Bioinformatics Analysis of Mutations Sheds Light on the Evolution of Dengue NS1 Protein With Implications in the Identification of Potential Functional and Druggable Sites

Abhishek Sharma,[1] Sudhir Krishna,[1,2] and Ramanathan Sowdhamini ⬥*,[1,3,4]

[1]National Centre for Biological Science, TIFR, Bangalore, India

[2]Department of School of Interdisciplinary Life Sciences, Indian Institute of Technology Goa, Farmagudi, Pond-403401, Goa, India

[3]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

[4]Computational Biology, Institute of Bioinformatics and Applied Biotechnology, Bangalore, India

*Corresponding author: E-mail: mini@ncbs.res.in.

Associate editor: Dr. Deepa Agashe

Abstract

Non-structural protein (NS1) is a 350 amino acid long conserved protein in the dengue virus. Conservation of NS1 is expected due to its importance in dengue pathogenesis. The protein is known to exist in dimeric and hexameric states. The dimeric state is involved in its interaction with host proteins and viral replication, and the hexameric state is involved in viral invasion. In this work, we performed extensive structure and sequence analysis of NS1 protein, and uncovered the role of NS1 quaternary states in its evolution. A three-dimensional modeling of unresolved loop regions in NS1 structure is performed. "Conserved" and "Variable" regions within NS1 protein were identified from sequences obtained from patient samples and the role of compensatory mutations in selecting destabilizing mutations were identified. Molecular dynamics (MD) simulations were performed to extensively study the effect of a few mutations on NS1 structure stability and compensatory mutations. Virtual saturation mutagenesis, predicting the effect of every individual amino acid substitution on NS1 stability sequentially, revealed virtual-conserved and variable sites. The increase in number of observed and virtual-conserved regions across NS1 quaternary states suggest the role of higher order structure formation in its evolutionary conservation. Our sequence and structure analysis could enable in identifying possible protein–protein interfaces and druggable sites. Virtual screening of nearly 10,000 small molecules, including FDA-approved drugs, permitted us to recognize six drug-like molecules targeting the dimeric sites. These molecules could be promising due to their stable interactions with NS1 throughout the simulation.

Key words: NS1 protein, molecular evolution, virtual saturation mutagenesis, single versus combination of mutations, protein stability, compensatory mutations, protein–protein interaction motifs, antiviral drug screening.

Introduction

Dengue is one of the most infectious and widespread vector-borne diseases, known to infect millions of people every year across the globe (Malavige et al. 2004). It is caused by single-stranded positive sense (+ssRNA) and non-segmented RNA virus belonging to the Flaviviridae family. Clinical symptoms range from mild to severe, which include dengue hemorrhagic fever and dengue shock syndrome (Malavige et al. 2004). Mild cases are treatable by available medication for the fever and fluid intake, but severe cases are still challenging and difficult to treat (Malavige et al. 2004). The main reason is the unavailability of effective vaccines, therapy, or antiviral drugs. Vaccines currently under clinical trials have under-performed on efficacy and effectiveness, which can be linked to the complex immune responses generated by heterotypic dengue

viruses, such as antibody-dependent enhancement (Boonnak et al. 2008; Beltramello et al. 2010; Screaton et al. 2015; Neufeldt et al. 2018) and cytokine storm. The size of the dengue virus genome is approx. ~11 kb and is composed of three structural (Capsid, PrM, and Envelope) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5) (Neufeldt et al. 2018).

Mutation rates in RNA viruses are generally high due to their error-prone RNA-dependent RNA polymerase-mediated replication (Drake 1993; Drake and Holland 1999; Duffy 2018), leading to high genetic diversity among RNA viruses (Sanjuán and Domingo-Calap 2021). Dengue serotypes share approx. 70% sequence similarity with each other (Weaver and Vasilakis 2009; Neufeldt et al. 2018). NS1 (Non-structural protein 1) is a 350 amino acids long (45–55 kD) glycoprotein, which exists in multiple

Open Access

structural states such as monomer, dimer, and hexamer (Gutsche et al. 2011; Akey et al. 2014; Watterson et al. 2016). Many non-structural proteins (NS3 and NS5), along with NS1, are highly conserved in nature (Neufeldt et al. 2018). The present study mainly focuses on the possible reasons for NS1 protein conservation.

NS1 is a critical protein that plays an essential role in viral replication and disease pathogenesis (Akey et al. 2014; Fan et al. 2014; Rastogi et al. 2016). Different states of NS1 are reported to perform a diverse range of functions. NS1 dimer is the predominant state within the host cell, which is critical for viral replication, and modulating viral production through their interaction with other dengue proteins [structural proteins PrM and PrE (Scaturro et al. 2015)]. It is reported to interact with several proteins (Khadka et al. 2011; Silva et al. 2013; Karyala et al. 2016; Hafirassou et al. 2017) belonging to various pathways, such as innate immunity, ERAD pathway, OST complex, and DNA repair. On the other hand, hexamer (sNS1) is secreted outside the cell as a lipid cargo (Screaton et al. 2015), where it is recognized by host immune system. Antibodies generated against sNS1 act as diagnostic markers. However, some of the NS1 antibodies have been identified as auto-antibodies. These auto-antibodies interact with the host proteins in endothelial cells and cause plasma leakage (Lin et al. 2003; Avirutnan et al. 2006; Cheng et al. 2009; Malavige and Ogg 2017; Jayathilaka et al. 2018). sNS1 interferes with complement systems proteins and activates CD4 monocyte, leading to the release of inflammatory cytokines (Modhiran et al. 2015; Neufeldt et al. 2018). Therefore, understanding the basis of NS1 sequence conservation would help in identifying immunogenic regions and druggable sites for therapeutic targeting.

Our initial hypothesis was based on multiple quaternary states of NS1, which might increase the number of residues that are not free to mutate. Such residues with functionally important regions could play an essential role in NS1 conservation. These residues can be created by forming new interfaces or structural rearrangement during the multimerization process. For more understanding, NS1 sequences from complete viral protein sequences available in two public databases: NCBI (Brister et al. 2015) and VIPER databases (Pickett et al. 2012) were extracted and analyzed. Shannon Entropy (SE) scores (Litwin and Jores 1992) for each residue position were calculated, and residue positions were divided into groups ("observed low SE positions" and "observed high SE positions") based on their SE scores. Mutations present in each position were identified from aligned sequences, and their effect on NS1 protein stability was predicted.

Further, virtual saturation mutagenesis was carried out where effect of every possible amino acid change on protein stability was predicted. Mutations were performed individually and sequentially (one at a time) for each residue position of NS1 protein. In order to evaluate the ability of amino acid position to mutate into the rest of the other 19 amino acids, a mutational capacity per residue position (MC/RP) score was calculated. The MC/RP score was then used to define residue positions which are recognized as virtual-conserved or virtual-variable sites. The distribution of virtual-conserved positions across NS1 states and their relationship with observed-conserved/variable regions was analyzed from patient samples. Our analysis also observed the effect of compensatory mutations that occur around singly destabilizing mutations in NS1. We finally showed the implication of our analysis in predicting possible functional (protein–protein interaction) sites used by NS1 protein and computationally validated them by blind docking. A solvent-exposed druggable site containing known critical residues was identified, and small molecule inhibitors were screened against them. MD simulations showed that some of these small molecules form stable interactions with NS1. The effect of mutation on other NS1 proteins (belonging to the family *flaviviridae* and the influenza virus) showed similar results compared to the Dengue NS1 protein. This provides a new direction to apply the afore-mentioned approach in predicting the druggable site on other viral proteins.

Results and Discussions

Generation of NS1 Monomer, Dimer, and Hexamer Models

The structure of the DENV2 NS1 dimer is available in the protein data bank, ID 4O6B. However, the structure has some unresolved loop regions, which were modeled using multi-template homology modeling (fig. 1a) and Zika virus NS1 structure (PDB ID 5GS6) as a template (see Methods for details). Predicted models were sorted according to the DOPE score, and then validated by two web servers: ProSA and SAVES 5.0 (data not shown). MD simulations were performed in triplicates for the best-predicted model in the OPLS force field for 100 nanoseconds (ns) duration. Root mean square deviation (RMSD) (fig. 1b) was calculated from the MD simulation trajectory. The model was found to be stable throughout the simulation run. Dengue sequences approved by WHO and CDC were used as references (Añez et al. 2016) to model NS1 of other dengue serotypes. The above-modeled DENV2 NS1 model was used as a template, and the protocol (Methods "Homology modeling of DENV 1, 3, and 4 NS1 proteins") was followed for the generation, validation, and selection of the best NS1 dimer model for each dengue serotype (supplementary fig. S1, Supplementary Material online). NS1 is majorly present in the dimer state within the cell, but for detailed analysis, a monomer unit from the dimer was also extracted. MD simulation run (100 ns) was performed to relax the dimer interface in the monomeric unit (supplementary fig. S1, Supplementary Material online). For the hexameric units, lipid-free models were constructed by applying the crystallographic 3-fold symmetry on the deposited DENV2 NS1 dimer coordinates (PDB ID 4O6B) (fig. 1c). A similar approach (Method "Homology modeling of DENV 1, 3, and 4 NS1 proteins") was used to construct the hexameric model for the rest of the
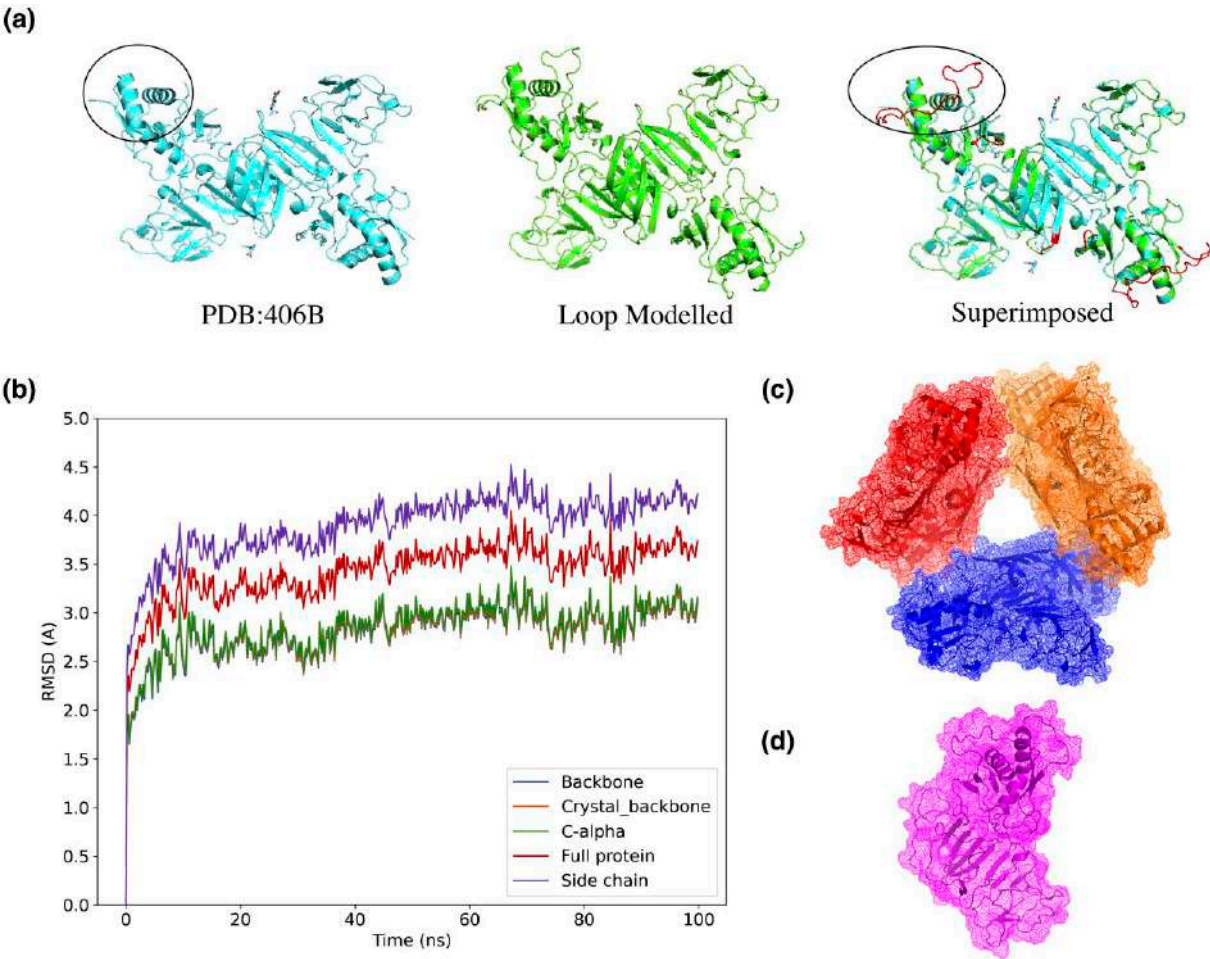
Fig. 1. Loop modeling of DENV2 NS1 followed by validation through MD simulation. (*a*) The unresolved loop regions in solved structure of the DENV2 (PDB:4o6B) (black circle) were modeled using Zika NS1 as template. Panel shows NS1 structure, loop modeled NS1, and superimposed structure between 4o6b and modeled NS1. Missing region (black circle) and modeled loops are shown in red. (*b*) The stability of modeled NS1 is validated by MD simulation run for 100 ns (*n* = 3). RMSD of modeled structure (Full protein) and structural components (Backbone, crystal backbone, C-alpha atoms, and side chains) for complete simulation trajectory was plotted, where *x*-axis represents time (ns) and *y*-axis represent RMSD (Å). Stabilization of RMSD trends suggests the modeled structure's stability. (*c*) The DENV2 hexamer model was constructed by applying the crystallographic 3-fold to the deposit DENV2 NS1 dimer coordinates. Each color represents one dimer unit. (*d*) Monomer unit was extracted from modeled NS1 structure, and simulated for 100 ns (*n* = 3) to relax NS1 dimer interface.

serotypes (supplementary fig. S1, Supplementary Material online).

Sequence Analysis

To perform sequence analysis, DENV1 (*n* = 2,555), DENV2 (*n* = 1,931), DENV3 (*n* = 1,181), and DENV4 (*n* = 420) sequences were extracted from NCBI and VIPER database, and used for analysis (table 1). Each sequence was compared with reference sequence and mutations per residue positions were identified. Number of mutations were found in DENV1 (*n* = 364), DENV2 (*n* = 741), DENV3 (*n* = 956), and DENV4 (*n* = 794) sequence dataset. Given the large dataset taken for DENV1, there was relatively fewer mutations due to the geographical and temporal confining of sequences. On the other hand, sequences of other dengue serotypes belonged to diverse geographical locations and periods (data not shown). Dengue NS1 is

Table 1. Sequences Used in the Analysis.

Serotypes	Sequences in the NCBI database	Sequences in the VIPER database	COMBINE (Unique accession Number)
DENV1	1,729	2,165	2,555
DENV2	1,401	1,589	1,931
DENV3	871	997	1,181
DENV4	344	254	420

Two publicly available databases (NCBI and VIPER) were used to extract and process sequences.

one of the most conserved non-structural proteins of the dengue virus. To understand the extent of diversity within NS1 protein, SE score per residue position (fig. 2*a* and *c*) was calculated. Similar SE scores among Dengue serotypes suggests similarity in their sequence conservation. Low SE scores reflect high conservation at that position. The DENV2 NS1 SE scores were represented on the DENV2 dimer model (fig. 2*b*).
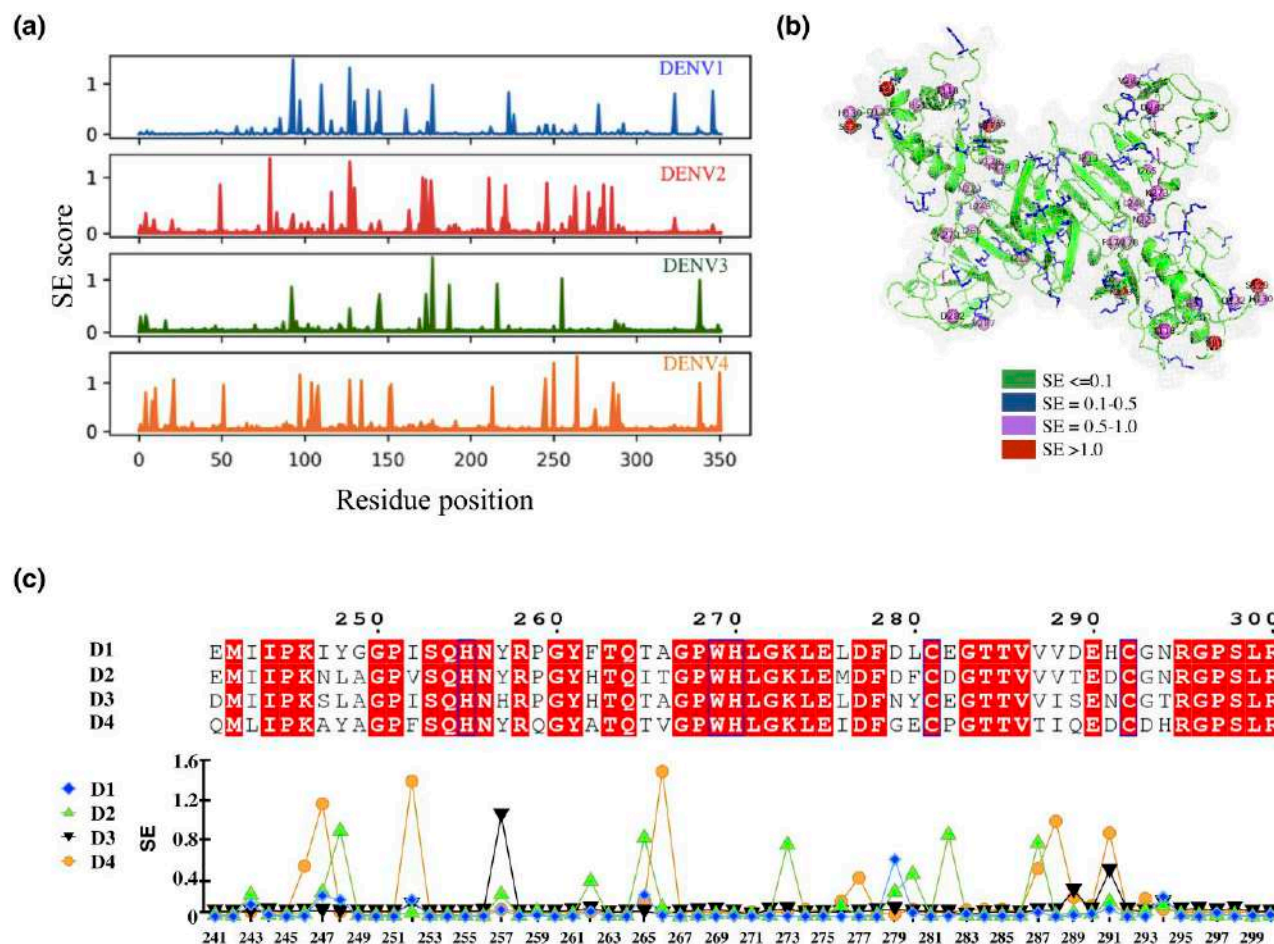
Fig. 2. Shannon entropy (SE) scores of each residue position of NS1 protein for all Dengue serotypes. (*a*) Shannon entropy (SE) scores (*y*-axis) per residue positions (*x*-axis) were calculated for all Dengue serotype NS1 proteins. SE score represents the frequency of different amino acid types at that position. It ranges from 0 (only one residue is present at that position) to 4.322 (all 20 residues are equally represented in that position). Majority of NS1 residues were observed to have low SE scores. Low SE scores suggest conserved residues. (*b*) The SE scores of DENV2 NS1 are represented in detail on modeled structure. Regions of structure are marked as per SE score <=0.1 (fully conserved), between 0.1 and 0.5 (Blue), between 0.5 and 1.0 (Magenta), and >0.1 (Red). SE score 0.1 was identified as cut-off, and positions were categorized into observed-conserved (SE < 0.1) and variable (SE >0.1) regions. (*c*) A segment of NS1 protein sequence (240aa-300aa) and respective SE score per position is shown. This represents the basis of our sequence analysis, where we attempt to understand possible reasons of varied SE score across whole protein. Here, red represents identical residues among serotypes. The Shannon entropy scores per residue position are shown as a line graph. D1 represents Dengue serotype 1, D2 represents Dengue serotype 2, D3 represents Dengue serotype 3, and D4 represents Dengue serotype 4.

A frequency distribution analysis of SE scores was performed. SE score of 0.1 was identified as the cut-off (supplementary fig. S2, Supplementary Material online) from distribution analysis to understand the diversity among residue positions. This led to the classification of residues into two categories: (a) Observed low SE (SE score ≤ 0.1), suggests conserved positions, (fig. 2*b*) and (b) Observed high SE (SE score > 0.1), represents comparatively variable positions. Residues with low SE were more in number as compared to diverse positions (fig. 2*a*). The possible reason behind residue conservation could be its role played in protein stability, functionality, or both. In this study, we focused on the role of structurally critical residues in NS1 sequence conservation. Therefore, we asked what will be the effect of observed mutations (obtained from NS1 sequences) on NS1 structure stability.

Majority of Single Mutations are Destabilizing in Nature

To predict the effect of identified mutations on protein stability, FoldX was used and free energy change between ($\Delta\Delta G$) wild-type and mutated protein structure ($\Delta\Delta G = \Delta G$ mutant—ΔG wild-type) was calculated. The $\Delta\Delta G$ score determines the effect of mutations on protein stability, and based on the score, each mutation was classified into three categories (neutral, stabilizing, and destabilizing). In the case of DENV2 dimer, 58.8% of mutations were predicted to be destabilizing in nature, further divided into subgroups (39%-highly destabilizing, 11.2% destabilizing, and 8.6%-slightly destabilizing) followed by neutral (20.9%). In case of stabilizing categories, 20.3% of mutations were further classified into subgroups, that is, 8.1%-slightly stabilizing, 7.7%-stabilizing and 4.5%-Highly
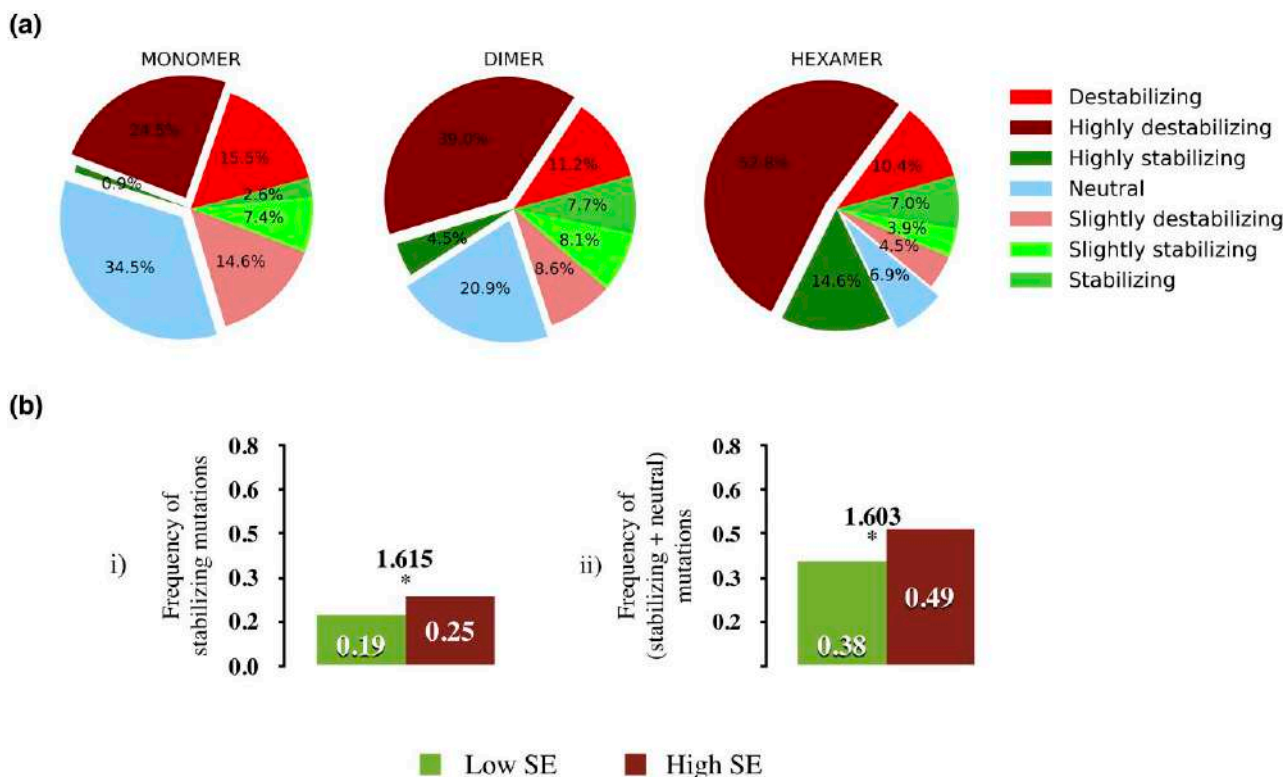
FIG. 3. The effect of know single mutation on NS1 protein stability. (a) Destabilizing mutations constitute the majority of known single mutations in DENV2 NS1. Mutations were classified based on free energy difference (ΔΔG) calculated by FoldX 5.0. Three pie charts representing the frequency of mutations in each category for NS1 state (Monomer, Dimer, and Hexamer) are shown. (b) The frequency of stabilizing and neutral mutations was significantly higher in observed-variable (high SE) versus conserved (low SE) residue positions. The numbers above pairs of bars are the odds ratios, and stars below them denote statistical significance (P-value) calculated according to one-tail Fisher's exact test (*0.01 < P < 0.05, **0.001 < P < 0.01, ***P < 0.0001 and no star for non-significant P-value).

stabilizing (fig. 3a). Similar analysis was performed on all NS1 states of each dengue serotype (supplementary fig. S3a, Supplementary Material online). The majority of the single mutations were predicted as destabilizing, followed by neutral and stabilizing in nature (fig. 3a and supplementary fig. S3a, Supplementary Material online)

Frequency distribution of ΔΔG of mutations within observed-conserved and variable regions was performed. The frequency of stabilizing mutations was higher in observed-variable region (high SE positions) with an odd ratio of 1.615 in DENV2 NS1 dimer (fig. 3a-i). High odd ratios suggest that the probability of finding a stabilizing mutation in the observed-variable (high SE) position is 1.6 times higher compared to observed-conserved (low SE) positions. The P-value for the DENV1 odd ratio was not significant. This could be due to the presence of less number of mutations in DENV1 dataset (supplementary table S1, Supplementary Material online). Stabilizing mutations suggest an advantageous structural change in the protein. Also, these mutations could be functionally important based on their presence or role in protein–protein interaction regions. Predicting the effect of these mutations on NS1 functionality is not considered for this study. Therefore, stabilizing and neutral fractions were considered as a complete set of functionally important

mutations. The probability of finding these (neutral + stabilizing) mutations in observed low versus high SE positions was calculated. Similar to stabilizing mutations, the frequency of stabilizing + neutral mutations was higher in the variable (high SE) positions with significant odd ratios (fig. 3b-ii). There was similar association between stabilizing, stabilizing + neutral mutations and variable positions but with a varied extent of enrichment (different odd ratio values) for other dengue serotypes (supplementary table S2, Supplementary Material online). This suggests that nature of effect of mutation on protein stability can be one of the factors result in formation of observed-conserved/variable positions.

On the other hand, destabilizing mutations were also present in observed-variable regions. Past studies suggest that the destabilizing effect of one mutation can be neutralized by a nearby compensatory substitution (Friedrich et al. 2004; Knies et al. 2008). Therefore, to understand the presence of destabilizing mutations in the observed-variable fraction, we hypothesized that the emergence of these mutations is due to the presence of some compensatory mutation which are nearby or compensatory effect by whole combination of mutations. A systematic study of each possible combination of mutations is computationally intensive. Therefore, we only

focused on the most frequent combinations present in our DENV2 NS1 sequence dataset (Method "NS1 Hexamer").

Structural Changes Caused by a Combination of Mutations Lead to the Selection of Destabilizing Mutations in the Viral Population

In the previous section, majority of observed single mutations were destabilizing in nature (fig. 3a). However, most of the combinations of mutations (which refer to simultaneous multiple mutations) were predicted as stabilizing in dimer and hexamer states (fig. 4a). This leads to the possibility of compensatory behavior within combination of mutations. Interestingly, this trend was reversed in the case of monomer, and majority of combinations were predicted as destabilizing in nature (fig. 4a). Furthermore, highly frequent mutation combinations in viral population tend to be stabilizing in dimer and hexamer (fig. 4a) states in comparison to single occurrence cases. This suggests that in most cases, combination of mutations which maintain structural integrity of NS1 functional states are selected. MD simulations were performed to further confirm the differences observed between the effect of combination of mutations on monomer versus dimer state.

Two of the high-frequency combinations (supplementary table S4, Supplementary Material online) (a) comb165 (H130Y, H51Q, S81A, T266A, V178A, and V287I) (b) comb167 (D282E, I213M, I265T, K173R, L248F, N223S, and T266A) within DENV2 dataset (method "Effect of mutations on NS1 stability") were selected for analysis (fig. 4b). Comb165 was predicted to be "slightly stabilizing" in monomer ($\Delta\Delta G = -0.806704$ kcal/mol) (fig. 4c-i) and "highly stabilizing" in both dimer ($\Delta\Delta G = -5.27766$ kcal/mol) (fig. 4d-i) and hexamer states ($\Delta\Delta G = -13.8669$ kcal/mol) (supplementary fig. S4d, Supplementary Material online). On the other side, comb167 was predicted as "highly destabilizing" in monomer ($\Delta\Delta G = 4.15687$ kcal/mol) (fig. 4c-ii), but "highly stabilizing" in both dimer ($\Delta\Delta G = -2.64617$ kcal/mol) (fig. 4d-ii) and hexamer ($\Delta\Delta G = -7.04378$ kcal/mol) (supplementary fig. S4d, Supplementary Material online) states. Individual mutations within these combinations had varied effects on NS1 stability (fig. 4c and d, supplementary table S5, Supplementary Material online).

MD simulation runs of 100 ns were performed in triplicate for each combination, and protein structure conformers at each nanosecond were sampled. RMSD between each conformer and reference structure was calculated (fig. 4c-iii and d-iii). The change in RMSD trend between wild-type and mutated structure was used to analyze the effect of mutation combinations on structure stability. Monomer with comb167 had higher RMSD in comparison to wild-type, and RMSD trajectory was not stabilized within 100 ns (fig. 4c-iii). Both observations suggested that comb167 was indeed destabilizing the monomer. On the other hand, stabilizing combination (comb165) showed similar RMSD trends compared to wild-type (fig. 4c-iii).

In the dimer case, both combination's RMSD trajectories were stabilized (fig. 4d-iii), and higher stabilization combination (comb165) had lower average RMSD (2.96 Å) in comparison to both wild-type (3.30 Å) and comb167 (3.02 Å). These results were in line with the FoldX predictions.

FoldX considers multiple parameters including various interactions, solvation energies, and entropies to predict the effect of mutation on protein structure. However, in this study, we only focused on the changes induced in amino acid interactions around a mutated position. In order to achieve this, individual mutations in each combination (monomer and dimer) were simulated for 100 ns as triplicate runs. RMSD trends were analyzed, and 60 ns was considered as uniform time-point for trajectory stabilization. Protein conformation at every nanosecond (60–100 ns) was extracted, and change in the amino acid interactions across mutated position were calculated using the Schrodinger suite (method "Amino acid interaction change analysis"). During analysis, changes in the number of hydrogen bonds (H-bond) were frequently observed, and in some cases, the formation of salt-bridge and pi-pi stacking was observed. For example, in case of K173R mutation, changes in the number of H-bond formations were evident in both monomer and dimer case (supplementary fig. S4-a, Supplementary Material online). Changes induced around mutated residue in a single chain of dimer were considered for analysis to keep uniformity during comparison between monomeric and dimeric states.

Comb165 consists of only one destabilizing mutation V178A (fig. 4c–i) and H51Q (fig. 4d-i) in monomer and dimer state respectively, while the rest were either neutral or stabilizing (supplementary table S5, Supplementary Material online). On the other side, comb167 showed interesting observation where, it is predicted as stabilizing in dimer but highly destabilizing in the monomer state. In monomer, comb167 consists of multiple individual destabilizing mutations (supplementary table S5, Supplementary Material online) such as N223S, L248F, I213, and I265T (fig. 4c-ii), whereas in dimer (fig. 4d–ii), two of these mutations (I213M and N223S) were stabilizing in nature (supplementary table S5, Supplementary Material online). A pair of mutations can be considered compensatory if individual mutations have the opposite effect on protein stability, and are spatially close to each other (Davis et al. 2009). In this context, two mutation pairs with possible compensatory behavior, (a) N223S (stabilizing) and L248F (destabilizing), (b) I265T (destabilizing) and T266A (stabilizing) were observed. The N223S-L248F pair was spatially close (∼4.6 Å apart), whereas I265T-T266A was linearly present one after another. Within comb167, the number of H-bonds present in L248F proximity decreases compared to wild and single mutation conditions (supplementary fig. S4b-i, Supplementary Material online), and an opposite trend was observed for N223S (supplementary fig. S4b-ii, Supplementary Material online). Here, bond formation
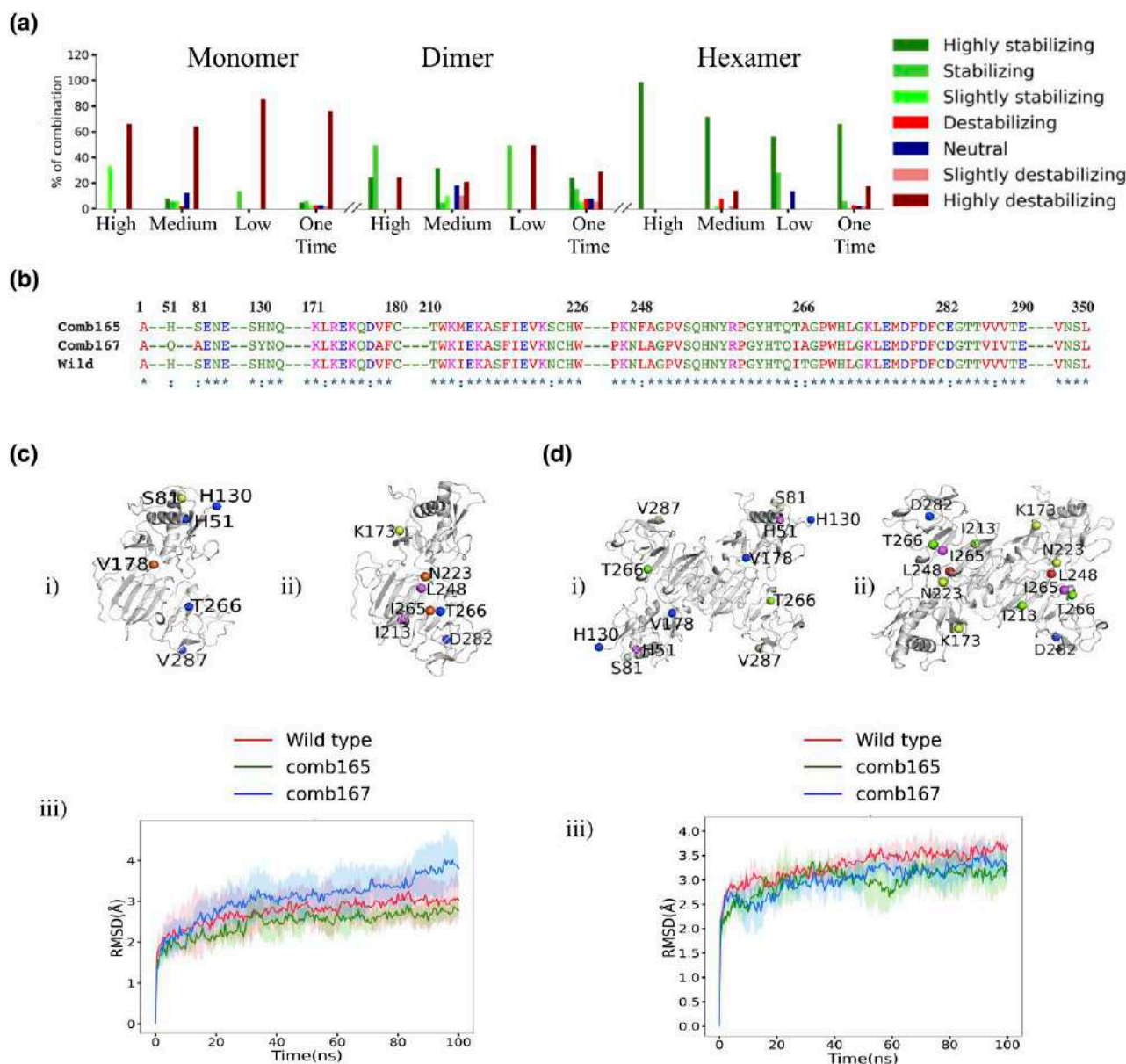
FIG. 4. Varied effect of combination of mutations on NS1 states (Monomer vs. Dimer) stability. (*a*) Most combinations of mutations were predicted as stabilizing in dimer and hexamer state, and destabilizing in monomer state. Mutation combinations were grouped into high (s > 50), medium (20 < s > 50), low (1 < s > 20) and one-time (s = 1) (*x*-axis) based on their observed frequency (cluster size(s)) in sequences from patient samples. *Y*-axis represents percentage of combinations having varied effect on stability (bar) within each group across NS1 states. Color of individual bars represents the effect of mutation combination on NS1 stability. (*b*) Multiple sequence alignment highlighting mutations present in comb165 and comb167 compared to wild-type amino acid sequence. "*" represents sequence conservation, ":" represents mutations. (*c*) and (*d*) MD simulation runs (*n* = 3) were performed to validate observed effect of combination of mutations, (*i*) comb165, and (*ii*) comb167 on DENV2 NS1 monomer and dimer state, respectively. Two mutation combinations are represented on DENV2 NS1 monomer (*c-i* & *ii*) and dimer state (*d-i* & *ii*) structures. Stabilizing (green), neutral (blue), and destabilizing (red) mutations are shown in different shades of colors. RMSD of protein structure conformer at every nanosecond was compared with starting structure of simulation. It is represented as line plot for both NS1 monomer (*c-iii*) and dimer (*d-iii*), where *x*-axis represents time (ns) and *y*-axis represents RMSD (Å). Solid line showed average RMSD after three runs, and the shaded region around the solid line represents standard deviation in RMSD value at that time point. RMSD values as well as trends for stabilizing mutation combination were comparable with wild-type. For destabilizing combination (comb165 in monomer), average RMSD was higher, and trend was not stabilized during whole simulation run.

across L248F-N223S suggested that the changes resulting in the loss of bonds around one mutated position (L248F) could be compensated by increasing the number of bonds around another position "N223S", present within the same region.

To further confirm compensatory behavior between these two mutations, a double mutant (L248F and N223S) was modeled and simulated for 100 ns as triplicate runs. The presence of L248F mutation in N223S background results in loss of H-bonds across residue position
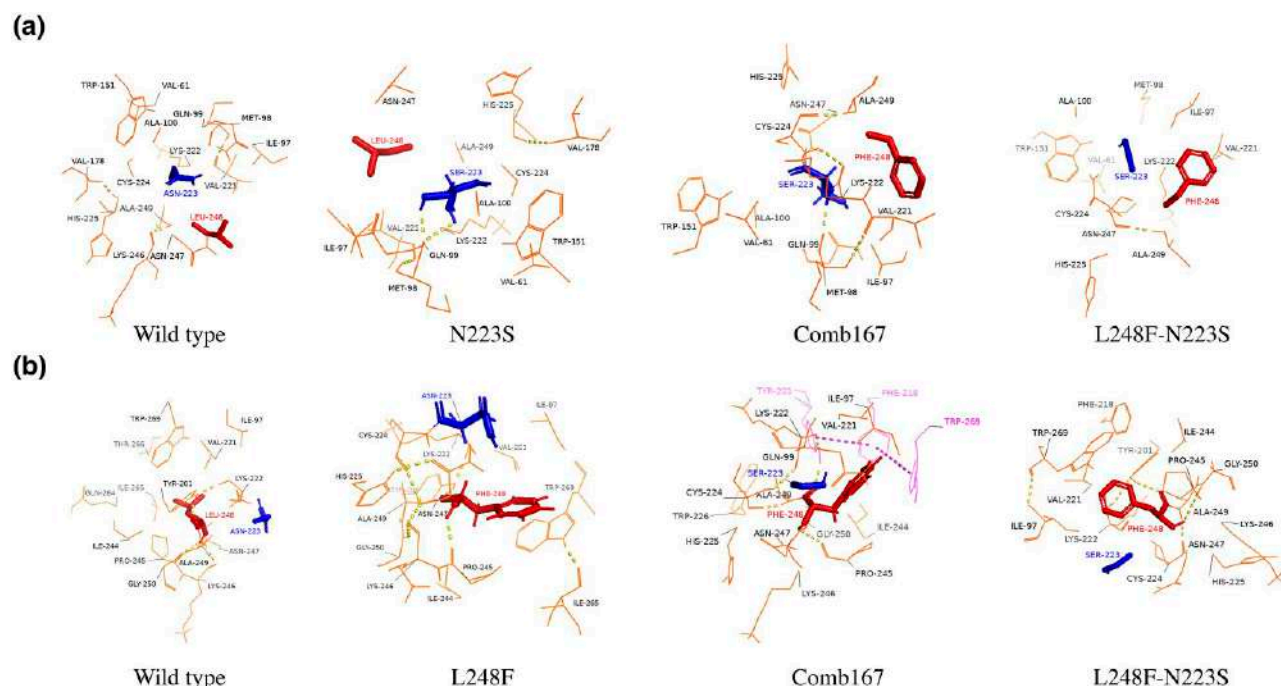
7

FIG. 5. Change in amino acid interaction of a mutated residue in single versus combination of mutations. (*a*) and (*b*) Changes in amino acid interaction around 5Å radius of residue (*a*) 223, and (*b*) 248 is shown during various conditions. Yellow dotted line represents H-bond and magenta lines represent pi-pi interaction. This is a snapshot of interactions at the end of the simulation run. Changes in type of residues, formation, and breaking of new bonds can be observed in the snapshots.

223 (supplementary fig. S4c-ii, Supplementary Material online). In the case of position 248 (supplementary fig. S4c-i, Supplementary Material online), increased retention in H-bond formation was observed in one simulation run, whereas for other two runs, changes were comparable to single mutation. This compensation in retention of H-bond formation within double mutants demonstrates the influence of mutations on interaction network of each other. Such influence could be compensatory over a broad range within combination of mutations. Figure 5 represents a snapshot of amino acid change around wild versus single versus double versus combination of mutations in 223 (fig. 5-*i*) and 248 (fig. 5-*ii*) residue position, respectively. A similar analysis was performed for I265T and T266A mutation pair (supplementary fig. S4c, Supplementary Material online). The effect of I265T mutation on the T266A environment was apparent from the difference in average number of H-bonds formed around T266A in I265T absent (~0.5) (comb165) versus present (~1.5) condition (comb167) (supplementary fig. S4c-i, Supplementary Material online). However, the effect of T266A mutation w.r.t to bond formation around I265T was not observed (supplementary fig. S4c-ii, Supplementary Material online). There are multiple factors which can contribute to mutation stabilization, therefore, compensatory behavior between this pair could be due to other mechanisms/parameters which are not considered in this study.

The observed results demonstrate that the reason for the varied effects between combination versus individual mutations on NS1 protein stability could be based on the changes observed in the local environment around the mutated residue (fig. 5a and b) (supplementary fig. S4, Supplementary Material online). Our results suggest that combinations of mutations play a role in the emergence and selection of destabilizing mutations in the viral population. Studies have shown that mutations have long-range effect on protein conformational change and stability (Unal et al. 2013; Bigman and Levy 2018). Therefore, the presence of compensatory mutations (immediately close to each other) can be one of the mechanisms to neutralize the destabilizing effect. Further investigation is required to uncover other probable mechanisms involved in neutralization of a destabilizing mutation in a combination of mutations.

During analysis of identified single mutations, an interesting observation was made in DENV2 NS1 dimer. Every mutation at several residue positions (such as A122, M98, M124) was predicted as destabilizing in nature. Buried regions of protein are more sensitive to mutations compared to solvent-exposed regions (Bhasin and Varadarajan 2021). Therefore, solvent accessibility status of these residues could explain the observation. To determine their accessibility status, relative solvent accessibility (RSA) was calculated (Method "Relative solvent accessibility calculation"). A122 (RSA = 30.9) and M124 (RSA = 62.6) were found to be solvent-exposed, whereas M98 was buried (RSA = 0.8). Similar residues (both exposed and buried) were also identified in other dengue serotypes. This suggests that irrespective of their accessibility, these
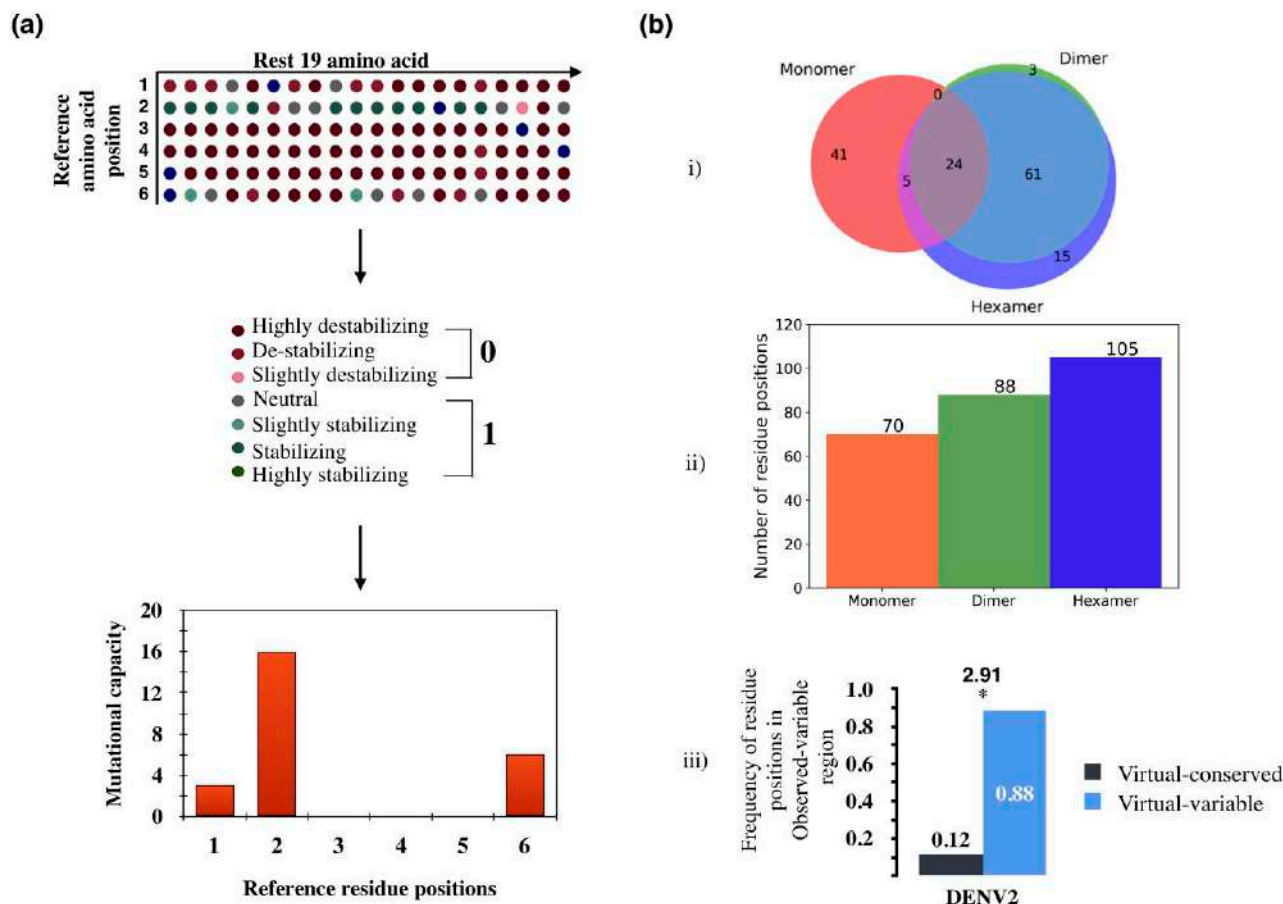
Fig. 6. Prediction and enrichment of virtual-conserved/variable positions in observed-conserved/variable position of DENV2 NS1 protein. (*a*) Workflow used to predict the mutational capacity of a reference residue position. Firstly, each reference residue position was mutated sequentially to the rest of the 19 amino acids, and mutation were classified based on their $\Delta\Delta G$ scores. To normalize the scores, destabilizing mutations were labeled as 0, whereas stabilizing and natural were labeled as 1, and then aggregated (each bar) to calculate MC/RP score. The x-axis represents reference residue positions, whereas *y*-axis represents the mutational capacity score (MC/RP). (*b*) The virtual-conserved residues were found to varying among different DENV2 NS1 states, represented by a Venn diagram (*b-i*). The number of such residues was observed to increase with increase in NS1 state (*b-ii*). The enrichment of virtual-conserved residues in observed-variable (high SE) versus observed-conserved (low SE) region was found through frequency distribution analysis (*b-iii*). Here, each bar represents the frequency of observed-variable residues in virtual-conserved versus virtual-variable positions. The number above the pair of bars is the odds ratio, and stars below them denote statistical significance (*P*-value) calculated according to one-tail Fisher's exact test (*$0.01 < P < 0.05$, **$0.001 < P < 0.01$, ***$P < 0.0001$ and no star for non-significant *P*-value).

residue positions could be structurally critical, resulting in every mutation (till date) to be destabilizing in nature. Further, figure 3*a* clearly shows the change in distribution of destabilizing and stabilizing mutations among different NS1 states. This could be due to NS1 state-dependent effect of mutations on stability. Therefore, understanding the distribution of such sites across NS1 states, and their association with observed-conserved versus observed-variable positions would unravel the role of NS1 protein states in its diversity.

Prediction of the Virtual-conserved and -variable Region in DENV NS1 and Their Association With Sequence Variability

Even if the trends of the multiple mutations from patient samples indicate an overall stabilizing effect on NS1, we

wanted to ensure that our observations are not limited or biased by the availability of patient samples. Hence, FoldX was used next to perform an extensive virtual saturation mutagenesis (Chauhan and Sowdhamini 2022), for predicting mutationally impermeable regions, where each residue position in the reference structure was sequentially and individually mutated to the rest of the 19 amino acids. The output of each substitution was characterized on their $\Delta\Delta G$ score (Method "Effect of mutations on NS1 stability"). For normalization, each destabilization mutation was labeled as 0, while neutral and stabilization mutations were labeled as 1. Normalized scores for every mutation on reference position were then aggregated, and a final score depicting MC/RP was obtained (fig. 6*a*). MC/RP score ranged from 0 to 19. MC/RP score equal to zero suggests all changes are destabilizing, and thereby, anticipated to not accommodate any mutation. Such

residues were termed as "virtual-conserved". On the other side, MC/RP score equal to 19 for a residue suggests all changes are either neutral or stabilizing or both, and were labeled as virtual-variable residues.

The Pearson correlation coefficient was calculated to check variation in MC/RP scores among NS1 states. The correlation coefficient of the MC/RP score of monomer with dimer and hexamer states was 0.13 and 0.07, respectively, whereas the correlation coefficient between dimer and hexamer state was 0.95. This correlation suggested variation in MC/RP scores of the same residue positions across NS1 states, especially from monomer to higher order. Therefore, a detailed comparison between DENV2 NS1 states was carried out. Out of 350 residues in NS1, 70, 88, and 105 residues, were predicted as virtual-conserved in monomer, dimer, and hexamer states, respectively (fig. 6b-ii). Twenty-four residues G4, C56, M68, D93, I97, I136, C144, G200, W211, W266, M242, G250, G267, P268, F280, G283, C292, L308, W312, C313, P320, P321, L322, G329 were common among all three states (fig. 6b-i). Both solvent-exposed (G4, C56, D93, I97, W211, F280, G283, L308, P320, G329) as well as buried (M68, I136, C144, G200, W211, W266, M242, G250, G267, P268, C292, W312, C313, P321, L322) residues were present in this fraction, which follows the observation found in Section "Structural changes caused by a combination of mutations lead to the selection of destabilizing mutations in the viral population". Comparable results were obtained for the rest of the Dengue serotypes (supplementary fig. S3b, Supplementary Material online).

The MC/RP scores of the NS1 dimer state were used for further analysis in this study. It was selected because (a) the dimer state of NS1 is predominant within the cell,

(b) we assumed that immediate dimerization of NS1 monomer could act as an initial purifying step, and any mutation resulting in misfolding of monomer will be instantly degraded, (c) any mutation in monomer affecting dimer formation will result in no viral replication, (d) most frequent mutation combinations were stabilizing in dimer (fig. 4a), and (e) dimer acts as the basic unit for hexamer formation. MC/RP scores of dimer and hexamer are also highly correlated (0.95). Therefore, if MC/RP score is really a good predictor of sequence diversity, then the probability of mutations occurring in NS1 protein would depend on the MC/RP score (virtual-conserved/variable) of the NS1 dimer state.

Validation of predicted MC/RP scores before their usage for further analysis is essential. Therefore, we assumed that the predicted MC/RP scores will align with observed-critical residues. Specifically, experimentally proved "structurally critical residues" were expected to be virtual-conserved type. An extensive literature survey was carried out and a set of known critical residues of the DENV2 NS1 protein (Pryor and Wright 1993, 1994; Wallis et al. 2004; Somnuke et al. 2011; Scaturro et al. 2015) were identified (table 2 and supplementary table S3, Supplementary Material online). MC/RP score of zero was observed for eight Cysteine residues important in structural stability and dimer formation. This observation suggests that other virtual-conserved residues might be structurally critical.

On the other hand, for functionally critical residues (effect only viral replication or viral production) (table 2), MC/RP scores were within the range of 0 and 19. Observed-conserved type and varied MC/RP scores of such positions were expected because of purifying selection toward maintaining their functionality. In this

Table 2. Some of the Known Critical Residues of DENV2 NS1 Protein and Their Respective Shannon Entropy (SE), Relative Solvent Accessibility (RSA), and Mutational Capacity Per Residue Position (MC/RP) Scores in the Dimer State.

Importance	Amino Acid residue	Shannon entropy (SE)	Relative solvent accessibility (RSA)	MC/RP scores (dimer)	Reference
Structure integrity	C5	0.01	55.2	0	Pryor and Wright (1993), Wallis et al. (2004),
	C56	0.01	22	0	Akey et al. (2014), Edeling et al. (2014) and
	C180	0.01	1.9	0	Scaturro et al. (2015)
	C292	0.01	2.2	0	
	C313	0.01	1.1	0	
	C314	0.012	0	0	
	C317	0.012	0	0	
	C330	0.012	12.8	0	
Critical for viral replication	Y159	0.00	43.4	2	Scaturro et al. (2015)
	G160	0.01	6.1	0	
	W8	0.03	88.9	5	
	F161	0.01	85.3	3	
	G162	0.02	123.4	0	
	P320	0.024	51.2	0	
	P321	0.012	0	0	
Less effect on viral replication but huge effect on viral production	S115	0.03	108.7	19	
	W116	0.01	63.3	9	Scaturro et al. (2015)
	D181	0.03	13.2	13	
	T302	0.02	44.2	14	
Important glycosylation site	N131	0.02	31.3	9	Pryor and Wright (1994)
	N208	0.03	66.5	17	

fraction, in addition, some residues (G160, G162, P320, and P321) were also grouped within virtual-conserved type. As the mechanism of action for these residues is unknown, it is quite possible that there is a replication inhibition by structural changes or backbone conformational constraints caused by them (Scaturro et al. 2015). The emergence of known structurally critical residues in our predicted set of virtual-conserved residues validates our approach.

To understand the role of the virtual-conserved residues in NS1 diversity, we hypothesized that there would be a significant enrichment of these residues in observed-conserved (low SE) regions. The frequency of virtual-conserved residues in observed-conserved regions, and virtual-variable residues in observed-variable regions, was higher with significant odd ratios (fig. 6b-iii and supplementary table S1, Supplementary Material online). Figure 6b-ii shows an increase in the formation of virtual-conserved residues across NS1 states (from monomer to hexamer). Therefore, the quaternary arrangement of NS1 affects its protein evolution by increasing the number of virtual-conserved sites, and thereby restricting overall mutational space.

There is a significant association between predicted (virtual-conserved or variable) and observed (low or high SE) sequence diversity. However, a few contradictory results were observed in analysis. The proportion of virtual-variable residues in observed-conserved regions was considerably high (supplementary table S1, Supplementary Material online). Such observations might be due to the reasons that (a) these residues are functionally critical (table 2), and (b) Dengue virus is in the early stage of evolution, and NS1 protein has not diverged significantly. Therefore, under the assumption that these residues have no functional importance, such positions would be likely to accumulate mutations, and can act as mutational hotspot regions in future. Our analysis also predicted a few observed-variable regions (high SE) as virtual-conserved residues (6 out of 52 high SE residues in DENV2 dimer). These two features are contradictory to each other. Possible reasons could be: (a) some of the mutations are functionally important, and therefore, selected at the expense of structure stability, (b) it can be considered as a limitation in FoldX prediction accuracy (Buß et al. 2018), and (c) the presence of compensatory behavior in combination of mutations (Section "Structural changes caused by a combination of mutations lead to the selection of destabilizing mutations in the viral population").

NS1 Protein Conservation

The sequence and structure analysis of the NS1 protein detailed above enabled us to outline possible structural factors involved in its evolution. One primary reason for NS1 protein conservation could be its resistance to change the structure, since most observed single mutations were destabilizing in nature (Section "Majority of single mutations are destabilizing in nature"). This restriction would be

enhanced by the formation of new virtual-conserved positions (a position which ideally can accumulate mutation on linear sequence) across multimeric states (dimer and hexamer) (Section "Prediction of the virtual-conserved and -variable region in DENV NS1 and their association with sequence variability"). Further analysis showed the role of a combination of mutations in NS1 diversity (Section "Structural changes caused by a combination of mutations lead to the selection of destabilizing mutations in the viral population"), allowing destabilizing mutations (which constitute a major proportion) to emerge along with neutral or stabilizing mutations. The requirement of compensatory mutations could be a rate-limiting step in protein divergence. In such cases, a protein requires a counterbalancing mutation to neutralize the effect of the destabilizing mutation. Also, mutations in high-frequency combinations were sparsely present in dimeric (fig. 4c-i-ii) or hexameric interfaces (supplementary fig. S4d-i-ii, Supplementary Material online). Out of 16 unique positions in the high combination group, only two mutations ("T165S", "V6I") were found in the dimer interface, while none of the positions were present in the hexamer interface. Interface residues for dimer and hexameric states were confirmed by PPCHECK (Sukhwal and Sowdhamini 2015).

Implementing Features of NS1 Protein Evolution in the Identification of Critical Residues for Protein Structure/Function and Druggable Sites

Studies indicate the role of understanding protein evolution in predicting or identifying functional and druggable sites (Ghadermarzi et al. 2019). Therefore, we applied the pattern of mutation accumulation and associations learnt from this study to predict critical residues involved in NS1 protein structure and function.

Prediction of Possible Critical Residues in Dimer Interface and Motifs for Protein–protein Interaction

The emergence of experimentally proved critical residues in the set of virtual-conserved residue positions (Section "Prediction of the virtual-conserved and -variable region in DENV NS1 and their association with sequence variability") suggested that MC/RP scores can be used to predict unknown critical residues in NS1 protein. We augmented our approach with literature studies and searched for regions with sequence conservation. Sequence conservation suggests that a residue is important structurally or functionally or both. Therefore, using MC/RP scores as an additional filter over sequence conservation will be beneficial in narrowing down the search. Figure 7a shows the complete set of virtual-conserved residues in DENV2 NS1.

Apart from known structurally critical Cysteine residues (Pryor and Wright 1993; Wallis et al. 2004; Scaturro et al. 2015), 39 observed and virtual-conserved positions (G4, G17, G19, P40, G54, G57, G92, D93, G96, G101, P107, G138, G160, G162, A187, A197, P227, G236, M242, P245, G250, P251, G260, G267, P268, W269, G272, T284, V286,
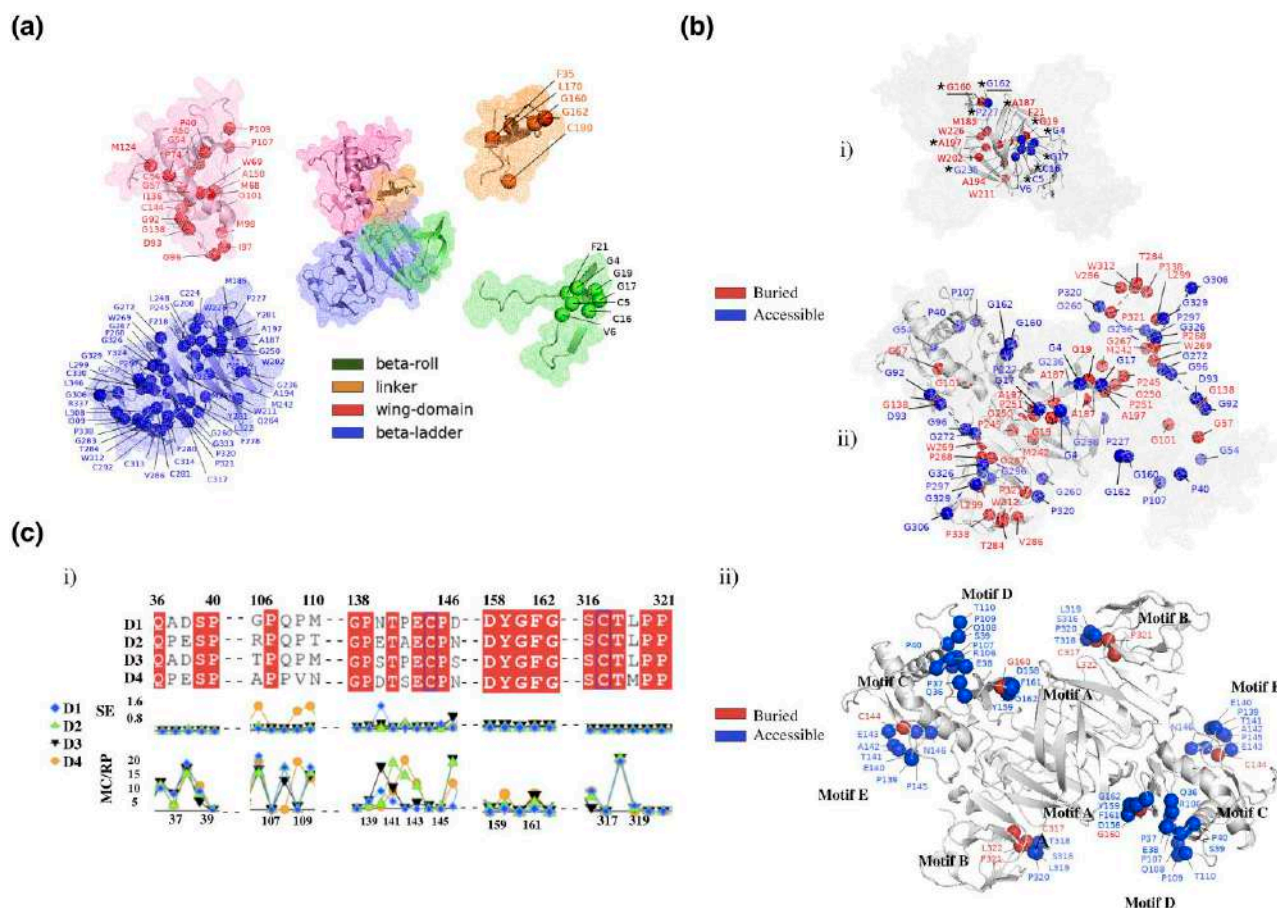
FIG. 7. Critical residues in the dimer interface of DENV2 NS1 protein and interaction motifs. (a) Representation of all virtual-conserved position in different domains of DENV2 NS1 protein in dimer state. (b) Representation of virtual-conserved residues in DENV2 NS1 dimer interface (b-i). "*" represents conserved position across all serotypes and "_" represents already known important residue position in DENV2 NS1 protein. Set of sequentially conserved, observed and virtually-conserved residue positions (b-ii) among all dengue serotypes are shown in NS1 dimer backbone. Red represents buried residues, whereas blue represents accessible residues. (c) Representation of predicted protein–protein interaction sites. Conservation and importance of residues (c-i) within each site is shown by sequence alignment (across dengue serotypes), Shannon entropy (SE), and MC/RP scores (line plots). Low SE (SE <0.1) indicates observed-conserved residues. MC/RP score equal to zero indicates virtual-conserved residue. Identified sites are shown on DENV2 dimer backbone (c-ii) where buried (red) and solvent-accessible (blue) residues are marked.

G296, P297, L299, G306, W312, P320, P321, G326, G329, P338; also see table 3) were identified in all four serotypes (fig. 7b-ii). NS1 dimer interface is critical for viral replication (Rastogi et al. 2016), therefore, interface residues were first identified, and virtual-conserved positions were filtered. Multiple such positions were found in DENV1 ($n = 16$), DENV2 ($n = 18$), DENV3 ($n = 21$), and DENV4 ($n = 16$) NS1 dimer interfaces (supplementary table S6, Supplementary Material online). Eleven residues (G4, C5, C16, G17, G19, G160, G162, A187, A197, P227, and G236) were conserved across all dengue serotypes (supplementary table S6, Supplementary Material online) (fig. 7b-i). G160 and G162 (fig. 7b-i [Underlined residues]) are critical for viral replication (table 2); other nine residue positions could have structural or functional importance. These residues can be considered as a starting point to identify protein–protein interaction motifs or druggable sites.

NS1 has a huge array of interacting host proteins (Karyala et al. 2016). A protein–protein interaction motif

should be stable and contain exposed and reactive amino acid residues. Therefore, we searched for sites with observed and virtual-conserved sites, high RSA score (solvent-exposed), and reactive residues. Motif A (158-DYGFG-162) on dimer interface satisfied the selection criteria. It contained reactive residues, virtual-conserved residues (D158, Y158, and F161), and three known functionally critical residues (G160, F161, and G162) (table 2). Interestingly, a number of Proline residues belonged to virtual-conserved set and were identified on the NS1 periphery (fig. 7b-ii). Past studies suggest Proline-rich motifs (Kay et al. 2000; Krieger et al. 2005; Pagano 2013) as interaction motifs for SH3-domain containing proteins and functionally important. SH3 domain containing Lyn (Li et al. 2020) and Src kinases (H Chu and Yang 2006; Pagano 2013) are involved in dengue and Zika virus maturation during infection. Therefore, NS1 protein might utilize these Proline-rich motifs to interact with SH3-domain-containing proteins. Motifs consisting of Proline, and other residues satisfying our

Table 3. Predicted Structurally Important and Conserved Residue Positions Across All Four Dengue Serotypes.

AA Num	Residue Name				Shannon Entropy (SE)				Relative solvent accessibility (RSA)			
	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
4	GLY	GLY	GLY	GLY	0.005	0.012	0	0	40.8	48.8	35.3	45.2
17	GLY	GLY	GLY	GLY	0	0.017	0.033	0.048	56.1	66	82	63.8
19	GLY	GLY	GLY	GLY	0	0.013	0.033	0.044	0.3	1.1	0.6	0.2
40	PRO	PRO	PRO	PRO	0.006	0.012	0.033	0.048	8.7	8.9	11.8	13.4
54	GLY	GLY	GLY	GLY	0.005	0.028	0	0.024	88.9	92.4	91.1	77.4
57	GLY	GLY	GLY	GLY	0.001	0.012	0.043	0.048	0	0	0	0
92	GLY	GLY	GLY	GLY	0.01	0.012	0.043	0.048	53.3	54.8	46.3	52.7
93	ASP	ASP	ASP	ASP	0.115	0.087	0.052	0.109	76.8	71.3	76.7	70.8
96	GLY	GLY	GLY	GLY	0.018	0.012	0.052	0.048	20	17.7	15.7	19.4
101	GLY	GLY	GLY	GLY	0	0.012	0.051	0.048	0	0	0	0
107	PRO	PRO	PRO	PRO	0	0.012	0.07	0.085	37.7	32.2	33.2	28.5
138	GLY	GLY	GLY	GLY	0	0	0.043	0.052	7	11.6	2.2	14.1
160	GLY	GLY	GLY	GLY	0.005	0.012	0.01	0.028	10.3	6.1	5.4	10.5
162	GLY	GLY	GLY	GLY	0	0.018	0.049	0.071	112.3	111	110.8	111.2
187	ALA	ALA	ALA	ALA	0.01	0.013	0.052	0.048	0.4	0.2	1.6	1.6
197	ALA	ALA	ALA	ALA	0	0.013	0.062	0.048	2.9	4.1	2.7	6.9
227	PRO	PRO	PRO	PRO	0.005	0.02	0.052	0.048	10.8	13.5	2.4	13
236	GLY	GLY	GLY	GLY	0	0.074	0.069	0.048	125.2	126.5	126	124.1
242	MET	MET	MET	MET	0	0.012	0.051	0.044	0.9	0.8	0.9	0.2
245	PRO	PRO	PRO	PRO	0	0.019	0.052	0.048	7	9.4	9.4	6.4
250	GLY	GLY	GLY	GLY	0	0.012	0.052	0.048	0	0	0	0
251	PRO	PRO	PRO	PRO	0	0.012	0.052	0.048	1.4	0.9	0.3	1.2
260	GLY	GLY	GLY	GLY	0.001	0.012	0.052	0.048	36.7	43.3	51.1	56.1
267	GLY	GLY	GLY	GLY	0	0	0.052	0.048	0	0	0	0
268	PRO	PRO	PRO	PRO	0	0.012	0.062	0.048	0	0	0	0.2
269	TRP	TRP	TRP	TRP	0	0.025	0.07	0.048	3.3	5.3	7.5	4.6
272	GLY	GLY	GLY	GLY	0.009	0.013	0.07	0.044	21.4	23.5	21.7	21.7
284	THR	THR	THR	THR	0	0.012	0.043	0.073	0	0	0	0
286	VAL	VAL	VAL	VAL	0.005	0.012	0.043	0.048	0	0	0	1.8
296	GLY	GLY	GLY	GLY	0	0.012	0.081	0.048	36.5	37.6	36.2	41.8
297	PRO	PRO	PRO	PRO	0.04	0.012	0.062	0.048	36	42.8	47.5	30.7
299	LEU	LEU	LEU	LEU	0.011	0	0.051	0.048	1.3	3.2	2.1	0
306	GLY	GLY	GLY	GLY	0.001	0	0.049	0.048	122.4	111.1	83.8	115.6
312	TRP	TRP	TRP	TRP	0.006	0.018	0.051	0.052	0	0	0	0
320	PRO	PRO	PRO	PRO	0.001	0.024	0.043	0.048	52.5	51.2	52.1	48.1
321	PRO	PRO	PRO	PRO	0.001	0.012	0.062	0.048	0	0	0	0.1
326	GLY	GLY	GLY	GLY	0.011	0.012	0.052	0.048	20	20	18.9	20
329	GLY	GLY	GLY	GLY	0.001	0.012	0.062	0.073	61.4	67.4	62.2	65.6
338	PRO	PRO	PRO	PRO	0	0.012	0.052	0.048	0.1	0.2	0	0

AA Num: Amino acid number, D1: Dengue serotype 1, D2: Dengue serotype 2, D3: Dengue serotype 3 and D4: Dengue serotype 4.
Residues having RSA score less than 7 are buried residues whereas positions with RSA score more than 7 are solvent-exposed. This table represents conserved positions (same sequence as well as less diversity [low SE]) across Dengue serotypes.

selection criteria were searched (supplementary table S7, Supplementary Material online), and four such Proline-rich motifs were identified (fig. 7c-i). Motif B (316-SCTLPPL-322) present on beta-ladder domain, was both evolutionarily (observed low SE) and sequentially conserved (expect for L319) among each serotype. It contains some of the known critical residues (P320, P321, and C317) (Scaturro et al. 2015). On the wing domain, motif C (36-Q(P/X)XSP-40) is an evolutionarily conserved site but with different composition in dengue serotypes (X represent variable residue). Motif D (106-XPX(P/X)X-110) is sequentially variable, however, it can be a potential interacting motif due to the presence of Proline residues and close localization with dimer interface. Motif E (139-PXT(P/X)ECPX-146) is present near the junction of the wing and beta-ladder domain. It consists of two Proline and one Cysteine residue in all four serotypes. All

the above sites are represented on DENV2 NS1 backbone in figure 7c-ii.

Computational validation of these motifs was carried out by protein–protein docking. SH3 domain kinases abl, cortactin, yes, Lyn and src were considered as ligand proteins. Profilin (Kay et al. 2000), an active actin-modulator which binds to Proline-rich motif, was also considered for docking (table 4). DENV2 NS1 dimer was used as a receptor protein, and blind docking was performed using the HDOCK server (Yan et al. 2017). Blind docking was done to check the unbiased preference of ligand proteins for NS1 regions. This approach enables one to ask how frequently motifs of interest will be sampled without any prior knowledge during docking. High sampling suggests a given motif would likely be an interacting motif within the host cell. In order to achieve this, top 100 models predicted by HDOCK were considered for analysis. Models were classified into

Table 4. Protein–protein Docking Against Some of the SH3 Domain Proteins and Profilin With DENV2 NS1 as a Receptor Protein.

| Protein Name | No Motif | Specific Motifs | | | | | Multiple Motifs | Predicted top model | | | | | | |
		A	B	C	D	E		Model No	Motif Description	HDOCK Score	PPCHECK (kcal/mol) TE	PPCHECK (kcal/mol) NE/R	MM-GBSA (kcal/mol)	Z-SCORE (DockScore)
Abl (5MO4)	17	21	12	1	1	8	40	3	Motif B & E	−254.11	−58.66	−0.55	−115.46	0.44
Cortactin (2DIX)	47	4	16	0	2	14	17	5	Motif E	−198.58	−138.33	−1.61	−63.15	2.19
Lyn (6NMW)	41	6	14	0	0	23	16	4	Motif E	−203.86	−127.60	−1.77	−92.98	1.46
Src (IFMK)	12	12	12	2	3	13	46	1	Motif B	−241.22	−45.59 & −65.87	−1.01 & −1.22	−124.34	0.40
Yes (2HDA)	33	3	15	0	0	21	18	96	Motif E	−166.21	−112.24	−1.68	−73.19	1.22
Profilin (2PBD)	23	6	14	2	2	25	28	9	Motif B	−226.07	−198.54	−2.33	−146.99	1.56

TE and NE/R represent calculated total and normalized energy per interface residues of protein–protein docked pose by PPCHECK, respectively.
"No motif" represents the total number of docked poses (out of 100) where no ligand residue was close within 5 Å distance of residues belonging to each motif. "Specific Motifs" and "Multiple Motifs" represents poses where the ligand's residues were within 5 Å distance of amino acid of a single motif and with multiple motifs at the same time, respectively. All distances were calculated w.r.t C-alpha atom of residues.
Src was found to interact with both chains ("A" and "B") of NS1. Therefore, two energy values are shown in PPCHECK columns.

no-motif, specific-motif, and multiple motif binders based on the type of NS1 residues presented within 5Å of the SH3 domain. The majority of models were motif binders where the ligand protein was either binding specifically to a single motif or simultaneously to multiple motifs (table 4).

In order to select the best-predicted model/pose, several factors like ligand's interaction with motifs, HDOCK score, total stabilization energy of complex (kcal/mol) calculated by PPCHECK, MM-GBSA score (Genheden and Ryde 2015) and Z-score predicted by DOCKSCORE (Malhotra et al. 2015) were checked. Models were ranked based on interaction with domain residue > HDOCK score > Stabilization energy ~MM-GBSA score > DOCKSCORE. Top poses (supplementary fig. S5, Supplementary Material online) and their respective scores are described in table 4. Motif B was identified as interacting motif for Src kinase and Profilin (table 4), whereas motif E was identified as an interacting motif for Cortactin, Lyn, and Yes kinase. Abl kinase was found to interact with both motifs B and E. Motif B contains some previously known critical residues; therefore, we propose motif E as a novel protein–protein interacting site for further experimental validation. Interestingly, motif A (without Proline residue) was not present in any top predicted poses, which is in accordance with the literature. Selecting a top pose is always challenging and could be easily mistaken for the correct pose. Therefore, our selected top poses should be considered with caution. The observations suggest that NS1 could have plasticity for the usage of interaction sites (table 4). This property could easily allow NS1 protein to modulate functionality or sequester multiple host proteins simultaneously.

The introduction of MC/RP score as a parameter in druggable site selection criteria could be important in drug discovery. This feature suggests that in cases of unnatural evolutionary pressure (small inhibitors), a druggable site would ideally have less probability of mutating at virtual-conserved positions. Therefore, along with protein–protein interaction motif prediction and identification of possible mutational hotspot regions (Section "Structural changes caused by a combination of mutations lead to the selection of destabilizing mutations in the viral population"), our sequence and structural analysis has direct application in identifying druggable sites.

Virtual Screening Against Identified Conserved Druggable Site Dengue NS1 Protein

The site 158-DYGFG-162 (fig. 8a) was identified as the target druggable site. This site was selected because of (a) evolutionary (Low SE and identical residues) conservation among all dengue serotypes (table 5), (b) it contains three known functionally critical residues G160, F161, and G162 (table 2), and two virtual-conserved residues (G160 and G162), (c) it is present on dimer interface and found as protein–protein interacting motif (Motif A) in many models during docking (table 4), and (d) it is solvent-accessible as RSA scores for most residues were more than 8 (table 5).
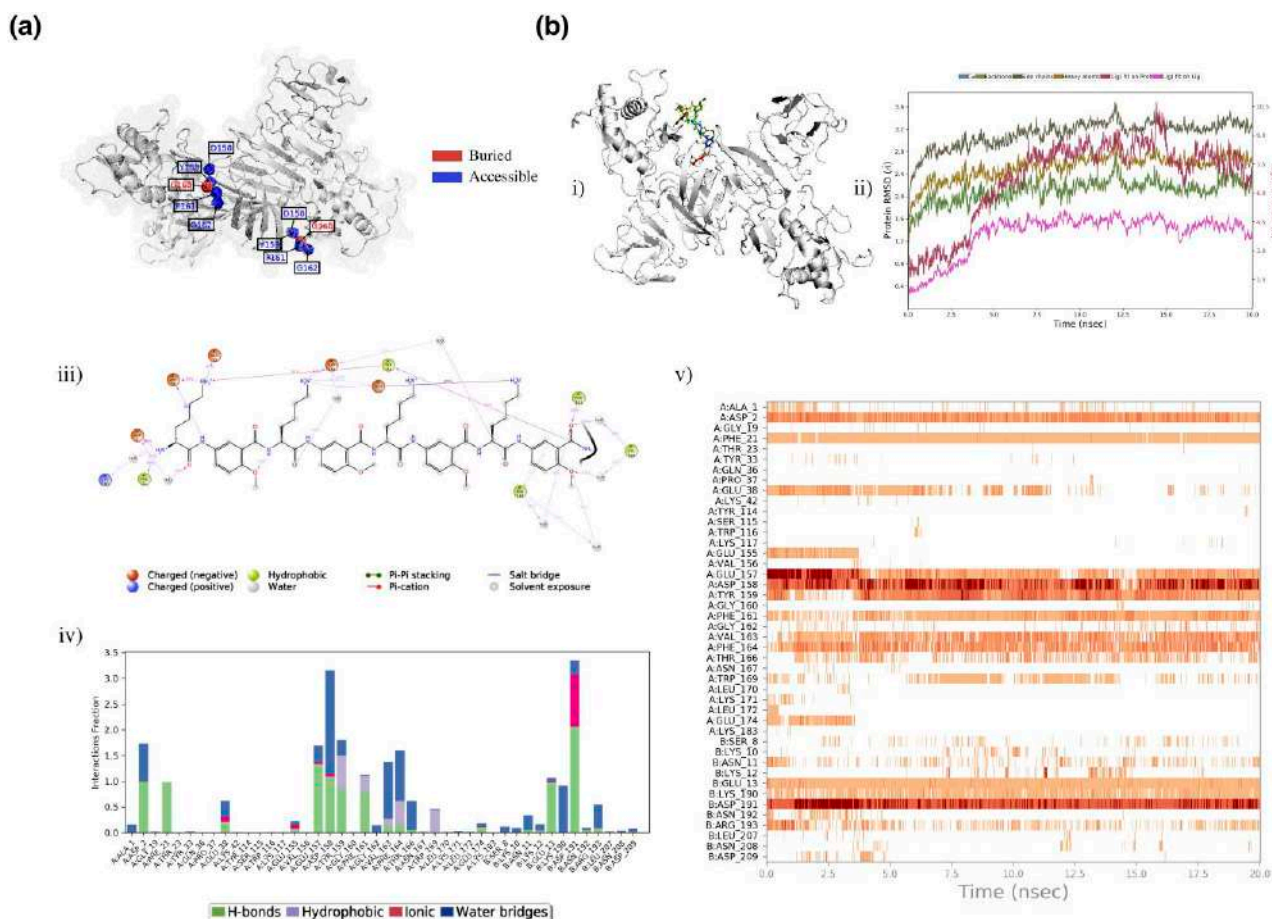
FIG. 8. High-throughput screening of small molecule inhibitors against predicted druggable site. (a) Representation of identified druggable site on the DENV2 dimer backbone. Red represents buried residues, and blue represents accessible residues. (b) High-throughput screening of small molecule inhibitors and validation through set of systemic analyses. (i) Represents protein–ligand complex of DENV2 NS1 and small molecule inhibitor Delparantag (ii) RMSD plot where the x-axis represents simulation time (ns) and the y-axis represents RMSD (Å). (iii) 2D-representation of the interaction of the ligand with NS1 residues throughout MD simulation (iv) Stacked bar plot represents the fraction of time (y-axis) as well as nature of interaction (different stacks in each bar) between ligand and NS1 residues (x-axis) (v) Protein–ligand interaction represented as a timeline where the x-axis represents interaction made by ligand (orange) during simulation run and the y-axis represents interacting residues. White space represents no interaction between ligand and protein residues at that time point.

An extensive virtual screening of small inhibitors (10,467 small molecule inhibitors from Drugbank) against the identified druggable site was carried out. Molecules were first prepared for docking using the LigPrep module. A receptor grid was generated over the site, and then docking was performed using a Glide dock program (Halgren et al. 2004). Virtual screening was performed in three phases: (a) High-throughput virtual screening (HTVS), (b) Standard precision (SP), and (c) Extra precision (XP), where molecules evaluated in one phase passed to the second phase (HTVS → SP → XP) for further evaluations. The top 10% of 9736 molecules screened in HTVS were passed to SP. The output of the SP step was further screened, and molecules with ΔG score less than −5 kcal/mol (522 molecules) were passed to the XP step. MM-GBSA calculations were performed on the top 20% of screened compounds (112) obtained from the XP step. To screen top hits, compounds showing interaction with at least three residues from the target site

(supplementary fig. S6a, Supplementary Material online) were retained for further analysis. Filtered compounds were clustered based on their structural fingerprints and Tanimoto coefficient (Halgren et al. 2004) (supplementary fig. S6a-b, Supplementary Material online). Sixteen clusters with unique backbone and functional groups were obtained. Compounds were then ranked (from different clusters) based on their cluster size, MM-GBSA, and XP scores, and top 6 ranked compounds were selected as top hits (table 6).

In order to check the ligand–protein complex stability of top hits, MD simulation run of 20 ns was performed using Desmond (Bowers et al. 2006). Figure 8b shows the analysis to check ligand–protein complex stability of one of the top hit "Delparantag" (an investigational drug for coronary artery disease) with DENV2 NS1. The interactions between protein and ligand (fig. 8b-i) were considered stable if (a) structural variations were less, (b) a high percentage of various interactions, such as hydrogen bonds,

15

Table 5. SE, RSA and MC/RP Score of the Identified Druggable Site.

AA Num	Residue Name				Shannon Entropy (SE)				Relative Solvent Accessibility (RSA)				Mutational capacity per residue position (MC/RP)			
	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
158	ASP	ASP	ASP	ASP	0.01	0.03	0.06	0.05	50.8	42.3	40.5	42.7	0	1	1	0
159	TYR	TYR	TYR	TYR	0.01	0.00	0.06	0.05	26.5	22.9	21.0	27.7	1	2	1	6
160	GLY	GLY	GLY	GLY	0.01	0.01	0.01	0.03	10.3	6.1	5.4	10.5	0	0	0	0
161	PHE	PHE	PHE	PHE	0.00	0.01	0.05	0.09	45.5	46.1	48.0	44.0	1	3	5	6
162	GLY	GLY	GLY	GLY	0.00	0.02	0.05	0.07	112.3	111.0	110.8	111.2	0	0	0	0

D1: Dengue serotype 1; D2: Dengue serotype 2; D3: Dengue serotype 3; D4:Dengue serotype 4.
MC/RP scores of NS1 dimer state of each dengue serotype were used for identifying the site.

Table 6. Top Hits Obtained After Virtual Screening Against a Conserved, Mutationally Impermeable, and Functionally Important Site in the DENV2 NS1 Protein.

Title	State	Function	XP score (Kcal/mol)	MMGBSA score (Kcal/mol)	Cluster —size	MD-simulation (20 ns)	Interaction with target residues
Brilacidin	Investigational	Brilacidin is under investigation for the supportive care of Mucositis, Stomatitis, Mouth Diseases, and Head and Neck Neoplasms.	−8.08	−58.00	8-2	Stable	throughout the simulation
Delparantag	Investigational	Delparantag has been used in trials studying the treatment of Angioplasty, Coronary Artery Disease (CAD), and Percutaneous Coronary Intervention.	−10.23	−54.09	11-3	Stable	throughout the simulation
Lividomycin A	Experimental	Antibiotic	−11.80	−39.97	7–20	Stable	throughout the simulation
Bradykinin	Investigational	Bradykinin has been investigated for the basic science and treatment of Hypertension and type 2 diabetes.	−10.87	−36.66	10–17	Stable	throughout the simulation
Spectinomycin	Approved	Antibiotic	−7.43	−33.52	18-3	Stable	With one residue for the majority of the simulation
Nadolol	Approved	Nadolol is a non-selective beta adrenal receptor blocker used to lower blood pressure.	−7.60	−31.82	15-2	Stable	weak

Ligand "Delparantag" (in bold) is explained in the text.

hydrophobic interactions, and salt-bridges, were formed during the run, and (c) protein–ligand complex is intact throughout the MD-trajectory. The RMSD plot assessed the stability of the complex (fig. 8b-ii), and fluctuations in the protein regions were analyzed by RMSF (root mean square fluctuation) calculations. Residues in loop regions are expected to fluctuate more than residues in secondary structures such as helix and beta-sheet. A similar trend was observed as NS1 loop regions and regions interacting with Delparantag were fluctuating more than others (supplementary fig. S6c, Supplementary Material online). Protein–ligand contacts (fig. 8b-iii) were analyzed to confirm interaction of ligand with critical residues in the target site. In the case of Delparantag, stable interactions with protein residues D158, Y159, F161, and G162 were found throughout the simulation run (fig. 8b-iv and b–v). A similar analysis was performed on other top hits, and summarized in table 6. Out of six, two drugs (Spectinomycin and Nadolol) are already FDA approved,

three (Brilacidin, Delparantag, and Bradykinin) were in the investigational stage, whereas one (Lividomycin A) was in the experimental stage (supplementary figs. S7–11, Supplementary Material online). Brilacidin, Lividomycin A, and Bradykinin were observed to maintain stable interaction with the target site throughout the simulation.

This study is focused on understanding the role of structural components of Dengue NS1 protein in its diversity. All the above results and correlations are in the context of Dengue NS1. Therefore, a similar analysis was next performed to test the effectiveness of our approach on NS1 proteins of other flaviviruses and influenza virus.

Extension of Dengue NS1 (Sequence and Structure) Analysis to Other NS1 Proteins Results in the Identification of Similar Evolutionary Trends

Complete NS1 protein sequences from other *flaviviruses* (Japanese encephalitis virus (JEV) = 386, West Nile virus

(WNV) = 2,672, Yellow fever virus (YFV) = 438, and Zika = 1,060) and influenza virus ($n = 13,095$) were retrieved from VIPER database. The sequence of each of these proteins of known structure was considered as a reference sequence. Structures were refined in case of presence of missing regions (Method "Model generations for NS1 protein of all Dengue serotypes"). SE scores per residue position were calculated and residue positions were categorized into observed low and high SE groups, respectively. The cut-off value for each protein was decided after analyzing the density distribution of their respective SE scores (supplementary fig. S12, Supplementary Material online). Further, single mutations present per residue position were identified, and their individual effect on protein stability was predicted (Method "Effect of mutations on NS1 stability"). Similar to the trend identified in Dengue virus, destabilizing mutations again dominated the spectrum, followed by neutral and stabilizing mutations (supplementary fig. S13a, Supplementary Material online). This similarity in trend suggested that NS1 proteins tends to remain conserved by occupying stable structural conformation. Such stable structural conformation might be vital for them to perform function and act as an antigenic protein.

Subsequently virtual-conserved and variable residues positions were predicted, and their association with low and high SE regions was determined. Except for Zika virus, odd ratios between observed-variable (high SE) and virtual-variable (supplementary table S2, Supplementary Material online) residue positions were positive with significant P-value for each NS1 protein (supplementary fig. S13b, Supplementary Material online). This enrichment was again similar to observation found in DENV NS1 (Section "Prediction of the virtual-conserved and -variable region in DENV NS1 and their association with sequence variability"). Further, the influenza virus NS1 had several variable positions with very high SE scores, and was thereby used to cross-validate the correlation between virtual-variable and observed-diverse regions (SE score > 1). Out of 40 high SE (diverse) positions only two residues were virtual-conserved (supplementary table S8, Supplementary Material online). These similar trends among Dengue NS1 and other NS1 proteins suggest that our analysis is applicable on other viral NS1 proteins, and can used to predict protein−protein interaction sites or screen/design drugs against druggable site. Interaction motifs will vary among different NS1 proteins. To check sequence conservation of our identified protein−protein interaction motif across *flaviviruses*, reference sequences of each NS1 were aligned, and motif conservation was analyzed. Interestingly, Proline residues within motifs were mostly conserved across flaviviruses (supplementary fig. S13c, Supplementary Material online). Motif B and Motif A which contain known critical residues are more conserved as compared to others.

Conclusion

The **s**equence analysis of NS1 protein distinguishes the residues into observed-conserved (low SE) and observed-variable (high SE) positions. The majority of the observed mutations were destabilizing with more enrichment in observed low SE regions (Section "Majority of single mutations are destabilizing in nature"). Virtual saturation mutagenesis was carried out to predict virtual-conserved (low MC/RP score) and virtual-variable positions (MC/RP score more than 0). The significant enrichment of virtual-conserved positions in observed-conserved (low SE) regions (Section "Prediction of the virtual-conserved and -variable region in DENV NS1 and their association with sequence variability") showed their role in defining NS1 diversity. The proportion of these positions increased with the multimeric state of NS1 (monomer to dimer to hexamer), in accordance with our hypothesis. This suggests the restriction of mutational space owing to higher order quaternary states which renders more surface area of the protein as solvent-buried.

Due to the high mutation rate, multiple mutations can be observed in the viral proteins. This led to the study of combinations of mutations present in high frequency (Section "Structural changes caused by a combination of mutations lead to the selection of destabilizing mutations in the viral population") in patient samples. The differential effect of mutation combinations versus individual mutations (Section "Structural changes caused by a combination of mutations lead to the selection of destabilizing mutations in the viral population") point to the possible mechanism for the emergence of destabilizing mutation in the viral population. The role of compensatory mutations was observed as a factor in the emergence of destabilizing mutations. Analysis of the combination of mutations showed avoidance of mutations at dimer and hexametric interfaces. Apart from understanding the trends of NS1 evolution, the application of our sequence and structure analysis is demonstrated by predicting possible protein−protein interaction (table 4) and druggable sites (table 5). Along with interacting motifs, a conserved, structurally and functionally important site near the dimer interface was selected followed by identification of six potential small inhibitors. A similar trend in other NS1 proteins suggests the importance of maintenance of structural integrity by NS1 proteins in acting as antigenic proteins.

We have considered some reference sequences, and the whole analysis was carried over them. These reference sequences are the building blocks of our analysis. Therefore, the repertoire of virtual-conserved/variable sites is subject to change with the reference sequence used. Unlike our virtual saturation mutagenesis, studying each combination is highly computationally extensive and challenging. This is another limitation of the study. Overcoming the limitations, our analysis conclusively shows the role of virtual-conserved and variable residues in Dengue NS1 evolution. Such residues could help to identify possible protein−protein interaction motifs. These regions can further be confirmed experimentally. The study also uncovers the nature of virtual-variable sites that can accommodate mutations during future viral evolution.

Methods

Data Retrieval, Processing, and Sequence Analysis:

Complete viral protein sequences were downloaded from two publicly available databases, NCBI (Brister et al. 2015) and Viper (Pickett et al. 2012), respectively. NS1 sequences were retrieved and filtered to obtain a set of complete and unique sequences. SE (Litwin and Jores 1992) was calculated to understand the diversity at each amino acid position. This study's reference strains were as follows: DENV1: Accession number KM204119.1, DENV2: Accession number U87411.1, DENV3: Accession number KU050695.1 and DENV4: Accession number KR011349.2 were retrieved from NCBI. Subsequently, an intense literature survey was done to identify structurally and functionally critical residues. A Python script was written to analyze the temporal and geographical distribution of NS1.

Model Generations for NS1 Protein of all Dengue Serotypes:

Homology Modeling for Disordered Loop Regions in DENV 2 NS1 Crystal Structure:

The solved structure of Dengue serotype 2 was downloaded from PDB (Berman et al. 2000) (PDB ID: 4O6B) (Akey et al. 2014). In this model, the loop from the wing domain was incomplete. Therefore, the complete model was generated by modeling loops using multi-template homology modeling by MODELLER 9.22 (Eswar et al. 2006). Blastp (Altschul et al. 1990) was performed on the PDB database to find the closest homolog. The closest predicted homolog "West Nile virus NS1" also had similar unresolved loop regions. Therefore, the second closest homologue, "Zika virus NS1" (PDB ID: −5GS6) (Xu et al. 2016), with query coverage 99% and sequence identity of 53.60%, was considered as template. Corresponding missing loop regions were selected from the Zika structure, and homology models were obtained. Individual sequence identity for loop1 (9–14), loop2 (109–132), and loop3 (160–167) were 40%, 35%, and 75%, respectively. Models were filtered according to the DOPE score. The top three predicted models were then validated using SAVES 5.0 and ProSA server (Wiederstein and Sippl 2007). The final selected model was simulated at 300 K temperature, 1 bar pressure for 100 ns in an OPLS force field (Jorgensen et al. 1996). GROMACS (Berendsen et al. 1995) was used for Molecular dynamic simulations.

Homology Modeling of DENV 1, 3, and 4 NS1 Proteins:

MODELLER 9.22 was used for homology modeling using the complete DENV2 NS1 model as a template. Obtained models were selected and validated as mentioned above.

NS1 Hexamer

The solved structure of DENV2 NS1 (PDB: 4O6B) is an asymmetric unit. Therefore, the hexameric conformation was constructed by applying the crystallographic 3-fold to PDB coordinates. For hexamer modeling of the rest of dengue serotype NS1, coordinates in 4O6B were used as a reference and the respective dimer model as a starting asymmetric unit. PyMol 2.4.0 was used to generate these bio-assemblies.

Relative Solvent Accessibility Calculation:

RSA of NS1 dimer for all four serotypes was calculated using JOY (Mizuguchi et al. 1998) from an in-house IWS server (http://caps.ncbs.res.in/iws/). Residues with a total side chain RSA value equal to or less than 7.0 were classified as buried residues, whereas residues with RSA value greater than 7.0 were considered accessible residues.

Effect of Mutations on NS1 Stability:

The effect of mutation on NS1 structure stability was predicted using FoldX 5.0 (Schymkowitz et al. 2005). Three types of mutations lists were created from the sequence dataset:

a) List of single mutations per amino acid position: This list consists mutations present in a residue position of all strains in comparison to reference strain.

b) Comprehensive list for virtual saturation mutagenesis: where each amino acid residue in the NS1 sequence was changed to the rest of 19 amino acid residues (no permutation and combinations has been tried).

c) List of combinations of mutations present in a strain in comparison to reference strain: To achieve this, NS1 sequences having metadata about sample type, isolation date, region, and disease status were filtered from the complete set of sequences. This ensured that sequences under analysis are representative of the viral population circulating in the human population. Identical sequences (100% sequence similarity) were clustered, and were categorized into high ($s > 50$), medium ($20 < s > 50$), low ($1 < s > 20$) and one-time ($s = 1$), based on their cluster size (s). Representative sequence from each cluster was compared with reference sequence and list of mutation combinations was generated.

To predict effect of mutations on protein stability, each mutation list was passed to FoldX. It first repairs the reference protein structure, and then calculates energy difference between wild-type and mutated structures ($\Delta\Delta G$). Calculation was performed in triplicates and mean $\Delta\Delta G$ score was used to classify mutations. Mutations are classified into three categories: stabilizing ($\Delta\Delta G < −0.46$ kcal), destabilizing ($\Delta\Delta G > −0.46$ kcal), and neutral ($−0.46$ kcal/mol $< \Delta\Delta G \leq +0.46$ kcal/mol). For detailed analysis, destabilizing and stabilizing groups were further divided into three subgroups based on $\Delta\Delta G$ scores. Highly stabilizing: $\Delta\Delta G < −1.84$ kcal/mol, stabilizing: $−1.84$ kcal/mol $\leq \Delta\Delta G < −0.92$ kcal/mol, slightly stabilizing: $−0.92$ kcal/mol $\leq \Delta\Delta G < −0.46$ kcal/mol, slightly destabilizing: $+0.46$ kcal/mol $< \Delta\Delta G \leq +0.92$ kcal/mol,

Destabilizing: $+0.92$ kcal/mol $< \Delta\Delta G \leq +1.84$ kcal/mol, highly destabilizing: $\Delta\Delta G > +1.84$ kcal/mol.

In case of virtual saturation mutagenesis, all stabilizing and neutral mutations were labeled as 1. On the other hand, all destabilizing mutations for a residue position were labeled as 0. Total aggregate score for each position was calculated and termed as Mutational capacity per residue positions (MC/RP). A Python script was used for calculation and plotting.

Molecular Dynamic Simulations:

MD simulation were used to validate some of the top predicted mutations/combinations. Simulation (in triplicate) were performed using GROMACS-5.2.2 (Berendsen et al. 1995). A standard procedure was adopted, including defining an orthorhombic box around the protein of interest with 10Å as a buffer distance. The box was solvated using a generic equilibrated 3-point solvent model (spc216.gro), and the system was neutralized by adding a counterpart ion. Both solvated and electro-neutral systems were minimized to avoid steric clashes energy before the production run, using the steepest descent minimization algorithm. Force $<1,000$ kJ/mol/nm was considered as the cut-off for the energy minimization step. It was followed by two-phase step where the temperature and pressure of assembly were brought sequentially in optimum to the main production run. In the first phase, short MD simulation run of 100 ps was done under an NVT ensemble using a modified Berendsen thermostat with constraints in bonds involving H-bond. This was followed by a 100 ps-run at 300 K and 1 atmospheric pressure using an NPT ensemble. Parrinello-Rahman barostat was used for pressure coupling. After relaxation, production MD was run in an NPT ensemble using an OPLS force field. For simulations, default parameters of leap-frog integrator were used. The temperature and pressure were kept at 300 K and 1 bar, respectively. The production MD was run for 100 nanoseconds. "rms" a utility of GROMACS was used to calculate RMSD.

Amino Acid Interaction Change Analysis

Each scenario (wild, single mutation, and combination of mutations) was simulated for 100 ns (method "Molecular dynamic simulations"), and the RMSD of each trajectory was calculated. A uniform stabilization time for all trajectories were determined, and trajectories were split from that point to the end of the simulation (100 ns). Conformation at every nanosecond was obtained, and amino acid interaction around residue/mutation of interest was identified using Schrödinger (Schrödinger release 2022–3: BioLuminate, Schrödinger, LLC, New York, NY) script:protein_interaction_analysis.py. Command \$SCHRODINGER/run protein_interaction_a-nalysis.py -structure {0} -group1 {1} -group2 {2} -outfile {3} was used to identify changes in amino acid interaction, where group1 represents residue of interest and group2 represents rest of the residues. Default parameters were used

to predict H-bonds, van der Waal collisions, pi-pi stacking, and salt-bridge formation. Changes induced by a mutation in a single chain of NS1 dimer were considered during analysis. This was done to uniformly compared changes between monomer and dimer state. In the dimer case, the exact number of any interaction formation can be obtained by multiplying the mentioned value by 2.

Protein−protein Docking

Blind docking was performed using a HDOCK server (Yan et al. 2017). It predicts the binding complexes between two proteins using template-based modeling and *ab-initio* docking as hybrid docking algorithms. It provides an HDOCK score which can be used to rank the models. Docking was performed using default parameters.

Virtual Screening of Inhibitors

Drugbank database (Wishart et al. 2018) was first downloaded in SDF format (Standard Delay Format). Virtual screening has been done for Dengue serotype 2 NS1 protein only. Downloaded compounds were prepared for screening using the Ligprep module in Schrödinger (Madhavi Sastry et al. 2013). OPLS3e (Harder et al. 2016) force field and pH 7.4 was used for ligand preparation. Specified chirality was retained, and one structure per ligand was specified as an output. Protein was prepared for docking using Protein-preparation wizard (Madhavi Sastry et al. 2013) in Maestro Schrödinger (Schrödinger Release 2022–3: Maestro, Schrödinger, LLC, New York, NY, 2021.) Docking is performed on a site having conserved solvent-exposed residues with zero mutational capacity and functional relevance. A receptor grid around the docking region was specified and generated using a receptor-grid generation module in Glide (Halgren et al. 2004). Residues from predicted sites were specified, and rotatable bonds across the site (if any) were enabled during grid generation. Protein−ligand docking was performed using a glide docking module in the Schrödinger suite. We followed 3-step docking as mentioned in suit: (a) HTVS to narrow the list of potential ligands. According to their docking score top, 10% ligands were taken further for (b) SP mode, and finally, the compound obtained from the SP step with a docking score less than 5 kcal/mol was screened by (c). Extra Precision mode (XP) (Friesner et al. 2006). MM-GBSA (Molecular Mechanics energies combined with Generalized Born and Surface Area continuum solvation) score was then calculated for the top 20% of compounds obtained from the XP step. After MM-GBSA calculations, the protein−ligand interaction fingerprint was generated using the SIFT module from Schrödinger. The screened hits were clustered based on the fingerprint and Tanimoto coefficient (Halgren 2007; Wang et al. 2015). Potential compounds were selected according to their interaction with critical residues, mode of action, MM-GBSA score, XP score, and cluster size. Top six compounds were identified as potential inhibitor candidates, and were taken forward for the MD simulation run. To

streamline the virtual screening pipeline, we have used Desmond, a utility from the Schrödinger suite, to perform the simulation runs. The pipeline for MD simulation was as followed in our previous study (Sharma et al. 2020).

Statistics

Odd ratios calculate any association between two categorical features (Szumilas 2010) and were calculated using a Python script for statistical correlation. The significance of the correlation was determined by P-value. High and positive odd ratios with significant P-values suggested a strong association between the two features.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

A.S. performed the entire work and wrote first draft of the manuscript. S.K., A.S., and R.S. conceptualized the ideas. S.K. and R.S. critically read and improved the manuscript.

Data Availability

The data used in this study are available from the corresponding author upon reasonable request.

References

Akey DL, Brown WC, Dutta S, Konwerski J, Jose J, Jurkiw TJ, DelProposto J, Ogata CM, Skiniotis G, Kuhn RJ, et al. 2014. Flavivirus NS1 structures reveal surfaces for associations with membranes and the immune system. *Science*. **343**:881–885.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. **215**:403–410.

Añez G, Heisey DA, Volkova E, Rios M. 2016. Complete genome sequences of dengue virus type 1 to 4 strains used for the development of CBER/FDA RNA reference reagents and WHO international standard candidates for nucleic acid testing. *Genome Announc*. **4**:1–2.

Avirutnan P, Punyadee N, Noisakran S, Komoltri C, Thiemmeca S, Auethavornanan K, Jairungsri A, Kanlaya R, Tangthawornchaikul N, Puttikhunt C, et al. 2006. Vascular leakage in severe dengue virus infections: a potential role for the nonstructural viral protein NS1 and complement. *J Infect Dis*. **193**:1078–1088.

Beltramello M, Williams KL, Simmons CP, MacAgno A, Simonelli L, Quyen NTH, Sukupolvi-Petty S, Navarro-Sanchez E, Young PR, De Silva AM, et al. 2010. The human immune response to dengue virus is dominated by highly cross-reactive antibodies endowed with neutralizing and enhancing activity. *Cell Host Microbe*. **8**:271–283.

Berendsen HJC, van der Spoel D, van Drunen R. 1995. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun*. **91**:43–56.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res*. **28**:235–242.

Bhasin M, Varadarajan R. 2021. Prediction of function determining and buried residues through analysis of saturation mutagenesis datasets. *Front Mol Biosci*. **8**:1–11.

Bigman LS, Levy Y. 2018. Stability effects of protein mutations: the role of long-range contacts. *J Phys Chem B*. **122**:11450–11459.

Boonnak K, Slike BM, Burgess TH, Mason RM, Wu S-J, Sun P, Porter K, Rudiman IF, Yuwono D, Puthavathana P, et al. 2008. Role of dendritic cells in antibody-dependent enhancement of dengue virus infection. *J Virol*. **82**:3939.

Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA, et al. 2006. Scalable algorithms for molecular dynamics simulations on commodity clusters. Proceedings of the 2006 ACM/IEEE Conference on Supercomputing; 2006 November 11–17; Tampa, Florida. New York (NY): Association for Computing Machinery. p. 84–es.

Brister JR, Ako-Adjei D, Bao Y, Blinkova O. 2015. NCBI Viral genomes resource. *Nucleic Acids Res*. **43**:D571–D577.

Buß O, Rudat J, Ochsenreither K. 2018. Foldx as protein engineering tool: better than random based approaches? *Comput Struct Biotechnol J*. **16**:25–33.

Chauhan PK, Sowdhamini R. 2022. LIM domain-wide comprehensive virtual mutagenesis provides structural rationale for cardiomyopathy mutations in CSRP3. *Sci Rep*. **12**:1–11.

Cheng H-J, Lin C-F, Lei H-Y, Liu H-S, Yeh T-M, Luo Y-H, Linà Y-S. 2009. Proteomic analysis of endothelial cell autoantigens recognized by anti-dengue virus nonstructural protein 1 antibodies. *Exp Biol Med*. **234**:63–73.

Davis BH, Poon AFY, Whitlock MC. 2009. Compensatory mutations are repeatable and clustered within proteins. *Proc R Soc B Biol Sci*. **276**:1823–1827.

Drake JW. 1993. Rates of spontaneous mutation among RNA viruses. *Proc Natl Acad Sci U S A*. **90**:4171–4175.

Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. *Proc Natl Acad Sci U S A*. **96**:13910–13913.

Duffy S. 2018. Why are RNA virus mutation rates so damn high? *PLoS Biol*. **16**:1–6.

Edeling MA, Diamond MS, Fremont DH. 2014. Structural basis of flavivirus NS1 assembly and antibody recognition. *Proc Natl Acad Sci U S A*. **111**:4285–4290.

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M, Pieper U, Sali A. 2006. Comparative protein structure modeling using modeller. *Curr Protoc Bioinforma* **5**:Unit-5.6.

Fan J, Liu Y, Yuan Z. 2014. Critical role of dengue virus NS1 protein in viral replication. *Virol Sin*. **29**:162–169.

Friedrich TC, Frye CA, Yant LJ, O'Connor DH, Kriewaldt NA, Benson M, Vojnov L, Dodds EJ, Cullen C, Rudersdorf R, et al. 2004. Extraepitopic compensatory substitutions partially restore fitness to simian immunodeficiency virus variants that escape from an immunodominant cytotoxic-T-lymphocyte response. *J Virol*. **78**:2581–2585.

Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT. 2006. Extra precision glide: docking

and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*. **49**:6177–6196.

Genheden S, Ryde U. 2015. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov*. **10**: 449–461.

Ghadermarzi S, Li X, Li M, Kurgan L. 2019. Sequence-derived markers of drug targets and potentially druggable human proteins. *Front Genet*. **10**:1–18.

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

Hafirassou ML, Meertens L, Umaña-Diaz C, Labeau A, Dejarnac O, Bonnet-Madin L, Kümmerer BM, Delaugerre C, Roingeard P, Vidalain PO, et al. 2017. A global interactome map of the dengue virus NS1 identifies virus restriction and dependency host factors. *Cell Rep*. **21**:3900–3913.

Halgren T. 2007. New method for fast and accurate binding-site identification and analysis. *Chem Biol Drug Des*. **69**:146–148.

Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. 2004. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem*. **47**:1750–1759.

Harder E, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, et al. 2016. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J Chem Theory Comput*. **12**:281–296.

H Chu JJ, Yang PL. 2006. c-Src protein kinase inhibitors block assembly and maturation of dengue virus. *Proc Natl Acad Sci U S A* **104**: 3520–3525.

Jayathilaka D, Gomes L, Jeewandara C, Jayarathna GSB, Herath D, Perera PA, Fernando S, Wijewickrama A, Hardman CS, Ogg GS, et al. 2018. Role of NS1 antibodies in the pathogenesis of acute secondary dengue infection. *Nat Commun*. **9**:1–15.

Jorgensen WL, Maxwell DS, Tirado-Rives J. 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc*. **118**: 11225–11236.

Karyala P, Metri R, Bathula C, Yelamanchi SK, Sahoo L, Arjunan S, Sastri NP, Chandra N. 2016. DenHunt - a comprehensive database of the intricate network of dengue-human interactions. *PLoS Negl Trop Dis*. **10**:e0004965.

Kay BK, Williamson MP, Sudol M. 2000. The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J*. **14**:231–241.

Khadka S, Vangeloff AD, Zhang C, Siddavatam P, Heaton NS, Wang L, Sengupta R, Sahasrabudhe S, Randall G, Gribskov M, et al. 2011. A physical interaction network of dengue virus and human proteins. *Mol Cell Proteomics*. **10**:M111.012187-1–M111.012187-16.

Knies JL, Dang KK, Vision TJ, Hoffman NG, Swanstrom R, Burch CL. 2008. Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. *Mol Biol Evol*. **25**: 1778.

Krieger F, Möglich A, Kiefhaber T. 2005. Effect of proline and glycine residues on dynamics and barriers of loop formation in polypeptide chains. *J Am Chem Soc*. **127**:3346–3352.

Li MY, Naik TS, Siu LYL, Acuto O, Spooner E, Wang P, Yang X, Lin Y, Bruzzone R, Ashour J, et al. 2020. Lyn kinase regulates egress of flaviviruses in autophagosome-derived organelles. *Nat Commun*. **11**:1–16.

Lin CF, Lei HY, Shiau AL, Liu CC, Liu HS, Yeh TM, Chen SH, Lin YS. 2003. Antibodies from dengue patient sera cross-react with endothelial cells and induce damage. *J Med Virol*. **69**: 82–90.

Litwin S, Jores R. 1992. Shannon information as a measure of amino acid diversity. *Theor Exp Insights into Immunol* **66**:279–287.

Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W. 2013. Protein and ligand preparation: parameters, protocols,

and influence on virtual screening enrichments. *J Comput Aided Mol Des*. **27**:221–234.

Malavige GN, Fernando S, Fernando DJ, Seneviratne SL. 2004. Dengue viral infections. *Postgrad Med J*. **80**:588–601.

Malavige GN, Ogg GS. 2017. Pathogenesis of vascular leak in dengue virus infection. *Immunology*. **151**:261–269.

Malhotra S, Mathew OK, Sowdhamini R. 2015. DOCKSCORE: a webserver for ranking protein-protein docked poses. *BMC Bioinformatics*. **16**:1–6.

Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP. 1998. JOY: protein sequence-structure representation and analysis. *Bioinformatics*. **14**:617–623.

Modhiran N, Watterson D, Muller DA, Panetta AK, Sester DP, Liu L, Hume DA, Stacey KJ, Young PR. 2015. Dengue virus NS1 protein activates cells via toll-like receptor 4 and disrupts endothelial cell monolayer integrity. *Sci Transl Med*. **7**:304ra142–304ra142.

Neufeldt CJ, Cortese M, Acosta EG, Bartenschlager R. 2018. Rewiring cellular networks by members of the Flaviviridae family. *Nat Rev Microbiol*. **16**:125–142.

Pagano MA. 2013. Viral proteins and Src family kinases: mechanisms of pathogenicity from a liaison dangereuse. *World J Virol* **2**:71.

Paranavitane SA, Gomes L, Kamaladasa A, Adikari TN, Wickramasinghe N, Jeewandara C, Shyamali NLA, Ogg GS, Malavige GN. 2014. Dengue NS1 antigen as a marker of severe clinical disease. *BMC Infect Dis*. **14**:570.

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, et al. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res*. **40**:D593.

Płaszczyca A, Scaturroid P, Neufeldtid CJ, Corteseid M, Cerikan B, Ferla S, Brancale A, Pichlmair A, Bartenschlager R. 2019. A novel interaction between dengue virus nonstructural protein 1 and the NS4A-2K-4B precursor is required for viral RNA replication but not for formation of the membranous replication organelle. *PLOS Pathog*. **15**:e1007736.

Pryor MJ, Wright PJ. 1993. The effects of site-directed mutagenesis on the dimerization and secretion of the NS1 protein specified by dengue virus. *Virology*. **194**:769–780.

Pryor MJ, Wright PJ. 1994. Glycosylation mutants of dengue virus NS1 protein. *J Gen Virol*. **75**(Pt 5):1183–1187.

Rastogi M, Sharma N, Singh SK. 2016. Flavivirus NS1: a multifaceted enigmatic viral protein. *Virol J*. **13**:131.

Sanjuán R, Domingo-Calap P. 2021. Genetic diversity and evolution of viral populations. *Encycl Virol*:53–61.

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**:1–32.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res*. **33**: W382.

Screaton G, Mongkolsapaya J, Yacoub S, Roberts C. 2015. New insights into the immunopathology and control of dengue virus infection. *Nat Rev Immunol*. **15**:745–759.

Sharma A, Tiwari V, Sowdhamini R. 2020. Computational search for potential COVID-19 drugs from FDA-approved drugs and small molecules of natural origin identifies several anti-virals and plant products. *J Biosci*. **45**:1–18.

Silva EM, Conde JN, Allonso D, Nogueira ML, Mohana-Borges R. 2013. Mapping the interactions of dengue virus NS1 protein with human liver proteins using a yeast two-hybrid system: identification of C1q as an interacting partner. *PLoS One*. **8**:1–9.

Somnuke P, Hauhart RE, Atkinson JP, Diamond MS, 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.

Sukhwal A, Sowdhamini R. 2015. PPCheck: a webserver for the quantitative analysis of protein-protein interfaces and prediction of residue hotspots. *Bioinform Biol Insights*. **9**:141–151.

MBE

Szumilas M. 2010. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*. **19**:227.

Unal H, Jagannathan R, Bhatnagar A, Tirupula K, Desnoyer R, Karnik SS. 2013. Long range effect of mutations on specific conformational changes in the extracellular loop 2 of angiotensin II type 1 receptor. *J Biol Chem*. **288**:540–551.

Wallis TP, Huang CY, Nimkar SB, Young PR, Gorman JJ. 2004. Determination of the disulfide bond arrangement of dengue virus NS1 protein. *J Biol Chem*. **279**:20729–20741.

Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, et al. 2015. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc*. **137**:2695–2703.

Watterson D, Modhiran N, Young PR. 2016. The many faces of the flavivirus NS1 protein offer a multitude of options for inhibitor design. *Antiviral Res*. **130**:7–18.

Weaver SC, Vasilakis N. 2009. Molecular evolution of dengue viruses: contributions of phylogenetics to understanding the history and epidemiology of the preeminent arboviral disease. *Infect Genet Evol*. **9**:523.

Wiederstein M, Sippl MJ. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*. **35**:W407–W410.

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al. 2018. Drugbank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. **46**:D1074–D1082.

Xu X, Song H, Qi J, Liu Y, Wang H, Su C, Shi Y, Gao GF. 2016. Contribution of intertwined loop to membrane association revealed by Zika virus full-length NS1 structure. *EMBO J*. **35**:2170–2178.

Yan Y, Zhang D, Zhou P, Li B, Huang SY. 2017. HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res*. **45**:W365–W373.

Computational search for potential COVID-19 drugs from FDA-approved drugs and small molecules of natural origin identifies several anti-virals and plant products

ABHISHEK SHARMA[†], VIKAS TIWARI[†] and RAMANATHAN SOWDHAMINI*

National Centre for Biological Sciences, GKVK Campus, Bengaluru 560 065, India

*Corresponding author (Email, mini@ncbs.res.in)

[†]*Joint first authors*

The world is currently facing the COVID-19 pandemic, for which mild symptoms include fever and dry cough. In severe cases, it could lead to pneumonia and ultimately death in some instances. Moreover, the causative pathogen is highly contagious and there are no drugs or vaccines for it yet. The pathogen, SARS-CoV-2, is one of the human coronaviruses which was identified to infect humans first in December 2019. SARS-CoV-2 shares evolutionary relationship to other highly pathogenic viruses such as Severe Acute Respiratory Syndrome (SARS) and Middle East respiratory syndrome (MERS). We have exploited this similarity to model a target non-structural protein, NSP1, since it is implicated in the regulation of host gene expression by the virus and hijacking of host machinery. We next interrogated the capacity to repurpose around 2300 FDA-approved drugs and more than 3,00,000 small molecules of natural origin towards drug identification through virtual screening and molecular dynamics. Interestingly, we observed simple molecules like lactose, previously known anti-virals and few secondary metabolites of plants as promising hits. These herbal plants are already practiced in Ayurveda over centuries to treat respiratory problems and inflammation. Disclaimer: we would not like to recommend uptake of these small molecules for suspect COVID patients until it is approved by competent national or international authorities.

Keywords. Anti-virals; drug design; herbal plants; repurposing drugs; SARS-CoV-2

Abbreviations: NSP1, Non-structural protein 1; SARS-CoV2, Severe acute respiratory syndrome coronavirus 2

1. Introduction

Coronavirus (CoV) belongs to the family *Coronaviridae* and the order *Nidovirales* (sharing with *Arteriviridae and Roniviridae*). Coronaviruses are enveloped, long positive-sense single-stranded RNA viruses, which are best known for causing mild to severe respiratory and enteric infection among a vast range of hosts (Masters 2006). These are further divided into 4 groups/genera named as Alphacoronavirus (α-CoV), Betacoronavirus (β-CoV), Gammacoronavirus (γ-CoV) and Deltacoronavirus (δ-CoV), based on sequence similarities and antigenic cross-reactivity. Human-CoV belongs to group I and group II of Betacoronavirus. HCoV-OC43, HCoV-229E, SARS and MERS are some examples of Human-CoV, out of which SARS and MERS are highly pathogenic in nature (Masters 2006; Narayanan *et al.* 2015). Recently, a new pathogenic Human-CoV strain known

This article is part of the Topical Collection: COVID-19: Disease Biology & Intervention.

as SARS-CoV2, spreading the COVID-19 infection, has emerged by December 2019 with Wuhan of Hubei province in China as the epicenter (Wu *et al.* 2020a, b). The origin of this virus is still under investigation but has been speculated as a zoonotic shift from bat to human. It has been shown that human ACE2 has a high predicted affinity than ACE2 from other species (Piplani *et al.* 2020). The outbreak of COVID-19 has spread across the globe and has taken the shape of a pandemic (Novel Coronavirus (2019-nCoV) situation reports, World Health Organization). America, Russia, United Kingdom, India, Italy, Spain, and France are among worst-hit countries. As of 14 June 2020, this has infected 7,891, 289 individuals and has caused more than 432,746 fatalities across the globe (Johns Hopkins Coronavirus Resource Center; Worldmeter. *https://www.worldometers.info/coronavirus/*; Novel Coronavirus (2019-nCoV) situation reports - World Health Organization). At present, there are no drugs or vaccines available against this, and patients are treated according to symptoms shown by them. Remdisivir (drug originally designed to treat Ebola), Colchicine (Deftereos *et al.* 2020), Chloroquine and Hydroxychloroquine (an antimalarial drug), Kevsara (an arthritis drug) and few other antiviral drugs are being considered for treatment. But they do not directly make use of the virome of SARS-CoV2 (Wang *et al.* 2020) (*https://www.nasdaq.com/articles/8-experimental-coronavirus-treatments-to-watch-2020-03-31*). Multiple vaccine trial has been going on worldwide, among which Bacillus Calmette Guerine (BCG) live attenuated vaccine and AZD1222 are in phase 2/3 of the clinical trials. (*https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker*)

The first genome of the COVID-19 strain was sequenced by Wu *et al.* (2020b) from a 41-year-old man and was found to be closely similar to SARS-CoV. The structural component of the virus consists of four proteins: Spike (S), Membrane (M), Envelope (E) and Nucleocapsid (N) protein respectively. S protein is critical for viral infection as it enables host-pathogen interaction and mediated viral entry into the host cell. M protein is a multipass transmembrane protein, a major constituent of virion envelope, and known to provide its shape. E protein, unlike the name, suggests it is a minor constituent of the envelope and 80–120 aa in length. N protein as the name suggests forms helical nucleocapsid of virion (Masters 2006).

The 5′-end of the genome encodes two open reading frames, ORF1a and ORF1b, respectively, which code for all non-structural proteins (NSP1-16) (Masters 2006; Narayanan *et al.* 2015). These proteins are essential for viral replication as well as infection, whereas the function of some is yet to be identified. Among these non-structural proteins within the CoV family, some are conserved in sequence, whereas others are highly diverged in nature. NSP1 (non-structural protein 1) is one of such diverged proteins which is encoded by ORF1a and varied in amino-acid length among CoV-groups (Narayanan *et al.* 2015). COVID-19 NSP1 consists of 180 and shows sequence similarity with SARS protein (Elbe and Buckland-Merrett 2017; Wu *et al.* 2020a). Despite differences in amino acid sequence and length, it has shown to be functionally highly conserved (Narayanan *et al.* 2015; Shen *et al.* 2019). SARS-CoV NSP1 is most well-studied amongst the viruses of the coronavirus family. NSP1 has shown to act as a virulence factor (Huang *et al.* 2011; Narayanan *et al.* 2015; Zst *et al.* 2007) and mutation in this protein results in the production of attenuated virus *in vitro* and *in vivo* (Zst *et al.* 2007). NSP1 deploys two strategies to inhibit host cell expression *viz.* Inhibition of host translation and Induction of host mRNA degradation. It inhibits host translation by forming a complex with 40s ribosome subunit, which prevents the formation of active polysome. Complex formation with 40s subunit also shown to inhibit its translational ability (Kamitani *et al.* 2009). It further affects host cell gene expression by inducing host mRNA degradation in a template-specific manner. Term template-specific does not imply its association with protein sequence but relates to the ability to specifically degrade capped host mRNA (Huang *et al.* 2011; Kamitani *et al.* 2006, 2009; Narayanan *et al.* 2008; Tanaka *et al.* 2012) as compared to SARS-mRNA (Kamitani *et al.* 2006). mRNA is hypothesized to get cleaved by unknown host endonuclease since NSP1 does not possess any endonucleolytic activity. Other than these, NSP1 is shown to cause chemokine dysregulation which correlates with high inflammation in severe patients (Law *et al.* 2007; Channappanavar and Perlman 2017; Wong *et al.* 2004). It suppresses innate immune response by degrading IFN-beta mRNA (Tanaka *et al.* 2012) and affecting antiviral signaling (Jauregui *et al.* 2013). Yeast-two hybrid assays have shown NSP1 to interact with multiple host proteins (Pfefferle *et al.* 2011). N-terminal region is shown to be important in immune response dysregulation (mutation studies) (Jauregui *et al.* 2013) and protecting viral RNA (R124, present in our dock site) (Tanaka *et al.* 2012). The C-terminal region is critical of transcriptional inhibition of host mRNA (Narayanan *et al.* 2015; Tanaka *et al.* 2012).

In this study, COVID-19 NSP1 is the target protein and we hypothesized that the inhibition of NSP1 can

potentially attenuate the virus and suppress adverse immune-pathology caused by it. We have generated a homology model of the NSP1 and used this model to carry out virtual screening to identify potential inhibitors and lead compounds. Our searches are directed within the database of FDA-approved drugs (DRUG-BANK) and those which are derived from molecules of natural origin (SUPERNATURALDB). Finally, we performed MD simulation to ensure that there are indeed stable protein-ligand interactions when the system is visualized to undergo limited conformational freedom. We find several anti-viral compounds, few secondary metabolites of plant origin and simple compounds (like lactose) to retain the high potential to act as NSP1 inhibitors.

2. Methods

2.1 *Sequence retrieval and analysis*

The full repository of COVID-19 protein sequences was downloaded from NCBI (Brister *et al.* 2015; Hatcher *et al.* 2017). Wuhan-Hu-1 strain [Accession number: NC_045512] was among the first to be sequenced from Wuhan of Hubei province. Hence this is considered as the 'reference genome' in this study. NSP1 protein sequences were extracted, incomplete sequences were removed and curated sequences were then passed to SNP analyzer (utility in ViPR Database) to understand variation among the NSP1 sequences (Pickett *et al.* 2012). A similar analysis had been done for the Indian sequences. To understand evolutionary pressure on NSP1 protein Shannon entropy (Shannon 1948) per residue has been calculated using a python script. A set of key residues, important in suppressing host gene expression and antiviral signaling were identified. A mutation study done by another group (Jauregui *et al.* 2013) was used as a reference.

2.2 *Homology modelling*

NSP1 protein sequence [Accession number: YP_009725297] was retrieved from NCBI for Homology modelling. Blastp (Camacho *et al.* 2009) was used to search for the nearest structural homologue in Protein Data Bank (PDB) (Berman *et al.* 2003) to serve as a template for modelling. Segments of NSP1 sequence, where the association with the template was unknown, were removed. Modeller 9.12 (Eswar *et al.* 2006) was used for Homology modelling. Homology

models were first filtered according to DOPE score. Top 3 predicted models were then subjected to structure validations (by using SAVES 5.0 (Laskowski *et al.* 1993) (*https://servicesn.mbi.ucla.edu/SAVES/*) and ProSA server (Wiederstein and Sippl 2007)). Based on DOPE score, Ramachandran plot and ProSA profile, the best predicted model was selected for virtual screening.

2.3 *Virtual screening of inhibitors*

FDA-approved drugs and Super Natural II database (database of natural products) were used for docking purposes. FDA-approved drugs were downloaded in SDF format (Standard Delay Format) from Drug-bank (Wishart *et al.* 2018) whereas supernatural compounds were obtained from supernatural database (Banerjee *et al.* 2015).

2.3.1 *Ligand and protein preparation:* Downloaded compounds were prepared for screening using Ligprep module in Schrodinger (Schrödinger Release 2019-4: LigPrep, Schrödinger, LLC, New York, NY, 2019). For FDA-approved drugs OPLS3e force field, targeted pH 7.4 +/−0.0, retain specified chiralities and 1 structure per ligand was specified during ligand preparation. For supernatural database we had specified pH range from 6.0 to 8.0 with maximum 32 structure per ligand. This was performed to scan and produce broad chemical and structural diversity from each molecule.

Protein was prepared for docking by using Protein-preparation wizard (Sastry *et al.* 2013) in Maestro Schrodinger (Schrödinger Release 2019-4: Maestro, Schrödinger, LLC, New York, NY, 2019.)

2.3.2 *Docking site prediction:* SiteMap (Halgren 2007; Halgren 2009) was used to predict the potential drugable deep and shallow sites on the target protein. The site with high S-score, as well as D-score, was selected for ligand docking.

2.3.3 *Receptor grid generation:* Receptor-grid around docking region on the protein was generated using receptor-grid generation module in Glide, Residues from top predicted deep and shallow sites were specified and rotatable bonds across the site (if any) were checked during grid generation.

2.3.4 *Protein–ligand docking:* Using the glide docking module, a library of prepared ligands and protein with prepared receptor binding grid were docked. First,

High-throughput virtual screening (HTVS) was performed. This narrowed down the list of potential ligands and the top 10 percent from this were then screened with Standard Precision (SP) mode. Finally, 10 percent of hits obtained from SP were passed to Extra precision (XP). Selection of top 10 percent compounds were done based on top dock score and binding energy (Friesner *et al.* 2004; Friesner *et al.* 2006; Halgren *et al.* 2004) (Schrödinger Release 2019-4: Glide, Schrödinger, LLC, New York, NY, 2019.)

Binding energy was calculated using MM-GBSA (Molecular Mechanics energies combined with the Generalized Born and Surface Area continuum solvation) tool of Schrodinger.

2.4   *MD simulations*

The conformer of protein–ligand complex, emerging from XP docking, was assembled using system builder and subject to Molecular Dynamics using the Desmond package of Schrodinger (Bowers *et al.* 2006). For water, the TIP4P model was specified and orthorhombic box shape was used having a buffer distance of 10 Å. Box volume was minimized. The system was neutralized and 150 mM salt (NaCl) was added. The output of the system builder was used for MD. The default relaxation protocol was used to relax the solvated system followed by production MD run for 20 nanoseconds (ns). The relaxation protocol involves energy minimization steps using the steepest descent method with a maximum of 2000 steps. The energy minimization was done with solute being restrained using 50 kcal/mol/Å force constant on all solute atoms and without restraints. Energy minimization was followed by short MD simulation steps which involve (1) Simulation for 12 picoseconds at 10 K in NVT ensemble using Berendsen thermostat with restrained non-hydrogen solute atoms, (2) Simulation for 12 picoseconds at 10 K and 1 atmospheric pressure in NPT ensemble using Berendsen thermostat and Berendsen barostat with restrained non-hydrogen solute atoms, (3) Simulation for 24 picoseconds at 300 K and 1 atmospheric pressure in NPT ensemble using Berendsen thermostat and Berendsen barostat with restrained non-hydrogen solute atoms, (4) Simulation for 24 picoseconds at 300 K and 1 atmospheric pressure in NPT ensemble using Berendsen thermostat and Berendsen barostat without restraints. After relaxation, production MD was run in NPT ensemble using OPLS 2003 force field (Harder *et al.* 2016). For

simulations, default parameters of RESPA integrator (Humphreys *et al.* 1994) (2 femtoseconds time step for bonded and near non-bonded interactions while 6 femtoseconds for far non-bonded interactions) were used. The temperature and pressure were kept at 300K and 1 bar using the Nose-Hoover chain method (Martyna *et al.* 1992) and the Martyna-Tobias-Klein method (Martyna *et al.* 1994) respectively. The production MD was run for 20 nanoseconds.

2.5   *Simulation analysis*

MD simulation analysis was performed using the Simulation interaction diagram (SID) module of the Desmond package. The entire range of simulation time was considered for all analyses. RMSD is calculated for each frame by aligning the complex to the protein backbone of the reference frame. Significantly higher values of 'Lig fit Prot' than protein RMSD signifies the diffusion of ligand away from its initial binding site. Lig fit lig RMSD is calculated by aligning the ligand on the reference ligand conformation and it indicates the internal fluctuation of the ligand. Along with RMSD, the RMSF (Root Mean Square Fluctuation) was also assessed for each MD run. Protein RMSF shows the fluctuation of protein residues, highlights secondary structure (Pink: α helix; Blue: β strand) and ligand interacting residues marked by green vertical lines. Protein-ligand interactions were also monitored throughout the simulation time. Different types of protein-ligand interactions measured. are H-bond, Hydrophobic interaction, ionic interaction and water bridges. Hydrophobic interaction also includes Π-Cation and Π-Π interactions. The normalized stacked bar charts suggest the fraction of simulation time for which interaction is maintained over the course of the simulation trajectory: for example, a value of 0.6 implies that a specific interaction is maintained for 60% of the simulation time. If a protein residue makes multiple interactions of the same type with ligand then values more than 1.0 are possible.

2.6   *ADME prediction*

ADME: 'absorption, distribution, metabolism, and excretion' properties for selected compounds from supernatural database has been done using Qikprop tool of Schrodinger suite. A star is assigned if the value of the query compound falls beyond the 95% range of similar values for known drugs. Therefore, a greater

number of stars indicates less drug-likeliness of the compound.

3.   Results

3.1   *Sequence Analysis identified COVID-19 NSP1 to be conserved*

Around 10,000 complete NSP1 sequences of SARS-CoV2 were available in the public domain (NCBI as of 14 June 2020) and downloaded. Out of these, 6383 were sequences deposited from USA, 1829 from Australia, 446 from India and196 from Greece. Within India, 202 were sequences deposited from Ahmedabad, 40 from Vadodara and 26 from Gandhinagar. Analysis of Indian sequences between two time points, 15 May 2020 and 10 June 2020, clearly shows NSP1 to be evolving. In the dataset of 15th May, only one residue mutation (S135N) was observed. However, in the dataset corresponding to the second time point, three additional residues were found to be mutated (V38F, D147E, V167A (supplementary figure 1). Mutation analysis of the entire set of around 10,000 sequences shows mutation at multiple residues. Shannon entropy is found to be close to 0 for most of the residues (with a maximum <0.2). This indicates NSP1 is not under huge selection pressure and can be considered as highly conserved till now. In the absence of the available structure of NSP1 of SARS-CoV2, we included SARS NSP1 as a template for modeling. The major reason for this assumption is 100% query coverage and 84.44% sequence similarity of SARS-CoV2 NSP1 with SARS NSP1 (figure 1a). We performed an extensive literature survey to identify a set of key residues, important in

suppressing host gene expression and antiviral signaling which are shown in figure 1b. Most of the residues among this set are found to be conserved between COVID and SARS NSP1.

3.2   *Model for virtual screening generated by homology modelling*

Blastp search of COVID-19 NSP1 sequence with PDB database enabled to identify 2hsx as the best template (with 68% query coverage and 86% identity). N- and C-terminal overhangs in SARS CoV-2 NSP1 have not been considered for modelling. Amino acid variations and key residues, important for function, are marked on the alignment (figure 1). Predicted models, derived using Modeller 9.22, were sorted according to the DOPE score and the top three models were validated using ProSA and SAVES5.0 sever. The best model from the above was chosen for virtual screening (figure 2a).

3.3   *List of potent inhibitors are identified by in silico screening of FDA-approved drugs and Supernatural Database compounds*

Three deep and five shallow ligand binding sites could be recognized on the surface of COVD19-NSP1 protein. Sites were ranked according to their ability to bind various ligands which were depicted as SITEMAP site score and D-Score (please see Methods). We selected Site 1 with a site score of 0.927 and D-score 0.791 among deep sites and site 3 with site score 0.883 and D-score 1.012 among shallow sites for ligand docking

Figure 1. Sequence analysis COVID-19 (Wuhan-Hu-1) Nsp1. Represents alignment between Wuhan-Hu-1 Nsp1 and SARS Nsp1 protein sequence. Red highlights consensus sequences whereas Blue highlights difference in amino-acid sequence. Important residues shown to play role in affecting host gene expression and anti-viral signaling are highlighted in green and pink color. Green highlighting similar residues whereas Pink highlighting residues which are different in COVID-19.
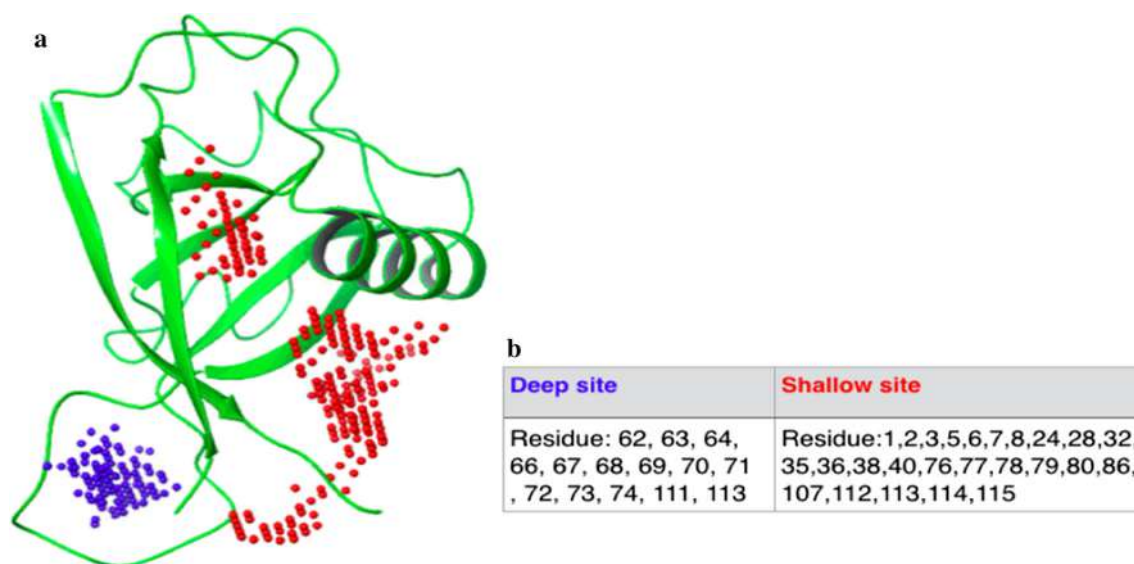
Figure 2. Model of COVID-19 (Wuhan-Hu-1) Nsp1 with Deep and shallow binding site predicted by SiteMap: (**a**) COVID-19 Nsp1 model derived using Modeller 9.22, using 2hsx as a template. Red dot represents Shallow binding site consisting region of alpha-helix and beta-sheets. Blue dots represent deep binding site present in mostly loop region. (**b**) Residues present in deep and shallow binding site respectively. Residue numbers are as per the structural model (Residue 1 of structure is residue 12 in the sequence)

(figure 2a). These sites also contain functionally important residues (figure 1) which showed their biological importance. Selected sites were then used to generate a receptor grid for molecular docking. Molecular docking for each site was carried out using a glide dock program with generated libraries of 2413 FDA-approved drugs and 3,25,287 natural compounds, respectively. The top hits from FDA-approved drug library were ranked according to their XP and MMGBSA scores. We have also considered ligands with well-known anti-viral and anti-inflammatory properties, along with top-ranked ones (entries 15–17 in table 1). The final list of compounds was taken further for the MD simulation run (table 1). The top hits from Supernatural Database compounds were ranked according to their MMGBSA score and were further selected for MD simulation runs. List of top hits, selected based on either binding energy or mode of action, for both deep and shallow binding sites are shown in table 1.

3.4 *MD simulation of protein–ligand complexes*

The best compounds from docking analysis were further subjected to 20 ns of MD simulation to assess the stability of protein–ligand complex. The interactions between protein and ligand were designated as stable if

there was less structural variations and a high percentage of hydrogen bonds or hydrophobic interactions with various residues of the protein at the docked site throughout the course of the simulations. Among the FDA-approved drugs docked at the deep site, Esculin is an example of stable complexes, while Zinc-gluconate is an example of an unstable complex. Figure 3a and supplementary figure 2a show the interaction of Esculin with NSP1 in the docking pose, where Esculin interacts mainly with Arg62, Ser63, Ala68 and His72. MD simulation of NSP1-deep-Esculin complex for 20 ns revealed the stability of the complex as assessed by the RMSD (root mean square deviation) plot. Residues in the secondary structure are expected to have fewer fluctuations than residues in the loop regions and the trend is followed for NSP1 which shows high RMSF between residues 62-76 which form a loop and also interact with Esculin (supplementary figure 2b). Arg62, Ser63, Ala68 and His72 (major interacting residues in the docking pose) interact mainly through H-bond interactions with Esculin. Met74 was also found to interact with Esculin mainly through H-bond (figure 3b). Few other residues interact with Esculin, but with less amount of simulation time. Further details of these interactions are provided in supplementary figure 2c and d.

Results of similarly detailed analysis for all the ligands, as in table 1, are provided in Supplementary

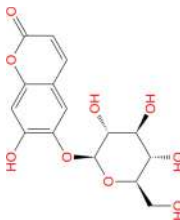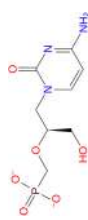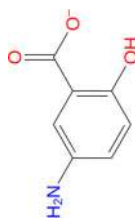**Table 1.** Top-ranking hits identified by the virtual screening and other promising small molecules
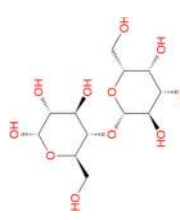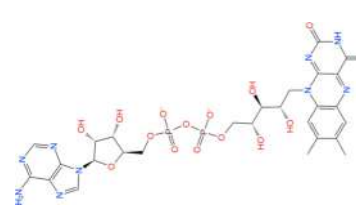
No.	Compound	Structure	Source	Deep binding site			Shallow binding site			Comment
				XP Score	MMGBSA score	MD	XP Score	MMGBSA score	MD	
1	**Esculin**		**Drugbank**	**− 6.883**	**− 29.24**	**Stable**				**Vast-Protective agent, antioxidant and anti-inflammatory** (Wishart *et al.* 2018)
2	**Cidofovir**		**Drugbank**	**− 5.776**	**− 16.53**	**Stable**				**Anti-viral** (Wishart *et al.* 2018)
3	Mesalazine		Drugbank	− 5.146	− 20.74	Not stable				Anti-inflammatory agent (Wishart *et al.* 2018)
4	**Lactose**		**Drugbank**	**− 11.16**	**− 40.72**	**Not stable**				**It is a disaccharide of glucose and galactose. Used as a nutrient supplement** (Wishart *et al.* 2018)
5	FAD		Drugbank	− 7.895	− 52.47	Not stable	− 5.643	− 40.25	Not stable	Used in ophthalmic treatment for vitamin B2 deficiency (Wishart *et al.* 2018)
6	Salmeterol		Drugbank	− 5.69	− 47.8	Not stable				Beta-2 adrenergic receptor agonist. Used in treatment of asthma and COPD (Wishart *et al.* 2018)
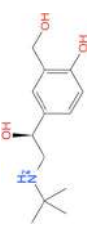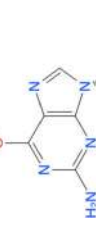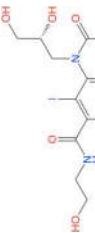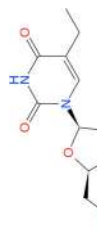
Table 1. (continued)

No.	Compound	Structure	Source	Deep binding site			Shallow binding site			Comment
				XP Score	MMGBSA score	MD	XP Score	MMGBSA score	MD	
7	Zinc gluconate		Drugbank	− 8.525	− 18.19	Not stable				Treating diarrheal episodes in children and reduced duration of common cold (Wishart et al. 2018)
8	Salbutamol		Drugbank	− 5.097	− 40.86	Not stable				A short-acting, beta-2 adrenergic receptor agonist (Wishart et al. 2018)
9	Fenoterol		Drugbank	− 6.543	− 45.18	Not stable				Adrenergic beta-2 receptor agonist (Wishart et al. 2018)
10	Nelarabine		Drugbank	− 6.307	− 36.86	Not stable				Anti-neoplastic agent (Wishart et al. 2018)
11	Ioxilan		Drugbank	− 8.451	− 35.28	Not stable				Tri-iodinated diagnostic contrast agent (Wishart et al. 2018)
12	**Edoxudine**		**Drugbank**	**− 5.941**	**− 33.63**	**Stable**				**Deoxy-thymidine analog and Anti-viral agent** [26]
13	Floxuridine		Drugbank	− 5.125	− 33.27	Not stable				Anti-neoplastic and antimetabolite agent (Wishart et al. 2018)
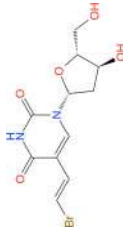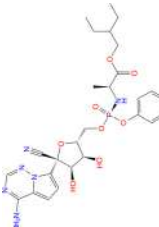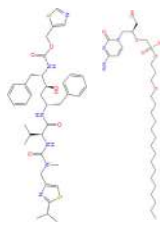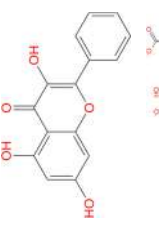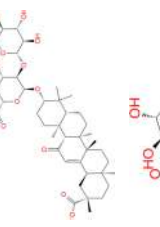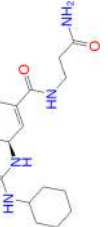
Table 1. (continued)

No.	Compound	Structure	Source	Deep binding site			Shallow binding site			Comment
				XP Score	MMGBSA score	MD	XP Score	MMGBSA score	MD	
14	**Brivudine**		**Drugbank**	**− 5.8**	**− 30.61**	**Stable**				**Anti-viral** (Wishart *et al.* 2018)
15	**Remdesivir**		**Drugbank**	**− 5.795**	**− 40.01**	**Stable**				**Anti-viral agent, a potential treatment for Ebola and shown to be effective against COVID-19** (Wang *et al.* 2020; Wishart *et al.* 2018)
16	Ritonavir		Drugbank	− 2.778	− 27.75	Not stable	− 2.25	− 55.25	Not stable	Anti-viral agent (Wishart *et al.* 2018)
17	Brincidofovir		Drugbank	− 2.024	− 27.75	Not stable				Anti-viral agent (Wishart *et al.* 2018)
18	Galangin		Natural product	− 3.532	− 17.69	Not stable	− 2.278	− 22.99	Not stable	Dietary flavonoid having anticancer properties (Chen *et al.* 2019)
19	**Glycyrrhizic Acid**		**Drugbank**	**− 4.609**	**− 23.59**	**Not stable**	**− 3.643**	**− 27.06**	**Stable**	**Anti-inflammatory, anti-diabetic, anti-oxidant, anti-tumor and anti-viral properties** (Ming and Yin 2013)
20	**SN00003849**		**Super natural database**	**− 6.295**	**− 51.08**	**Stable**				**Plant product**
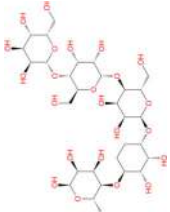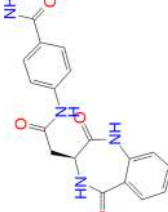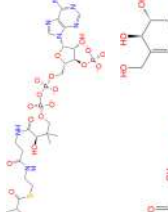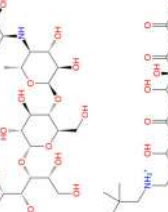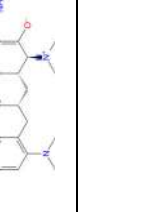
Table 1. (continued)

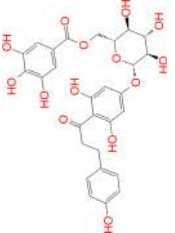| No.	Compound	Structure	Source	Deep binding site			Shallow binding site			Comment
				XP Score	MMGBSA score	MD	XP Score	MMGBSA score	MD	
21	SN00220639		Super natural database	− 12.25	− 50.79	Not stable	− 10.59	− 67.94	Not stable	Plant product
22	SN00103215		Super natural database	− 5.264	− 47.66	Not stable				Plant product
23	**SN00003832**		**Super natural database**	**− 6.308**	**− 47.03**	**Stable**				**Plant product**
24	**SN00216190**		**Super natural database**	**− 7.78**	**− 46.11**	**Stable**				**Plant product**
25	**Acarbose**		**Drugbank**				**− 8.365**	**− 28.35**	**Stable**	**Anti-diabetic** (Wishart *et al.* 2018)
26	Omadacycline		Drugbank				− 6.025	− 35.52	Not stable	Antibiotic (Wishart *et al.* 2018)
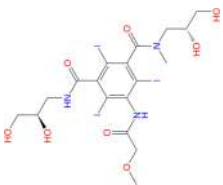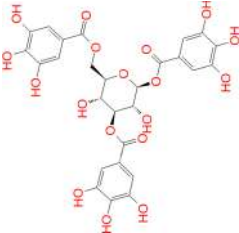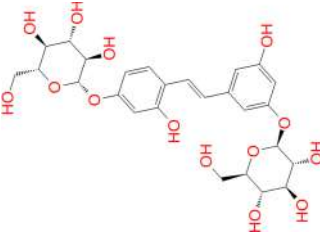
Table 1. (continued)

No.	Compound	Structure	Source	Deep binding site			Shallow binding site			Comment
				XP Score	MMGBSA score	MD	XP Score	MMGBSA score	MD	
27	**Iopromide**		**Drugbank**				**− 4.73**	**− 31.58**	**Stable**	**Used as a contrast agent** (Wishart *et al.* 2018)
28	SN00037405		Super natural database				− 9.177	− 61.45	Not stable	Plant product
29	SN00161170		Super natural database				− 6.676	− 57.94	Not stable	Plant product
30	SN00038342		Super natural database				− 6.066	− 51	Not stable	Plant product

Table 1. (continued)

No.	Compound	Structure	Source	Deep binding site			Shallow binding site			Comment
				XP Score	MMGBSA score	MD	XP Score	MMGBSA score	MD	
31	**SN00156190 (Gingerenone)**		**Super natural database**	**− 4.39**	**− 22.27**	**Not done**	**− 2.578**	**− 36.38**	**Not done**	**Anti-obesity, Anti-inflammatory and antibiotic (Suk et al. 2017; Rampogu et al. 2018)**
32	**SN00002189 (Shogaol)**		**Super natural database**	**− 2.64**	**− 36.18**	**Not done**				**Anticancer, anti-oxidant, antimicrobial, anti-inflammatory anti-allergic and antibiotic (Rampogu et al. 2018; Semwal et al. 2015)**

Molecules marked in **bold** have been discussed in the text.

figures S1–S35. Other promising lead compounds, among the FDA-approved ligands docked at the deep site of NSP1, are Cidofovir (supplementary figure 3), Remdesivir (the drug under investigatory group; supplementary figure 17), Brivudine (supplementary figure 16) and Edoxudine (supplementary figure 14). In the case of FDA-approved drugs docked at the shallow site, acarbose was found to be the most stable ligand. It interacts mainly with Arg32, Leu77 and Asn115 through H-bond and water bridge interactions (supplementary figure 31).

Amongst compounds from SuperNatural database docked at the deep binding site of NSP1, SN00003849 interacts mainly with Arg62, Arg66, Ala68, Gly71, His72 and Met74 (figure 4a). Further, the NSp1-SN00003849 complex was found to be stable, as suggested by RMSD plot of 20ns MD simulation (supplementary figure 25a). Residues interacting with SN00003849 are similar to that of Esculin (supplementary figure 25b). These include Arg62, Arg66, Gly71, His72 and Met74 interacting mainly through H-bond and water bridge interactions (figure 4b). Arg61, Gly71 and Met74 interact with the same atom of SN00003849 for more than 80% of simulation time (supplementary figure 25c). At any point during the simulation, the minimum number of contacts between SN00003849 and NSP1 is more than four, suggesting the strong interaction at the binding site (supplementary figure 25d). SN00003849 also has the highest binding energy as per MM-GBSA calculation (table 1). SN00003849 is structurally similar to terpene/steroid and can be classified as proto and pseudo alkaloids. SN00003832 and SN00216190 also form a stable complex with NSP1 at the deep site (supplementary figures 29 and 30).

For shallow binding sites, none of the compounds in supernaturaldb form complex that are as stable as that for the deep binding site. Natural compounds (entries 18, 19, 32 and 33) are derived from herbal plants, well-known to treat coughs and viral fevers. Another FDA-approved compound is Glycyrrhizic acid which was ranked a bit lower for the deep site, as well as the shallow binding site during docking. The MD simulation was run for deep as well as shallow site complex of NSP1 with Glycyrrhizic acid. Glycyrrhizic acid bound at the shallow site interacts mainly with Arg32, Lys36, Arg113 and Asn115 in the docked pose (figure 5a).

The MD simulation of 20 ns suggested that the complex is stable as per RMSD plot (supplementary figure 24a). NSP1-rmsf plot indicates that the few residues of α helix along with residues at the N and
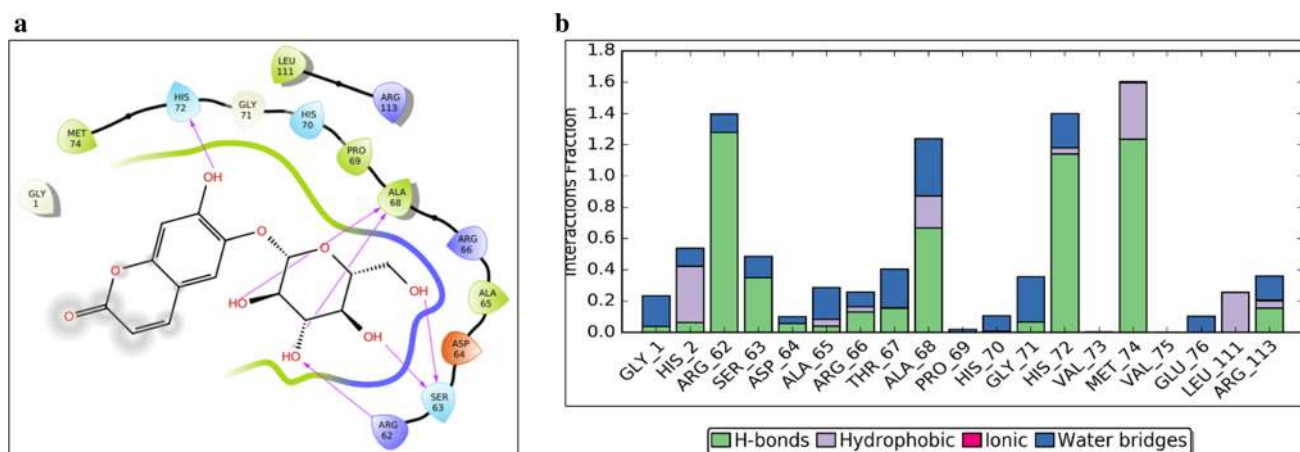
Figure 3. Docking and MD simulation results for NSP1-deep-Esculin. (**a**) Esculin-NSP1 interactions after XP docking. (**b**) Interaction types and Interacting residues of NSP1 with Esculin over simulation time. Normalized stacked bars indicate the fraction of simulation time for which a particular type of interaction was maintained. Values more than 1.0 suggest that the residue forms multiple interactions of the same subtype with ligand (Esculin).

Figure 4. Docking and MD simulation results for NSP1-deep-SN00003849. (**a**) SN00003849-NSP1 interactions after XP docking. (**b**) Interaction types and Interacting residues of NSP1 with SN00003849 over simulation time. Normalized stacked bars indicate the fraction of simulation time for which a particular type of interaction was maintained. Values more than 1.0 suggest that the residue forms multiple interactions of the same subtype with ligand (SN00003849).
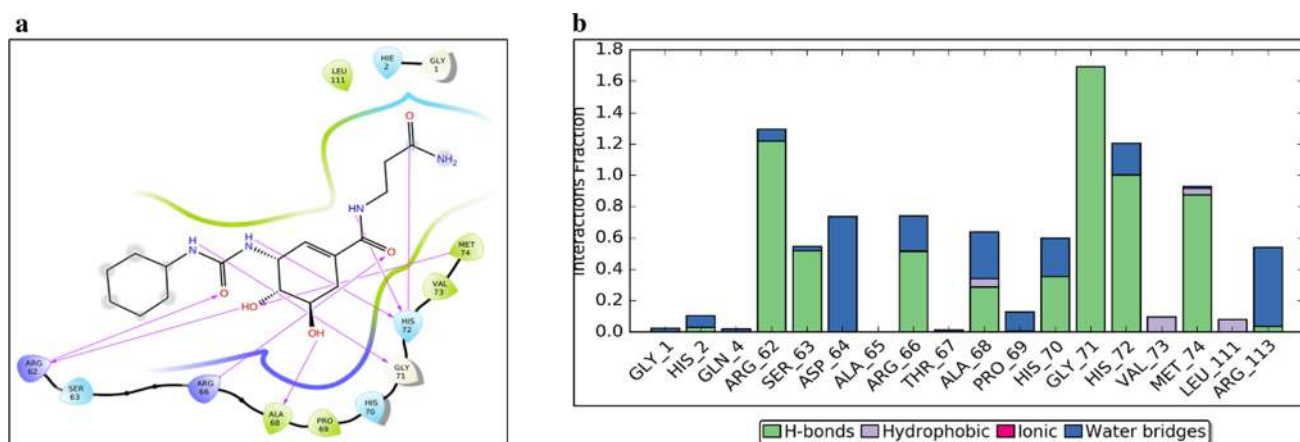
C-termini are also involved in an interaction with Glycyrrhizic acid (supplementary figure 24b). Major interacting residues of NSP1 are the same as those in the docking pose (figure 5b). Atom wise interactions of Glycyrrhizic acid with NSP1 has been shown in figure 24c. Similar to SN00003849, Glycyrrhizic acid also maintains at least 4 contacts with NSP1 over the entire course of simulation time (figure 24d). Glycyrrhizic acid bound at the deep site is not stable (-supplementary figure 23). Interestingly, Glycyrrhizic acid is from the plant Mulethi or Liquorice (also referred to as Yashtimadhura (*Glycyrrhiza glabra*), which is a natural herb for cough and has expectorant properties. It can also reduce infection of the upper

respiratory tract. It may reduce throat irritation and helps cases of a chronic cough.

The ADME related properties for compounds like Gingerenone, Shogaol and SN00103215 follow Lipinski's rule of five, while others violate either one or 3 of the 4 rules of Lipinski's rule of five (supplementary table 1). Qikprop also summarizes the drug-likeness of compounds by comparing the properties of query compounds with known drugs. Gingerenone, Shogaol and SN00103215 are observed to not retain any star (please see Methods) suggesting strong drug-likeliness of these compounds. The water solubility, a key parameter required for absorption and distribution of the compounds, ranges from − 5.101 to 0.261 and it
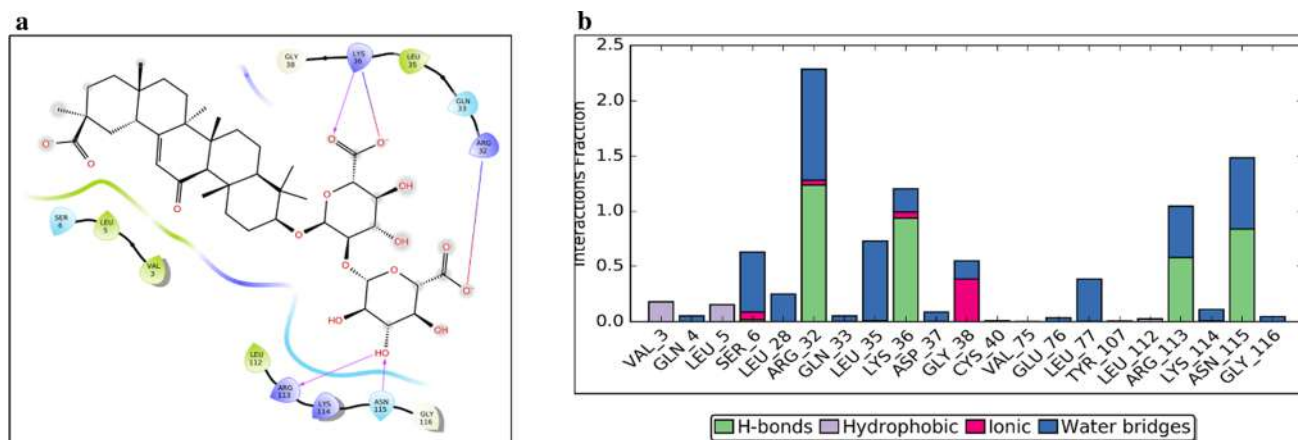
Figure 5. Docking and MD simulation results for NSP1-shallow-Glycyrrhizic acid. (**a**) Glycyrrhizic acid-NSP1 interactions after XP docking. (**b**) Interaction types and Interacting residues of NSP1 with Glycyrrhizic acid over simulation time. Normalized stacked bars indicate the fraction of simulation time for which a particular type of interaction was maintained. Values more than 1.0 suggest that the residue forms multiple interactions of the same subtype with ligand (Glycyrrhizic acid).

falls within the acceptable range. Cell permeability is important for metabolism and it was found that most of the compounds have poorly predicted cell permeability. However, the cell permeability predictions are for non-active transport. Gingerenone and Shogaol also show a high percentage of oral absorption (supplementary table 1).

4. Discussion

COVID-19 outbreak has turned into a pandemic, which makes the identification of new target molecules, repurposing of drugs and designing vaccine an imminent necessity. Since the outbreak, many studies have been conducted along these lines (Chakraborti *et al.* 2020; Gordon *et al.* 2020; Narayanan and Nair 2020; Wu *et al.* 2020a) (Manfredonia *et al.* 2020; Quimque *et al.* 2020). We used NSP1 protein as our target protein. It shows 86 % identity with SARS NSP1. A model of COVID-19 NSP1 was made using SARS NSP1 as a template. Please note: During the submission process for this manuscript, the structure of NSP1 with the ribosome has been solved by another group (Thoms *et al.* 2020). This is not published yet, nor any mention of PDB id submissions. This preprint showed the role of NSP1 in translational shutdown and innate immune evasion). Understanding the genetic diversity of a viral gene is key in understanding evolutionary pressure and add one more dimension to virtual screening (Kasibhatla *et al.* 2020; Somasundaram *et al.* 2020). NSP1 is evolving with key residues being conserved. Virtual screening, against NSP1 protein, suggests a list of FDA-approved drugs and natural compounds against

the deep and shallow binding site on NSP1. Deep and shallow binding sites include functionally important residues such as H81, H83, R124 and R43, K47, E91, R124, K125 respectively (AR Jauregui *et al.* 2013). R124 has shown to be important for NSP1 to interact with viral mRNA 5′-UTR region which prevents viral mRNA from NSP1 mediated mRNA degradation (Kamitani *et al.* 2006) (Note: Residue number in modelled structure starts with 12th residue of the sequence). Docking and MMGBSA scores suggest the binding potential of these compounds towards NSP1. Further, MD simulation of the selected compounds in complex with NSP1 ensures that some of these hits form stable interactions with NSP1.

Esculin, Cidofovir, Edoxudine, Brivudine and Remdesivir were found to form a stable complex with NSP1, among FDA-approved drugs binding at deep site of NSP1. Esculin is a glucoside and naturally occurs in barley, horse chestnut, etc. It is given to improve capillary permeability and fragility and has been reported to inhibit collagenase and hyaluronidase enzymes. This molecule has been shown to have antioxidant and anti-inflammatory activity (Wishart *et al.* 2018). This suggests the ability of esculin to not only inhibit NSP1 activity but also being effective against secondary symptoms such as inflammation. Cidofovir is a known anti-viral agent against CMV infection and acts via inhibition of CMV DNA Polymerase. Edoxudine is a deoxy-thymidine analog shown to be effective against herpes simplex virus type 1 and type 2. It acts as a competitive inhibitor of viral DNA polymerase in its phosphorylated form. Edoxudine is initially phosphorylated by viral thymidine kinas and it is specifically incorporated in viral DNA. Edoxudine

has been discontinued. Brivudine is an organic compound and a pyrimidine 2′-deoxyribonucleosides analog. This is used in the treatment of herpes zoster, results from reactivation of varicella-zoster virus. Remdesivir is proposed as a potential antiviral drug against Ebola (Wishart *et al.* 2018). However, this molecule appears within the Investigational group of DRUGBANK. It is an adenosine-triphosphate analog and has shown effectivity against coronaviruses. A recent publication on COVD-19 treatment, shows it to be a potential drug along with chloroquinone (Wang *et al.* 2020). Remdesivir is an RNA polymerase inhibitor. Hence our study suggests an additional mechanism of action for this drug. An interesting and unexpected molecule among this list is lactose. Lactose is a disaccharide of glucose and galactose and used as nutrient supplement. Derivatives of lactose, 3′-sialyllactose have been shown to have broad-spectrum neutralization activity against avian influenza viruses in chickens (Pandey *et al.* 2018). Further investigation is necessary to check the anti-viral property of lactose against coronavirus.

Acarbose, Iopromide and Glycyrrhizic Acid form stable interactions with the shallow binding site of NSP1. Acarbose is an alpha-glucosidase inhibitor and administered to patients with non-insulin-dependent diabetes mellitus (Wishart *et al.* 2018). As the death rate among COVD-19 patients with diabetes is high, the anti-diabetic nature of acarbose can be highly useful in the treatment regime. Iopromide is a contrast agent, used in radiographic studies. Glycyrrhizic acid is a plant product obtained form Mulethi or Liquorice (also referred as Yashtimadhura (*Glycyrrhiza glabra*)). It has been shown to have anti-inflammatory, anti-diabetic, anti-oxidant, anti-tumor and anti-viral properties (Ming and Yin 2013). These properties suggest Glycyrrhizic acid to be of high importance in COVID-19 treatment.

We next pursued virtual screening against supernaturaldb – a database of 3,25,287 natural small molecules (giving rise to 5,03,604 confirmations). Virtual screening for the shallow site also predicted natural products with high medicinal value such as Gingerenone A (SN00156190) and Shogaol (SN00002189), but with lower docking score (table 1). Gingerenone A has anti-obesity, anti-inflammatory and antibiotic properties (Rampogu *et al.* 2018; Suk *et al.* 2017), whereas Shogaol is anticancer, anti-oxidant, antimicrobial, anti-inflammatory anti-allergic and antibiotic in nature (Rampogu *et al.* 2018; Semwal *et al.* 2015). MD simulation was not performed for the above two because of their lower rank but can be tested

further. Molecules like Galangin, Gingerenone and Shaogaol are reported in high quantities in the medicinal plant, Sitharathai (*Alpinia Officinarum*; a form of ginger, also referred as 'Kulanjan' (Chen *et al.* 2019) which has been used for bronchial infections, as a carminative and recently recognized for its antiviral properties (Pillai and Young 2017). Extracts from herbal plants provide a host of secondary metabolites which could have a combinatorial effect to reduce the viral load, once consumed in the proper manner.

Other hits from supernaturaldb include compounds SN00003849, SN00003832 and SN00216190, which were found to have stable interactions with the deep binding site of NSP1 as suggested by docking and MD simulation. Therefore, along with FDA-approved drugs which will constitute the treatment by repurposing, these new natural compounds can also be tested for their activity against COVID-19.

5. Conclusion

Virtual screening helps in the identification of novel drug candidates and repurposing of known drugs. The current pandemic caused by SARS-Cov2. In order to assist in the development of a cure, we have targeted NSP1 protein of this virus and screened known drugs and natural compounds against it. In this effort, we have identified known antiviral compounds like Remdesivir and Edoxudine. Other drugs, like Esculin and Acarbose which are not antiviral, but are used as anti-inflammatory and antidiabetic (respectively) were also identified. These FDA-approved drugs can be considered as potential candidates for drug repurposing. Natural compounds like Glycyrrhizic acid (entry 19 in table 1) from Liquorice and Galangan, Gingeronone and Shogaol (entries 18, 32 and 33 in table 1) from Sitharathai, were also found to be interacting with NSP1. These compounds can be considered as novel drug candidates against COVID-19. We find these results to be encouraging and hopefully useful immediately to the community and follow-up validation by other researchers.

References

Banerjee P, Erehman J, Gohlke B-O, Wilhelm T, Preissner R, *et al.* 2015 Super Natural II—a database of natural products. *Nucleic Acids Res.* **43** D935–D939

Berman H, Henrick K and Nakamura H 2003 Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **10** 980. *https://doi.org/10.1038/nsb1203-980*

Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, *et al.* 2006 Scalable algorithms for molecular dynamics simulations on commodity clusters. SC '06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing *https://doi.org/10.1145/1188455.1188544*

Brister JR, Ako-Adjei D, Bao Y and Blinkova O 2015 NCBI viral Genomes resource. *Nucleic Acids Res.* **43** D571–D577

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al.* 2009 BLAST+: Architecture and applications. *BMC Bioinform.* **10** 421

Channappanavar R and Perlman S 2017 Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. *Semin. Immunopathol.* **39** 529–539

Chakraborti S, Bheemireddy S and Srinivasan N 2020 Repurposing drugs against main protease of SARS-CoV-2: mechanism based insights supported by available laboratory and clinical data. *https://doi.org/10.26434/chemrxiv.12057846.v2*

Chen CY, Lin CL, Kao CL, Yeh HC, Li HT *et al.* 2019 Secondary Metabolites from the Rhizomes of Alpinia officinarum. *Chem. Nat. Compounds* **55** 1176–1178

Deftereos S, Giannopoulos G, Vrachatis DA, Siasos G, Giotaki SG *et al.* 2020 Colchicine as a potent anti-inflammatory treatment in COVID-19: can we teach an old dog new tricks? *Eur Heart J Cardiovasc. Pharmacother. https://doi.org/10.1093/ehjcvp/pvaa033*

Elbe S and Buckland-Merrett G 2017 Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1** 33–46

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D *et al.* 2006 Comparative protein structure modeling using modeller. *Curr. Protocols Bioinform. https://doi.org/10.1002/0471250953.bi0506s15*

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ *et al.* 2004 Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47** 1739–1749

Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR *et al.* 2006 Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **49** 6177–6196

Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K *et al.* 2020 A SARS-CoV-2-Human protein-protein interaction map reveals drug targets and potential drug-repurposing. *bioRxiv https://doi.org/10.1101/2020.03.22.002386*

Halgren T 2007 New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Design* **69** 146–148

Halgren TA 2009 Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inform. Model.* **49** 377–389

Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL *et al.* 2004 Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **47** 1750–1759

Harder E, Damm W, Maple J, Wu C, Reboul M *et al.* 2016 OPLS3: A force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **12** 281–296

Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP *et al.* 2017 Virus Variation Resource-improved response to emergent viral outbreaks. *Nucleic Acids Res.* **45** D482–D490

Huang C, Lokugamage KG, Rozovics JM, Narayanan K, Semler BL *et al.* 2011 SARS coronavirus nsp1 protein induces template-dependent endonucleolytic cleavage of mRNAs: Viral mRNAs are resistant to nsp1-induced RNA cleavage. *PLoS Pathogens* **7** e1002433

Humphreys DD, Friesner RA and Berne BJ 1994 A multiple-time-step molecular dynamics algorithm for macromolecules. *J. Phys. Chem.* **98** 6885–6892

Jauregui AR, Savalia D, Lowry VK, Farrell CM and Wathelet MG 2013 Identification of residues of SARS-CoV nsp1 that differentially affect inhibition of gene expression and antiviral signaling. *PLoS ONE* **8** 1–11

Kamitani W, Huang C, Narayanan K, Lokugamage KG and Makino S 2009 A two-pronged strategy to suppress host protein synthesis by SARS coronavirus NSP1 protein. *Nat. Struct. Mol. Biol.* **16** 1134–1140

Kamitani W, Narayanan K, Huang C, Lokugamage K, Ikegami T *et al.* 2006 Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proc. Nat. Acad.Sci. USA* **103** 12885–12890

Kasibhatla SM, Kinikar M, Limaye S, Kale MM and Kulkarni-Kale U 2020 Understanding evolution of SARS-CoV-2: A perspective from analysis of genetic diversity of RdRp gene. *J. Med. Virol.* https://doi.org/10.1002/jmv.25909

Laskowski RA, MacArthur MW, Moss DS and Thornton JM 1993 PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26** 283–291

Law AHY, Lee DCW, Cheung BKW, Yim HCH and Lau ASY 2007 Role for nonstructural protein 1 of severe acute respiratory syndrome coronavirus in chemokine dysregulation. *J. Virol.* **81** 416–422

Manfredonia I, Nithin C, Ponce-Salvatierra A, Ghosh P, Wirecki TK *et al.* 2020 Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. *bioRxiv https://doi.org/10.1101/2020.06.15.151647*

Martyna GJ, Klein ML and Tuckerman M 1992 Nos \' e - Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.*

Martyna GJ, Tobias DJ and Klein ML 1994 Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **101** https://doi.org/10.1063/1.467468

Masters PS 2006 The molecular biology of coronaviruses. *Adv. Virus Res.* **65** 193–292

Ming LJ and Yin ACY 2013 Therapeutic effects of glycyrrhizic acid. *Nat. Prod. Commun.* **8(3)** 415–418

Narayanan K, Huang C, Lokugamage K, Kamitani W, Ikegami T *et al.* 2008 Severe Acute Respiratory Syndrome Coronavirus nsp1 suppresses host gene expression, including that of type i interferon, in infected cells. *J. Viro.* **82** 4471–4479

Narayanan K, Ramirez SI, Lokugamage KG and Makino S 2015 Coronavirus nonstructural protein 1: Common and distinct functions in the regulation of host and viral gene expression. *Virus Res.* **202** 89–100

Narayanan N and Nair DT 2020 Vitamin B12 may inhibit RNA-dependent-RNA polymerase activity of nsp12 from the SARS-CoV-2 virus. *Preprints* https://doi.org/10.20944/preprints202003.0347.v1

Pandey RP, Kim DH, Woo J, Song J, Jang SH *et al.* 2018 Broad-spectrum neutralization of avian influenza viruses by sialylated human milk oligosaccharides: In vivo assessment of 3′-sialyllactose against H9N2 in chickens. *Sci. Rep.* **8** 2563

Pfefferle S, Schpf J, Kgl M, Friedel CC, Mller MA *et al.* 2011 The SARS-Coronavirus-host interactome: Identification of cyclophilins as target for pan-Coronavirus inhibitors. *PLoS Pathogens* **7** https://doi.org/10.1371/journal.ppat.1002331

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB *et al.* 2012 ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40** D593–D598

Pillai MK and Young DJ 2017 Therapeutic potential of *Alpinia officinarum*. *Mini Rev. Med. Chem.* **18** 1220–1232

Piplani S, Singh PK, Winkler DA and Petrovsky N 2020 In silico comparison of spike protein-ACE2 binding affinities across species; significance for the possible origin of the SARS-CoV-2 virus. *arXiv: 2005.06199*

Quimque MTJ, Notarte KIR, Fernandez RAT, Mendoza MAO, Liman RAD *et al.* 2020 Virtual screening-driven drug discovery of SARS-COV2 enzyme inhibitors targeting viral attachment, replication, post-translational modification and host immunity evasion infection mechanisms. *J. Biomol. Struct. Dyn.* 10.1080/07391102.2020.1776639

Rampogu S, Baek A, Gajula RG, Zeb A, Bavi RS *et al.* 2018 Ginger (*Zingiber officinale*) phytochemicals-gingerenone-A and shogaol inhibit SaHPPK: Molecular docking,

molecular dynamics simulations and in vitro approaches. *Ann. Clin. Microbiol. Antimicrob.* **17** 16

Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W 2013 Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design* **27** 221–234

Semwal RB, Semwal DK, Combrinck S and Viljoen AM 2015 Gingerols and shogaols: Important nutraceutical principles from ginger. *Phytochemistry* **117** 554–568

Shannon CE 1948 A mathematical theory of communication. *The Bell System Technical Journal* **27** 379–423

Shen Z, Wang G, Yang Y, Shi J, Fang L *et al.* 2019 A conserved region of nonstructural protein 1 from alphacoronaviruses inhibits host gene expression and is critical for viral virulence. *Journal of Biological Chemistry* **294** 13606–13618

Somasundaram K, Mondal M and Lawarde A 2020 Genomics of Indian SARS-CoV-2: Implications in genetic diversity, possible origin and spread of virus. https://doi.org/10.1101/2020.04.25.20079475

Suk S, Kwon GT, Lee E, Jang WJ, Yang H *et al.* 2017 Gingerenone A, a polyphenol present in ginger, suppresses obesity and adipose tissue inflammation in high-fat diet-fed mice. *Mol. Nutr. Food Res.* **61** https://doi.org/10.1002/mnfr.201700139

Tanaka T, Kamitani W, DeDiego ML, Enjuanes L and Matsuura Y 2012 Severe Acute Respiratory Syndrome Coronavirus nsp1 facilitates efficient propagation in cells through a specific translational shutoff of host mRNA. *J. Virol.* **86** 11128–11137

Thoms M, Buschauer R, Ameismeier M, Koepke L, Denk T *et al.* 2020 Structural basis for translational shutdown and immune evasion by the NSP1 protein of SARS-CoV-2. *bioRxiv* https://doi.org/10.1101/2020.05.18.102467

Wang M, Cao R, Zhang L, Yang X, Liu J *et al.* 2020 Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* **30** 269–271

Wiederstein M and Sippl MJ 2007 ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35** W407–W410

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A *et al.* 2018 DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46** D1074–D1082

Wong CK, Lam CWK, Wu AKL, Ip WK, Lee NLS *et al.* 2004 Plasma inflammatory cytokines and chemokines in severe acute respiratory syndrome. *Clin. Exp. Immunol.* **136** 95–103

Wu C, Liu Y, Yang Y, Zhang P, Zhong W *et al.* 2020a Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods.

Acta Pharmaceut. Sinica B https://doi.org/10.1016/j.apsb.2020.02.008

Wu F, Zhao S, Yu B, Chen Y-M, Wang W, *et al.* 2020b A new coronavirus associated with human respiratory disease in China. *Nature* **579** 265–269

Zst R, Cervantes-Barragn L, Kuri T, Blakqori G, Weber F, *et al.* 2007 Coronavirus non-structural protein 1 is a major pathogenicity factor: Implications for the rational design of coronavirus vaccines. *PLoS Pathogens* **3** e109

Corresponding editor: Sᴦᴇᴇɴɪᴠᴀs Cʜᴀᴠᴀʟɪ