It was decided to explore the point mutation effect directly on this structure with the required preparation steps mentioned below.

The structure preparation was performed using the Protein Preparation Wizard tool available in Prime within the Schrodinger suite. This tool uses as input the original pdb file and perform a series of tasks as elimination of the water molecules between the ligand and the protein, deletion of duplicated chains on the file, filling missing atoms and residues, elimination of spurious bonds with metallic ions. In addition, the tool assigns bond orders, formal charges and corrects the orientation of certain groups if needed. Finally, it runs computational algorithms to release tensions created during adjustments, examines the refined structure to identify the need of final adjustments and checks the orientation of water molecules and other groups present.[24]

#### *Structure-based analysis*

Reactive residues identification was performed on the structure. In this analysis each region or residue is compared with certain established patterns from the program default script, then each residue is assigned the reaction that is most prone to suffer (only reactive residues are reported by the software). The reactions the software can identify are deamination, oxidation, glycosylation and proteolysis.[25].

In addition we calculated some physical and chemical properties of each residue using the Bioluminate Residue Analysis Tool.[25] Specifically we calculated for all residues on the NS2B and NS3Pro the Solvent Accessible Surface Area (SASA), the Hydropathy, and the residue charge.

#### *Phyre 2 - Conservation analysis*

The residue sequence of NS2B and NS3Pro were analyzed individually using Phyre2. This online tool compares the residue sequence and the folding pattern with other reported structures. In this case, the 3U1I DENV3 (Dengue Virus Serotype 3) Protease was compared with protease structures from the other three Dengue Virus Serotypes. In addition, it was compared with Proteases from the West Nile Virus, Murray Encephalitis Virus and Japanese Encephalitis Virus. Using these data each residue was classified as highly-conserved (folding and sequence), moderately-conserved (only sequence); and not conserved. These clas-

sification are reported as 2, 1 and 0 respectively as presented on tables S1, S2, and S3. In order to combine all the conservation analysis for all DV serotypes, we established two classes, True (T) and False (F). When the residue on serotype 3 was highly conserved for all serotypes was assigned as T and when not as F. (See Table 1)

#### *Raptor X - Active site prediction*

The residues that are more likely to be part of the active site were identified using Raptor X.[] In table 1 and Tables S1, S2 and S3 these are reported as 1 and 0 to denote if the residue was predicted as part of the active site or if was not respectively.

## 2.4 Representations and diagrams

All representations and diagrams were created using Maestro (Schrödinger LLC) or Visual Molecular Dynamics Software, version 1.9.1.[27] The detailed color coding and symbols used can be found in the supporting information Figure S1 and S2.

## 2.5 Classification

In order to generate an integral interpretation of all the previous analysis, we used each of them as a feature to train a machine learning algorithm in order to classify if a certain residue if mutated, and possibly if targeted by a "drug-like molecule" would eliminate the function of the enzyme. We perform a literature search for experimental mutations where the loss, partial loss or no loss of the activity was reported. With this data we established first five classes (described above), and later we condensed the five classes into three classes, class A (Total loss or high loss of activity), class C (moderate loss of activity), and WT (slight loss of activity or as much activity as the wild type protein). As features we selected fourteen properties: 1) residue type, 2) activity (NS3 or NS2B), 3)  $\Delta$  affinity (as defined above), 4)  $\Delta$  stability solvated (as defined above), 5) Solvent accessible surface area (SASA), 6) hydrophathy, 7) protease's amino acid sequence conservation with all DV serotypes (true or false), degree of protease's amino acid sequence sequence conservation with 8) DENV1 M, 9) DENV 4 pH 8.5, 10) DENV2, 11) West Nile Virus, 12) Murray 2 Valley EV, 13) Japanese Encephalitis Virus, 14) RaptorX prediction if the residue is on the active site.

Two main machine learning algorithms were tested Random Forest and a Multilayer Perceptron algorithm. The Multilayer perceptron algorithms is a feedforward neural network trained with the backpropagation learning algorithm. The neural network was automatically build by the default algorithm provided by Waikato Environment for Knowledge Analysis (WEKA) v. 3.6.13 (The University of Waikato, Hamilton, New Zealand).[28] The parameters and conditions used for the algorithm were a learning rate of 0.3, momentum term 0.2, a nominal to binary filter. All the attributes were normalized as well as the the numerical classes. We did not used a decay scheme. For the validation, a 10 fold cross-validation protocol was used. For testing these algorithms the software WEKA was used.[28] The size of the training plus validation set was 40 residues (5 of NS2B and 35 of NS3); a summary of the this set is presented on Table 1. After the chosen model was built,

we applied it over all the NS2B and NS3 residues to classify the data within class A, class C or class WT. To achieve this, we used a two step approach. First with a Multilayer Perceptron algorithm we classified some residues of the original training/validation set and from the Test set onto the class A. These residues for the second phase were removed from the training/validation set and from the Test set in order to apply a second, best suited, multilayer perceptron model (MP model 2) in order to classify the remaining residues between class C and class WT. In summary, the parameters for generating this second model (MP model 2) were the same as for the first, but we used a modified training/validation set without the data classified as class "A" for the first Multilayer Perceptron model (MP model 1).

### 3 Results

A multi step approach has being followed in this study in order to bridge atomistic details with overall properties evaluated by homology based predictions, geometry based analysis, and conservation information. We aim to find novel ligandability sites or ligandability features on the DV protease. For "bindability" we mean a definition in the same spirit as it is used by Sheridan et al (2010) [29] therefore here we report measurements as tools to asses "bindable sites" for "drug-like molecules".

Detailed descriptions of the active site of a DV protease has being already reported.[1, 19, 30–32] Nevertheless, a very detailed description of binding pockets has being proposed to not necessarily be the best practical approach to identify "bindable sites" because the molecular interactions *in vivo* show flexibility and diversity rather than a static nature.[33] Here we report features and analysis intended to give a more general description that could provide guidelines or insights to researchers for the design of novel drug-like molecules for inhibiting this enzyme. Briefly, first we report a set of analysis based on properties readily derived from geometry alone including computations based on computer based alanine scanning mutagenesis analysis, and hydrophathy Second, we report predictions based only on sequence of residues on the active site.[34] Third, a literature search of mutations on several DV protease and its effects was used as the class for the training and validation set for a machine learning model that uses the features calculated by us in order to classify if a residue on the NS2B or the NS3-Pro if mutated potentially will cause a loss of the activity of the enzyme.

The significant correlation between the predicted and the experimental data (directed-site mutagenesis) shows that the model performs well at predicting activity hot spots (regions or single residues where mutations have significant impact on the enzyme activity) of the complex and it shows to be a viable methodology for identification of bindability sites for drug-like molecules.

#### 3.1 Geometry based analysis

For all the analysis we used the geometry of the DV serotype 3, structure PDB 3U11 (chains A and B), that is folded on the functional conformation. With the residue analysis tool we calculated SASA for each residue, and the Hydrophathy profile both with Maestro. All the NS2B-NS3Pro residues were mutated with the Alanine

Scanning Mutagenesis Bioluminate Tool, except for the outermost residues (NS2B: Asp50, Asp88; NS3: Gly1, Gly0, Ser1, Gly2, Val3, Leu4, Trp5, Asp6, Gln167, Thr168, Asn169, Ala170, Glu171). All mutations and calculations were performed by mutating each residue one by one for Alanine. The properties obtained from this analysis were  $\Delta$  SASA (total),  $\Delta$  SASA (not polar),  $\Delta$  SASA (polar),  $\Delta$  pKa,  $\Delta$ affinity,  $\Delta$  hydrophathy, total rotatable bonds,  $\Delta$  stability (gas) and  $\Delta$  stability (solvated). From all these variables, we used as features for the classifier model only  $\Delta$  affinity and  $\Delta$  stability (solvated).

According to the MM-GBSA method[35] for the alanine scanning mutagenesis, the first step is to calculate individually the energies of the NS3Pro system, the NS2B system and finally of the NS2B-NS3Pro system for the wild type; followed by the calculation of the energies of the same three systems for the mutant type. The score of  $\Delta$  affinity is obtained by comparing the energies of the systems for the wild type to the systems of the mutant type, for each residue mutation, all based in a thermodynamic cycle described elsewhere.[35] Here we use a cutoff of 4 kcal/mol to define a hotspot, referred later as a "high"  $\Delta$  affinity or "high"  $\Delta$  stability. These results are presented on Table 1 with the "high" marked on red color.

The  $\Delta$  affinity was chosen as a feature for all residues since it will allow identifying if a residue is important for the interaction between NS2B and NS3Pro. In addition, the parameter  $\Delta$  stability was also selected as feature. These value give insights about the importance of a residue in the overall stability of the active conformation. In the case of His51, known as part of the catalytic triad, showed a high  $\Delta$  stability energy. Considering that His51 could be positively charged and NS2B-Asp81 negatively charged, they may form a salt bridge that will explain the high  $\Delta$  stability of the complex. In addition His51 was predicted susceptible to undergo oxidation reactions. Ser135, another of the triad catalytic residues, did not show a high  $\Delta$  stability energy. These results suggest that there is not a simple correlation between  $\Delta$  stability energy with the importance of the residue for the activity of the enzyme. This justify our choice of these variables as some of the features for the classifier, with the same weight as other properties, since several properties may characterize a residue as important and not just a single feature.

A summary of the results for the calculations on each residue mutated for Ala are presented on Table 1 if the experimental site mutagenesis information is included. Later this information will be assigned as the class for the training and validation sets. If the experimental site mutagenesis information was not available or taken into consideration as a class for the classifier, the results for those particular residues are reported on Tables S1, S2 and S3. The results of Reactive Residue Identification are presented on the Supplementary Information, Table S6. Not all residues are reactive and as a consequence there is not a entry for all residues, this parameter was not chosen as a feature to be used for classification. Nevertheless, it may be useful for the reader interested on a particular residue. For example, Asn152 that is part of the active site, is located in contact with NS2B and has a high  $\Delta$  affinity, was predicted to be susceptible to deamination. Indeed, this residue is interacting with NS2B-Gly82 via a hydrogen bond, in agreement with the calculated interaction energy.



### 3.2 Homology based analysis

In order to facilitate a generalization of the methodology presented here, we explore Raptor X as a tool to predict if a residue is part of the active site. This prediction tool may be used for proteins where not such detailed knowledge of the residues on the catalytic site is available. With the Raptor X algorithm of active site prediction, we analyzed all residues in the DV protease sequence. The results are summarized on Table 1 and Tables S1, S2 and S3. These results were in good agreement with the residues identified experimentally and reported previously in the literature[1, 31, 36, 37]. Twenty two residues were predicted to be part of the active site. Asp75, His51 and Ser135, the three residues of the catalytic triad, were correctly predicted by raptor X to be part of the active site. On the other hand, geometry based analysis did not showed neither a high  $\Delta$  stability energy neither a high  $\Delta$  affinity energy for the key residue SER135. Another example, in the case of Tyr161, even when it showed to be just moderately conserved, Raptor X predicted the residue as part of the active site. For this residue, Tyr 161, a high  $\Delta$  stability energy was found as well as to be susceptible to oxidation. These examples show that any particular attribute by themselves is not enough to assign if a residue is important for the enzyme activity. Therefore, this justify our multi step approach, since it shows that this particular Raptor X algorithm classifies the residue correctly (a part or not of the active site), adding a parameter (feature) suggesting that the residue is important for the enzyme activity even when others of our main parameters derived from geometry based analysis alone do not suggest its importance.

#### *homology based conservation analysis*

Protease inhibitors for HIV and HCV have shown rapid emergence of resistant viral strains.[9] In order to lead the search in the direction to overcome this pitfall, we have used several features to train the classifier based on conservation criteria. The assumption is that the most a residue is conserved the least likely is to be mutated without an important cost for the virus because of the importance of that residue for the enzyme activity. In addition, considering that NS2B actively participates in the formation of sub-pockets in the protease active site[9], our current model worked with a structure that contained NS2B when cooperating for the binding with a ligand. Conservation studies from the point of view of NS2B cofactor could not be made with all serotypes because of lack of structures. (The current results reflect conservation analysis only with serotypes 1 and 4). Previously for NS3, it has being reported that the degree of aminoacid sequence conservation is between 63% to 74% [9] what is positive for the search of drug-like molecules able to inhibit the protease on all four serotypes. On the other hand, this high degree of conservation makes more difficult to identify the non obvious key residues for the enzyme. In order to overcome this, we have used as features for the classifier also conservation predictions in comparison with West Nile Virus, Murray 2 Valley EV, and Japanese Encephalitis Virus.

All the results on the present section were obtained with Phyre 2 for the DV serotype 3. The previous residues identified as part of the active site by Raptor X were analyzed with Phyre 2 and showed that 12 of the 22 residues were highly conserved between the four main DV serotypes, 6 residues showed a moderate degree of conservation and 4 of them were not conserved with the other three DV serotypes. Figure 3 shows the degree of conservation for the residues that Phyre 2 identified as part of the active site. In the case of His 51, known as one of

the catalytic triad residues, a high conservation between the four main serotypes was found (shown in red). In general, the residues on the NS2B cofactor are not highly conserved between the DV serotypes studied. A moderate (sequence only) conservation is observed only with serotypes 1 and 4. Comparisons could not be made with serotype 2 because there was not an available structure of DV serotype 2 NS2B as mentioned previously on the Phyre2 database. Considering only the analysis between Serotypes 1, 3 and 4, the conservation degree obtained is in agreement with previous reports (64.1 % vs 68.1 % [37]) for the NS2B domain. In addition, there is a high degree of conservation at the sequence level between DV serotype 3 and the West Nile Virus.

Since high  $\Delta$  affinity energy may suggest an important role in the interaction between NS2B and NS3, we clustered those residues that are highly conserved and that have a high  $\Delta$  affinity energy (a value greater than 4 kcal/mol). It was found that only six NS3 residues that comply with this criteria were conserved on all four DV serotypes. These were residues Gly21, Tyr23, Ile25, Phe46, Leu58 and Asn152. Figure 9 presents the interaction diagram of some of these residues. The residue Gly21 presents hydrophobic interactions with NS2B-ALA 57. Tyr23 is in close proximity to residues NS2B-ALA56 and NS2B-Val 53. In addition Tyr23 showed a high  $\Delta$  stability that may be explained by hydrogen bonding with other residues within the same chain.(Figure 9)

In a similar trend of thought as for  $\Delta$  affinity, high  $\Delta$  stability energy may suggest an important role in the conservation of the functional structure for the complex NS2B and NS3. For this reason we clustered those residues that are highly conserved and that have a high  $\Delta$  stability energy (a value greater than 4 kcal/mol). In overall, all residues that complied with these criteria were neutral amino acids. Some of the residues found were His51, Val52, Thr53, Tyr150, Gly151, Gly153. Figure S4 presents the interaction diagram of some of these residues. SASA shows evidence that all six residues are accessible to the solvent. Gly133 and Ser 135 also comply with the criteria and as mentioned previously Ser 135 is one of the residues of the catalytic triad.

### 3.3 Machine Learning classification

High-throughput computational screening, and *in vitro* biochemical assays have being used to identify possible lead compounds.[11] Anthracene based lead compounds have shown to interact with the active site and P1 pocket of the NS3Pro. A variety of other compounds of very different structures have being identified using these methods but many were weakly active or they have too low therapeutic index.[9] Some protease inhibitors have being tested for both, West Nile Virus protease and DV, it was found that they inhibit both enzymes but with a lower efficacy for DV protease.[9] These results suggested subtle differences on susceptibility between these proteases.[11] Part of the challenge for developing the inhibitors has being the shallow nature of the binding site.[9]. (See Figure 3) Another problem is that certain inhibitors will recognize the Arg in P1 pocket that is also present in human proteases such as trypsin, thrombin and elastase.[9] This difficulty finding promising clinically feasible drug-like compounds suggest the need for inhibitors that target other sites of the enzyme than the active site

and sites close to the active site. This constitutes one of the main motivations for the search of other potentially bindability sites outside the active site.

The search for non competitive inhibitors has also being pursued actively. For example, Lys74, Leu149 and Asn152 have being previously identified to be on a non active site pocket.[38] Lys74 is directly bonded to Asp75 so it was suggested that the formation of a hydrogen bond between Lys74 and one of the tested inhibitors may have caused a conformational change on Asp75 and therefore an effect on the catalytic activity.[38] Neither of these three residues was identified by our classifier since the three residues were already included in our training/validation set where Lys74 and Leu149 were assigned as residues where a very important change in enzyme activity is observed when mutated. Asn152 showed a moderate effect (Table 1). Recently, another inhibitor has being found for this pocket close to the active site.[14] Other inhibitors directed to pockets close to the active site have proven to inhibit all four serotypes suggesting that it is possible to develop drug-like molecules against the four serotypes.[39] Within this context we have made use of our classifier model in order to corroborate previous findings and suggest other pockets.

A random forest algorithm and a multilayer perceptron algorithm were tested on their ability to classify the set of data available. The first attempts to classify on the five categories assigned from literature resulted in classifiers that mostly assigned all the residues to the category with the largest number of data available. The performance indexes were poor. We concluded that the classes were unbalanced since we have more data for certain categories than for others. We decided that there was not enough data for five categories and decided to merge categories into three categories. Categories A and B were merged in class A, category C was left as class C and categories D and WT were merged on class WT. For these new data set for validation/training we found that the best performance was obtained by the multilayer perceptron executed in a two phases approach. Here we report only the algorithm found to be the best choice to classify the data. All the data to be classified is presented on the Supplementary Information, Tables. S1, S2 and S3.

The two phases consisted on first running all the training/validation set in order to get a first classification. In this first phase, residues classified to class A were obtained with good performance parameters (Recall 0.722). With this validated model we proceeded to run the test set to be classified (Tables S1, S2 and S3). In addition, we also run the validated model on the training/validation set again in exactly the same form as for the test set. We got as a result each residue classified into categories A, C or WT.(Supplementary information Fig. S5-8). All data that was classified as A was taken out of the validation/training set for phase two. With this approach we were trying to help the phase two algorithm trainin for an improved classification into classes C and WT that on the first round were not clearly separated. On the phase two, we proceeded to train a new multilayer perceptron model and we obtained good performance parameters for this second classifier (Recall for class C = 0.786). In this way we were able to classify all the data with high level of confidence within class A, or class C. Residues classified on class WT on the second phase still did not give as high confidence as for the residues classified as A on phase one, or the residues classified as C on phase two. We decided to leave all residues not classified as A or C, on their own category of "non classified". The confusion matrix and the detailed accuracy by class for both

phases can be found on the supplementary information, Table S4 and Table S5. The class assigned for each residue on phase one is reported in figure S5-8.

One of the aims of the current work was to identify new possible sites on the DV protease. As a working hypothesis we explored interactions sites between the NS2B and the NS3 chains. As described previously one of the analysis was to calculate the  $\Delta$  affinity and we identified three regions with residues with high  $\Delta$  affinity energies. (See Figure 4 - CASM). Nevertheless, the results of the classifier showed that none of the residues on these regions was classified as important but NS2B-Met 84. Surprisingly, close-by residues on only one of the three regions that showed higher  $\Delta$  affinity values were identified as likely to cause lost of enzymatic activity if changed (Class A). Those were NS2B-Thr 77 and NS2B-Thr83.(See Figure 4 In addition, other three residues were clustered in a region not characterized by high  $\Delta$  affinity binding energy (NS2B-Thr 68, NS2B-Gly69, and NS2B-His72 as shown in figure 5). NS2B-Met84 was identified as important by the classifier with a probability to belong to class A 0.985, the first highest ranked residue other than a Thr for the NS2B cofactor. NS2B-His72 and NS2B-Gly69 were the next two highest ranked residues. All residues reported here were identified with a confidence value by the classifier higher than 0.969). These results suggest other reasons behind the importance of these clusters different from exclusively their contribution to the binding between NS2B and NS3.

On the DV NS3 protease the classifier identified on class A the residues Phe46, Thr48, His51, Thr53, Asp75, Trp89, and Thr156 as the seven highest ranked residues. (See Figure 6) Additional residues were classified as class A, all the 25 residues identified from both chains with a confidence higher than 0.95 are reported on the Supplementary Information (Pages S2-S5). To these overall results we can try to find the rational behind the classification by analyzing some of the features used on the classification of these residues. For example, Phe46 presents a high  $\Delta$  affinity energy and presents hydrophobic interactions with NS2B-Val53. NS2B-Val53 is not conserved between the DV serotypes so a better target for a bindable site is Phe46 on the NS3 rather than the residue on NS2B. In the case of Thr53, it was found that the residue is at certain proximity to NS2B-His72 also classified on class A. In addition Thr 53 was calculated to have a high  $\Delta$  affinity but it was not conserved on all four serotypes. NS2B-His72 is moderately conserved between the DV serotypes. This suggest that this pair may constitute a possible bindability site. Residues Val52 and Thr53 are residues exposed to solvent (SASA 15.55 and 8.18  $\text{\AA}^2$  respectively), both in close proximity to the catalytic site. The calculations showed both residues to contribute to the overall stability of the complex. Thr53 presents a hydrogen bond with Tyr79 (See Figure 9)

Furthermore, as expected several important residues are clustered on the experimental active site. Additionally, it was found a cluster of residues classified as important (Class A) close to the beta sheet-loop-beta sheet of the NS2B cofactor.(See Fig. 4 Another region on NS3-PRO was constituted by residues 58, 59 and 65 on a beta sheet loop - beta sheet structure as shown in Figure 6. This one seems to be another important region where residues Leu58 and Thr59 may contribute to the molecular recognition of NS2B and residues Leu65 and Tyr79 to the stability of the secondary structure of that region. Phe189 (NS3), NS2B-Met 84 are also in close proximity to this region and and may contribute to the molecular recognition between the two peptides. This is supported from the observation thatt  $\Delta$  affinity for NS2B-Met 84 was found to be one of the highest. Finally, a

possible NS2B-NS3Pro important recognition site is constituted by NS2B-His 72, PHE116 and THR156 as shown in Figure 7.

#### 4 Discussion

The finding of novel sites on the DV protease that can be targeted by drug-like molecules constitutes an important step towards the development of a clinically effective inhibitor since traditional targets on the binding site and/or close to the binding site have not yet generated viable compounds. Since tools to correlate multiple characteristics of the amino acids constituting both NS2B cofactor and NS3Pro domain may help to unveil these possible sites, we have used a two phases- multilayer perceptron algorithm to classify residues unto three groups. Those residues likely to cause a major change in activity, those likely to cause a moderate change in activity, and residues that have not being able to be classified. A similar approach has being used before to help in complex protein problems such as to assist the design and engineering of new proteins by studying the effect of introducing certain mutations with comparable or better results than for previously published results with other methods.[40] In our case we can not effectively make comparison with experimental results for all the test set since no experimental values exist for most of these residues. Nevertheless, we have data for some key residues such as for His51 that was not included on the training/validation set. The multilayer perceptron phase 1, classified this residue on the A class what is in very good agreement with the experimental fact that His51 constitutes a key residue (part of the catalytic triad). Furthermore, in overall for all cluster of residues classified on class A, it is possible to find sound chemical reasons for such classification by analyzing their different attributes. No chemical contradiction has being found to their automatic assignment by the classifier model.

There was not a big overlap with previously identified binding sites since most of the residues on those sites were used for the training/validation set. Authors found that site formed by residues Leu58 and nearby residues may be a target to inhibit binding of NS2B to NS3Pro and possibly disrupting NS3Pro activity. Something similar can be suggested for residues NS2B-His72, Phe116 and Thr156. Our model has predicted that one of the bindability sites is precisely close to the active site (residues NS2B-Met84, NS2B-Thr83 and NS2B-Thr77) where NS2B forms a beta sheet-loop-beta sheet structure.

In addition, we confirmed the importance of a previously reported site using our multi step *in silico* approach. The region comprising the residues Val72, Tyr150, Gly151, Asn152, Gly153 is located between the catalytic triad and the NS2B beta hairpin [41]. All these residues are highly conserved. Residues Tyr150, Gly151 and Gly153 have been experimentally determined to be class A and had a high  $\Delta$  stability. Asn152 is a class B residue and has a high  $\Delta$  affinity. Also, unising an interaction diagram we determined that Asn152 interacts with NS2B-Gly82 through a hidrogen bond. Residue Val72 had no experimental data available but showed a high  $\Delta$  stability. Even though Val72 has a low  $\Delta$  affinity it is very close to NS2B-Asp80, NS2B-Asp81 and NS2B-Gly82 residues and showed weak charge interactions with them. Moreover, Val72 and Gly153 were identified by RaptorX as part or the active site.

In a future work the full NS3 including the helicase domain can be considered in case the interface between the protease domain and helicase domain may reveal novel bindability sites. In addition exploration of the additional conformation for the DV serotype 3, 3UI1 structure, may reveal new features overlooked in this study. In addition, other structures that have being resolved in the presence of inhibitors may be used for the methodology described in this article and some new consensus between the different structures may allow to have more data to run a classifier able to assign residues to more informative categories that we were not able to assign first because the categories were unbalanced and/or with too little data points. Indeed, due to the amount of features available we needed to collapse five categories unto three, and the multi step approach was able to assign residues to two categories leaving many residues without a class. Nevertheless, the two classes that our model was able to classify were assigned with high level of confidence and contain the classes of greater interest for development of drug-like molecules.

The decision of using a methodology employing only the default minimization steps on the Prime MM-GBSA is supported by the careful comparison reported by Beard et al[35] that found that minimization of only the side chain of interest produced the best correlation between the calculated and the experimental results of protein-protein binding affinity when using point mutations and Prime MM-GBSA.[35] This observation was explained on the base that mutating an amino acid by alanine that is always a smaller residue, will avoid clashes with neighbouring atoms so a moderate minimization step is enough. Therefore, though other methods may be used to characterize the effect of a point mutation on the binding affinity between two proteins, the Prime MM-GBSA method was shown to perform very well at a fraction of the computational cost. In addition, the correct protonation state of the amino acids will have a great impact on the binding energy. PROPKA on the protein preparation Wizard was used for the correct assignment of these protonation states. Previously, it has been shown to be very reliable for assignments on entire proteins.[35] We found from the predictions made by our methodology that the use of Prime MM-GBSA for a computational alanine scanning mutagenesis provides great insights to characterize a protein though by itself is not enough. This is the reason why other attributes such as conservation-based were also added to the model. The suggested methodology reported here should be tested on other systems likely providing useful insights for drug discovery.

## 5 List of abbreviations

DV: Dengue virus, NS2B:Non Structural protein 2B from DV, NS3: Non Structural Protein 3 from DV.

## 6 Competing interests

## 7 Funding

This research was partially funded by Chancellor Grant - USFQ 2014-2015 (granted to CS, JT and MM).

## 8 Authors contributions

## 9 Acknowledgements

Authors thank Universidad San Francisco de Quito for the use of the High Performance Computing-USFQ. Authors also thank CECIRA program from CEDIA-Ecuador.

## 10 Authors information

## References

1. B Falgout, M Pethel, Y M Zhang, and C J Lai. Both nonstructural proteins NS2B and NS3 are required for the proteolytic processing of dengue virus nonstructural proteins. *Journal of virology*, 65(5):2467–2475, 1991. ISSN 0022-538X.
2. C.J. Lai, M. Pethel, L.R. Jan, H. Kawano, A. Cahour, and B. Falgout. Processing of dengue type 4 and other flavivirus nonstructural proteins. *Archives of Virology, Supplementum*, No. 9:359–368, 1994.
3. Hussin a Rothan, Ammar Y Abdulrahman, Pottayil G Sasikumer, Shatrah Othman, Noorsaadah Abd Rahman, and Rohana Yusof. Protegrin-1 inhibits dengue NS2B-NS3 serine protease and viral replication in MK2 cells. *Journal of biomedicine & biotechnology*, 2012: 251482, 2012. ISSN 1110-7251. doi: 10.1155/2012/251482. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3470887&tool=pmcentrez&rendertype=abstract>.
4. Christoph Nitsche, Steven Holloway, Tanja Schirmeister, and Christian D Klein. Biochemistry and Medicinal Chemistry of the Dengue Virus Protease. *Chemical Reviews*, 114(22):11348–11381, 2014. ISSN 0009-2665. doi: 10.1021/cr500233q. URL <http://pubs.acs.org/doi/abs/10.1021/cr500233q>.
5. Linfeng Li, Chandrakala Basavannacharya, Kitti Wing Ki Chan, Luqing Shang, Subhash G Vasudevan, and Zheng Yin. Structure-guided Discovery of a Novel Non-peptide Inhibitor of Dengue Virus NS2B-NS3 Protease. *Chemical biology & drug design*, (1): 1–10, 2014. ISSN 1747-0285. doi: 10.1111/cbdd.12500. URL <http://www.ncbi.nlm.nih.gov/pubmed/25533891>.
6. Mirta Roses Periago and María G. Guzmán. Dengue y dengue hemorrágico en las Américas. *Revista Panamericana de Salud Pública*, 21(3):187–191, 2007. ISSN 1020-4989. doi: 10.1590/S1020-49892007000300001.
7. R Aruna. Review on Dengue viral Replication , assembly and entry into the host cells. 3(11):1025–1039, 2014.
8. Dahai Luo, Subhash G Vasudevan, and Julien Lescar. The flavivirus NS2B-NS3 protease-helicase as a target for antiviral drug development. *Antiviral research*, 118(APRIL):148–158, 2015. ISSN 1872-9096. doi: 10.1016/j.antiviral.2015.03.014. URL <http://www.ncbi.nlm.nih.gov/pubmed/25842996>.
9. Siew Pheng Lim, Qing Yin Wang, Christian G. Noble, Yen Liang Chen, Hongping Dong, Bin Zou, Fumiaki Yokokawa, Shahul Nilar, Paul Smith,

- David Beer, Julien Lescar, and Pei Yong Shi. Ten years of dengue drug discovery: Progress and prospects. *Antiviral Research*, 100(2):500–519, 2013. ISSN 01663542. doi: 10.1016/j.antiviral.2013.09.013. URL <http://dx.doi.org/10.1016/j.antiviral.2013.09.013>.
10. Choon Han Heh, Rozana Othman, Michael J C Buckle, Yusrizam Shari-fuddin, Rohana Yusof, and Noorsaadah Abd Rahman. Rational discovery of dengue type 2 non-competitive inhibitors. *Chemical biology & drug de-sign*, 82(1):1–11, 2013. ISSN 1747-0285. doi: 10.1111/cbdd.12122. URL <http://www.ncbi.nlm.nih.gov/pubmed/23421589>.
  11. Hemalatha Beesetti, Navin Khanna, and Sathyamangalam Swami-nathan. Drugs for dengue: a patent review (2010–2014). *Expert Opinion on Therapeutic Patents*, 24(11):1171–1184, 2014. ISSN 1354-3776. doi: 10.1517/13543776.2014.967212. URL <http://informahealthcare.com/doi/abs/10.1517/13543776.2014.967212>.
  12. Maria Rosario Capeding, Ngoc Huu Tran, Sri Rezeki S Hadinegoro, Hus-sain Imam HJ Muhammad Ismail, Tawee Chotpitayasunondh, Mary Noreen Chua, Chan Quang Luong, Kusnandi Rusmil, Dewa Nyoman Wirawan, Re-vathy Nallusamy, Punnee Pitisuttithum, Usa Thisyakorn, In-Kyu Yoon, Di-ane van der Vliet, Edith Langevin, Thelma Laot, Yanee Hutagalung, Ca-rina Frago, Mark Boaz, T Anh Wartel, Nadia G Tornieporth, Melanie Saville, and Alain Bouckennooghe. Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, ran-domised, observer-masked, placebo-controlled trial. *The Lancet*, 384(9951): 1358–1365, 2014. ISSN 01406736. doi: 10.1016/S0140-6736(14)61060-6. URL <http://linkinghub.elsevier.com/retrieve/pii/S0140673614610606>.
  13. Luis Villar, Gustavo Horacio Dayan, José Luis Arredondo-García, Doris Mari-bel Rivera, Rivaldo Cunha, Carmen Deseda, Humberto Reynales, Maria Selma Costa, Javier Osvaldo Morales-Ramírez, Gabriel Carrasquilla, Luis Carlos Rey, Reynaldo Dietze, Kleber Luz, Enrique Rivas, Maria Consuelo Mi-randa Montoya, Margarita Cortés Supelano, Betzana Zambrano, Edith Langevin, Mark Boaz, Nadia Tornieporth, Melanie Saville, and Fer-nando Noriega. Efficacy of a Tetravalent Dengue Vaccine in Chil-dren in Latin America. *New England Journal of Medicine*, 372(2): 141103114505002, 2014. ISSN 0028-4793. doi: 10.1056/NEJMoa1411037. URL <http://www.ncbi.nlm.nih.gov/pubmed/25365753>.
  14. Hongmei Wu, Stefanie Bock, Mariya Snitko, Thilo Berger, Thomas Wei-dner, Steven Holloway, Manuel Kanitz, Wibke E. Diederich, Holger Steuber, Christof Walter, Daniela Hofmann, Benedikt Weißbrich, Ralf Spannaus, Eliana G. Acosta, Ralf Bartenschlager, Bernd Engels, Tanja Schirmeister, and Jochen Bodem. Novel Dengue Virus NS2B/NS3 Pro-tease Inhibitors. *Antimicrobial Agents and Chemotherapy*, 59(2):1100–1109, 2015. ISSN 0066-4804. doi: 10.1128/AAC.03543-14. URL <http://aac.asm.org/lookup/doi/10.1128/AAC.03543-14>.
  15. U S F Tambunan, N Apriyanti, a a Parikesit, W Chua, and K Wuryani. Computational design of disulfide cyclic peptide as potential inhibitor of complex NS2B-NS3 dengue virus protease. *African Journal of Biotechnology*, 10(57):12281–12290, 2011. ISSN 1684-5315. doi: 10.5897/AJB11.1837. URL [http://www.academicjournals.org/AJB/abstracts/abs2011/28Sep/Tambunan et al.htm](http://www.academicjournals.org/AJB/abstracts/abs2011/28Sep/Tambunan%20et%20al.htm).



16. Allan Bastos Lima, Mira A M Behnam, Yasmin El Sherif, Christoph Nitsche, Sergio M Vecchi, and Christian D Klein. Dual inhibitors of the dengue and West Nile virus NS2B-NS3 proteases : Synthesis , biological evaluation and docking studies of novel peptide-hybrids. *Bioorganic & medicinal chemistry*, 23(17):5748 – 5755, 2015.
17. Christian G Noble, Cheah Chen Seh, Alexander T Chao, and Pei Yong Shi. Ligand-bound structures of the dengue virus protease reveal the active conformation. *Journal of virology*, 86(1): 438–46, 2012. ISSN 1098-5514. doi: 10.1128/JVI.06225-11. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3255909&tool=pmcentrez&rendertype=abstract>.
18. Lawrence A Kelley, Stefans Mezulis, Christopher M Yates, Mark N Wass, and Michael J E Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6): 845–858, 2015. ISSN 1754-2189. doi: 10.1038/nprot.2015.053. URL <http://dx.doi.org/10.1038/nprot.2015.053> \n10.1038/nprot.2015.053\nh<http://www.nature.com/doifinder/10.1038/nprot.2015.053>
19. D. Luo, T. Xu, C. Hunke, G. Gruber, S. G. Vasudevan, and J. Lescar. Crystal Structure of the NS3 Protease-Helicase from Dengue Virus. *Journal of Virology*, 82(1):173–183, 2008. ISSN 0022-538X. doi: 10.1128/JVI.01788-07. URL <http://jvi.asm.org/cgi/doi/10.1128/JVI.01788-07>.
20. Juan Du, Huijun Sun, Lili Xi, Jiazhong Li, Ying Yang, Huanxiang Liu, and Xiaojun Yao. Molecular modeling study of checkpoint kinase 1 inhibitors by multiple docking strategies and prime/MM-GBSA calculation. *Journal of Computational Chemistry*, 32(13):2800–2809, 2011. ISSN 01928651. doi: 10.1002/jcc.21859.
21. Mani Srivastava, Harvinder Singh, and Pradeep Kumar. Naik. Molecular modeling evaluation of the antimalarial activity of artemisinin analogues: molecular docking and rescoring using prime/MM-GBSA approach. *Current Research Journal of Biological Sciences*, 2(2):83–102, 2010. ISSN 2041-0778. URL <http://maxwellsci.com/print/crjbs/v2-83-102.pdf>.
22. Irina S. Moreira, Pedro A. Fernandes, and Maria J. Ramos. Unravelling Hot Spots: a comprehensive computational mutagenesis study. *Theoretical Chemistry Accounts*, 117(1):99–113, 2006. ISSN 1432-881X. doi: 10.1007/s00214-006-0151-z. URL [http://apps.webofknowledge.com/full\\_record.do?product=UA&search\\_mode=GeneralSearch&qid=24&SID=U2KI47df](http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=24&SID=U2KI47df)
23. D S Gestó, N M F S a Cerqueira, M J Ramos, and P a Fernandes. Discovery of new druggable sites in the anti-cholesterol target HMG-CoA reductase by computational alanine scanning mutagenesis. *Journal of molecular modeling*, 20(4):2178, 2014. ISSN 0948-5023. doi: 10.1007/s00894-014-2178-8. URL <http://www.ncbi.nlm.nih.gov/pubmed/24671303>.
24. Schrödinger Press. *Protein Preparation Guide*. Schrödinger, New York, 2009.
25. User Manual. BioLuminate 1.0.
26. Centro de Transferencia y Desarrollo de Tecnologías Universidad San Francisco de Quito. 2014.
27. William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.
28. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software. *ACM SIGKDD Explorations*, 11(1):10–18, 2009. ISSN 19310145. doi: 10.1145/1656274.1656278. URL

- <http://portal.acm.org/citation.cfm?doid=1656274.1656278\npapers2://publication/doi/10.1145/1656274.1656278>
29. Robert P Sheridan, Vladimir N Maiorov, M Katharine Holloway, Wendy D Cornell, and Ying-Duo Gao. Drug-like density: a method of quantifying the bindability of a protein target based on a very large set of pockets and drug-like ligands from the protein data bank. *Journal of chemical information and modeling*, 50(11):2029–2040, 2010.
  30. C.J. Lai, M. Pethel, L.R. Jan, H. Kawano, A. Cahour, and B. Falgout. Processing of dengue type 4 and other flavivirus nonstructural proteins. *Archives of Virology, Supplementum*, No. 9:359–368, 1994.
  31. Barry Falgout, Roger H Miller, and Ching-juh Lai. nonstructural protein NS2B : identification of a domain required for NS2B-NS3 protease Deletion Analysis of Dengue Virus Type 4 Nonstructural Protein NS2B : Identification of a Domain Required for NS2B-NS3 Protease Activity. 67(4):2034–2042, 1993.
  32. Wan Na Chen, Karin V. Loscha, Christoph Nitsche, Bim Graham, and Gottfried Otting. The dengue virus NS2B-NS3 protease retains the closed conformation in the complex with BPTI. *FEBS Letters*, 588(14):2206–2211, 2014. ISSN 18733468. doi: 10.1016/j.febslet.2014.05.018. URL <http://dx.doi.org/10.1016/j.febslet.2014.05.018>.
  33. Britta Nisius, Fan Sha, and Holger Gohlke. Structure-based computational analysis of protein binding sites for function and druggability prediction. *Journal of biotechnology*, 159(3):123–34, 2012. ISSN 1873-4863. doi: 10.1016/j.jbiotec.2011.12.005. URL <http://www.sciencedirect.com/science/article/pii/S0168165611006614>.
  34. Jian Peng and Jinbo Xu. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins*, 79 Suppl 1 (Suppl 10):161–71, 2011. ISSN 1097-0134. doi: 10.1002/prot.23175. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3226909&tool=pmcentrez&rendertype=abstract>.
  35. Hege Beard, Anuradha Cholleti, David Pearlman, Woody Sherman, and Kathryn a. Loving. Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. *PLoS ONE*, 8(12):1–11, 2013. ISSN 19326203. doi: 10.1371/journal.pone.0082849.
  36. J F Bazan and R J Fletterick. Detection of a trypsin-like serine protease domain in flaviviruses and pestiviruses. *Virology*, 171(2):637–639, 1989. ISSN 00426822. doi: 10.1016/0042-6822(89)90639-9. URL <http://www.ncbi.nlm.nih.gov/pubmed/2548336>.
  37. R P Valle and B Falgout. Mutagenesis of the NS3 protease of dengue virus type 2. *Journal of virology*, 72(1):624–632, 1998. ISSN 0022-538X.
  38. Rozana Othman, Tan Siew Kiat, Norzulaani Khalid, Rohana Yusof, E. Irene Newhouse, James S. Newhouse, Masqudul Alam, and Noorsaadah Abdul Rahman. Docking of noncompetitive inhibitors into dengue virus type 2 protease: Understanding the interactions with allosteric binding sites. *Journal of Chemical Information and Modeling*, 48(8):1582–1591, 2008. ISSN 15499596. doi: 10.1021/ci700388k.
  39. Rajendra Raut, Hemalatha Beesetti, Poornima Tyagi, Ira Khanna, Swatantra K Jain, Variam U Jeankumar, Perumal Yogeewari, Dharmarajan Sriram, and Sathyamangalam Swaminathan. A small molecule inhibitor of dengue virus type 2 protease inhibits the replication of all four dengue virus serotypes in cell culture. *Virology journal*, 12(1):

- 16, 2015. ISSN 1743-422X. doi: 10.1186/s12985-015-0248-x. URL <http://www.virologyj.com/content/12/1/16>.
40. Majid Masso and Iosif I. Vaisman. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics*, 24(18):2002–2009, 2008. ISSN 13674803. doi: 10.1093/bioinformatics/btn353.
41. Pornwaratt Niyomrattanakit, Pakorn Winoyanuwattikun, Santad Chanpraph, and Chanan Angsuthanasombat. Identification of Residues in the Dengue Virus Type 2 NS2B Cofactor That Are Critical for NS3 Protease Activation Identification of Residues in the Dengue Virus Type 2 NS2B Cofactor That Are Critical for NS3 Protease Activation. *Journal of virology*, 78(24):13708–13716, 2004. doi: 10.1128/JVI.78.24.13708.
42. Muslum Yildiz, Sumana Ghosh, Jeffrey a. Bell, Woody Sherman, and Jeanne a. Hardy. Allosteric inhibition of the NS2B-NS3 protease from dengue virus. *ACS Chemical Biology*, 8(12):2744–2752, 2013. ISSN 15548929. doi: 10.1021/cb400612h.
43. Wanisa Salaemae, Muhammad Junaid, Chanan Angsuthanasombat, and Gerd Katzenmeier. Structure-guided mutagenesis of active site residues in the dengue virus two-component protease NS2B-NS3. *Journal of biomedical science*, 17:68, 2010. ISSN 1423-0127. doi: 10.1186/1423-0127-17-68.
44. Zhili Zuo, Oi Wah Liew, Gang Chen, Pek Ching Jenny Chong, Siew Hui Lee, Kaixian Chen, Hualiang Jiang, Chum Mok Puah, and Weiliang Zhu. Mechanism of NS2B-mediated activation of NS3pro in dengue virus: molecular dynamics simulations and bioassays. *Journal of virology*, 83(2):1060–70, 2009. ISSN 1098-5514. doi: 10.1128/JVI.01325-08. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2612360&tool=pmcentrez&rendertype=abstract>.

50	ASP	NS2B	-	-	C	F	0	2	c
61	TRP	NS2B	30,83	-5,99	A	F	0	7	d
74	LEU	NS2B	7,49	11,81	A	F	0	4	d
76	ILE	NS2B	11,82	12	C	F	0	1	d
78	VAL	NS2B	2,24	5,41	A	F	0	3	d
19	GLU	NS3	1,93	-6,22	D	F	1	8	a
27	GLN	NS3	0,44	3,13	C	F	1	7	c
31	PHE	NS3	0	0,12	D	F	1	8	a
35	GLN	NS3	0,27	23,46	C	T	1	7	c
54	ARG	NS3	1,28	6,7	C	F	1	7	c
83	TRP	NS3	-0,52	40,33	A	T	0	1	a
86	SER	NS3	-0,05	-0,95	D	F	1	8	a
105	ASN	NS3	0,02	-0,2	D	F	1	7	a
111	THR	NS3	6,52	8,14	A	F	0	1	a
115	THR	NS3	-1,68	3	A	F	0	8	a,b
125	ALA	NS3	-	-	C	T	0	7	a
126	ILE	NS3	-0,16	14,65	C	T	0	7	e
129	ASP	NS3	-1,64	-2,97	B	T	0	7	b,e
130	PHE	NS3	-0,55	20,13	B	F	0	2	e
131	LYS	NS3	1,3	-4,52	WT	F	0	5	e
133	GLY	NS3	0,02	15,39	B	T	0	6	b,e
134	THR	NS3	-0,29	-1,92	A	F	1	6	b,e
135	SER	NS3	0,15	1,17	A	T	0	8	e
136	GLY	NS3	0,1	59,52	B	T	0	7	e
139	ILE	NS3	-0,21	29,09	C	T	0	7	e
140	ILE	NS3	8,51	17,35	C	F	1	7	e
141	ASN	NS3	-0,44	8,66	WT	F	0	7	e
142	ARG	NS3	9,38	0,27	WT	F	0	7	e
143	GLU	NS3	-1,57	0,28	WT	F	0	5	e
144	GLY	NS3	-4,64	7,72	C	T	0	6	e
148	GLY	NS3	-0,06	55,79	A	T	1	8	e
149	LEU	NS3	-0,14	21,96	A	T	0	5	e
150	TYR	NS3	-0,32	25,45	A	T	0	5	b,e
151	GLY	NS3	-0,13	52,7	A	T	0	7	b,e
152	ASN	NS3	4,5	-2,28	B	T	0	4	b,e
153	GLY	NS3	-0,04	42,23	A	T	1	4	e
154	VAL	NS3	10,58	5,15	C	F	0	7	e
155	VAL	NS3	0,02	4,66	C	F	0	8	e
163	SER	NS3	0,36	6,85	C	T	0	7	b
165	ILE	NS3	-0,27	21,54	C	F	0	8	b

a) Mishum Yildiz, 2013

b) Salaemae, Junaid, Angsuthanasombat, &amp; Katzenmeier, 2010

c) Zhili Zuo, 2009

d) Niyomrattanakit et al., 2004

e) Valle &amp; Falgout, 1998

Table 1: Summary of the Training/Cross validation Set. References correspond to [37, 41–44]

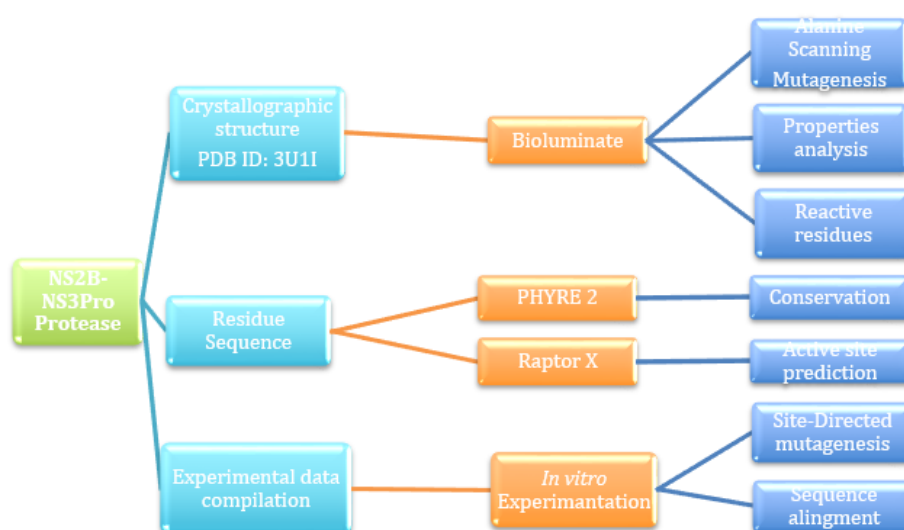


Fig. 1: Flow chart of the methods used for generating the features and class for the machine learning classifier. The categories for the experimental inhibitor result were A, total loss of activity; B, major loss of activity; C, moderate-activity lost; D slight activity lost and WT, same as wild type.

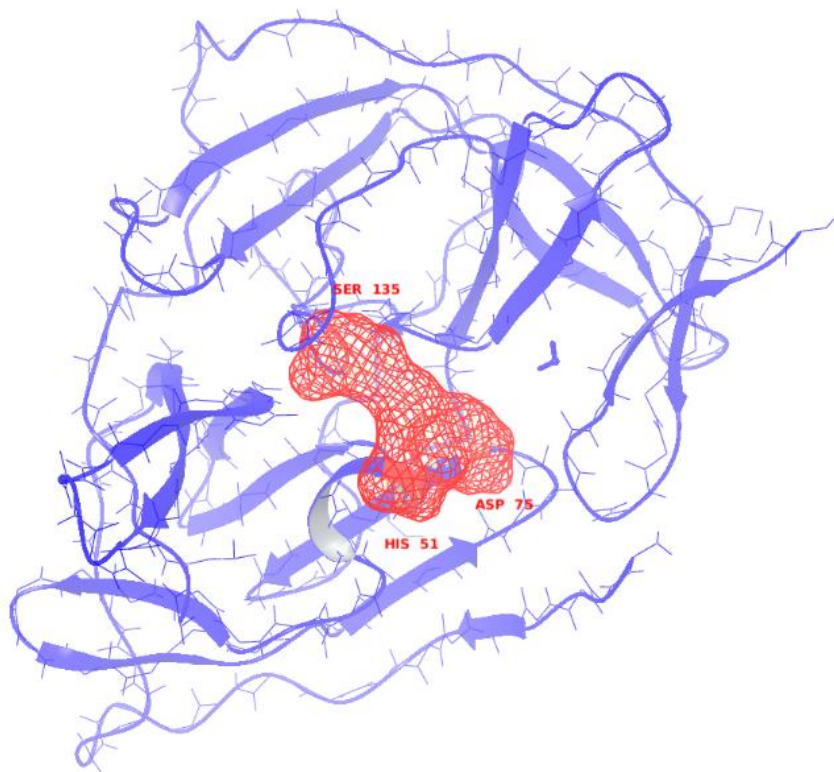


Fig. 2: Structure, including the NS2B cofactor and the protease region of NS3, used for the computational alanine scanning mutagenesis (3U11). Surrounded by a red mesh are the most important residues for the catalytic activity (Ser135, Asp75 and His 51); referred here as the catalytic triad

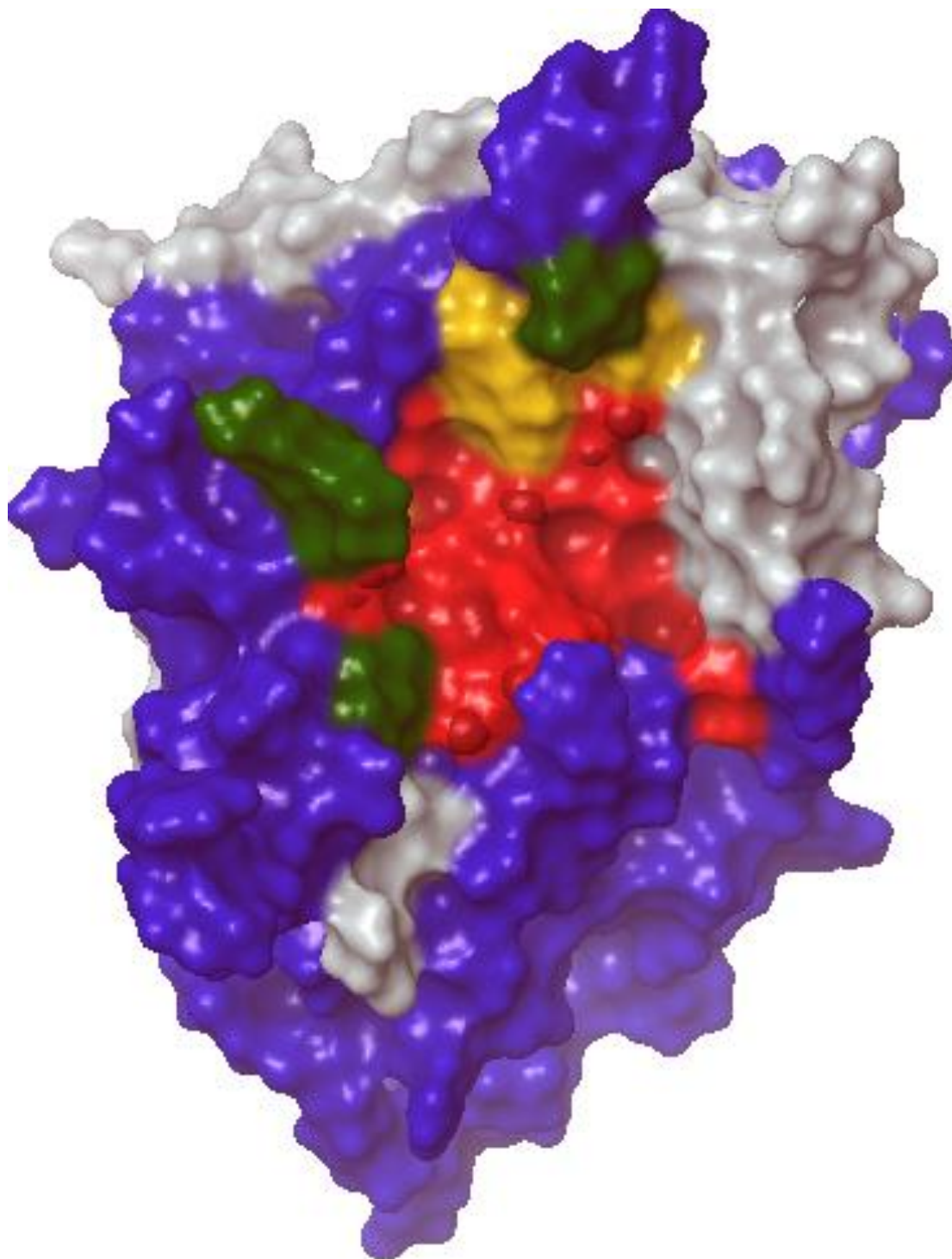


Fig. 3: Conservation degree according to Phyre2 of the residues identified as part of the Active Site by Raptor X. The NS2B chain is gray and the NS3 chain is blue. Residues identified by raptor as part of the active site are colored according to their conservation level. Highly-conserved residues are red, moderately-conserved residues are yellow and not conserved residues are green. It is also evident here that the catalytic site is very shallow. This view is in approximately the same orientation as on Figure 2. View is in surface representation.

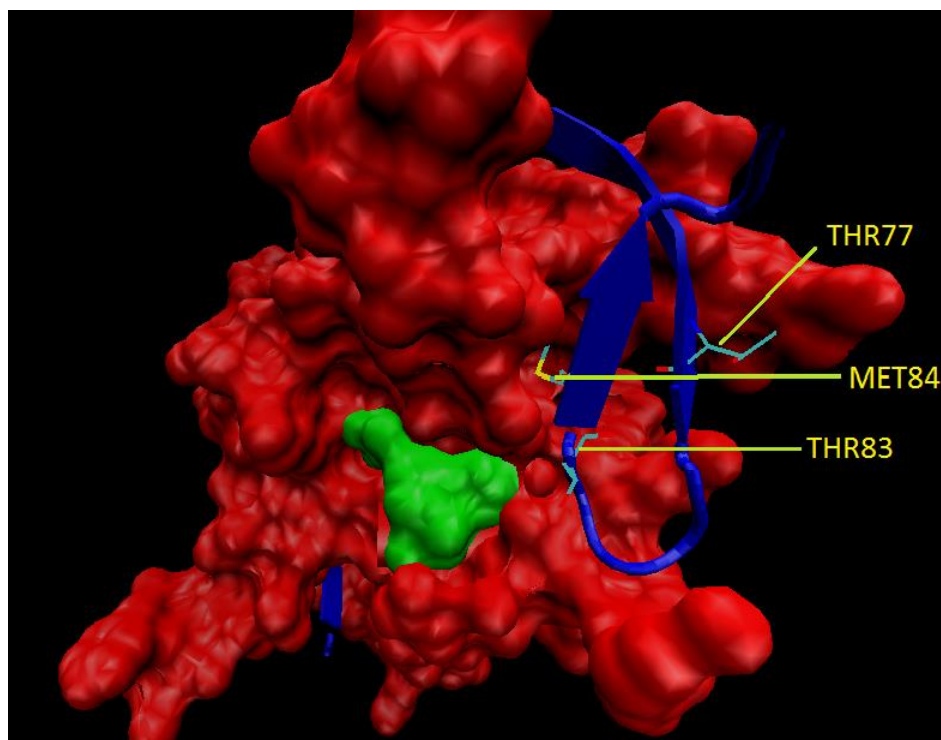


Fig. 4: NS2B Residues labeled in yellow were classified as class A, likely to be fundamental on the enzyme activity. The catalytic triad residues are shown in green only as reference points for the reader.



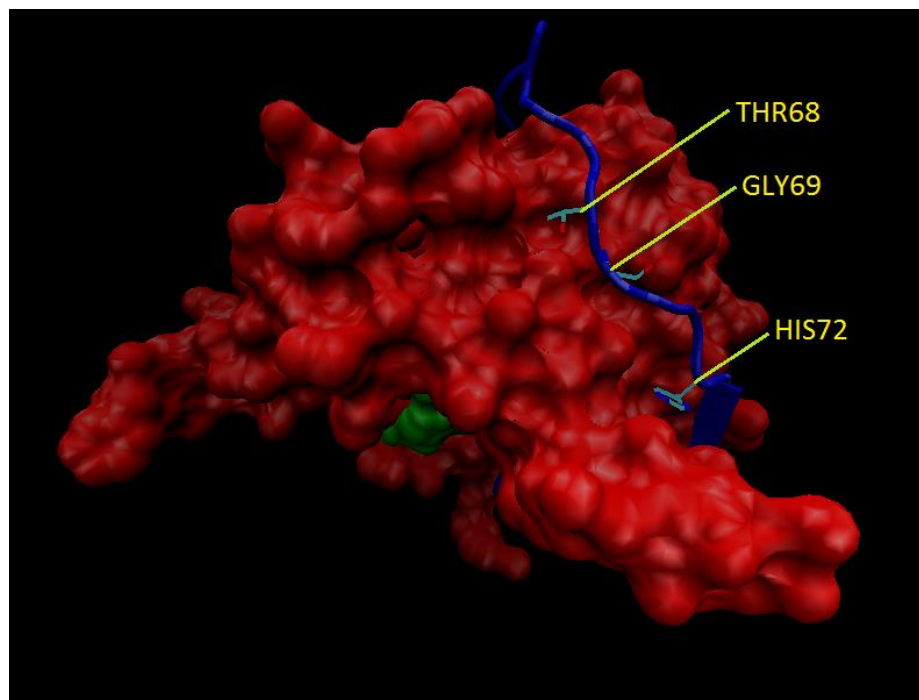


Fig. 5: NS2B Residues labeled in yellow were classified as class A, likely to be fundamental on the enzyme activity. The catalytic triad residues are shown in green only as reference points for the reader.

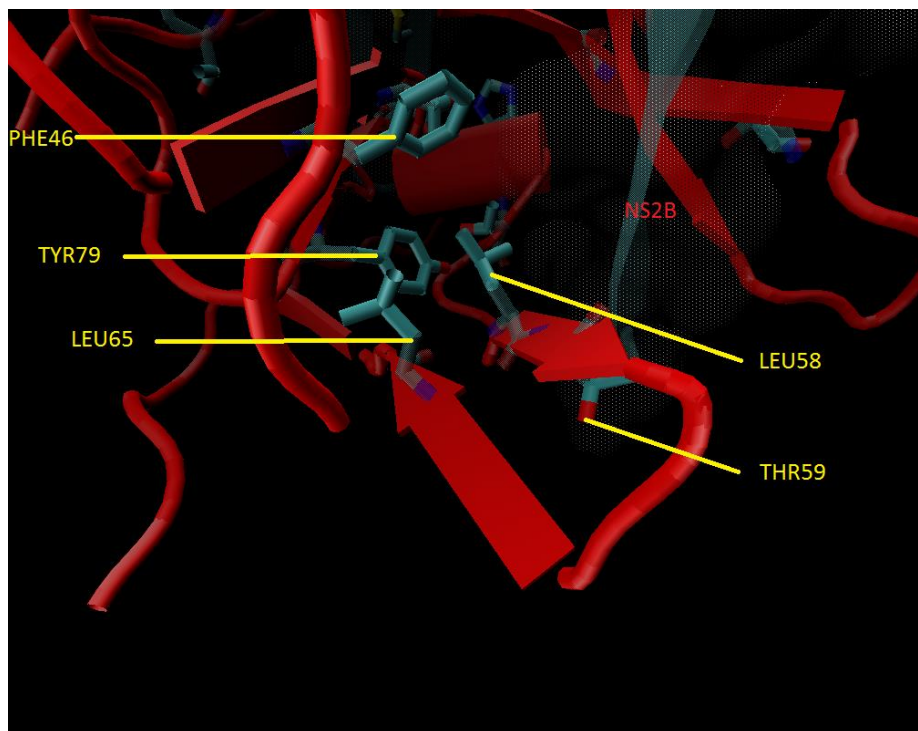


Fig. 6: NS3 Residues labeled in yellow were classified as class A, likely to be fundamental on the enzyme activity. NS2B is shown in cartoon visualization surrounded by a surface calculation both on a transparent material. Bindability pocket shown here seems to be important for the NS2B's binding to NS3.

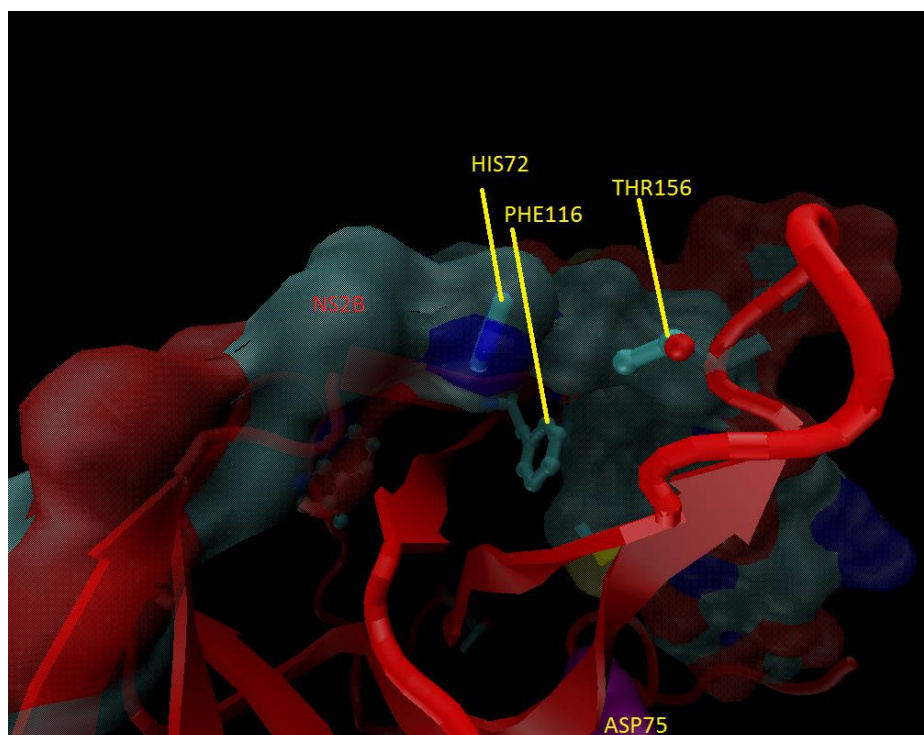


Fig. 7: NS3 Residues labeled in yellow were classified as class A, likely to be fundamental on the enzyme activity. NS2B-HIS72, NS3-PHE116 and THR156 cluster on a possible bindability pocket. NS2B is shown in surface visualization on a transparent material. Bindability pocket shown here seems to be important for the NS2B's binding to NS3. ASP75 in purple shown only as a reference point for the reader.

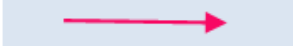

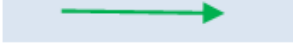

Interactions	Representation
H-bonds backbone	
H-bonds side chains	
pi-pi stacking	
pi-cation interaction	

Fig. 8: Interactions representations' symbols as found for Phe46 and Tyr79 (NS3Pro residues) and Supplementary Information.

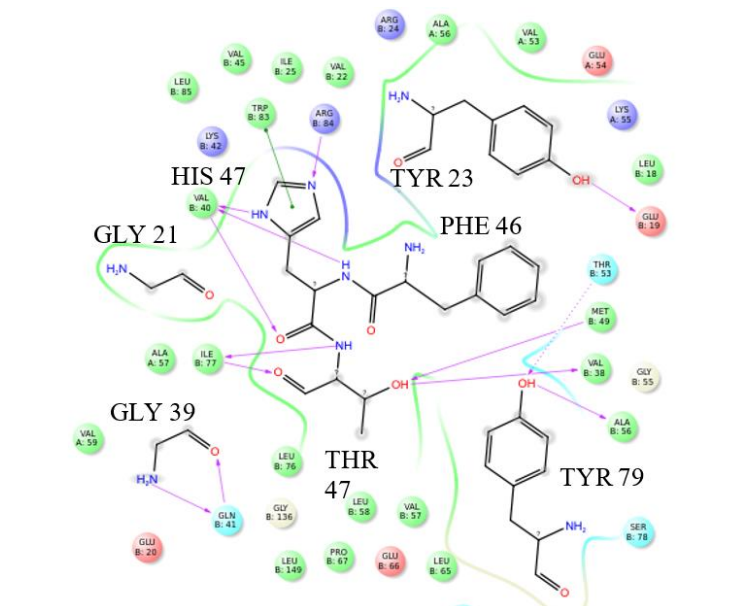


Fig. 9: Interactions diagram of Phe46 and Tyr79 (NS3Pro residues) classified as class A, likely to be fundamental on the enzyme activity.

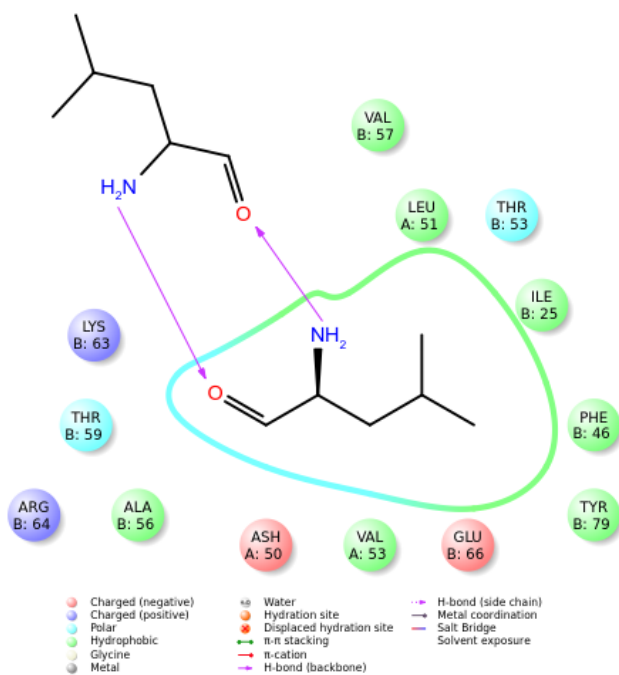


Fig. 10: Interactions diagram of Leu58 (NS3Pro residue) classified as class A, likely to be fundamental on the enzyme activity. Here Leu 58 (center) is showing a backbone hydrogen bonding interaction with Leu 65 (also classified as A)

