COMPUTATIONAL APPROACHES TO UNDERSTAND MOLECULAR MECHANISMS OF SARS-COV-2

EDITED BY: Arvind Ramanathan, S. Chakra Chennubhotla,
Nadine Kabbani, James Leland Olds and Shantenu Jha

PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-545-6
DOI 10.3389/978-2-88974-545-6

COMPUTATIONAL APPROACHES TO UNDERSTAND MOLECULAR MECHANISMS OF SARS-COV-2

Topic Editors:
Arvind Ramanathan, Argonne National Laboratory (DOE), United States
S. Chakra Chennubhotla, University of Pittsburgh, United States
Nadine Kabbani, George Mason University, United States
James Leland Olds, George Mason University, United States
Shantenu Jha, Brookhaven National Laboratory (DOE), United States

Citation: Ramanathan, A., Chennubhotla, S. C., Kabbani, N., Olds, J. L., Jha, S., eds. (2022). Computational Approaches to Understand Molecular Mechanisms of SARS-CoV-2. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-545-6

Table of Contents

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 16 October 2020
doi: 10.3389/fmolb.2020.583523

Investigation of the Effect of Temperature on the Structure of SARS-CoV-2 Spike Protein by Molecular Dynamics Simulations

Soumya Lipsa Rath[1]* *and Kishant Kumar*[2]

[1] Department of Biotechnology, National Institute of Technology Warangal, Warangal, India, [2] Department of Chemical Engineering, National Institute of Technology Warangal, Warangal, India

OPEN ACCESS

Edited by:
Arvind Ramanathan,
Argonne National Laboratory (DOE),
United States

Reviewed by:
Debsindhu Bhowmik,
Oak Ridge National Laboratory (DOE),
United States
Christopher Stanley,
Oak Ridge National Laboratory (DOE),
United States

Correspondence:
Soumya Lipsa Rath
slrath@nitw.ac.in

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *15 July 2020*
Accepted: *22 September 2020*
Published: *16 October 2020*

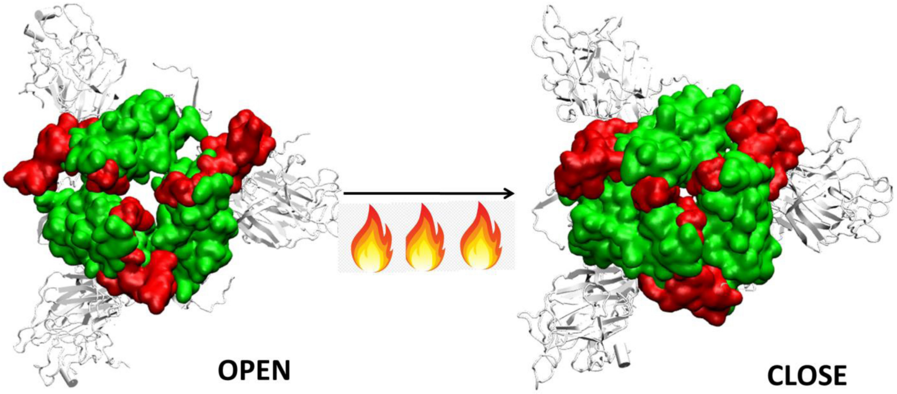
Citation:
Rath SL and Kumar K (2020)
Investigation of the Effect
of Temperature on the Structure
of SARS-CoV-2 Spike Protein by
Molecular Dynamics Simulations.
Front. Mol. Biosci. 7:583523.
doi: 10.3389/fmolb.2020.583523

Statistical and epidemiological data imply temperature sensitivity of the SARS-CoV-2 coronavirus. However, the molecular level understanding of the virus structure at different temperature is still not clear. Spike protein is the outermost structural protein of the SARS-CoV-2 virus which interacts with the Angiotensin Converting Enzyme 2 (ACE2), a human receptor, and enters the respiratory system. In this study, we performed an all atom molecular dynamics simulation to study the effect of temperature on the structure of the Spike protein. After 200 ns of simulation at different temperatures, we came across some interesting phenomena exhibited by the protein. We found that the solvent exposed domain of Spike protein, namely S1, is more mobile than the transmembrane domain, S2. Structural studies implied the presence of several charged residues on the surface of N-terminal Domain of S1 which are optimally oriented at 10–30°C. Bioinformatics analyses indicated that it is capable of binding to other human receptors and should not be disregarded. Additionally, we found that receptor binding motif (RBM), present on the receptor binding domain (RBD) of S1, begins to close around temperature of 40°C and attains a completely closed conformation at 50°C. We also found that the presence of glycan moieties did not influence the observed protein dynamics. Nevertheless, the closed conformation disables its ability to bind to ACE2, due to the burying of its receptor binding residues. Our results clearly show that there are active and inactive states of the protein at different temperatures. This would not only prove beneficial for understanding the fundamental nature of the virus, but would be also useful in the development of vaccines and therapeutics.

Keywords: structural protein, receptor binding motif, N-terminal domain, closed conformation, temperature-sensitive

INTRODUCTION

Severe Acute Respiratory Syndrome Coronavirus 2 or SARS-COV-2, attacks the cells of the human respiratory system. Recent studies have found that the virus also interacts with the cells of the digestive system, renal system, liver, pancreas, eyes and brain (Gordon et al., 2020). It is known to cause severe sickness and is fatal in many cases (World Health Organization [WHO], 2020). It is believed that the virus originated in bats, which act as the natural reservoir;



GRAPHICAL ABSTRACT | Closure of the receptor binding motif of SARS-COV-2 Spike protein at high temperatures.

subsequently it got transmitted to human. It then gradually spread across almost all the nations through aerial transmission resulting in one of the worst known global pandemic of this century (MacKenzie and Smith, 2020).

SARS-COV-2 is one of the seven forms of coronaviruses that affect the human population. The other known coronaviruses include HCoV-229E, HCoV-OC43, SARS-CoV, HCoV-NL63, HCoV-HKU1, and MERS-CoV (Gao et al., 2016; Su et al., 2016). Their infection varies from common cold to SARS, MERS or Covid19 (Su et al., 2016). These viruses have been observed to affect the human population predominantly during a particular season. For instance, the 2002 SARS infections began during the cold winters of November and after 8 months, the number of reported cases became almost negligible (Su et al., 2016). Statistics show that countries with hot and humid weather conditions had lesser number of infectious cases of SARS (Chan et al., 2011). However, MERS-COV, which was identified in Middle East regions, affected individuals during the summer (Su et al., 2016). Thus, the disease epidemiology suggests that the virus is found to be prominent in certain climatic conditions only.

The viability of SARS-COV-2 was measured on different surfaces by Chin et al. (2020), who found that the virus droplets survived at 4°C but quickly deactivated at elevated temperatures of 50°C. Smooth surfaces, plastics and iron show greater viability of the virus compared to that of paper, tissue, wood or cloth. Surgical masks had detectable viruses even on 7th day (Casanova et al., 2010; Van Doremalen et al., 2020). Soaps and disinfectants which disintegrate the virus membrane and structural proteins are a potent example of how the modulation of atmospheric conditions can affect the virus viability. Statistical reports by Cai et al. (2007), and several others had shown that tropical countries like Malaysia, Indonesia or Thailand with high temperature and high relative humidity did not have major community outbreaks of SARS (Tan et al., 2005; Chan et al., 2011). Although viruses cannot be killed like bacteria by autoclaving, temperature sensitivity of virus have been reported several times in the past. Seasonal Rhinoviruses could not replicate at 37°C, whereas 33–35°C is ideal for their survival in nasal cavity (Foxman et al., 2015). Influenza was found to be effective at a temperature around 37°C, whereas higher temperatures of 41°C resulted in clumping of viruses on cell surfaces (Ishida et al., 2002; Pelletier et al., 2011; Lowen and Steel, 2014). Similarly, the viability of SARS virus that persisted for 5 days at temperatures ranging between 22–25°C and 40–50% humidity, was lost when the temperature was raised to 38°C and 95% humidity (Chan et al., 2011).

When the virus is exposed to different temperature conditions, the initial interactions of the atmosphere occur with the structural proteins. There are four major structural proteins present on the virus, the Spike glycoprotein, the Envelope protein, the Membrane protein and the Nucleocapsid. Each of the proteins performs specific functions in receptor binding, viral assembly and genome release (Astuti and Ysrafil, 2020). One of the first and largest structural proteins of the Coronavirus is the Spike glycoprotein (Li, 2016). The protein exists as a homotrimer where each monomer consists of 1,273 amino acid residues (**Figure 1**) and is intertwined with each other. Each monomer has two domains, namely S1 and S2 (Walls et al., 2020). The S1 and S2 domains are cleaved at a furin site by a host cell protease (Belouzard et al., 2009; Walls et al., 2020). The S1 domain lies predominantly above the lipid bilayer. The S2 domain, which is a class I transmembrane domain, travels across the bilayer and ends toward the inner side of the lipid membrane (Walls et al., 2020). **Figure 1** shows the two domains of the Spike glycoprotein.

The S1 domain comprises of mostly beta pleated sheets. It can be further classified into Receptor Binding domain (RBD) and N-terminal Domain (NTD). The RBD binds to Angiotensin Converting Enzyme 2 (ACE 2) on the host cells (Wang et al., 2020). It lies on the top of the complex, where around 14 residues from the RBD domain bind to the ACE2 receptor on the host protein (Yuan et al., 2017; Lan et al., 2020). The NTD is the outermost domain that is relatively more exposed and lies on the three sides giving a triangular shape to the protein when viewed from top (**Figure 1**). The NTD has a galectin fold and is known to bind to the sugar moieties (Yuan et al., 2017). The S2 domain on the other hand is a transmembrane region with strong interchain bonding between the residues. It is mostly α-helical

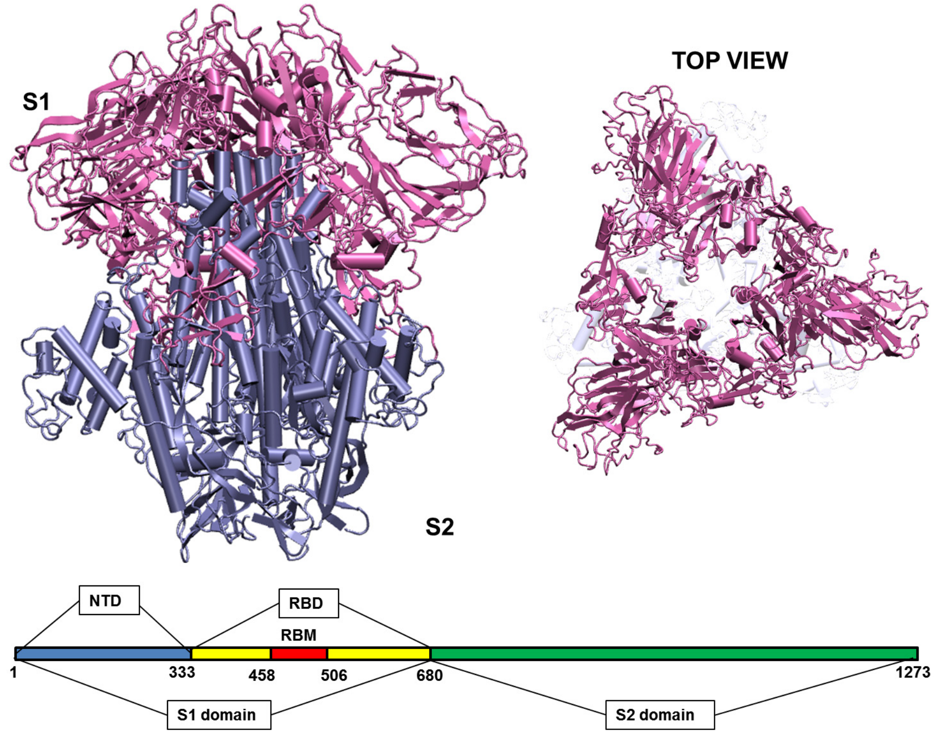


FIGURE 1 | Structure of the Spike glycoprotein. The Spike glycoprotein, S1 and S2 domains, in the absence of glycan residues and lipid bilayer. The S1 domain is shown in pink and S2 in iceblue color. The top view of the protein shows a triangular arrangement of the S1 domain. Below the structure is a schematic showing the location of important regions on the S1 and S2 domains of the protein. The abbreviations in the S1 and S2 domains are: – NTD, N-terminal domain; RBD, receptor binding domain; RBM, receptor binding motif. The starting and ending residues are numbered.

and forms a triangle when viewed from bottom, though there is no overlapping of the top and bottom triangles.

Temperature is a very significant variable parameter for proteins because proteins respond differently in high and low temperature conditions. Many proteins have high thermal stability while others can unfold or even denature at high temperatures (Dong et al., 2018; Julió Plana et al., 2019). Experimental studies by Xiong et al. (2020), where they generated a mutated Spike protein that exhibited high level of thermal stability. Further studies had shown that the activity of the SARS-CoV-2 reduced significantly at high temperatures, when compared to that of SARS-CoV. It was concluded that the comparatively lower energy barrier of SARS-CoV-2 would result in higher transmission rate of the virus (Ou et al., 2020). MD studies also spoke about the temperature sensitivity of RBD of the SARS-CoV-2 compared to its predecessor SARS-CoV (He et al., 2020). During November, 2019, when the first outbreak of Covid19 was reported, the temperature in Wuhan, China was around 17°C in the morning and 8°C at night. Tropical countries such as India, where a large number of cases still persist, had over 40°C of temperature (Bukhari and Jameel, 2020; Wu et al., 2020). Although statistical and experimental evidence show that temperature influences the activity and virulence of the virus, we still lack the understanding of the molecular level changes that are taking place in the virus due to the different weather conditions. Till date, there is no concrete evidence on whether atmospheric conditions actually influence the structure of the virus.

Here, by using all atom molecular dynamics (MD) simulations we explore the dynamics of the Spike glycoprotein of SARS-COV-2 at different temperatures. This is the first molecular study on the environmental influence on the protein structure. Results suggest that S1 domain is more flexible than S2. In the S1 domain, we observed the sensitivity of the receptor binding motif to different temperatures. We also found that the N-terminal domain of the protein has the potential of binding to different human receptors. The study will not only help us in understanding the nature of the virus but is also useful to design effective therapeutic strategies.

MATERIALS AND METHODS

The complete model of the Spike glycoprotein of SARS-COV-2 was obtained from Zhang lab (GenBank: QHD43416.1). This model was considered because it had modeled the missing 871 residues that were absent in the crystal structure (PDB ID 6VXX). It had a Template Modeling (TM) score of 0.6 (Xu and Zhang, 2010). The initial Root Mean Square Deviation (RMSD) between the model and the closed crystal structure of Spike glycoprotein (PDB ID 6VXX) was found to be 1.54 Å. The model was devoid of N-acetyl glycosamine (NAG) glycan residues and consisted of the glycoprotein trimer where each monomer had amino acids ranging from 1 to 1,273. The structure was initially solvated with a TIP3P water box having a cubic box of size 17.9 × 17.9 × 17.9 nm and 569,293 atoms with water and ions (Jorgensen et al., 1983). The minimum distance between the protein and the edge of the water box was fixed at 13 Å. Particle-Mesh Ewald (PME) method was used for electrostatic interactions using a grids pacing of 0.16 nm and a 1.0 nm cutoff. After energy minimization and equilibration, by maintaining harmonic restraints on the protein heavy atoms, the system was heated to 300 K in a canonical ensemble. The harmonic restraints were gradually reduced to zero and solvent density was adjusted under isobaric and isothermal conditions at 1 atm and 300 K. This was followed by 500 ps NVT and 500 ps NPT equilibration with harmonic restraints of 1,000 kJ mol^{-1} nm^{-2} on the heavy atoms. Production run for all the systems was carried out for 200 ns till it reached a stable RMSD. All simulations were carried out in GROMACS (2020) with AMBERff99SB-ILDN force field for proteins (Lindorff-Larsen et al., 2010; GROMACS, 2020). The long-range electrostatic interactions were treated by using Particle-Mesh Ewald sum and SHAKE was used to constrain all bonds involving hydrogen atoms. After equilibration, systems were heated or cooled at different temperatures (**Supplementary Table S1**) and simulated for 200ns. All analyses were carried out using Gromacs analysis tools (Lindorff-Larsen et al., 2010). Protein Blast was used to search similar sequences in the human proteome. The Blast Tree View widget helped us generate the phylogenetic tree which is a simple distance based clustering of the sequences based on pairwise alignment results of Blast relative to the query sequence (Sayers et al., 2009). VMD was used for visualization of results and generation of figures (Humphrey et al., 1996). Principal component analysis was carried out in Gromacs. Pymol was used for generation of porcupine plots (The PyMOL Molecular Graphics System, 2019). We used the glycan bound Spike protein deposited by Woo et al. (2020) in CHARMM GUI as starting structures, for understanding the difference in dynamics of Spike protein in the presence of carbohydrates (Jo et al., 2008).

RESULTS AND DISCUSSION

The crystal structure of the Spike glycoprotein (PDB: 6VXX) was found to have 871 missing residues. Thus, for our study we considered the complete model of the trimeric Spike protein generated by Xu and Zhang (2010) and had a Template modeling score of 0.6. The model was devoid of N-acetyl glucosamine (NAG) sugar moieties which are known to bind and stabilize the protein. The envelope lipid bilayer was not considered in the work to avoid large system size in atomistic simulations. After initial minimization and equilibration, we generated five different systems having temperatures ranging from 10 to 50°C at an interval of 10 degrees. This was done to maintain the uniformity of the simulations, where temperature was the only variable that was different. It should be noted that when a temperature is raised, the electronic distribution of the atoms undergo change. However, classical force fields like AMBER-ff99SB-ILDN are based on certain approximations and the changes in temperature in a particular force field do not impact the results very significantly (Lindorff-Larsen et al., 2013). In addition, a temperature of 70°C was also imposed on the system to observe any possible deformation in the structure of spike protein, although this high temperature is not realistic to imitate the environmental condition (**Supplementary Table S1**). Production run for 200 ns was carried out in isothermal isobaric (NPT) ensemble. To understand the impact of glycans on the dynamics, subsequently three additional simulations were run at 10, 30, and 50°C.

Spike Glycoproteins Are Sensitive to Temperature

After performing 200 ns of classical Molecular dynamics simulations, the root mean square deviation (RMSD) of the trajectory, with respect to the starting structure, was calculated to check if the systems have attained stability. **Supplementary Figure S1** shows the complete RMSD of all the systems at different temperatures. It can be seen that the stability was attained within the first 50 ns of the simulation time, thus, indicating that the systems are well equilibrated. The RMS values lie between 0.6 and 0.7 nm for all the systems with an exception at 40°C where a marginally higher RMSD was seen after 100 ns of simulation time. At temperatures 20 and 30°C, a small rise in RMSD curves after 100 ns of simulation time was observed. This implies that the Spike protein was more stable at temperatures 10 and 50°C. The overall energy of the systems were monitored and it was also found to have attained stability (data not shown).

Since, the protein comprises of two distinct domains S1 and S2, we checked the RMSD of S1 and S2 domains individually, with respect to the starting structure, to understand the cause for higher RMSD values observed at 20, 30, and 40°C (**Figure 2**). The RMS values of S1 domain at 20, 30, and 40°C were found to be around 0.7 nm, nearly 0.5 nm more than simulations at 10 and 50°C, respectively. A similar trend was observed in the RMSD of S2 domain, but, the difference in values was only 0.15 nm. Although, in this study, we haven't considered the bilayer lipid membrane of the SARS-COV-2 envelope inside which the Spike glycoprotein resides, the S2 domain shows remarkable stability in its RMSD values (**Figure 2**). The stability of the S2 domain can be conferred to the strong interchain interactions among the highly α-helical S2 domain.

Since the Spike protein is a homotrimer, the S1 of individual domains was checked to account for the difference in fluctuations. **Figures 2C–E** shows the RMSD of S1 domain of chains A, B and C at different temperatures. The Spike glycoprotein comprises of homotrimeric chain. Hence, to nullify the signal to noise ratio of individual chains, we also plotted average RMSD of the three chains at different temperatures (**Figure 2F**). The figure clearly shows that at 30 and 40°C, the RMSD is higher (∼0.5 nm) when compared to the average RMSD at 10, 20, and 50°C (∼0.42 nm). The above data indicates that the protein chains, especially the S1 domains are quite flexible around the temperatures of 30–40°C in comparison to low temperatures of 10°C or high 50°C of simulation temperature. Irrespective of the presence of the bilayer

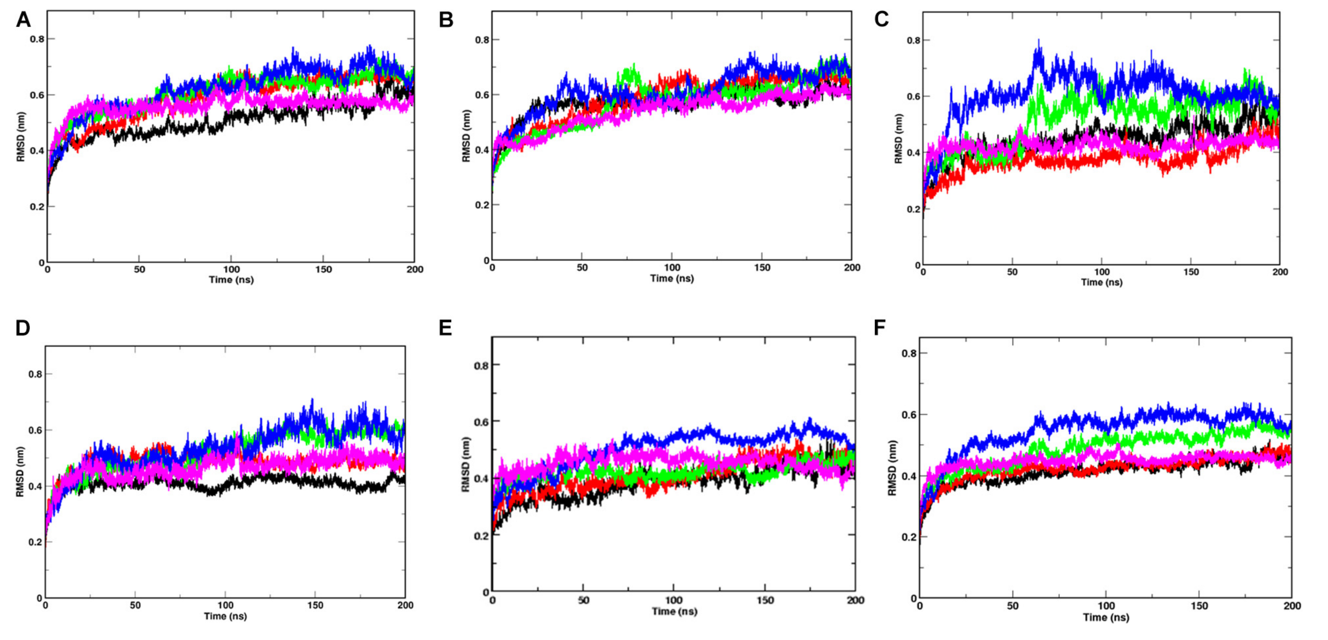


FIGURE 2 | Temperature sensitivity of the S1 domain of Spike protein. RMSDs of Spike glycoprotein **(A)** S1 and **(B)** S2 domain showing stability of the S2 chains. The differential fluctuations of chains **(C)** A, **(D)** B, **(E)** C, and the average RMSD of the three chains **(F)** of S1 domain at 10°C (black), 20°C (red), 30°C (green), 40°C (blue), and 50°C (magenta) implying effect of temperature of the chain stability.

membrane, at different temperature conditions, the stalk of the Spike protein remains stable.

Domain Flexibility of S1 Is More Pronounced

In order to identify the region on the Spike protein that causes the deviations in RMSDs, we plotted the root mean square fluctuation (RMSF) of CA atoms of both S1 and S2 domains separately (**Figure 3**) at different temperatures. Each plot shows the RMSF of each individual chain at different temperatures. The RMSF of individual chains of S1 domain at different temperatures show that the residues ranging from 1 to 333 which constitute the N-Terminal Domain (NTD) of S1, show greater fluctuations compared to the Receptor binding Domain (RBD) ranging from residues 334 to 680. In the NTD, three distinct peaks could be seen, viz:- residues 85–90, 100–200, and 240–260. The first peak in the NTD was observed around residues 85–90 (β4–β5), which is a loop directed inwards to the S2 domain (**Supplementary Figure S2**). The peak was found to be highest in chain A at 40°C (∼0.8 nm), however, at other temperatures all the chains have approximately 0.5 nm RMS fluctuation of its CA atoms. The residues 100–200 constitute the solvent exposed β sheet (β6–β12) of the NTD of S1 domain (**Supplementary Figure S2**). The crystal structure (PDB: 6VXX) had shown as many as three glycosylated groups adjacent to this region of the protein (**Supplementary Figure S3** and Berman et al., 2000). The residues 240–260 are solvent exposed loop around β14–β15. No glycan binding sites were observed in the crystal structure. The RBD domain consists of a receptor binding motif (RBM) ranging from residues 458 to 506 that show flexibility in all the systems. The lowest flexibility was observed at 10°C. At 30°C, the peaks were found for a wider range of residues. This indicates differential flexibility of the RBM at different temperatures. Since, the RBM is involved directly in binding to the ACE2 human receptor; its altered behavior at different temperatures would affect the protein-protein interaction. However, the average of the root mean square fluctuation of the three chains at different temperatures exhibit similar level of fluctuations (**Supplementary Figure S2B**). This indicates that although, movements of individual chains vary, the overall protein structure does not undergo major changes at different temperatures.

The RMSF of S2 domain on the other hand shows marked stability compared to domain S1 (**Supplementary Figure S4**). This is in good agreement to our earlier observations of the RMSD of the S2 domain. Since it is a triple helical coil, the coiled-coil motif of the S2 domain which is further supported by three shorter helices supports domain stability (Walls et al., 2016). However, the C-terminal residues 1,125–1,273 show greater flexibility compared to the rest of the domain. It should be noted that the C-terminal region of the Spike glycoprotein is exposed toward the inner side of the envelope bilayer and does not participate in the interchain interactions. It also has a more relaxed packing compared to the rest of the S2 (Berman et al., 2000; Guillén et al., 2005).

NTD of the Spike Protein Could Act as a Receptor Binding Site

The NTD is relatively more exposed to solvents and more susceptible to external environmental conditions. However,

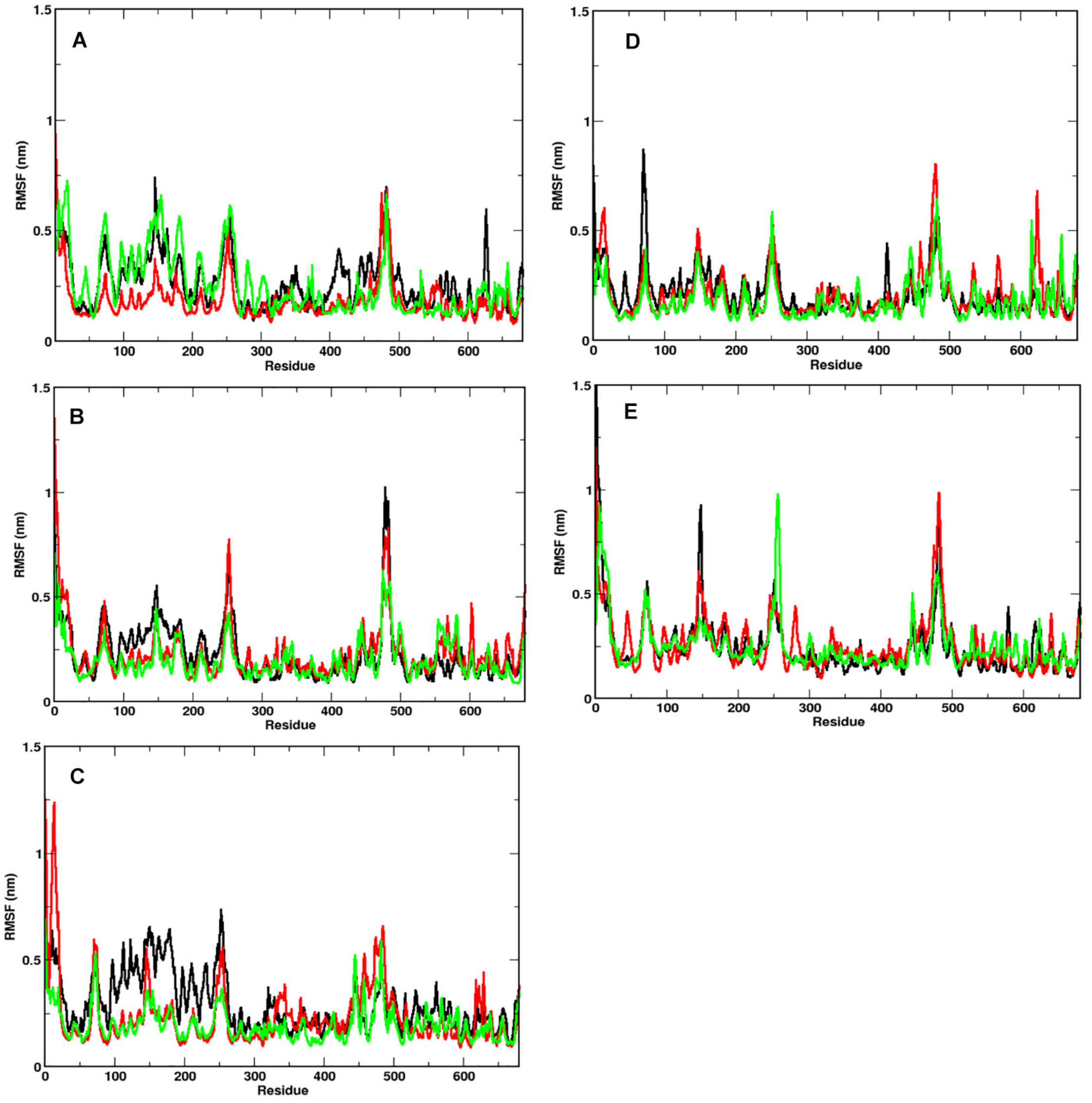


FIGURE 3 | Fluctuation of CA of individual chains at different temperatures. RMSF of the CA atoms of the S1 domain for chains A (in black), B (in red), and C (in green) is shown. At temperatures **(A)** 10°C, **(B)** 20°C, **(C)** 30°C, **(D)** 40°C, and **(E)** 50°C, N-terminal domain (residues 1–333) have higher mobility than the receptor binding domain (residues 334–680).

unlike RBD, the NTD doesn't have a defined open or closed conformation. The coronavirus NTD is composed of three layered beta-sheet sandwich with 7, 3 and 6 antiparallel β strands in each layer making it a total of 16 beta stranded sheet with 5 prominent β hairpin loops (**Supplementary Figure S5**). The crystal structures of Mouse Hepatitis Coronavirus (MHC) Spike protein and its receptor shows that the β1 and β6 of the NTD are the binding motif for CECAM1a protein (Shang et al., 2020). However, unlike the MHC NTD, the arrangement of strands in SARS-CoV-2 is in opposite direction. The upper layer of the beta sandwich is composed of beta strands β4, β6, β7, β8, β9, β10, β14 (**Supplementary Figure S4**). The three prominent regions which are exposed to the solvent and capable of interacting with potential receptors are regions N-terminal β strand, β8–β9, β9–β10, and β14–β15 loop.

Comparison of the NTD at different temperatures (**Figure 4**) show differential arrangement of the solvent exposed loops. The loops are formed by residues from N-terminal β strand, β8–β9, β9–β10, and β14–β15. The time averaged conformation of the loops after 200 ns of simulation show that the loops are oriented close to each other at temperatures 10–30°C, however at 40 and 50°C, they move farther away from each other. Moreover, comparison between Bovine coronavirus and Bovine hemagglutinin-esterase enzyme indicated close evolutionary link between the virus and the host proteins, which could facilitate attachment in the host cells (Li et al., 2013). Since, there was

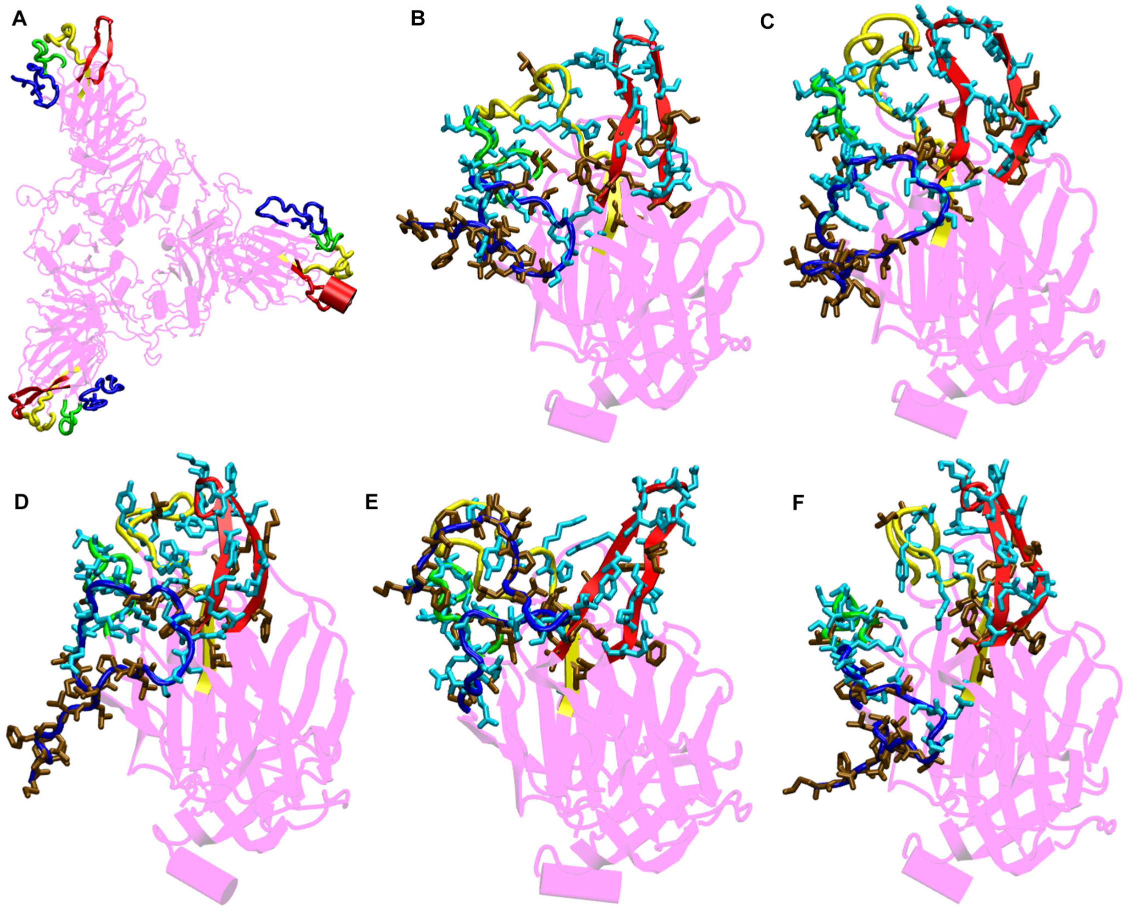


FIGURE 4 | Structures of the N-terminal domain of Spike protein after 200 ns of simulation showing the relative orientation of solvent exposed loops. **(A)** The solvent exposed loops of NTD; the N-terminal β strand, β8–β9, β9–β10, and β14–β15 are shown in red, blue, green and yellow colors, respectively. Time-averaged conformation of N- terminal domain of SARS-CoV-2 Spike protein at, **(B)** 10°C, **(C)** 20°C, **(D)** 30°C, **(E)** 40°C, and **(F)** 50°C showing the relative orientation of the polar and hydrophobic residues. The residues are shown in licorice. Polar resides are colored in light blue and hydrophobic in brown colors, respectively.

a similarity of NTD with the Ehprin A proteins (that binds to the Ephrin A receptors) we compared the residues involved in protein-protein interaction in the crystal structure of the human EphA4 ectodomain in complex with human Ephrin A5 for comparison (PDB ID: 4BKA) (**Supplementary Figure S6**). There are three salt bridges and seven hydrogen bonds between the Ephrin protein and its receptor. Moreover, it can be clearly seen that the NTD loops host a large number of polar residues (**Figure 4**). These residues form a stable motif at temperatures 10–30°C, primarily due to the stability between the loops. At 40 and 50°C, hydrophobic patch from N-terminal β strand is exposed toward the solvent. The polar residues from β9–β10 to β14–β15 move away from the N-terminal β strand and the β8–β9 loop, reducing the possibility of protein-protein interaction. Hence, a strong possibility exists for the NTD to act as a protein binding site at lower temperature ranges. There is also an uncanny correlation between the prevailing literature where scientists claim the virus affecting different parts of the human body and the similarity of NTD with human proteins (Gordon et al., 2020; Zaim et al., 2020).

From the bioinformatics and structural analyses (**Figure 4** and **Supplementary Figure S6**), we observed that the NTD not only acts as a glycan binding site but can also as a site for binding of several human proteins. The motif formed out of several polar residues on the solvent exposed loops at 10–30°C could form salt-bridges and hydrogen bonds with partner proteins. At higher temperatures, the propensity of forming such interactions would be lost owing to the differential orientation of the loops. Nonetheless, the NTD could act as a possible target for development of vaccines and inhibitors. Similar vaccines developed against the NTD of Spike protein in mice, had earlier shown that NTD could act as a potential therapeutic target (Coleman et al., 2014; Jiaming et al., 2017).

The RBD Behaves Differently at Higher Temperatures

The receptor binding domain (RBD) of the Spike glycoprotein is a potential target for vaccine and drug development (He et al., 2004; Tai et al., 2020). It is highly conserved among the human coronaviruses and binds to ACE2 receptor present on the lung tissues (Li, 2016). Residues 458–506 of the RBD domain comprises of the receptor binding motif (RBM). The RBM has 8 residues which are identical and 5 residues with similar biochemical properties between SARS, MERS and SARS-COV-2. This conserved region primarily interacts with the ACE2 receptor and hence, often scientists target the RBD domain of for developing therapeutic agents (He et al., 2004; Li, 2015; Tai et al., 2020). Earlier in **Figure 2**, we saw that the RBD domain spanning from residues 333–680 shows higher stability when compared to the NTD of the S1 domain at different range of temperatures.

We compared the time averaged conformation of the RBD generated from the last 10ns of the simulation time at different temperatures (**Figure 5**). The core β pleated sheet was very stable demonstrating no lack of secondary structures at higher temperatures. However, the RBM motif (highlighted in magenta in **Figure 5**) shows a very dynamic conformation across different temperature ranges. The dynamics was more pronounced at 10, 20, and 30°C whereas at 40 and 50°C of temperature, the RBM had a more confined conformation. The RBD flexibility was more apparent at 20 and 30°C where the three chains moved further away from each other. However, a tighter and well packed structure was found for the protein at 50°C. The figures suggest that although residue wise movements in RBD were not visible in RMSF (**Figure 2**), the RBD domains and motifs show intrinsic flexibility along particular temperature ranges. Previous studies have indicated that the RBD domain can adopt either an open or a closed conformation in the virus (Walls et al., 2020). We compared the conformation of the Spikeprotein-ACE2 crystal structure and found that in the open conformation, the RBD exposes its RBM residues Phe456, Ala475, Phe486, Asn487, Tyr489, Gln493, Gly496, Gln498, Thr500, Asn501, Gly502, and Tyr505 to facilitate the binding of the ACE2 receptors. It is fascinating to see that at 40°C and more interestingly at 50°C, the RBM motif is in a closed loop conformation and very compact which hinders its association with the partner proteins.

Spike Protein Adopts a Closed Conformation at Higher Temperature

Principal Component Analysis (PCA) was carried out to study the different conformations generated during the simulations. Principal components help us in identifying the most essential motions in complex systems (Berendsena and Haywardb, 2000; Meyer et al., 2006). To get a more pronounced picture, we studied the principal components of only the S1 domain. As seen in **Supplementary Figure S6**, the first 5 eigenvectors capture nearly 50% of the dynamics of the protein. We then went on to project the principal components along the first and second eigenvectors (**Supplementary Figure S6B**). The X axis represents the PC1

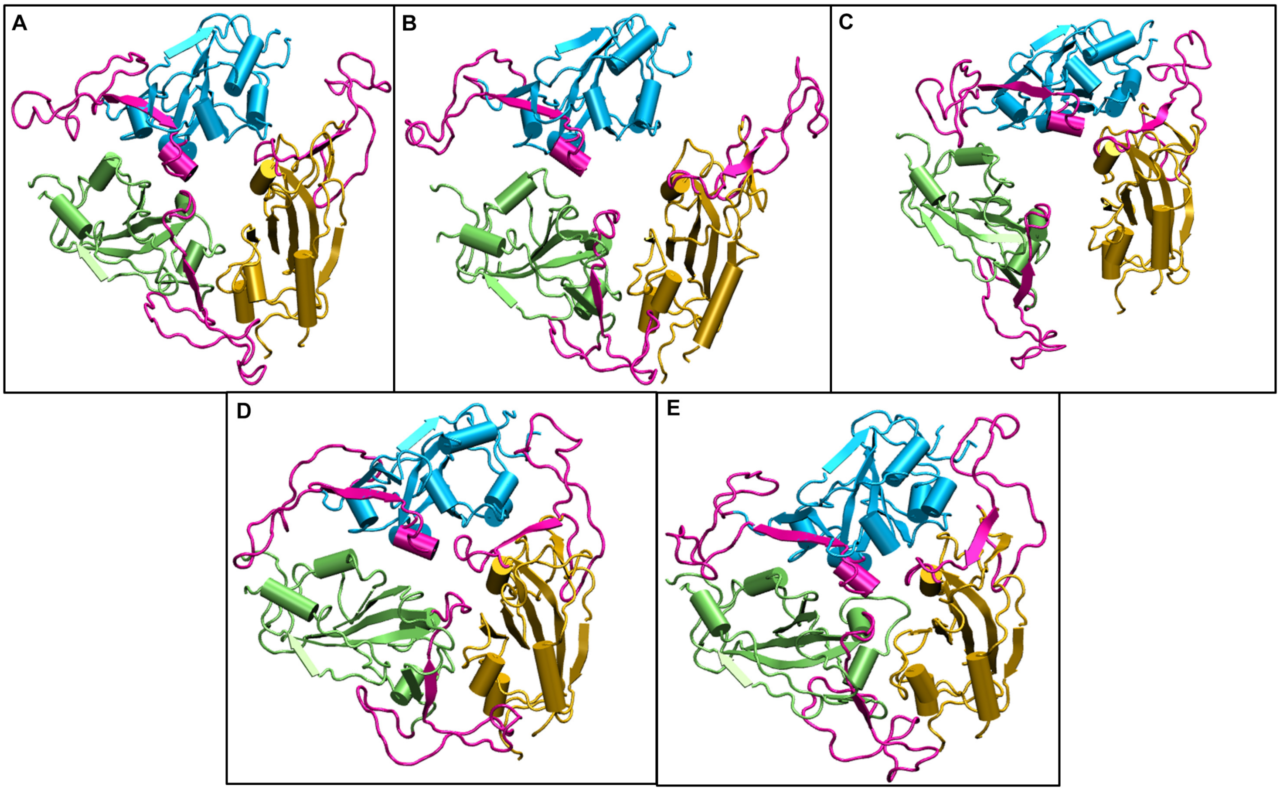


FIGURE 5 | Structures of the receptor binding domain of Spike protein after 200 ns of simulation at different temperatures exhibit diverse structural dynamics. Time-averaged conformations of RBD of SARS-CoV-2 Spike protein at, **(A)** 10°C, **(B)** 20°C, **(C)** 30°C, **(D)** 40°C, and **(E)** 50°C. The three chains are colored in lime, cyan, and orange. The receptor binding motif (shown in magenta) is oriented in a confined conformation at higher temperatures.

where maximum fluctuations could be seen. It can be seen that the data fluctuates between -18 to 15 in X axis and -7 and 12 in Y axis. At lower temperatures (10 and 20°C), the fluctuations varied between -8 and 12 at X axis, at 30°C the fluctuations were relatively higher varying from -10 to 15. However, at increased temperature of 40°C the data varied between -12 and 8 drifting more toward the left. At 50°C it further shifted toward the left varying from -18 to 4. In the Y axis at temperatures 10, 20, and 30°C, the data varied between -9 and 10. At 40°C it varied from -12 to 9. Surprisingly at 50°C, it only fluctuated between -10 and 5, indicating marked restrictions in movements. This clearly indicates larges conformational changes in the S1 domain at higher temperatures. The flexibility of the protein was found to have reduced with the increase in temperature.

Subsequently, we went on to check the extreme movements of the CA atoms of the S1 domain along the first principal component (**Figure 6**). The figure clearly shows that the arrows at 10, 20, and 30°C the three chains do not point toward each other, although the dynamics was high at 10 and 30°C. This would facilitate a more open conformation. The degree of movement was found to be more at 30°C which corresponds well with **Figure 5** as described above. Surprisingly, at 40 and 50°C, opposite domain movements were observed. The arrows point toward each other in these systems. Additionally, the three chains come close to each other and reduce the accessible area.

To further validate our findings, we ran another simulation of the Spike protein at a higher temperature of 70°C. After 100 ns of simulation, we found that significant similarity between the closed conformation observed at 50°C and the conformation at 70°C. The RBM residues, specifically Phe456, Ala475, Phe486, Asn487, Tyr489, Gln493, Gly496, Gln498, Thr500, Asn501, Gly502, and Tyr505 were found to be clearly buried between the interchain subunits at 70°C (**Figure 7** and **Supplementary Figure S7**). However, when compared to the orientation at 30°C the residues are directed toward the solvent. Thus, the reason for very stable RMSD observed in **Figure 1**, is largely due to the confined architecture of the receptor binding domain at 50°C and higher temperatures. The unavailability of RBM residues to bind to ACE2 receptor would nonetheless destabilize virus-protein interactions at higher temperatures.

Since the Spike protein is surrounded by glycans, it is of utmost importance to know if the carbohydrates cause change in the dynamics of the Spike protein. We ran three additional simulations with the glycans one at 10°C, one at 30°C and one at high temperature of 50°C for 200 ns. We found that irrespective of the presence of carbohydrates, the dynamics of the protein remains similar although the magnitude is diminished due to the large number of glycan chains. **Figure 8** shows the average conformation of the Spike protein RBD domain in the presence of carbohydrate molecules. At lower temperature of 10 and 30°C, the carbohydrate moieties are interspersed between the protein chains and the complex is open, similar to the structure without carbohydrates (**Figure 5**). The closed conformation at 50°C can be clearly observed in **Figure 8**, where now the carbohydrates

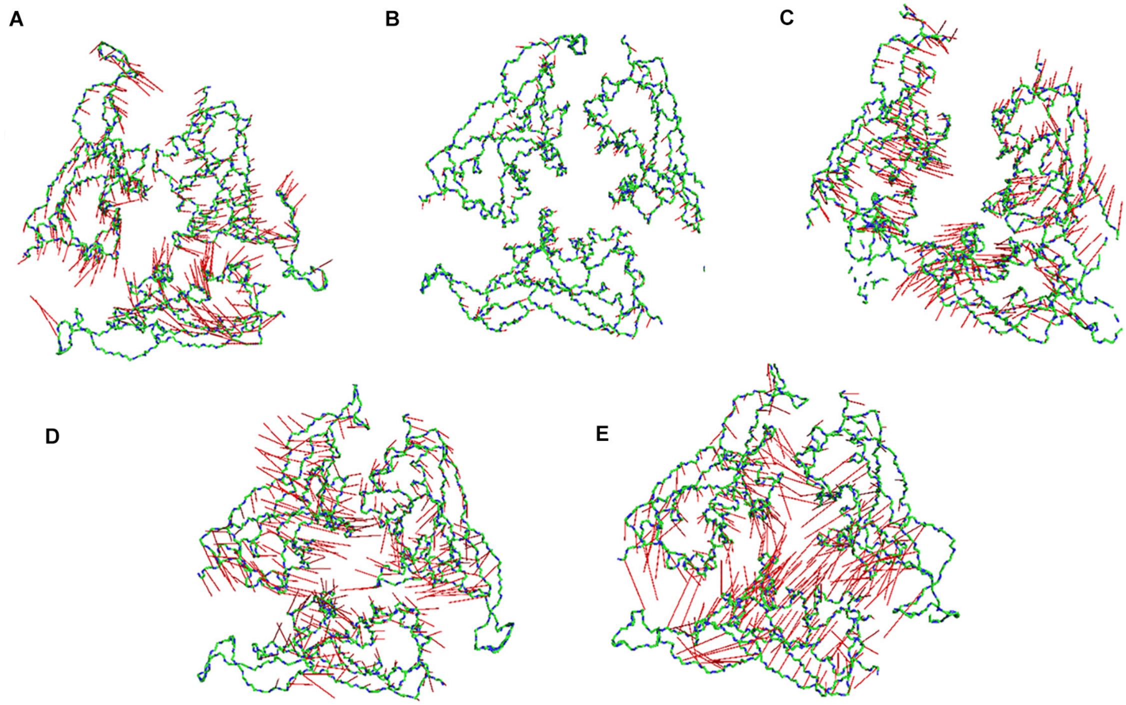


FIGURE 6 | Porcupine plots of the RBD of SARS-CoV-2 Spike protein. Porcupine plots generated from the extreme conformations of the RBD of the Spike protein at **(A)** 10°C, **(B)** 20°C, **(C)** 30°C, **(D)** 40°C, and **(E)** 50°C showing difference in dynamics of the protein at different temperature. The arrows are colored in red and the protein backbone is shown as sticks colored by CPK. More arrows are pointed toward the protein core at 40 and 50°C implying a closed conformation.

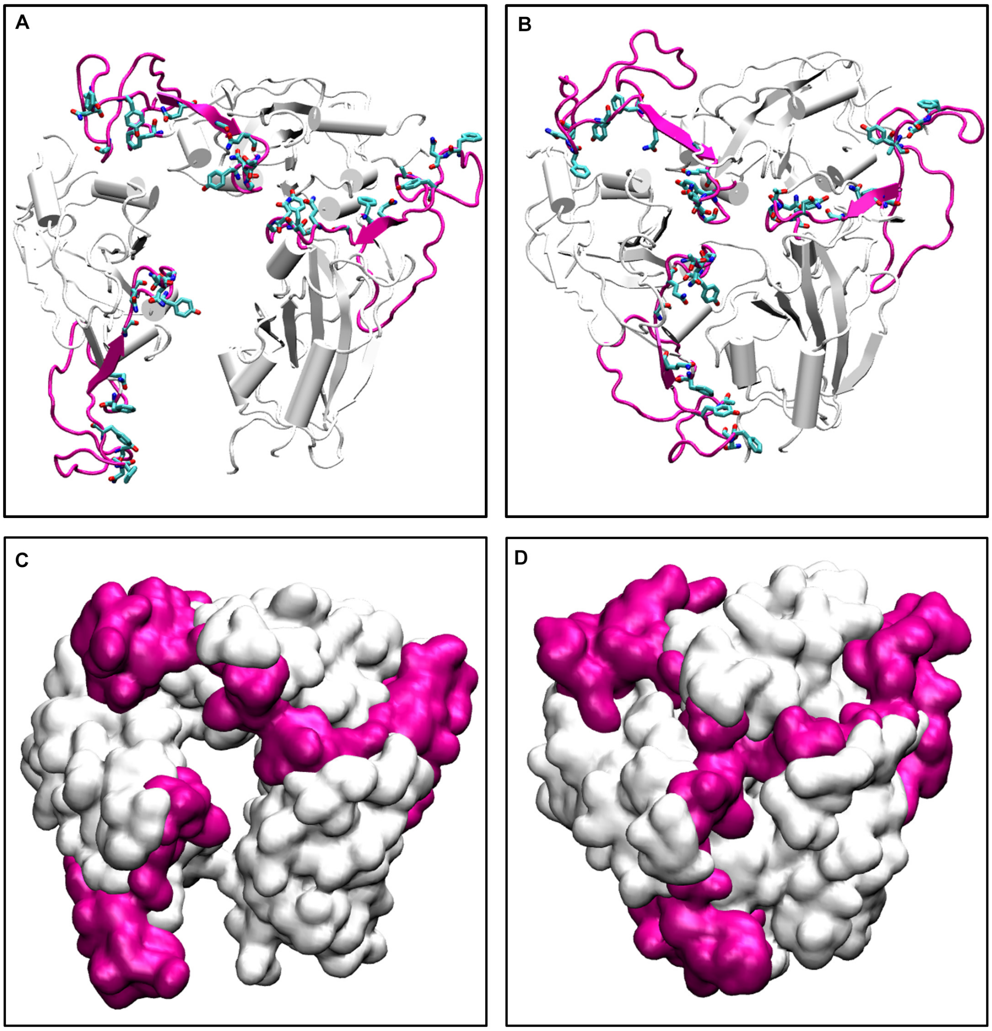


FIGURE 7 | The conformations of receptor binding motif. The time averaged structures of the Spike protein showing the open and closed conformations of the receptor binding domain at **(A)** 30°C and **(B)** 70°C. The residues on the Receptor binding motif (shown as sticks and colored by CPK) are buried at the interchain interface at high temperatures but readily exposes its residues at 30°C. **(C,D)** show the open and closed conformations of the receptor binding domain at **(A)** 30°C and **(B)** 70°C in surf mode. The receptor binding motif is colored in magenta and the receptor binding domain is shown in white.

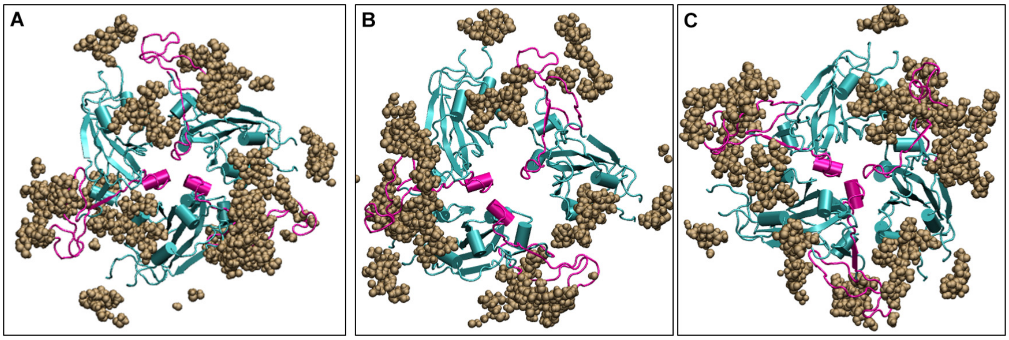


FIGURE 8 | Structures of the receptor binding domain of glycan bound Spike protein after 200 ns of simulation displays similar behavior to that protein in the absence of glycans. Time-averaged conformations of RBD of SARS-CoV-2 Spike protein at, **(A)** 10°C, **(B)** 30°C, and **(C)** 50°C. The glycan moieties are colored in brown and shown in Vander Waal's representation. Protein is shown in cyan color and the RBM motif in magenta. RBD, receptor binding domain; NTD, N –terminal domain; RBM, receptor binding motif; ACE2, angiotensin converting enzyme 2.

appear to be surrounding the protein while the protein closes its RBM. Thus, although carbohydrates are very important for the immunogenicity of the protein, the temperature dependent conformation of the spike glycoprotein is largely due to the protein dynamics.

Thus, although the RBM stays largely in open conformation state, surprisingly, from around the temperature of 40°C, a closed conformation of the motif was observed. In this conformation, the RBM from the three chains come very close to each other sealing the visibility of the trimeric pore. At temperatures >50°C, the Spike RBM is completely closed (**Figure 7**). The closing of the RBM buries the receptor binding residues inside trimer abolishing the possibility of contacts with the ACE2 receptor and making the Spike protein inactive. Our results clearly show that the activity of the Spike protein is dependent on the external temperatures where a higher temperature renders it completely inactive.

CONCLUSION

The SARS-CoV-2 has severely affected the human population with large number of infected individuals around world. The propensity of virus to survive in cold and dry climatic conditions have been speculated by researchers and supported by the statistical evidence from earlier SARS epidemic of 2002. However, it is still unclear how the virus undergoes changes at the structural level in different environments. The Spike protein of the virus helps in the attachment and entry of the coronaviruses inside the host cells. It exists as a homotrimer and is partly exposed to the outer environment and partly immersed inside the lipid bilayer of the viral envelope. Here, we studied the differential response of the Spike protein at different temperature conditions.

Our results show that the S2 transmembrane domain remains stable even without the bilayer membrane, whereas the solvent exposed S1 domain is quite flexible. Moreover, the S1 comprises of two subdomains, namely N-terminal domain (NTD) and the receptor binding domain (RBD). The simulations results show that the RBD is relatively less mobile. Its flexibility is limited only to the receptor binding motif or RBM which interacts with the Angiotensin Converting enzyme 2 (ACE2), its human receptor. However, the NTD was found to be quite mobile.

Although, the NTD doesn't directly interact with the ACE2 receptor in humans, it has been found to bind to receptors in other mammals (Humphrey et al., 1996). The flexible NTD hosts a large number of charged residues on the top layer of its tri-layered beta sandwich architecture. However, at 40–50°C of temperature, the polar residues were found to be less solvent exposed. The similarity of the NTD sequence with the several human receptors such as Ephrins, Briakunumab, anti-TSLP, etc. indicated a possibility of the subdomain to be involved in binding to alternate human proteins.

The RBM present on the RBD is very crucial in initial protein-protein interaction between the host and virus. We found that this domain is largely in an open conformation which enables receptor binding at lower temperatures. Surprisingly, from ∼40°C, a closed conformation of the motif was observed. In this conformation, the RBM from the three chains come very close to each other sealing the visibility of the trimeric pore. At temperatures >50°C, the Spike RBM is completely closed. This was also evident from the dynamics obtained from principle component analysis. The closing of the RBM buries the receptor binding residues inside the trimer abolishing the possibility of contacts with the ACE2 receptor and making the Spike protein inactive. The same phenomena was observed in the presence of glycan moieties.

Our results have shown for the first time that the Spike protein has the possibility to stay in an active and inactive state based on the external temperature. They corroborate very well with the experimental observations and observations from simulation studies which talks about thermal stability and cold-induced destabilization of the protein (Edwards et al., 2020; He et al., 2020; Ou et al., 2020; Xiong et al., 2020). Moreover, since no visible loss of secondary structure was observed at higher temperatures (**Supplementary Figure S8**), it would be interesting to know if the conformational change is reversible in nature. Nevertheless, this work would prove very beneficial in the development of vaccines as well as development of therapeutic strategies that target not only the receptor binding domain but also the N-terminal domain of the Spike protein.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

SR and KK conceived and designed the experiments. SR performed the experiments, analyzed the data, contributed reagents, materials, analysis tools, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

This research used resources of the National Energy Research Scientific Computing Center of the Ernest Orlando Lawrence Berkeley National Laboratory, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and used the Extreme Science and Engineering Discovery Environment (XSEDE). We are thankful to the Covid19 HPC Consortium for providing resources and helping researchers work for a noble cause. We are also thankful to Drs. Suchetana Gupta, Debakanta Tripathy, and Chockalingam S. for critically proofreading the manuscript. We are also grateful to National Institute of Technology Warangal for providing facilities. This manuscript has been released as a pre-print at BioRxiv (Rath and Kumar, 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.583523/full#supplementary-material

REFERENCES

Astuti, I., and Ysrafil (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): an overview of viral structure and host response. *Diabetes Metab. Syndr. Clin. Res. Rev.* 14, 407–412. doi: 10.1016/j.dsx.2020.04.020

Belouzard, S., Chu, V. C., and Whittaker, G. R. (2009). Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5871–5876. doi: 10.1073/pnas.0809524106

Berendsena, H. J. C., and Haywardb, S. (2000). Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.* 10, 165–169. doi: 10.1016/s0959-440x(00)00061-0

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank nuc. *Acids Res.* 28, 235–242.

Bukhari, Q., and Jameel, Y. (2020). Will coronavirus pandemic diminish by summer?. *SSRN Electron. J.*

Cai, Q. C., Lu, J., Xu, Q. F., Guo, Q., Xu, D. Z., Sun, Q. W., et al. (2007). Influence of meteorological factors and air pollution on the outbreak of severe acute respiratory syndrome. *Public Health* 121, 258–265. doi: 10.1016/j.puhe.2006.09.023

Casanova, L. M., Jeon, S., Rutala, W. A., Weber, D. J., and Sobsey, M. D. (2010). Effects of air temperature and relative humidity on coronavirus survival on surfaces. *Appl. Environ. Microbiol.* 76, 2712–2717. doi: 10.1128/AEM.02291-09

Chan, K. H., Peiris, J. S. M., Lam, S. Y., Poon, L. L. M., Yuen, K. Y., and Seto, W. H. (2011). The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Adv. Virol.* 2011:734690. doi: 10.1155/2011/734690

Chin, A. W. H., Chu, J. T. S., Perera, M. R. A., Hui, K. P. Y., Yen, H.-L., Chan, M. C. W., et al. (2020). Stability of SARS-CoV-2 in different environmental conditions. *Lancet Microbe* 1:e10. doi: 10.1016/s2666-5247(20)30003-3

Coleman, C. M., Liu, Y. V., Mu, H., Taylor, J. K., Massare, M., Flyer, D. C., et al. (2014). Purified coronavirus spike protein nanoparticles induce coronavirus neutralizing antibodies in mice. *Vaccine* 32, 3169–3174. doi: 10.1016/j.vaccine.2014.04.016

Dong, Y. W., Liao, M. L., Meng, X. L., and Somero, G. N. (2018). Structural flexibility and protein adaptation to temperature: molecular dynamics analysis of malate dehydrogenases of marine molluscs. *Proc. Natl. Acad. Sci. U.S.A.* 115, 1274–1279. doi: 10.1073/pnas.1718910115

Edwards, R., Mansouri, K., Stalls, V., Manne, K., Watts, B., and Parks, R. (2020). Cold sensitivity of the SARS-CoV-2 spike ectodomain. *bioRxiv* [Preprint]. doi: 10.1101/2020.07.12.199588

Foxman, E. F., Storer, J. A., Fitzgerald, M. E., Wasik, B. R., Hou, L., and Zhao, H. (2015). Temperature-dependent innate defense against the common cold virus limits viral replication at warm temperature in mouse airway cells. *Proc. Natl. Acad. Sci. U.S.A.* 112, 827–832. doi: 10.1073/pnas.1411030112

Gao, H., Yao, H., Yang, S., and Li, L. (2016). From SARS to MERS: evidence and speculation. *Front. Med.* 1, 377–382. doi: 10.1007/s11684-016-0466-7

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., and White, K. M. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468. doi: 10.1038/s41586-020-2286-9

GROMACS (2020). *GROMACS Development Team, Gromacs Documentation Release 2020.1*.

Guillén, J., Pérez-Berná, A. J., Moreno, M. R., and Villalaín, J. (2005). Identification of the membrane-active regions of the severe acute respiratory syndrome coronavirus spike membrane glycoprotein using a 16/18-mer peptide scan: implications for the viral fusion mechanism. *J. Virol.* 79, 1743–1752. doi: 10.1128/JVI.79.3.1743-1752.2005

He, J., Tao, H., Yan, Y., Huang, S.-Y., and Xiao, Y. (2020). Molecular mechanism of evolution and human infection with SARS-CoV-2. *Viruses* 12:428. doi: 10.3390/v12040428

He, Y., Zhou, Y., Liu, S., Kou, Z., Li, W., Farzan, M., et al. (2004). Receptor-binding domain of SARS-CoV spike protein induces highly potent neutralizing antibodies: implication for developing subunit vaccine. *Biochem. Biophys. Res. Commun.* 324, 773–781. doi: 10.1016/j.bbrc.2004.09.106

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD - visual molecular dynamics. *J. Mol. Graphics* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Ishida, Y. I., Hiraki, A., Hirayama, E., Koga, Y., and Kim, J. (2002). Temperature-sensitive viral infection: inhibition of hemagglutinating virus of japan (sendai virus) infection at 41°. *Intervirology* 45, 125–135. doi: 10.1159/000065865

Jiaming, L., Yanfeng, Y., Yao, D., Yawei, H., Linlin, B., Baoying, H., et al. (2017). The recombinant N-terminal domain of spike proteins is a potential vaccine against middle east respiratory syndrome coronavirus (MERS-CoV) infection. *Vaccine* 35, 10–18. doi: 10.1016/j.vaccine.2016.11.064

Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* 29, 1859–1865. doi: 10.1002/jcc.20945

Jorgensen, W. L., Chandrasekhar, J., and Madura, J. D. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926. doi: 10.1063/1.445869

Julió Plana, L., Nadra, A. D., Estrin, D. A., Luque, F. J., and Capece, L. (2019). Thermal stability of globins: implications of flexibility and heme coordination studied by molecular dynamics simulations. *J. Chem. Inf. Model.* 59, 441–452. doi: 10.1021/acs.jcim.8b00840

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., and Fan, S. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 581, 215–220. doi: 10.1038/s41586-020-2180-5

Li, F. (2015). Receptor recognition mechanisms of coronaviruses: a decade of structural studies. *J. Virol.* 89, 1954–1964. doi: 10.1128/jvi.02615-14

Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. *Annu. Rev. Virol.* 3, 237–261. doi: 10.1146/annurev-virology-110615-042301

Li, J., Ulitzky, L., Silberstein, E., Taylor, D. R., and Viscidi, R. (2013). Immunogenicity and protection efficacy of monomeric and trimeric recombinant SARS coronavirus spike protein subunit vaccine candidates. *Viral Immunol.* 26, 126–132. doi: 10.1089/vim.2012.0076

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., Shaw, D. E., et al. (2013). Systematic validation of protein force fields against experimental data. *PLoS One* 8:e32131. doi: 10.1371/journal.pone.0032131

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., and Dror, R. O. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinform.* 78, 1950–1958. doi: 10.1002/prot.22711

Lowen, A. C., and Steel, J. (2014). Roles of humidity and temperature in shaping influenza seasonality. *J. Virol.* 88, 7692–7695. doi: 10.1128/jvi.03544-13

MacKenzie, J. S., and Smith, D. W. (2020). COVID-19: a novel zoonotic disease caused by a coronavirus from china: what we know and what we don't. *Microbiol. Aust.* 41, 45–50. doi: 10.1071/MA20013

Meyer, T., Ferrer-Costa, C., Perez, A., Rueda, M., Bidon-Chanal, A., Luque, F. J., et al. (2006). Essential dynamics: a tool for efficient trajectory compression and management. *J. Chem. Theory Comput.* 2, 251–258. doi: 10.1021/ct050285b

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., et al. (2020). Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat. Commun.* 11:1620.

Pelletier, I., Rousset, D., Enouf, V., Colbère-Garapin, F., van der Werf, S., and Naffakh nadianaffakh, N. (2011). Highly heterogeneous temperature sensitivity of 2009 pandemic influenza A(H1N1) viral isolates, northern france. *Euro Surveill.* 16:19999.

Rath, S. L., and Kumar, K. (2020). Investigation of the effect of temperature on the structure of SARS-Cov-2 spike protein by molecular dynamics simulations. *bioRxiv* [Preprint]. doi: 10.1101/2020.06.10.145086

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., and Chetvernin, V. (2009). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 37, 5–15. doi: 10.1093/nar/gkn741

Shang, J., Wan, Y., Liu, C., Yount, B., Gully, K., Yang, Y., et al. (2020). Structure of mouse coronavirus spike protein complexed with receptor reveals mechanism for viral entry. *PLoS Pathog.* 16:e1008392. doi: 10.1371/journal.ppat.1008392

Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C. K., and Zhou, J. (2016). Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol.* 1, 490–502. doi: 10.1016/j.tim.2016.03.003

Tai, W., He, L., Zhang, X., Pu, J., Voronin, D., Jiang, S., et al. (2020). Characterization of the Receptor-Binding Domain (RBD) of 2019 novel coronavirus: implication for development of rbd protein as a viral attachment inhibitor and vaccine. *Cell. Mol. Immunol.* 17, 613–620. doi: 10.1038/s41423-020-0400-4

Tan, J., Mu, L., Huang, J., Yu, S., Chen, B., and Yin, J. (2005). An initial investigation of the association between the SARS outbreak and weather: with the view of the environmental temperature and its variation. *J. Epidemiol. Community Health* 59, 186–192. doi: 10.1136/jech.2004.020180

The PyMOL Molecular Graphics System (2019). *The PyMOL Molecular Graphics System, Version 1.2r3pre*. Schrödinger: LLC.

Van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., and Williamson, B. N. (2020). Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *N. Engl. J. Med.* 16, 1564–1567. doi: 10.1056/NEJMc2004973

Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181, 281.e6–292.e6. doi: 10.1016/j.cell.2020.02.058

Walls, A. C., Tortorici, M. A., Bosch, B. J., Frenz, B., Rottier, P. J. M., DiMaio, F., et al. (2016). Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* 531, 114–117. doi: 10.1038/nature16988

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., and Zhang, Z. (2020). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* 181, 894.e9–904.e9. doi: 10.1016/j.cell.2020.03.045

Woo, H., Park, S.-J., Choi, Y. K., Park, Y., Tanveer, M., and Cao, Y. (2020). Developing a fully-glycosylated full-length SARS-CoV-2 spike protein model in a viral membrane. *J. Phys. Chem. B* 124, 7128–7137. doi: 10.1021/acs.jpcb.0c04553

World Health Organization [WHO] (2020). *WHO Coronavirus Disease (COVID-19) Dashboard*. Geneva: World Health Organization.

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., et al. (2020). Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *Sci. Total Environ.* 729:139051. doi: 10.1016/j.scitotenv.2020.139051

Xiong, X., Qu, K., Ciazynska, K. A., Hosmillo, M., Carter, A. P., and Ebrahimi, S. (2020). A thermostable, closed SARS-CoV-2 spike protein trimer. *Nat. Struc. Mol. Biol.* [Epub ahead of print].

Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5?. *Bioinformatics* 26, 889–895. doi: 10.1093/bioinformatics/btq066

Yuan, Y., Cao, D., Zhang, Y., Ma, J., Qi, J., and Wang, Q. (2017). Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat. Commun.* 8:15092. doi: 10.1038/ncomms15092

Zaim, S., Chong, J. H., Sankaranarayanan, V., and Harky, A. (2020). COVID-19 and multiorgan response. *Curr. Probl. Cardiol.* [Epub ahead of print]. doi: 10.1016/j.cpcardiol.2020.100618

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 24 November 2020
doi: 10.3389/fmolb.2020.590165

Computational Investigation of Structural Dynamics of SARS-CoV-2 Methyltransferase-Stimulatory Factor Heterodimer nsp16/nsp10 Bound to the Cofactor SAM

Md Fulbabu Sk[†], *Nisha Amarnath Jonniya*[†], *Rajarshi Roy, Sayan Poddar and Parimal Kar*[*]

Discipline of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Khandwa, India

Recently, a highly contagious novel coronavirus disease 2019 (COVID-19), caused by SARS-CoV-2, has emerged, posing a global threat to public health. Identifying a potential target and developing vaccines or antiviral drugs is an urgent demand in the absence of approved therapeutic agents. The 5′-capping mechanism of eukaryotic mRNA and some viruses such as coronaviruses (CoVs) are essential for maintaining the RNA stability and protein translation in the virus. SARS-CoV-2 encodes S-adenosyl-L-methionine (SAM) dependent methyltransferase (MTase) enzyme characterized by nsp16 (2′-O-MTase) for generating the capped structure. The present study highlights the binding mechanism of nsp16 and nsp10 to identify the role of nsp10 in MTase activity. Furthermore, we investigated the conformational dynamics and energetics behind the binding of SAM to nsp16 and nsp16/nsp10 heterodimer by employing molecular dynamics simulations in conjunction with the Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA) method. We observed from our simulations that the presence of nsp10 increases the favorable van der Waals and electrostatic interactions between SAM and nsp16. Thus, nsp10 acts as a stimulator for the strong binding of SAM to nsp16. The hydrophobic interactions were predominately identified for the nsp16-nsp10 interactions. Also, the stable hydrogen bonds between Ala83 (nsp16) and Tyr96 (nsp10), and between Gln87 (nsp16) and Leu45 (nsp10) play a vital role in the dimerization of nsp16 and nsp10. Besides, Computational Alanine Scanning (CAS) mutagenesis was performed, which revealed hotspot mutants, namely I40A, V104A, and R86A for the dimer association. Hence, the dimer interface of nsp16/nsp10 could also be a potential target in retarding the 2′-O-MTase activity in SARS-CoV-2. Overall, our study provides a comprehensive understanding of the dynamic and thermodynamic process of binding nsp16 and nsp10 that will contribute to the novel design of peptide inhibitors based on nsp16.

Keywords: SARS-CoV-2, nsp16/nsp10, molecular dynamics, PCA, MM-PBSA

OPEN ACCESS

Edited by:
Arvind Ramanathan,
Argonne National Laboratory (DOE),
United States

Reviewed by:
Shruthi Viswanath,
National Centre for Biological Sciences, India
Om Choudhary,
University of Pittsburgh, United States
Filippo Pullara,
SpIntellx Inc., Pittsburgh,
United States

****Correspondence:***
Parimal Kar
parimal@iiti.ac.in

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *31 July 2020*
Accepted: *23 October 2020*
Published: *24 November 2020*

Citation:
Sk MF, Jonniya NA, Roy R, Poddar S and Kar P (2020) Computational Investigation of Structural Dynamics of SARS-CoV-2 Methyltransferase-Stimulatory Factor Heterodimer nsp16/nsp10 Bound to the Cofactor SAM. Front. Mol. Biosci. 7:590165. doi: 10.3389/fmolb.2020.590165

INTRODUCTION

Coronaviruses (CoVs) are considered as an etiological agent for causing severe acute respiratory syndrome (SARS) in humans. In the past two decades, SARS-CoV and MERS-CoV (middle east respiratory syndrome coronavirus) were responsible for the epidemic in 2003 and 2012, respectively. Recently, in December 2019, the novel coronavirus (COVID-19) pandemic, caused by SARS-CoV-2, first broke out in Wuhan city of China and has been spreading worldwide (Bogoch et al., 2020; Guan et al., 2020; Lin X. et al., 2020). So far, ∼42 million people worldwide are infected by SARS-CoV-2, including ∼1.1 million deaths. In India, the COVID-19 tally has already crossed 7.7 million, including ∼0.1 million deaths. However, neither prophylactic vaccines nor any direct antiviral drugs are available to effectively treat the human and animal coronavirus disease (Lu, 2020; Pillaiyar et al., 2020; Sheahan et al., 2020).

CoVs are single-stranded positive-sense RNA viruses (Eckerle et al., 2010; Fehr and Perlman, 2015) belonging to the family *Coronaviridae* and possess the largest genome (26.4–31.7 kb) (Woo et al., 2010). They are classified into four genera, namely *Alphacoronaviruses*, *Betacoronaviruses*, *Gammacoronaviruses,* and *Deltacoronaviruses* (King et al., 2011). CoVs can infect both humans and animals (Lun and Qu, 2004; Coleman and Frieman, 2014), and can cause diseases like Hepatitis and Pneumonitis in mouse (Weiss and Leibowitz, 2011) and neurologic & respiratory diseases in humans (Arabi et al., 2015; Huang et al., 2020). So far, seven distinctive strains of human coronaviruses (HCoV) have been disclosed that includes 229E and NL63 (*Alphacoronaviruses*), and OC43, HKU1, SARS-CoV (2002), MERS-CoV (2012), and SARS-CoV-2 (2019) (*Betacoronaviruses*) (Weiss and Navas-Martin, 2005; Zeng et al., 2018; Singh et al., 2020). SARS-CoV-2 shares a strong correlation with the bat CoV RaTG13 having a 96.2% genome sequence identity, suggesting its possible evolution from bat CoVs (Wu et al., 2020; Zhou P. et al., 2020). SARS-CoV-2 displays a sequence similarity of 80% with SARS-CoV, whereas only 50% with MERS-CoV (Lu et al., 2020; Wu et al., 2020; Zhou Y. et al., 2020). Currently, most of the therapeutic options that are available for controlling COVID-19 is based on the previous knowledge and information gathered from SARS-CoV and MERS-CoV.

Most viruses or eukaryotic cellular mRNA possess the 5′-end capping mechanism that plays a vital role in mRNA splicing, translation initiation, stability, and intracellular RNA transport (Furuichi and Shatkin, 2000). The capping of the 5′-end occurs through a sequential enzymatic process. This involves three enzymes, such as RNA guanylyltransferase (GTase), RNA triphosphatase (TPase), and RNA guanine-N7-methyltransferase (N7-MTase). This generates a cap-0 structure (m7GpppN). It is further methylated at the 2′-O position of mRNA by 2′-O- methyltransferase (2′-O-MTase) and generates the cap-1 (m7GpppNm) and cap-2 (m7GpppNmNm) structures. This mimicking of the eukaryotic mRNA capping mechanism ultimately helps the virus evade the host innate immune system (Furuichi and Shatkin, 2000; Wang et al., 2015). Both the MTase uses S-adenosyl-L-methionine (SAM or AdoMet) as a donor of methyl and gives a by-product, S-adenosyl-L-homocysteine (SAH or AdoHcy).

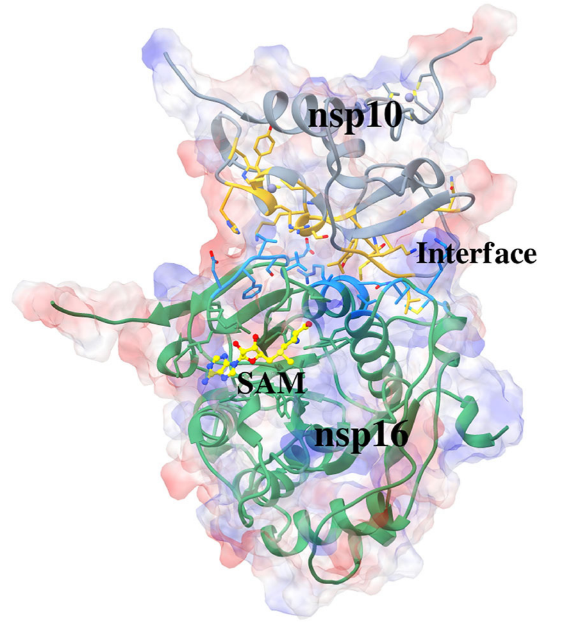


FIGURE 1 | Crystal structure of SARS-CoV-2 nsp10-nsp16 complexed with cofactor S-Adenosyl Methionine (SAM) (PDB: 6W4H). The nsp10, interacting part of nsp10, interacting part of nsp16, nsp16, and the cofactor SAM in a complex is shown in color teal, gold, dark sky blue, light green, and yellow, respectively. The surface represents an electrostatic surface.

The genome of SARS-CoV-2 is comprised of 10 ORFs (Open Reading Frame). ORF1ab encodes the replicase polyprotein 1ab (PP1ab), which gets cleaved by the two viral proteases, PL^{pro} (papain-like protease) and 3CL^{pro} (3-C like protease) at the N-terminus and C-terminus, respectively, to form all the 16 non-structural proteins (nsp1, nsp2, nsp3 by PL^{pro}, and nsp4-nsp16 by 3CL^{pro}). The remaining ORFs are associated with encoding structural proteins, such as spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins, as well as other accessory proteins (Harcourt et al., 2004; Yang et al., 2005; Chen et al., 2020; Wu et al., 2020). Previous structural and biochemical characterization studies on SARS-CoV (2002) showed that in the nsp10-nsp16/SAM complex, nsp10 enhances the methylase activity of nsp16 by increasing the stability of the SAM-binding pocket (Chen et al., 2011; Wang et al., 2015).

The present study involves analyzing protein-protein interactions in the heterodimeric structure formed by nsp10 and nsp16 of SARS-CoV-2. The recently solved crystal structure of SARS-CoV-2 nsp16/nsp10 heterodimer bound to SAM (PDB: 6W4H) (Minasov et al., 2020) is used in the current study. The three-dimensional structure of the complex is shown in **Figure 1**. Here, we have used the standard molecular dynamics (MD) simulations in conjunction with the molecular-mechanics Poisson-Boltzmann surface area (MM-PBSA) method to identify the critical residues involved in the complex formation. Further, we conducted MD simulations of nsp16/SAM, elucidating the importance of nsp10 in the binding of SAM to nsp16. The detailed structural analysis and inter-molecular interactions reveal the interface of nsp16/ns10, which provides a distinctive feature for coronaviruses, can be exploited as an attractive target in developing specific antiviral drugs for controlling COVID-19. MD simulations were conducted to elucidate the role of nsp10 in the interaction of SAM to nsp16. Also, the conformational changes in nsp10 and nsp16 due to their association were investigated.

MATERIALS AND METHODS

We retrieved the experimental coordinate of the SARS-CoV-2 nsp16/nsp10 complex from Protein Data Bank (PDB), which was crystallized at 1.8 Å (PDB ID 6W4H) (Minasov et al., 2020). The complex structure includes S-adenosyl-L-methionine (SAM). The protonation states were determined using PROPKA 3.1 (Olsson et al., 2011), and the corresponding residues were modified accordingly. The following three systems were simulated for the current study: complex (nsp16/nsp10/SAM), nsp16$_{\mathrm{SAM}}$ (nsp16/SAM), and nsp10$_{\mathrm{apo}}$ (apo nsp10). The starting configurations of nsp16$_{\mathrm{SAM}}$ and nsp10$_{\mathrm{apo}}$ were constructed manually from nsp16/nsp10/SAM (PDB: 6W4H).

Missing hydrogens in the crystal structure were built using the *Leap* module of AmberTools19 (Case et al., 2018), and a proper amount of Na$^+$ was added to neutralize the system. All the systems were solvated in a periodic octahedron TIP3P water box (Price and Brooks, 2004) with a 10 Å buffering distance from all directions. The proteins (nsp10 and nsp16) were assigned the Amber ff14SB force field parameters (Maier et al., 2015) while the cofactor SAM was modeled using the updated Generalized Amber Force Field (GAFF2) (Wang et al., 2004) and AM1-bcc (Jakalian et al., 2002) charges. The bond lengths having hydrogen atoms were kept fixed using the SHAKE algorithm (Kräutler et al., 2001), which allowed 2 fs time-step for MD simulations. The Particle-Mesh Ewald (PME) summation (Darden et al., 1993) scheme was employed to compute the long-range interactions. The non-bonded cut-off was set to 10 Å.

Firstly, the solvated systems were subjected to an energy minimization using 500 steps of steepest descent, followed by another 500 steps of the conjugant gradient algorithm. During the minimization, the solute atoms were kept fixed with a restraint force of 2.0 kcal mol^{-1}Å^{-2}. The second stage of minimization was carried out without any restraint force. After the minimization, each system was gradually heated from 0 to 300 K in the NVT ensemble, and the solute atoms were restrained with force constant of 2.0 kcal mol^{-1}Å^{-2}. The temperature was maintained using a Langevin thermostat (Pastor et al., 1988; Loncharich et al., 1992). Subsequently, each system was simulated for 50 ps at 300 K at a constant pressure of 1 atm using the Berendsen barostat (Berendsen et al., 1984). In this stage also, the same restraint force was applied to the solute atoms. Before the production run, we equilibrated each system by conducting 1 ns MD simulation in the NPT ensemble without restraining the system. Finally, the production simulation was carried out for 1 μs using the *pmemd.cuda* module of AMBER18. Overall, 100,000 snapshots were generated and used for the analysis using the *Cpptraj* module (Roe and Cheatham, 2013) of AMBER18 (Case et al., 2018).

To study the effect of salt concentration on the binding, we simulated the complex system at 300 K for 1 μs at two different salt concentrations (0.15 and 0.25 M). The desired salt concentration was achieved by adding an appropriate number of monovalent ions (Na$^+$ and Cl$^-$) obtained with the protocol suggested by Machado and Pantano (2020). Similarly, we have also investigated the effect of temperature on the binding by conducting MD simulation of the complex at a relatively elevated temperature of 310 K.

Trajectory Analysis

All analyses, including root-mean-square deviations (RMSDs), root-mean-square fluctuations (RMSFs), radius of gyration (R$_g$), solvent accessible surface area (SASA), etc. were performed using the *Cpptraj* module (Roe and Cheatham, 2013) of AMBER18. A distance of $\leq$ 3.5 Å and an angle cut-off of $\geq$120° were used for hydrogen bond calculations. The same criterion was used in earlier studies (Jeffrey, 1997; Hu et al., 2016; Shi et al., 2018; Sanachai et al., 2020).

Principal Component Analysis

One of the widely used unsupervised data reduction schemes is principal component analysis (PCA) (Ichiye and Karplus, 1991). First, a covariance matrix (C) is constructed from the atomic fluctuations of C$_\alpha$-atoms of each residue (Amadei et al., 1996). The elements C$_{\mathrm{ij}}$ of the covariance matrix *C* are defines as

$$C_{ij} = < ((x_i - < x_i >)\,(x_j - < x_j >)\,) > \tag{1}$$

where x_i and x_j denotes the instant coordinates of the *i*th or *j*th atoms, while $< x_i >$ and $< x_j >$ denote the mean of *i*th or *j*th atoms over the ensemble. The diagonalization of the covariance matrix yields orthogonal eigenvectors and the corresponding eigenvalues. The resulting eigenvectors are termed as principal components (PCs) and indicate the direction of the movement, while the corresponding eigenvalue describes the amplitude of motions. We adopted the same methodology for PCA, as discussed in our earlier studies (Jonniya et al., 2019; Sk et al., 2020c).

The free energy landscape (FEL) was generated based on the following Equation (Frauenfelder et al., 1991):

$$G_i = -k_B T \ln\left(\frac{N_i}{N_m}\right) \tag{2}$$

where k_B represents the Boltzmann constant, and T is the absolute temperature. N_{i} is the population of the *i*th bin, and N_{m} is the population of the most populated bin. The 2-dimensional FEL was constructed using PC1 and PC2 as the reaction coordinate.

Residual Network Analysis of Protein

The residual network analysis approach is widely used to explore the viral fitness and resistance development of protein structure. The network analysis of protein structures (NAPS) (Chakrabarty et al., 2019) server (http://bioinf.iiit.ac.in/NAPS/) was used to identify key residue interactions in the residual network and the network-based hydrophobic contacts from the simulation trajectories. Here, we used the protein-protein complex option, followed by the C_α network type and unweighted edge weight. An edge represented the distance between a pair of C_α atoms within the lower and upper thresholds (default upper threshold = 7 Å; lower threshold = 0 Å). We considered the default parameters for the long-range interaction networks and minimum residue separation of 1. The hydrophobicity indices of 20 amino acids range from −4.5 to 4.5 (Kyte and Doolittle, 1982). These values were colored in gradients of red to show the hydrophobic residues, while the gradients of blue were used to display the hydrophilic residues.

Binding Free Energy and Alanine Scanning

The interaction energy between nsp10 and nsp16, and between the cofactor SAM and nsp16 in its both monomer and the complex were computed using the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) methodology (Kollman et al., 2000; Wang et al., 2006; Kar et al., 2007a,b, 2011, 2013; Hou et al., 2011). MMPBSA.py script available in the AmberTools19 was used for the analysis. Details of the MM-PBSA protocol were provided in our previous studies (Kar and Knecht, 2012a,b,c,d; Kar et al., 2013; Jonniya et al., 2019; Jonniya and Kar, 2020; Roy et al., 2020; Sk et al., 2020a,d), and the same protocol was adopted here.

The binding free energy is estimated by using the following equations of the MM-PBSA scheme:

$$\Delta G_{\text{bind}} = \Delta H - T\Delta S \approx \Delta E_{\text{internal}} + \Delta G_{\text{solv}} - T\Delta S \quad (3)$$

$$\Delta E_{\text{internal}} = \Delta E_{\text{covalent}} + \Delta E_{\text{elec}} + \Delta E_{\text{vdW}} \quad (4)$$

$$\Delta G_{\text{solv}} = \Delta G_{\text{pol}} + \Delta G_{\text{np}} \quad (5)$$

where $\Delta E_{\text{internal}}$, ΔG_{solv}, and $T\Delta S$ represent the total internal energy, desolvation free energy, and conformational entropy, respectively. Further, the internal energy is composed of $\Delta E_{\text{covalent}}$ (bond, dihedral, and angle), ΔE_{elec} (electrostatic) and ΔE_{vdW} (van der Waals) and the desolvation free energy is composed of polar (ΔG_{pol}) and non-polar solvation energies (ΔG_{np}). Here, ΔG_{pol} was estimated from the Poisson-Boltzmann (PB) equation. The dielectric constant of the solute and solvent was set to 1.0 and 80.0, respectively. ΔG_{np} was estimated from the following Equation (6),

$$\Delta G_{np} = \gamma\ (SASA) + b \quad (6)$$

Here, *SASA* represented the solvent-accessible surface area and was estimated using the LCPO (linear combination of pairwise overlap) algorithm (Weiser et al., 1999) with a probe radius of 1.4 Å. The surface tension coefficient, γ and offset (b) value were set to 0.00542 kcal.mol^{-1}.Å^{-2} and 0.92 kcal.mol^{-1}, respectively (Gohlke et al., 2003). For the MM-PBSA calculation, 2000 snapshots selected from the last 300 ns trajectory with a frequency of 15 ps were employed.

The configurational entropy was calculated using the normal mode analysis (NMA) method (Karplus and Kushick, 1981; Rempe and Jónsson, 1998; Xu et al., 2011), and ∼300 snapshots were used in the calculation due to high computational cost. Each configuration was energy minimized in an implicit solvent (nmode_igb = 1) using a maximum of 50,000 steps, and a target root-mean-square gradient of 10^{-4} kcal mol^{-1}Å^{-1} via *mmpbsa_py_nabnmode* (Hawkins et al., 1995, 1996; Kar and Knecht, 2012a,b,c; Kar et al., 2013). NMA was applied to calculate the vibrational entropy (S_{vib}) using the following equation (Carlsson and Åqvist, 2006);

$$S_{vib} = R[\frac{x}{(e^x - 1)} - \ ln(1 - e^{-x})];\ x = \frac{h\nu}{kT} \quad (7)$$

Where *R* denotes the universal gas constant, *k* is the Boltzmann constant, and *T* is the absolute temperature. The Planck constant and the vibrational frequency are denoted as *h* and ν, respectively.

Further, the decomposition of the total binding free energy at the residue level was conducted by using the Molecular Mechanics Generalized Born Surface Area (MM-GBSA) scheme (Kar and Knecht, 2012a,b,c; Kar et al., 2013; Jonniya and Kar, 2020) as per the following equation.

$$\Delta G_{\text{residue}} = \ \Delta E_{\text{vdW}} + \ \Delta E_{\text{ele}} + \ \Delta \nu G_{\text{pol}} + \ \Delta G_{\text{np}} \quad (8)$$

The total contribution from each residue can also be defined as the sum of van der Waals (ΔE_{vdW}), electrostatic (ΔE_{ele}), polar (ΔG_{pol}), and non-polar (ΔG_{np}) terms. This method was proposed by Gohlke et al. (2003).

Finally, the Computational Alanine Scanning (CAS) was performed for some essential residues. This method yields the energy difference between the wild type and mutant (alanine) variants.

$$\Delta\Delta G_{\text{bind}} = \Delta G_{\text{mutant}} - \Delta G_{\text{wild}} \quad (9)$$

The basic principle involves in AS (alanine scanning) is the substitution of a residue with alanine that has an impact on the side-chain beyond C_β and not in the main-chain. *In vitro*, AS has been proven an advantageous mutagenesis method in finding hotspot residues in protein-protein interfaces. CAS is an excellent alternative approach to the *in vitro* experimental alanine scanning (Massova and Kollman, 1999; Moreira et al., 2007). Therefore, in this study, CAS was applied to illustrate further the importance of specific residues except alanine mentioned in the binding decomposition free energy of nsp16 of COVID-19.

RESULTS AND DISCUSSION

To explore the mechanism underlying the dimerization of nsp16 and nsp10 as well as the preferential binding of the cofactor SAM to the nsp16/nsp10 heterodimer compared to nsp16 (nsp16$_{\text{SAM}}$), a conformational free energy landscape (FEL) and binding free energy calculations were performed using the MD/MM-PBSA scheme.

Overall Structural Dynamic Features

To explore the thermodynamic stability of each system and to ensure the rationality of the sampling method, we monitored

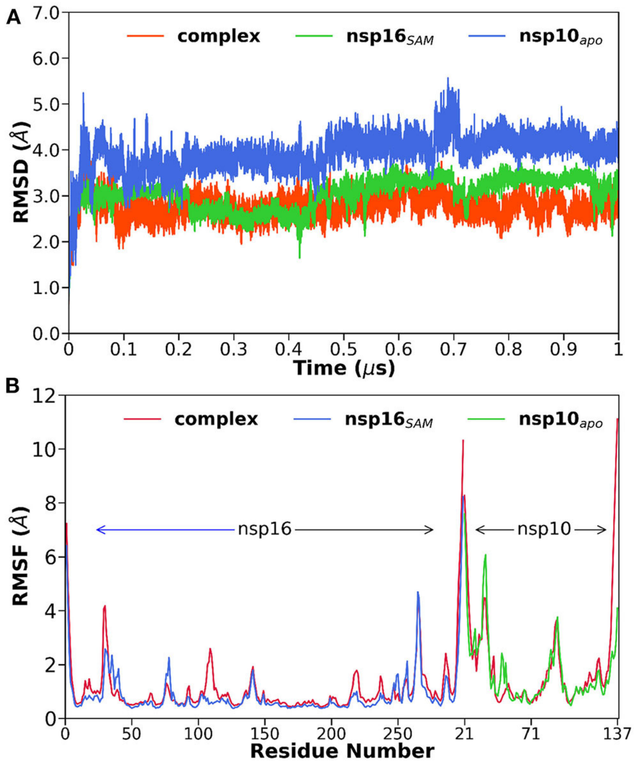


FIGURE 2 | (A) The root-mean-square deviations (RMSDs) of the backbone atoms relative to their initial coordinates as a function of simulation time, **(B)** the root-mean-square fluctuations (RMSFs) of C_α atoms for each residue in the complex and monomers of nsp16$_{SAM}$ and nsp10$_{apo}$.

TABLE 1 | The average value of protein backbone RMSD and solvent-accessible surface area (SASA) obtained from MD simulations.

System	RMSD (Å)	SASA (Å^2)
Complex	2.8 (0.1)	18928.1 (103.3)
nsp16$_{SAM}$	3.1 (0.3)	13467.3 (140.2)
nsp10$_{apo}$	4.0 (0.2)	6663.8 (145.7)

Standard deviations are given in parentheses.

the structural and energetic properties during the entire 1 μs production simulation. The time evolution of the root-mean-square deviations (RMSDs) of the protein backbone atoms, which reflects the stability of the system, were calculated relative to the initial configuration and shown in **Figure 2A**. It is worth mentioning here that a stable RMSD does not always provide stable energy profiles; hence we have also verified the potential and total energy of each system from the respective MD trajectory (data not shown).

The RMSD plots indicated that all the studied systems reached equilibrium after ∼100 ns. The average RMSD value varies between 2.8 and 4.0 Å for all systems (see **Table 1**). The highest and lowest deviations were obtained for nsp10$_{apo}$ and complex, respectively. An intermediate RMSD value of 3.1 Å was obtained for nsp16$_{SAM}$. Overall, the nsp16/nsp10/SAM complex showed more stability throughout the simulations than nsp16$_{SAM}$ or nsp10$_{apo}$. This suggests that the dimerization leads to the overall stabilization of the complex.

Moreover, we also calculated the temporal RMSDs of heavy atoms of the cofactor SAM (see **Supplementary Figure 1A**) and the backbone atoms of residues within 5 Å around SAM in the binding pocket (see **Supplementary Figure 1B**). It is evident from **Supplementary Figure 1B** that the RMSD values of SAM and the binding pocket for the complex were stable up to 850 ns of the simulation. After that, in the last 150 ns, a sudden increase in RMSD was observed. Overall, the average RMSD value was 0.7 Å and 0.8 Å for SAM and the binding pocket, respectively (see **Supplementary Table 1**). It indicates that SAM binds strongly onto the cofactor binding cavity of the heterodimer nsp16/nsp10. However, in the case of nsp16$_{SAM}$, the RMSD values of SAM and the binding cavity were relatively stable up to 100 ns. After that, the RMSD value increased by ∼3 times compared to the complex system (see **Supplementary Table 1** and **Supplementary Figure 1**). These high values of RMSDs suggest that SAM does not bind to nsp16 monomer (nsp16$_{SAM}$), which is in agreement with the experimental study by Chen and coworkers for SARS-CoV (Chen et al., 2011).

Additionally, three loops, namely 71–79, 100–108, and 130–148, which comprises the SAM binding pocket as shown in the crystal structure, were also analyzed and found to be stable for both complex and nsp16$_{SAM}$ (see **Supplementary Figures 2A–C**). However, the cap-binding groove of nsp16 in SARS-CoV, as well as SARS-CoV-2, is mainly composed of two flexible loops, comprised of residues 26–38 and 130–148, while in the case of flavivirus, the NS5 MTase was relatively stable with α-helices (A1, A2, and half of αD) along the cap-binding groove (Egloff et al., 2002). Among the two loops involved in the cap-binding, only the loop 26–38 exhibited differences in the complex and nsp16$_{SAM}$ system, as revealed from our simulations. The 26–38 loop is located near the interface of nsp16 and nsp10. On the other hand, the 130–148 loop is found near the SAM binding pocket, away from the nsp16/nsp10 interface. As shown in **Supplementary Table 1**, the average RMSD value of this loop was higher in nsp16$_{SAM}$ (3.1 Å) compared to the complex (1.4 Å). The time evolution of RMSD of the 26–38 loop and the corresponding potential mean force (PMF) were displayed in **Supplementary Figures 3A,B**. From **Supplementary Figure 3A**, it was observed that in the case of nsp16$_{SAM}$, the RMSD value of the loop 26–38 increased during the initial 300 ns and remained stable thereafter. On the other hand, in the complex case, the RMSD of the loop was found to increase during the initial 300 ns and gradually decreased up to 600 ns of the simulation and finally reached equilibrium. We computed the potential of mean force (PMF), taking the loop's RMSD as a reaction coordinate and shown in **Supplementary Figure 3B**. The primary low energy structure of the loop was observed at ∼3.2 Å for nsp16$_{SAM}$. However, for the complex system, the global energy minimum was found at a relatively lower RMSD (1.4 Å). Besides, we detected two secondary minima (∼1.9 Å, and ∼2.8 Å) for the 26–38 loop in the complex system, and the energy barriers of the adjacent states were ∼1.1 and 1.2 kcal/mol, respectively. Overall, our results suggest that the dimerization resulted in the loop rearrangement near the interface.

Next, we measured the root-mean-square fluctuations (RMSFs) of Cα atoms for all systems (see **Figure 2B**) to elucidate the effect of dimerization on the residual flexibility. It is evident from **Figure 2B** that all three systems displayed a similar trend in RMSF patterns. However, we found some dynamic fluctuations in different loop regions, including the N- and C-terminals. A relatively large fluctuation was observed for various loop regions located at residues 26–38, 74–80, 104–115, and 130–148. The loop regions 26–38 and 104–115, which contribute to the interface of the heterodimer, showed higher fluctuations in complex compared to nsp16SAM, which suggests that the dimerization leads to the loop rearrangements. In contrast, the 74–80 loop region, located near SAM binding pocket, stabilized after the heterodimerization. However, the cap-binding groove of the 130–148 loop region exhibited no differences in both complex and apo forms, which agrees with the RMSD analysis (**Supplementary Figure 2C**). Overall, it depicts the binding of nsp10 to nsp16 favors and stabilizes the binding pocket of SAM, leading to its high activity.

Finally, we also measured the solvent-accessible surface area (SASA) of all systems and reported in **Table 1**. The average SASA values were found to be 18928.1, 13467.3, and 6663.8 Å^2, for the complex, nsp16$_{SAM,}$ and nsp10$_{apo}$, respectively. This suggested that the dimerization of nsp16 and nsp10 covered ∼1,203 Å^2 surface area in total, indicating a very stable interaction. Our simulation results also favor an experimental study that showed the value of the solvent-exposed surface area for the heterodimer complex of nsp16/nsp10 as 19710 Å^2 (Rosas-Lemus et al., 2020).

Free Energy Landscape (FEL) of SAM Binding

The RMSD profile of SAM and the binding pocket (**Supplementary Figure 1**) for the complex and nsp16$_{SAM}$ systems suggested that SAM binds more strongly to the heterodimer nsp16/nsp10 than the monomer nsp16. The 2-dimensional free energy landscape (FEL) was generated for the complex and nsp16$_{SAM}$ systems to explore the strong and weakly bound states of the SAM molecule. The RMSD of heavy atoms of SAM and the center of mass (CoM) distance between SAM and nsp16, which characterizes the displacement of SAM from the substrate-binding pocket of nsp16, were used as reaction coordinates for the construction of FEL. The value of a CoM distance < 14.5 Å denotes the strong binding of SAM toward nsp16. The two-dimensional FEL of the complex (nsp16/nsp10/SAM) and monomer (nsp16$_{SAM}$) systems were shown in **Figures 3A,B**, and the corresponding positions of SAM in the binding cavity were depicted in **Figures 3C,D**. FEL of the complex and monomer systems suggested that the conformational state of SAM in two systems prefer disparate configurations. In the case of the heterodimeric complex system, the underlying FEL was characterized by a single global minimum (see **Figure 3A**), which corresponds to the strongly bound state of SAM. On the other hand, in the case of nsp16$_{SAM}$, the global free energy minimum was obtained at a CoM distance of ∼17.5 Å, which corresponds to a weakly or unbound state of SAM. The secondary minimum was detected at a CoM distance of ∼13.5 Å, which corresponds to a strongly bound state of SAM to nsp16$_{SAM}$. Overall, our results suggested that although nsp16 monomer (nsp16$_{SAM}$) favored two binding states, the weakly bound or unbound state is energetically more favorable than the strongly bound state. These results illustrated a stronger binding of SAM with the nsp16/nsp10 heterodimer than nsp16 alone (nsp16$_{SAM}$), which agrees with the experimental results obtained for SARS-CoV (Chen et al., 2011).

Principal Component Analysis of nsp16/nsp10

The principal component analysis (PCA) was carried out for both nsp16 and nsp10 when they are in complex and monomeric states. FELs of sub-units nsp16 (nsp16$_{complex}$) and nsp10 (nsp10$_{complex}$) in the complex were compared with the corresponding monomer simulations (nsp16$_{SAM}$ and nsp10$_{apo}$). Each eigenvector and eigenvalue was plotted in the decreasing order, as shown in **Supplementary Figure 4**. For all cases, the first few eigenvectors describe the collective motion of local fluctuations. Comparing the four systems indicated that the first few PCs that describe the properties of movements were not the same. The first two eigenvectors encapsulated for 42, 67, 46, and 70% of overall movements in nsp16$_{complex}$ (nsp16 in complex), nsp10$_{complex}$ (nsp10 in complex), nsp16$_{SAM}$, and nsp10$_{apo}$, respectively. Similarly, the first ten eigenvectors accounted for 80–90% of the total motions in all four systems.

Next, we constructed the 2-dimensional FEL at 300 K for each system using the first two principal components (PC1 and PC2) as reaction coordinates and shown in **Figures 4A–D**. The sampling space of four systems was different, as evident from **Figure 4**. The conformational sampling space of nsp16$_{complex}$ (nsp16 in nsp16/nsp10/SAM), depicted in **Figure 4A**, was found to be restricted compared to the other three systems, which sampled more expansive conformational space. From **Figure 4A**, we observed a global minimum of 62.7% occupancy and a secondary minimum of 37.3% occupancy, which suggested that the nsp16$_{complex}$ system is more stabilized in the presence of nsp10. In the monomeric form of nsp16$_{SAM,}$ a wider conformational basin was sampled (see **Figure 4C**), and the energy barriers between adjacent minima were ∼1.5 kcal/mol. On the other hand, nsp10 sampled more phase space than nsp16 in the complex as well in its monomer apo form, as shown in **Figures 4B,D**. However, nsp10$_{apo}$ has three distinct conformations with an energy barrier of 4.0 kcal/mol suggesting structural disparity in both nsp10$_{complex}$ and nsp10$_{apo}$ as compared to nsp16.

Residual Network Analysis

The Network Analysis of Protein Structure (NAPS) server provided a visual examination of sub-network based on the physicochemical properties of the protein residues to find more details about the interaction between nsp16 and nsp10 (see

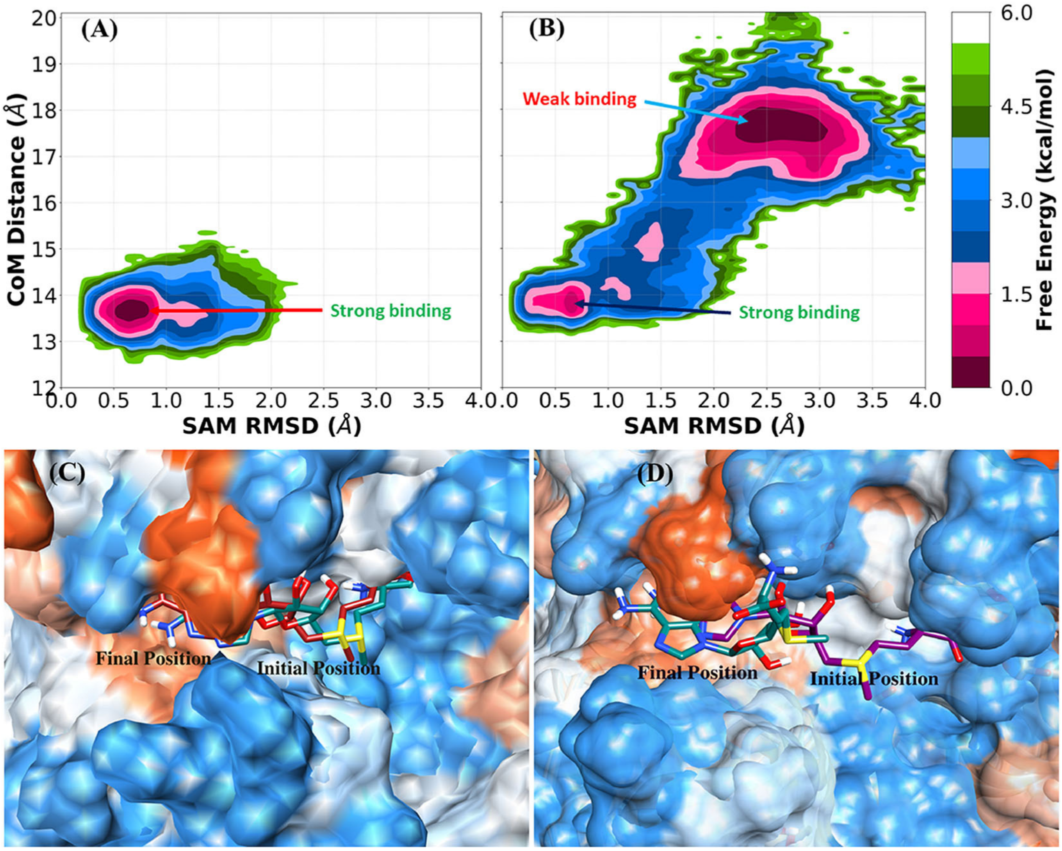


FIGURE 3 | Free energy landscape (FEL) and representative SAM position in the complex (nsp16/nsp10/SAM) and nsp16$_{SAM}$ (nsp16/SAM). **(A)** FEL of SAM bound with nsp16 in the complex, **(B)** FEL of SAM bound with monomer nsp16$_{SAM}$. Representative structures of **(C)** SAM with nsp16 in the complex and **(D)** SAM with monomer nsp16$_{SAM}$.

Figure 5). Herein, we explored the 3D interaction network of hydrophobic residues of nsp16 and nsp10. The results showed that the number of hydrophobic interaction networks between nsp16 and nsp10 was initially very high, and after 100 ns, the number of networks reduced to ∼2–3. The hydrophobic interaction networks, such as (V104, **A71**) and (P80, **V42**), were found as strong and stable contacts throughout the simulation time (see **Supplementary Table 2**). We also calculated the total number of interactions and noticed that the number of interaction networks was more or less conserved throughout the first 500 ns (see **Supplementary Figure 5**). In general, it can be concluded from our analyses that nsp16 and nsp10 interact with a very high affinity.

Binding Free Energy of SAM Bound to nsp16/nsp10 Heterodimer and nsp16 Alone

The differences in binding affinity of the SAM molecule toward nsp16 alone (nsp16$_{SAM}$) and nsp16/nsp10 heterodimer (complex) were calculated by utilizing the MM-PBSA scheme. The MM-PBSA scheme provides various components contributing to the total binding energy (ΔG_{bind}), such as van der Waals interactions (ΔE_{vdW}), electrostatic interactions (ΔE_{ele}), polar solvation energy (ΔG_{pol}), non-polar solvation free energy (ΔG_{np}), and configurational entropy ($T\Delta S$). All these components of the binding free energy of SAM to nsp16$_{SAM}$ or nsp16/nsp10 (complex) were shown in **Table 2**. The binding free energy of SAM to the heterodimer nsp16/nsp10 was estimated as −6.8 (± 0.9) kcal/mol, which agrees with the experimental result (−7.4 ± 0.1 kcal/mol) (Lin S. et al., 2020). On the other hand, SAM was found to bind very weakly ($\Delta G_{bind} = -0.6 \pm 0.8$ kcal/mol) to nsp16 monomer (nsp16$_{SAM}$), as evident from **Table 2**. This is consistent with **Figures 3A,B**. A similar mode of SAM binding was observed for SARS-CoV nsp16/nsp10 or nsp16 alone. For both cases, the intermolecular van der Waals interactions (ΔE_{vdW}), electrostatic interactions (ΔE_{ele}), and non-polar solvation free energy (ΔG_{np}) favored the binding of SAM. In contrast, polar solvation energy (ΔG_{pol}) and entropy ($T\Delta S$) disfavored the complexation. The electrostatic interaction energy was more favorable than the van der Waals interactions for both cases. In nsp16$_{SAM}$, although ΔE_{ele} was much more favorable (−76.8 kcal/mol) than ΔE_{vdW} (−28.6 kcal/mol). However, the disfavorable polar solvation energy ($\Delta G_{pol} = 84.7$ kcal/mol) overcompensated the intermolecular electrostatic interaction energy. Hence, the overall polar contribution ($\Delta E_{ele} + \Delta G_{pol}$) was found to be unfavorable (7.9 kcal/mol) to the SAM binding. In contrast, the overall non-polar contribution

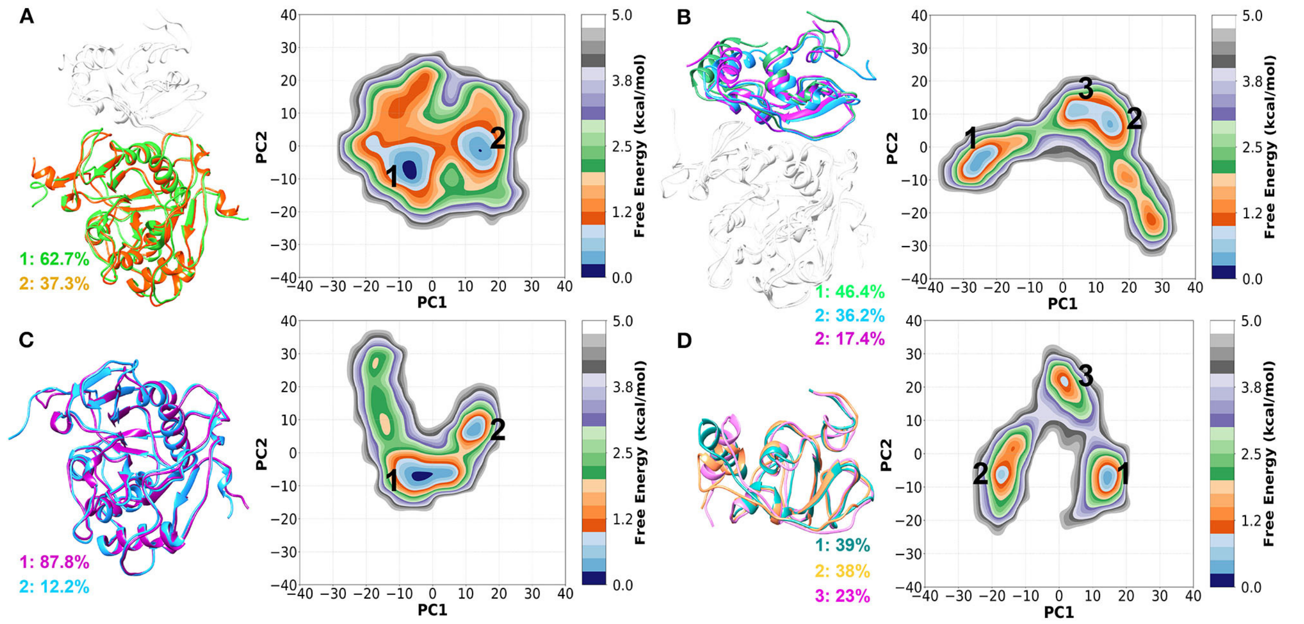


FIGURE 4 | Two-dimensional FEL generated by projecting the first two principal components, PC1 and PC2 for **(A)** nsp16$_{\mathrm{complex}}$ (nsp16 in nsp16/nsp10/SAM), **(B)** nsp10$_{\mathrm{complex}}$ (nsp10 in nsp16/nsp10/SAM), **(C)** nsp16$_{\mathrm{SAM}}$ (nsp16/SAM), and **(D)** nsp10$_{\mathrm{apo}}$. The representative structures are shown on the left panel.

($\Delta E_{\mathrm{vdW}} + \Delta G_{\mathrm{np}}$) was found to be −32.1 kcal/mol. It implies that hydrophobic interactions drive the binding. In the case of the complex system (nsp16/nsp10/SAM), the binding affinity of SAM to nsp16 increases, as evident from **Table 2**. Here, the electrostatic interactions energy ($\Delta E_{\mathrm{ele}} = -181.1$ kcal/mol) was more favorable than the van der Waals energy ($\Delta E_{\mathrm{vdW}} = -40.4$ kcal/mol). The total polar contributions ($\Delta E_{\mathrm{ele}} + \Delta G_{\mathrm{pol}}$) value was found to be favorable (−2.1 kcal/mol) to the binding of SAM, which is in contrast to what had been observed for nsp16$_{\mathrm{SAM}}$. Further, the overall polar energy was less favorable than the net non-polar contribution ($\Delta E_{\mathrm{vdW}} + \Delta G_{\mathrm{np}} = -44.5$ kcal/mol). Hence, in the complex, the main force behind the binding of SAM to nsp16 is hydrophobic interactions. Overall, the binding free energy analysis showed that the binding affinity of the SAM molecule to nsp16 increases with the presence of nsp10. This implies that nsp10 acts as a stimulator to bind SAM to nsp16/nsp10 of SARS-CoV-2, which agrees with the experiment (Viswanathan et al., 2020). Wang and coworkers obtained a similar result for SARS-CoV and MERS-CoV (Wang et al., 2015).

Further, to explore the significant residues of nsp16 in the binding of SAM and to evaluate the differences in the complex (nsp16/nsp10) and monomer (nsp16$_{\mathrm{SAM}}$), the per-residue decomposition of the binding free energy was performed using the MM-GBSA approach. All the residues having > 1.5 kcal/mol of energetic contributions were considered important and shown in **Figure 6**. As seen in **Table 3**, the number of amino acids contributing to binding with SAM is high in the complex as compared to nsp16$_{\mathrm{SAM}}$. Residues such as Leu100, Asp99, Cys115, Met131, Phe149, and Asp114 are common in cases of complex and nsp16$_{\mathrm{SAM}}$. Apart from these residues, additional residues, such as Asn43, Tyr47, Gly71, Tyr132, and Ser74, also played a significant role in the binding of SAM to the complex heterodimer. These results are consistent with the higher binding affinity of SAM in the complex than the nsp16 monomer. All these residues were also considered as SAM-engaging in an experimental study by Lin S. et al. (2020). It is worth noting here that all these SAM-interacting residues are conserved both in SARS-CoV and MERS-CoV, suggesting a conserved binding mode of SAM in these viruses. The conserved SAM-binding mechanism to nsp16/nsp10 also opens the possibility of developing a pan-CoV inhibitor by targeting this SAM-binding pocket.

Hydrogen Bond Interaction Between nsp16 and SAM

Next, we investigated the hydrogen bond (H-bond) interactions between SAM and nsp16 in the complex and apo forms (see **Table 4**). The % occupancy calculated from the respective MD trajectories reflected the stability of H-bonds. As seen in **Table 4**, residues with > 50% H-bond occupancy in the MD simulations were recognized for the complex (nsp16/nsp10) compared to monomer nsp16$_{\mathrm{SAM}}$. The residues, including Asp130, Tyr47, Gly71, and Asp99, showed more H-bond stability in the complex. Overall, a stable H-bond is formed between nitrogen and oxygen atoms of SAM with Asp130, Tyr47, Gly71, and Asp99 residues of nsp16, respectively. Hence, the identified significant residues like Asp130, Tyr47, Gly71, and Asp99 of nsp16 may aid in the development of SAM competitive inhibitors.

The changes in the distance of the atoms forming H-bond between SAM and nsp16 were also determined and shown in **Supplementary Figure 6**. In the complex, the distance between

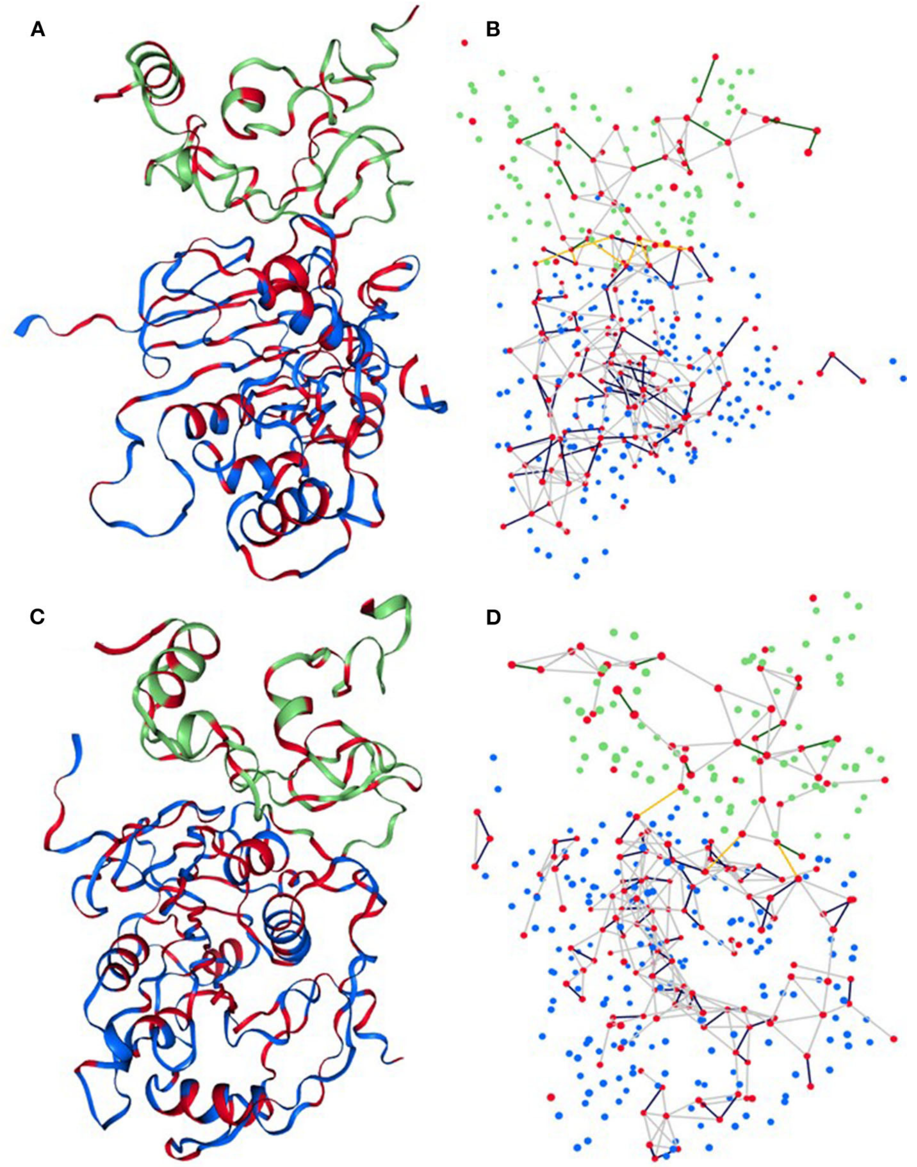


FIGURE 5 | Sub-network representation of network 3D view of **(A,B)** initial structure and **(C,D)** final structure of the complex. Each residual C_α atom is represented by a sphere where the red sphere indicates hydrophobic residues. Blue and green color spheres correspond to residues of nsp16 and nsp10, respectively. The yellow line edges represent the contact network of hydrophobic residues between nsp16 and nsp10.

TABLE 2 | Energetic components of the total binding free energy of SAM bound to the heterodimer nsp16/nsp10 (complex), and nsp16 alone ($nsp16_{SAM}$).

System	ΔE_{vdW}	ΔE_{elec}	ΔG_{pol}	ΔG_{np}	ΔH	$\Delta -TS$	ΔG_{bind}
Complex	−40.4 (0.1)	−181.1 (0.5)	179.0 (0.4)	−4.1 (0.0)	−46.6 (0.2)	39.8 (0.9)	−6.8 (0.9)
$nsp16_{SAM}$	−28.6 (0.1)	−76.8 (0.6)	84.7 (0.6)	−3.5 (0.0)	−24.2 (0.1)	23.6 (0.8)	−0.6 (0.8)

All values are given in kcal/mol. The standard error of the mean is provided in parentheses.

the oxygen atom of Asp130, Tyr47, and Gly71 of nsp16 and the nitrogen atom of SAM illustrated the average distance of 3.01, 2.89, and 3.01 Å, respectively, suggesting strong H-bond interactions. In the case of $nsp16_{SAM}$, the average distances for these atoms were 12.48, 14.64, and 11.06 Å, respectively, indicating a lack of formation of H-bonds in the nsp16 monomer. However, for Asp99, the distance between its oxygen atom (OD1 and OD2) and the oxygen atom (O3) of SAM illustrated

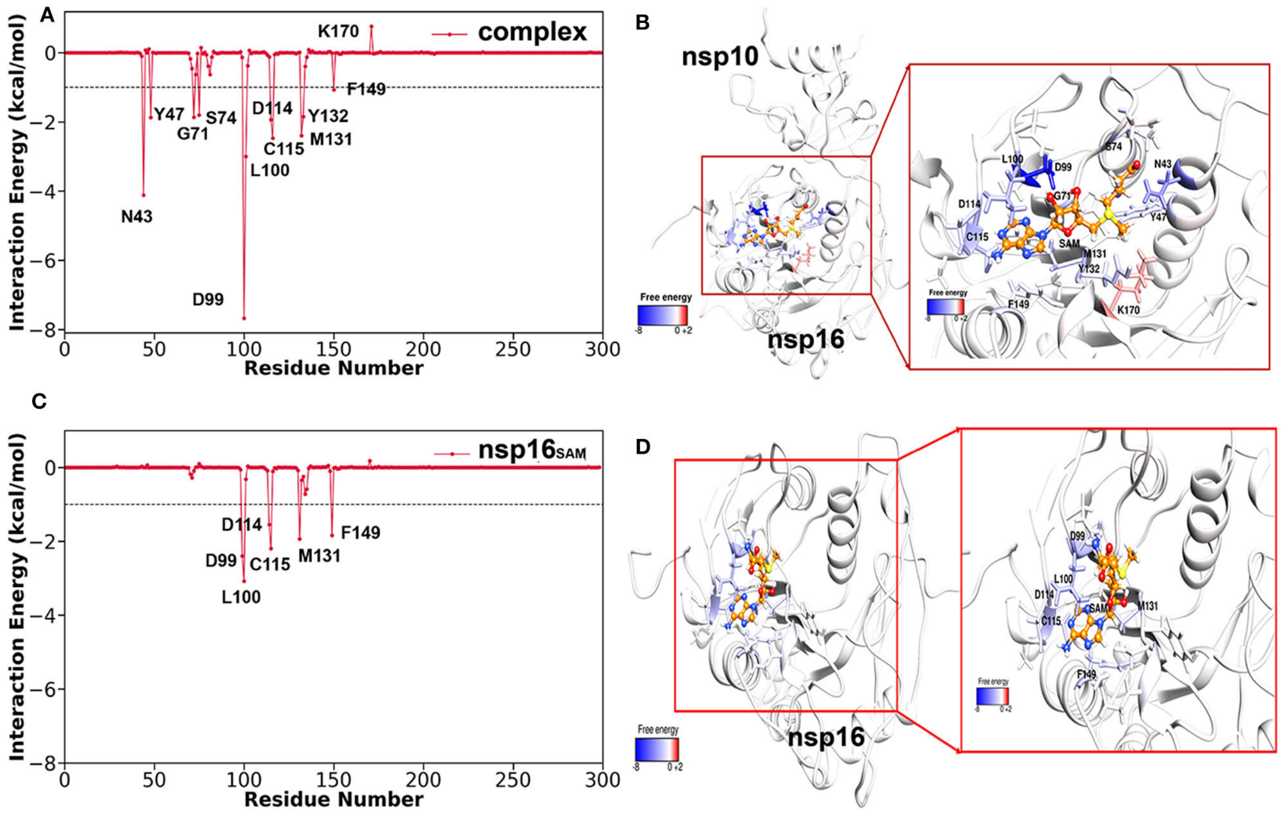


FIGURE 6 | Per-residue decomposition free energy of nsp16 for the binding of SAM (in the presence of nsp10 and without nsp10) and respective binding pocket drawn from MD snapshots. **(A,B)** nsp16 with SAM in the presence of nsp10, **(C,D)** nsp16 with SAM without nsp10.

an average distance of 2.9 Å and 3.2 Å for the complex and monomer, respectively. In contrast, the average distance between Asp99 (OD2) and SAM (O2) was higher for monomer (4.2 Å) than the complex (3.1Å). All these results were consistent with the occupancy analysis. Further, they emphasized that Asp130, Tyr47, and Gly71 of nsp16 were key residues in the binding of SAM and for the 2′-O-MTase activity of nsp16.

Energetics of nsp16/nsp10 Complexation

To evaluate further the binding free energy of the heterodimer (nsp16/nsp10), ΔG_{bind} was estimated using the MD/MM-PBSA approach and reported in **Table 5**. From **Table 5**, the binding free energy (ΔG_{bind}) between nsp16 and nsp10 was found to be −47.4 kcal/mol. The various components of the total binding free energy indicate that the intermolecular van der Waals interactions (ΔE_{vdW}), electrostatic interactions (ΔE_{ele}), and non-polar solvation free energy (ΔG_{np}) favors the binding of nsp10 and nsp16. While polar solvation free energy (ΔG_{pol}) disfavor the binding. The electrostatic interactions (ΔE_{ele}) energy is higher (−429.4 kcal/mol) than the van der Waal interactions (ΔE_{vdW}) (−90.4 kcal/mol) (see **Supplementary Figure 7**). However, the disfavouring components of polar solvation energy (ΔG_{pol}) compensate for the ΔE_{ele} being a value of 481.7 kcal/mol. Hence, the total polar interactions ($\Delta E_{\text{ele}} + \Delta G_{\text{pol}}$) disfavor the binding between nsp10 and nsp16 with a value of 52.3 kcal/mol. In contrast, the total non-polar contributions ($\Delta E_{\text{vdW}} + \Delta G_{\text{np}}$) favor the binding between nsp16 and nsp10 (−100.0 kcal/mol). Therefore, the binding between nsp10 and nsp16 is mainly driven by hydrophobic interactions. Overall, the binding free energy analysis depicts, the binding affinity of the SAM in the complex nsp16/nsp10 ($\Delta G_{\text{bind}} = -46.6$ kcal/mol) is similar to that between nsp10 and nsp16 ($\Delta G_{\text{bind}} = -47.4$ kcal/mol) of SARS-CoV-2. These observations were in agreement with its similar homology virus, SARS-CoV (2002). A similar binding affinity was seen between nsp10 and nsp16, as well as between SAM and the nsp16/nsp10 complex of SARS-CoV (Chen et al., 2011).

To further investigate the critical residues involved in the binding between nsp10 and nsp16, the binding free energy was decomposed at the individual residue level using the MM-GBSA approach and shown in **Figure 7**. Here, we considered only residues having the energy value ≥ 1.0 kcal/mol and listed in **Supplementary Table 3**. The key residues from nsp10 include Leu45, Ala71, Val42, Met44, Tyr96, Gly69, Thr47, Arg78, Gly70, Gly94, and Pro59. Similarly, the critical residues from nsp10 include Ile40, Val104, Ala83, Val78, Met247, Val44, Gln87, Arg86, Lys76, Lys38, Val84, and Met41. It indicates that hydrophobic interactions significantly control the binding between nsp10 and nsp16. Our findings agree with a recent experimental study (Lin S. et al., 2020) where it was reported that a total of 31

TABLE 3 | Per-residue decomposition of the total binding free energy (kcal/mol) of SAM to the complex (nsp16/nsp10) and monomer (nsp16$_{SAM}$) systems.

Residue	T_{vdW}	T_{ele}	T_{pol}	T_{np}	T_{back}	T_{side}	T_{total}
COMPLEX							
Asp99	0.71	−29.08	20.84	−0.14	−0.73	−6.94	−7.67
Asn43	−0.01	−9.90	5.89	−0.12	−0.30	−3.84	−4.14
Leu100	−2.42	−0.19	−0.10	−0.28	−0.87	−2.12	−2.99
Cys115	−0.74	−2.01	0.33	−0.05	−1.21	−1.26	−2.47
Met131	−2.72	0.56	−0.12	−0.13	−0.47	−1.94	−2.41
Asp114	−0.30	−8.98	7.41	−0.07	−0.77	−1.17	−1.94
Tyr47	0.07	−5.46	3.56	−0.04	−0.01	−1.86	−1.87
Gly71	−1.35	−3.27	2.84	−0.08	−1.36	−0.50	−1.86
Tyr132	−2.15	−2.23	2.86	−0.32	−0.93	−0.91	−1.84
Ser74	−0.95	−4.31	3.54	−0.08	−0.45	−1.35	−1.80
Phe149	−1.20	0.12	0.12	−0.13	−0.16	−0.93	−1.09
nsp16$_{SAM}$							
Leu100	−3.04	1.61	−1.17	−0.47	−0.18	−2.89	−3.07
Asp99	0.62	−19.83	16.98	−0.16	−0.15	−2.24	−2.39
Cys115	−0.70	−1.73	0.30	−0.06	−1.16	−1.03	−2.19
Met131	−1.72	−0.02	−0.05	−0.15	−0.55	−1.39	−1.94
Phe149	−1.86	−0.33	0.54	−0.18	−0.20	−1.63	−1.83
Asp114	−0.28	−7.47	6.32	−0.10	−0.73	−0.80	−1.53

The van der Waals (T_{vdW}), electrostatic interactions (T_{ele}), polar (T_{pol}), and non-polar solvation energy (T_{np}) and the total residual energy (T_{total}) are shown. T_{back} and T_{side} represent the backbone and side-chain parts, respectively.

TABLE 4 | Hydrogen bonds formed between SAM and receptor (nsp16/nsp10 or nsp16).

Binding couples		Molecular dynamics	
Acceptor	**Donor…H**	**Distance (Å)**	**Occupancy[a] (%)**
COMPLEX			
Asp130@OD2	SAM@N…HN1	2.84	71.69
SAM@N	Tyr47@OH…HH	2.79	64.95
Gly71@O	SAM@N…HN2	2.86	55.16
Asp99@OD2	SAM@O2′…HO2′	2.62	53.90
Asp99@OD1	SAM@O3′…HO3′	2.66	53.80
Asp99@OD2	SAM@O3′…HO3′	2.66	46.27
SAM@HN1	Tyr47@OH…HH	2.78	45.41
Asp99@OD1	SAM@O2′…HO2′	2.63	45.15
SAM@O	Asn43@ND2…HD22	2.84	42.20
SAM@OXT	Asn43@ND2…HD22	2.84	37.76
SAM@N1	Cys115@N…H	2.92	33.86
Asp114@OD2	SAM@N6…HN61	2.80	18.10
Asp114@OD1	SAM@N6…HN61	2.80	15.11
SAM@O2′	Asn101@ND2…HD22	2.89	12.21
nsp16$_{SAM}$			
Asp99@OD1	SAM@O3′…HO3′	2.64	48.80
Asp99@OD2	SAM@O3′…HO3′	2.63	45.78
SAM@N1	Cys115@N…H	2.92	28.66
Asp114@OD2	SAM@N6…HN61	2.80	26.44
Asp114@OD1	SAM@N6…HN61	2.80	24.80
Asp114@OD2	SAM@N6…HN62	2.80	19.03
Asp114@OD1	SAM@N6…HN62	2.80	17.95
Asp99@OD1	SAM@O2′…HO2′	2.65	16.42
Asp99@OD2	SAM@O2′…HO2′	2.65	14.70
SAM@N	Tyr47@OH…HH	2.92	11.31

[a]*only H-bonds with more than 10% occupancy are listed.*
The corresponding average distance and percentage occupancy are also provided.

TABLE 5 | Energetic components of the dimerization energy (G_{bind}) between nsp16 and nsp10.

System	ΔE_{vdW}	ΔE_{elec}	ΔG_{pol}	ΔG_{np}	ΔG_{bind}	ΔG^{a}_{bind}
WT	−90.4 (0.2)	−429.4 (3.0)	481.7 (3.2)	−9.6 (0.0)	−47.7 (0.4)	
I40A	−86.0 (0.2)	−429.2 (3.0)	480.5 (3.2)	−9.5 (0.0)	−44.2 (0.4)	3.5 (1.2)
V44A	−88.0 (0.2)	−429.2 (3.0)	481.4 (3.2)	−9.6 (0.0)	−45.5 (0.4)	2.1 (1.5)
V78A	−87.5 (0.2)	−429.5 (3.0)	481.0 (3.2)	−9.5 (0.0)	−45.5 (0.4)	2.2 (1.5)
R86A	−87.5 (0.2)	−522.6 (3.0)	574.5 (3.3)	−9.5 (0.0)	−45.1 (0.4)	2.6 (2.5)
Q87A	−86.8 (0.2)	−416.9 (3.0)	466.8 (3.2)	−9.4 (0.0)	−46.2 (0.4)	1.5 (1.6)
V104A	−86.9 (0.2)	−429.7 (3.0)	481.11 (3.2)	−9.4 (0.0)	−44.9 (0.4)	2.8 (1.5)
M247A	−87.1 (0.2)	−428.5 (3.0)	479.0 (3.2)	−9.6 (0.0)	46.1 (0.4)	1.5 (0.8)

Along with the Computational Alanine-Scanning (CAS) mutagenesis results of the complex nsp16/nsp10. The binding energy is given in kcal/mol, and the standard error of the mean is provided in parentheses.
$\Delta\Delta G^{a}_{bind} = \Delta G^{mut}_{bind} - \Delta G^{WT}_{bind}$.

residues (Asn40, Val42, Lys43, Met44, Leu45, Cys46, Thr47, Pro59, Gly70, Ala71, Cys77, Arg78, Lys93, Gly94, and Tyr96 in nsp10 and Lys38, Gly39, Ile40, Met41, Val44, Lys76, Val78, Pro80, Ala83, Arg86, Gln87, Val104, Ser105, Asp106, Leu244, and Met247 in nsp16) mediates the intimate nsp16-nsp10 interaction. The per-residue decomposition of the binding free energy was further validated by other computational approaches, such as FTmap (Kozakov et al., 2015). The final complex structure obtained from the MD trajectory was analyzed using FTmap, and Gln87, Arg86, Val104, Met247 of nsp16, and Leu45, Met44, Pro59, Arg78, Tyr96 of nsp10 were identified as the hotspot residues, which again agrees with the MM-GBSA analysis.

Interaction Analysis Between nsp16 and nsp10

In the above finding of binding free energy analysis, we have seen that nsp10 acts as a stimulator in the binding of the SAM to nsp16. Hence, to further explore the binding interactions between nsp16 and nsp10, H-bond and hydrophobicity analysis was calculated. As seen in **Table 6**, the H-bond occupancy reflects the stability of H-bond formation between protein-protein in the MD simulations. The strong H-bonds were formed between residues Ala83 (nsp16) and Tyr96 (nsp10) and between Gln87 (nsp16) and Leu45 (nsp10). Other residues, including Asp106 (nsp16), formed two H-bonds with Ala71 and Gly94 of nsp10, Lys38 (nsp16) to Lys43 (nsp10), and Ser105 (nsp16) to Lys93 (nsp10). Hydrophobic interactions also played an important role in protein-protein or protein-ligand interactions (Jonniya et al., 2020; Roy et al., 2020; Sk et al., 2020d). Different H-bonds and hydrophobic interactions from the stable structure of the

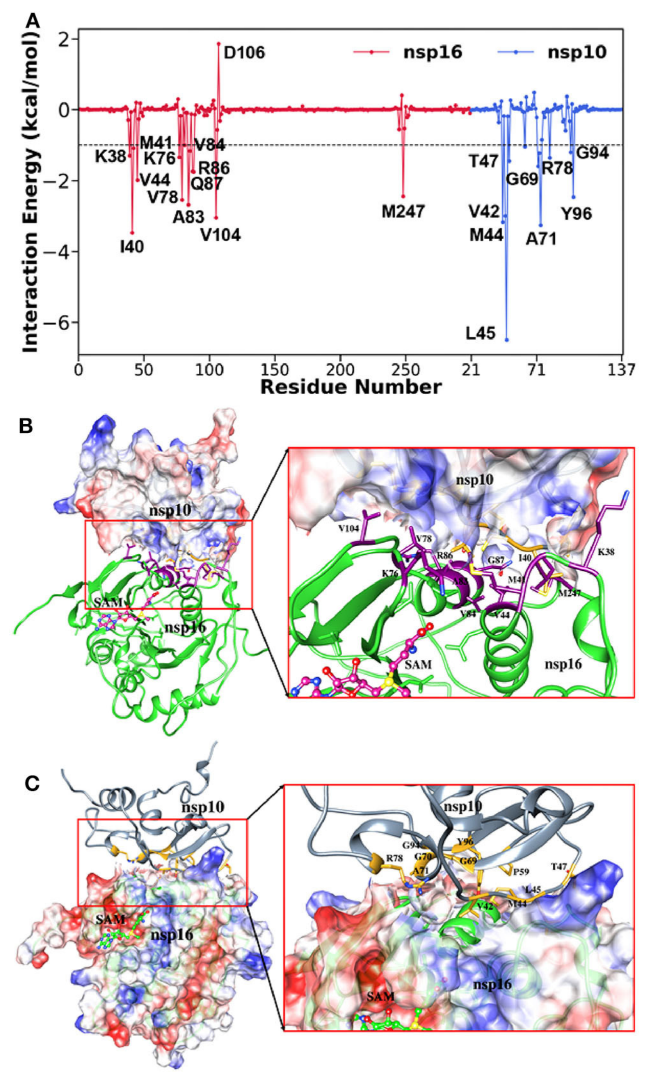


FIGURE 7 | (A) Decomposition of Δ*G* on a per-residue basis for the complex, **(B)** corresponding residual position of nsp16 in complex, and **(C)** residual position of nsp10 in the complex.

TABLE 6 | The hydrogen bonds formed between nsp16 and nsp10* in the complex and the corresponding average distance and percent determined using the production trajectories in the MD simulations.

Binding couples		Molecular dynamics	
Acceptor	Donor…H	Distance (Å)	Occupancy (%)
Ala83@O	**Tyr96@OH…HH**	2.73	90.80
Leu45@O	Gln87@NE2…HE21	2.85	67.83
Asp106@OD1	**Ala71@N…H**	2.87	29.23
Asp106@OD1	**Gly94@N…H**	2.84	24.23
Lys38@O	**Lys43@NZ…HZ2**	2.80	25.28
Lys38@O	**Lys43@NZ…HZ3**	2.80	25.01
Lys38@O	**Lys43@NZ…HZ1**	2.80	23.63
Asp106@OD2	**Gly94@N…H**	2.84	17.58
Ser105@O	**Lys93@NZ…HZ3**	2.79	12.65
Ser105@O	**Lys93@NZ…HZ2**	2.79	12.35
Ser105@O	**Lys93@NZ…HZ1**	2.79	12.25

**Bold letters belong to the nsp10 structure. Only H-bonds with more than 10% occupancy are listed.*

nsp16-nsp10 obtained from the MD simulations were plotted via Ligplot (Wallace et al., 1995) and shown in **Figure 8**.

The percentage occupancy of the residual contacts with the cut-off 3.9 Å for the protein-protein is also listed in **Supplementary Table 4**. Overall, these results highlighted the significant residues forming strong interactions between nsp16 and nsp10, which may aid in the development and design of inhibitors that could block these protein-protein interactions by inhibiting the 2′-O-MTase activity.

Computational Alanine Scanning

The changes in the binding free energy were computed as in Equation (4) (see Method section) after replacing the residue of WT to alanine. The impact of a single mutation in the protein-protein interaction is mainly reflected by the more positive value of the $\Delta\Delta G_{bind}$. In our study, we have conducted CAS for nsp16 by considering those residues with the decomposition free energy $>$ 1.5 kcal/mol, as given in **Supplementary Table 3**. The binding free energy components calculated from the CAS mutagenesis for residues I40A, V104A, R86A, V78A, V44A, M247A, and Q87A of nsp16 are listed in **Table 5** and compared with the WT. As seen in **Supplementary Table 3**, although the ΔG_{bind} of WT is higher than mutants, different energy components of mutants follow the same trend as WT. The ΔE_{ele} is higher than $\Delta E_{\mathrm{vdW},}$ but disfavouring polar solvation energy ΔG_{pol} compensates ΔE_{ele}. As shown in **Figure 9A**, for all the systems, the binding free energy is mainly coming from the hydrophobic interactions, where the total non-polar energy ($\Delta E_{\mathrm{vdW}} + \Delta G_{\mathrm{np}}$) is higher than the total polar energy ($\Delta E_{\mathrm{ele}} + \Delta G_{\mathrm{pol}}$). **Figure 9B** reflects significant residues in the protein interface calculated from CAS mutagenesis. It shows that $\Delta\Delta G_{bind}$ for mutants, I40A, V104A, and R86A are comparatively high, suggesting that primarily these residues play a significant role in the heterodimer formation, which is in accordance with the decomposition of energy. The CAS mutagenesis results depict that in addition to hydrophobic residues I40 and V104, the hydrophilic residue R86 also plays a vital role in the binding of nsp10-nsp16.

Temperature and Salt Concentration Effect on the Binding Affinity

Next, we investigated the effect of temperature and salt concentration in the binding of nsp16/nsp10 and nsp16/nsp10/SAM using the MM-PBSA method. This method presents itself to be a perfect model to study the effect of conformational dynamics of the same protein evolved in different environmental conditions to the fate of the ligand, ultimately decided by the binding affinity or dissociation.

Supplementary Tables 5, **6** displayed the binding free energy components for nsp16-nsp10 and nsp16/nsp10/SAM, respectively, at two different temperatures. As shown in

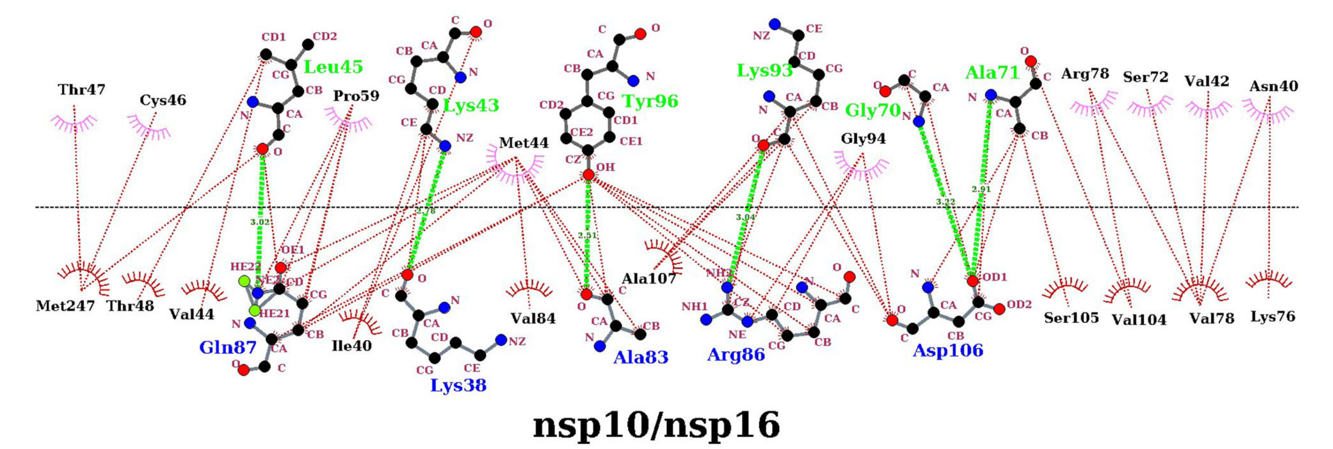


FIGURE 8 | Protein-protein interaction diagrams for the nsp16/nsp10 dimerization. The upper panel corresponds to nsp10, and the lower panel corresponds to nsp16 in the complex. The plot was generated with the dimplot module of LigPlot+. Hydrogen bonds are shown as a lime green dotted line, and hydrophobic bonds are represented in red.

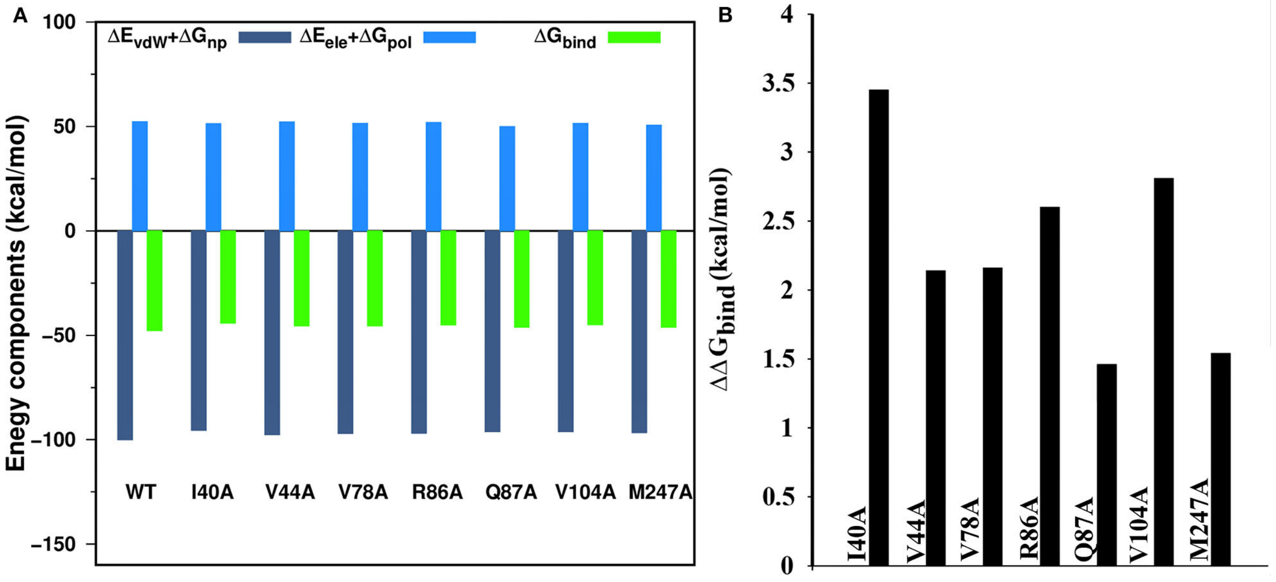


FIGURE 9 | (A) Binding free energy components of the wild type and seven mutations (alanine scanning) in the complex, **(B)** alanine scanning mutagenesis analysis of complex.

Supplementary Table 5, the predicted binding free energies between nsp10 and nsp16 (ΔG_{bind}) was found to be −47.7 kcal/mol, and −52.2 kcal/mol at 300 and 310 K, respectively. It suggests that nsp16 binds more strongly to nsp10 at 310 K than 300 K. It can also be noted from **Supplementary Table 5** that at the elevated temperature (310 K), the magnitude of total polar interactions ($\Delta E_{\text{ele}} + \Delta G_{\text{pol}}$) decreases (52.3 to 49.6 kcal/mol) and oppose less unfavorably to the protein-protein association. At a higher temperature, the non-polar interactions ($\Delta E_{\text{vdW}} + \Delta G_{\text{np}}$) decreased from −100 to −102 kcal/mol and contributed more favorably to the association of nsp16-nsp10. Similarly, **Supplementary Table 6** suggested that the estimated binding free energies (without entropy) of the cofactor SAM with nsp16/nsp10 (ΔG_{bind}) were −46.6 and −13.2 kcal/mol at 300 and 310 K, respectively. It was further evident from **Supplementary Table 6** that after increasing the temperature, the binding affinity of SAM toward nsp16/nsp10 decreased. At room temperature (300 K) both total polar ($\Delta E_{\text{ele}} + \Delta G_{\text{pol}} = -2.1$ kcal/mol) and non-polar ($\Delta E_{\text{vdW}} + \Delta G_{\text{np}} = -44.5$ kcal/mol) interactions favored the binding of SAM with nsp16. After increasing the temperature to 310 K, the total polar interaction ($\Delta E_{\text{ele}} + \Delta G_{\text{pol}}$) became unfavorable to the binding of SAM to nsp16/nsp10. However, the total non-polar contribution ($\Delta E_{\text{vdW}} + \Delta G_{\text{np}}$) still favored the binding. As compared to room temperature, the strength of the non-polar ($\Delta E_{\mathrm{vdW}} + \Delta G_{\mathrm{np}}$) interactions decreases significantly (−44.5 to −23.5 kcal/mol). At 310 K, we have noticed that up to 400 ns, SAM binds strongly to nsp16. After that, SAM got detached from the binding cavity (see **Supplementary Figure 8**).

The salt concentration also influenced the protein-protein interactions as well as the binding of SAM to nsp16/nsp10, which are evident from **Supplementary Tables 5**, **6**. As shown in **Supplementary Table 5**, the predicted binding free energies of nsp10 with nsp16 (ΔG_{bind}) were found to be −44.8 and −37.7 kcal/mol, for the salt concentration of 0.15 and 0.25 M, respectively. After increasing the salt concentration, it was observed that the total polar ($\Delta E_{\mathrm{ele}} + \Delta G_{\mathrm{pol}}$) and non-polar interactions ($\Delta E_{\mathrm{vdW}} + \Delta G_{\mathrm{np}}$) increased from 54.1 to 55.3 kcal/mol and −98.9 to −93 kcal/mol, respectively. The rise in unfavorable interactions caused the destabilization of the protein-protein interactions. Similarly, the binding of SAM to the heterodimer was found to be affected by the increased salt concentration (see **Supplementary Table 6**). The estimated binding free energy of SAM (ΔG_{bind}) was −23.7 and −22.1 kcal/mol for the salt concentration of 0.15 and 0.25 M, respectively. These values are significantly higher than what was obtained for the neutral salt concentration (−46.6 kcal/mol). ΔE_{vdW} and ΔE_{elec} were found to contribute less favorably to the binding of SAM to nsp16/nsp10 in comparison to neutral simulations. A similar trend was observed when the temperature was increased from 300 to 310 K. The time evolution of CoM distance between nsp16 and SAM was displayed in **Supplementary Figure 9**. It is evident from **Supplementary Figure 9** that with the increased salt concentration, the CoM distance between nsp16 and SAM increased and remained stable around 17 Å and 18 Å for 0.15 and 0.25 M, respectively. This implies that the cofactor SAM moves away from the binding cavity with the increased salt concentration.

CONCLUSIONS

Herein, we have employed extensive MD simulations of 1 μs along with the molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) method to study the binding mechanism of the cofactor SAM to the nsp16/nsp10 heterodimer and sub-unit nsp16 alone ($\mathrm{nsp16_{SAM}}$) of SARS-CoV-2. Our MD/MMPBSA calculations suggested that SAM binds strongly to the nsp16/nsp10 heterodimer and fails to bind to the nsp16 monomer. It emphasized that nsp10 helps in the strong interaction between SAM and nsp16 to execute the 2′-O-MTase activity. This finding agrees well with the experimental study, and a similar observation was made for other coronaviruses, including SARS-CoV and MERS-CoV. We have also investigated the interactions between nsp16 and nsp10. Our study revealed that the binding of nsp10 stabilizes the complex (nsp16/nsp10) structure. Further, our study showed that hydrophobic interactions are critical for the heterodimer association. Thus, apart from the active site of nsp16, the interface of nsp16/nsp10 can also be considered as a potential target site in the design of antiviral drugs such as peptide inhibitors. It includes Ile40, Val104, Ala83, Val78, Met247, Val44, Gln87, Arg86, Lys76, Lys38, Val84, and Met41 from nsp16, and Leu45, Ala71, Val42, Met44, Tyr96, Gly69, Thr47, Arg78, Gly70, Gly94, and Pro59 from nsp10. Besides, the stable hydrogen bond between Ala83 (nsp16) and Tyr96 (nsp10), and between Gln87 (nsp16) and Leu45 (nsp10) were important in the nsp16-nsp10 interface. The CAS study reveals that residues I40A, V104A, R86A, V78A, V44A, M247A, and Q87A of nsp16 were considered as hot spot residues for the association of nsp16-nsp10. $\Delta\Delta G_{bind}$ for mutants, I40A, V104A, and R86A are comparatively high, suggesting the significance of these residues. Finally, we investigated the effect of temperature and salt concentration on the binding of SAM to the heterodimer. Our study revealed that the SAM binding was impaired due to an increase in temperature or salt concentration. Finally, our study provides a comprehensive understanding of the dynamic and thermodynamic process of binding nsp16 and nsp10 that will contribute to the rational design of inhibitors targeting nsp16/nsp10.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Project conceived and supervised by PK. Simulations and analyses were performed by MS, NJ, RR, and SP. Manuscript was written by MS, NJ, and PK. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially supported by the Department of Biotechnology, Govt. of India (Grant number BT/RLF/RE-entry/40/2014, DBT Ramalingaswami Re-entry Fellowship), and Department of Science and Technology (DST), Govt. of India (Grant number ECR/2017/000010). MS would like to thank DST, Govt. of India, for providing fellowship under the INSPIRE Fellowship Scheme (DST/INSPIRE Fellowship/2017/IF170145).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at ChemRxiv (Sk et al., 2020b).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.590165/full#supplementary-material

REFERENCES

Amadei, A., Linssen, A., de Groot, B., van Aalten, D., and Berendsen, H. (1996). An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dyn.* 13, 615–625. doi: 10.1080/07391102.1996.10508874

Arabi, Y., Harthi, A., Hussein, J., Bouchama, A., Johani, S., Hajeer, A., et al. (2015). Severe neurologic syndrome associated with Middle East respiratory syndrome corona virus (MERS-CoV). *Infection* 43, 495–501. doi: 10.1007/s15010-015-0720-y

Berendsen, H. J., Postma, J. V., van Gunsteren, W. F., Dinola, A., and Haak, J. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81, 3684–3690. doi: 10.1063/1.448118

Bogoch, I. I., Watts, A., Thomas-Bachli, A., Huber, C., Kraemer, M. U., and Khan, K. (2020). Potential for global spread of a novel coronavirus from China. *J Travel Med.* 27:taaa011. doi: 10.1093/jtm/taaa011

Carlsson, J., and Åqvist, J. (2006). Calculations of solute and solvent entropies from molecular dynamics simulations. *Phys Chem. Chem. Phys.* 8, 5385–5395. doi: 10.1039/B608486A

Case, D. A., Huang, Y., Ben-Shalom, S. R., Brozell, D. S., Cerutti, T. E., Cheatham, I., et al. (2018). *AMBER 2018*. San Francisco, CA: University of California.

Chakrabarty, B., Naganathan, V., Garg, K., Agarwal, Y., and Parekh, N. (2019). NAPS update: network analysis of molecular dynamics data and protein–nucleic acid complexes. *Nucleic Acids Res.* 47, W462–W470. doi: 10.1093/nar/gkz399

Chen, Y., Su, C., Ke, M., Jin, X., Xu, L., Zhang, Z., et al. (2011). Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2′-O-methylation by nsp16/nsp10 protein complex. *PLoS Pathog.* 7:e1002294. doi: 10.1371/journal.ppat.1002294

Chen, Y. W., Yiu, C.-P. B., and Wong, K.-Y. (2020). Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research* 9:129. doi: 10.12688/f1000research.22457.2

Coleman, C. M., and Frieman, M. B. (2014). Coronaviruses: important emerging human pathogens. *J. Virol.* 88, 5209–5212. doi: 10.1128/JVI.03488-13

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi: 10.1063/1.464397

Eckerle, L. D., Becker, M. M., Halpin, R. A., Li, K., Venter, E., Lu, X., et al. (2010). Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog.* 6:e1000896. doi: 10.1371/journal.ppat.1000896

Egloff, M. P., Benarroch, D., Selisko, B., Romette, J. L., and Canard, B. (2002). An RNA cap (nucleoside-2′-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *EMBO J.* 21, 2757–2768. doi: 10.1093/emboj/21.11.2757

Fehr, A. R., and Perlman, S. (2015). "Coronaviruses: an overview of their replication and pathogenesis," in *Coronaviruses. Methods in Molecular Biology*, Vol. 1282, eds H. Maier, E. Bickerton, and P. Britton (New York, NY: Humana Press), 1-23. doi: 10.1007/978-1-4939-2438-7_1

Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science* 254, 1598–1603. doi: 10.1126/science.1749933

Furuichi, Y., and Shatkin, A. J. (2000). Viral and cellular mRNA capping: past and prospects. *Adv. Virus Res.* 55:135. doi: 10.1016/S0065-3527(00)55003-9

Gohlke, H., Kiel, C., and Case, D. A. (2003). Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the Ras–Raf and Ras–RalGDS complexes. *J. Mol. Biol.* 330, 891–913. doi: 10.1016/S0022-2836(03)00610-7

Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q., He, J. X., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med.* 382, 1708–1720. doi: 10.1056/NEJMoa2002032

Harcourt, B. H., Jukneliene, D., Kanjanahaluethai, A., Bechill, J., Severson, K. M., Smith, C. M., et al. (2004). Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity. *J. Virol.* 78, 13600–13612. doi: 10.1128/JVI.78.24.13600-13612.2004

Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1995). Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* 246, 122–129. doi: 10.1016/0009-2614(95)01082-K

Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1996). Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* 100, 19824–19839. doi: 10.1021/jp961710n

Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* 51, 69–82. doi: 10.1021/ci100275a

Hu, G., Ma, A., Dou, X., Zhao, L., and Wang, J. J. (2016). Computational studies of a mechanism for binding and drug resistance in the wild type and four mutations of HIV-1 protease with a GRL-0519 inhibitor. *Int. J. Mol. Sci.* 17:819. doi: 10.3390/ijms17060819

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506. doi: 10.1016/S0140-6736(20)30183-5

Ichiye, T., and Karplus, M. (1991). Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* 11, 205–217. doi: 10.1002/prot.340110305

Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* 23, 1623–1641. doi: 10.1002/jcc.10128

Jeffrey, G. A. (1997). *An Introduction to Hydrogen Bonding*. New York, NY: Oxford University Press.

Jonniya, N. A., and Kar, P. (2020). Investigating specificity of the anti-hypertensive inhibitor WNK463 against With-No-Lysine kinase family isoforms via multiscale simulations. *J. Biomol. Struct. Dyn.* 38, 1306–1321. doi: 10.1080/07391102.2019.1602079

Jonniya, N. A., Sk, M. F., and Kar, P. (2019). Investigating phosphorylation-induced conformational changes in WNK1 kinase by molecular dynamics simulations. *ACS Omega* 4, 17404–17416. doi: 10.1021/acsomega.9b02187

Jonniya, N. A., Sk, M. F., and Kar, P. (2020). A comparative study of structural and conformational properties of WNK kinase isoforms bound to an inhibitor: insights from molecular dynamic simulations. *J. Biomol. Struct. Dyn.* doi: 10.1080/07391102.2020.1827035. [Epub ahead of print].

Kar, P., and Knecht, V. (2012a). Energetic basis for drug resistance of HIV-1 protease mutants against amprenavir. *J. Comput. Aided Mol. Des.* 26, 215–232. doi: 10.1007/s10822-012-9550-5

Kar, P., and Knecht, V. (2012b). Energetics of mutation-induced changes in potency of lersivirine against HIV-1 reverse transcriptase. *J. Phys. Chem. B* 116, 6269–6278. doi: 10.1021/jp300818c

Kar, P., and Knecht, V. (2012c). Origin of decrease in potency of darunavir and two related antiviral inhibitors against HIV-2 compared to HIV-1 protease. *J. Phys. Chem. B* 116, 2605–2614. doi: 10.1021/jp211768n

Kar, P., and Knecht, V. (2012d). Mutation-induced loop opening and energetics for binding of tamiflu to influenza N8 neuraminidase. *J. Phys. Chem. B* 116, 6137–6149. doi: 10.1021/jp3022612

Kar, P., Lipowsky, R., and Knecht, V. (2011). Importance of polar solvation for cross-reactivity of antibody and its variants with steroids. *J. Phys. Chem. B* 115, 7661–7669. doi: 10.1021/jp201538t

Kar, P., Lipowsky, R., and Knecht, V. (2013). Importance of polar solvation and configurational entropy for design of antiretroviral drugs targeting HIV-1 protease. *J. Phys. Chem. B* 117, 5793–5805. doi: 10.1021/jp3085292

Kar, P., Seel, M., Hansmann, U. H., and Höfinger, S. (2007a). Dispersion terms and analysis of size-and charge dependence in an enhanced poisson– Boltzmann approach. *J. Phys. Chem. B* 111, 8910–8918. doi: 10.1021/jp072302u

Kar, P., Wei, Y., Hansmann, U. H., and Höfinger, S. (2007b). Systematic study of the boundary composition in Poisson Boltzmann calculations. *J. Comput. Chem.* 28, 2538–2544. doi: 10.1002/jcc.20698

Karplus, M., and Kushick, J. N. (1981). Method for estimating the configurational entropy of macromolecules. *Macromolecules* 14, 325–332. doi: 10.1021/ma50003a019

King, A. M. Q., Adams, M. J., Carstens, E. B., and Lefkowitz, E. J. (2011). *Virus Taxonomy: 9th Report of the International Committee on Taxonomy of Viruses*. San Diego, CA: Academic Press

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* 33, 889–897. doi: 10.1021/ar000033j

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* 10, 733–755. doi: 10.1038/nprot.2015.043

Kräutler, V., van Gunsteren, W. F., and Hünenberger, P. H. (2001). A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* 22, 501–508. doi: 10.1002/1096-987X(20010415)22:5<501::AID-JCC1021>3.0.CO;2-V

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132. doi: 10.1016/0022-2836(82)90515-0

Lin, S., Chen, H., Ye, F., Chen, Z., Yang, F., Zheng, Y., et al. (2020). Crystal structure of SARS-CoV-2 nsp10/nsp16 2′-O-methylase and its implication on antiviral drug design. *Signal Transduct Target Ther.* 5:131. doi: 10.1038/s41392-020-00241-4

Lin, X., Gong, Z., Xiao, Z., Xiong, J., Fan, B., and Liu, J. (2020). Novel coronavirus pneumonia outbreak in 2019: computed tomographic findings in two cases. *Korean J. Radiol.* 21, 365–368. doi: 10.3348/kjr.2020.0078

Loncharich, R. J., Brooks, B. R., and Pastor, R. W. (1992). Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N′-methylamide. *Biopolymers* 32, 523–535. doi: 10.1002/bip.360320508

Lu, H. (2020). Drug treatment options for the 2019-new coronavirus (2019-nCoV). *Biosci. Trends* 14, 69–71. doi: 10.5582/bst.2020.01020

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395, 565–574. doi: 10.1016/S0140-6736(20)30251-8

Lun, Z.-R., and Qu, L.-H. (2004). Animal-to-human SARS-associated coronavirus transmission? *Emerg Infect Dis.* 10:959. doi: 10.3201/eid1005.040022

Machado, M. R., and Pantano, S. (2020). Split the charge difference in two! a rule of thumb for adding proper amounts of ions in MD simulations. *J. Chem. Theory Comput.* 16, 1367–1372. doi: 10.1021/acs.jctc.9b00953

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* 11, 3696–3713. doi: 10.1021/acs.jctc.5b00255

Massova, I., and Kollman, P. A. (1999). Computational alanine scanning to probe protein– protein interactions: a novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* 121, 8133–8143. doi: 10.1021/ja990935j

Minasov, G., Shuvalova, L., Rosas-Lemus, M., Kiryukhina, O., Wiersum, G., Godzik, A., et al. (2020). *1.80 Angstrom Resolution Crystal Structure of NSP16 - NSP10 Complex from SARS-CoV-2.* Available online at: http://www.rcsb.org/structure/6W4H (accessed March 18, 2020).

Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2007). Computational alanine scanning mutagenesis—an improved methodological approach. *J. Comput. Chem* 28, 644–654. doi: 10.1002/jcc.20566

Olsson, M. H., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions. *J. Chem. Theory Comput.* 7, 525–537. doi: 10.1021/ct100578z

Pastor, R. W., Brooks, B. R., and Szabo, A. (1988). An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol. Phys.* 65, 1409–1419. doi: 10.1080/00268978800101881

Pillaiyar, T., Meenakshisundaram, S., and Manickam, M. (2020). Recent Discovery and Development of Inhibitors Targeting Coronaviruses. *Drug Discov. Today.* 25, 668–688. doi: 10.1016/j.drudis.2020.01.015

Price, D. J., and Brooks, C. L. III. (2004). A modified TIP3P water potential for simulation with Ewald summation. *J. Chem. Phys.* 121, 10096–10103. doi: 10.1063/1.1808117

Rempe, S. B., and Jónsson, H. (1998). A computational exercise illustrating molecular vibrations and normal modes. *Chem. Educ.* 3, 1–17. doi: 10.1007/s00897980231a

Roe, D. R., and Cheatham, T. E. III. (2013). PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* 9, 3084–3095. doi: 10.1021/ct400341p

Rosas-Lemus, M., Minasov, G., Shuvalova, L., Inniss, N.L., Kiryukhina, O., Wiersum, G., et al. (2020). The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. *bioRxiv*. doi: 10.1101/2020.04.17.047498

Roy, R., Ghosh, B., and Kar, P. (2020). Investigating conformational dynamics of lewis Y oligosaccharides and elucidating blood group dependency of cholera using molecular dynamics. *ACS Omega* 5, 3932–3942. doi: 10.1021/acsomega.9b03398

Sanachai, K., Mahalapbutr, P., Choowongkomon, K., Poo-Arporn, R. P., Wolschann, P., and Rungrotmongkol, T. (2020). Insights into the binding recognition and susceptibility of tofacitinib toward janus kinases. *ACS Omega* 5, 369–377. doi: 10.1021/acsomega.9b02800

Sheahan, T., Sims, A., Leist, S., Schäfer, A., Won, J., Brown, A., et al. (2020). Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV. *Nat. Commun.* 11:222. doi: 10.1038/s41467-019-13940-6

Shi, S., Zhang, S., and Zhang, Q. (2018). Insight into binding mechanisms of inhibitors MKP56, MKP73, MKP86, and MKP97 to HIV-1 protease by using molecular dynamics simulation. *J. Biomol. Struct. Dyn.* 36, 981–992. doi: 10.1080/07391102.2017.1305296

Singh, S., Sk, M. F., Sonawane, A., Kar, P., and Sadhukhan, S. (2020). Plant-derived natural polyphenols as potential antiviral drugs against SARS-CoV-2 via RNA-dependent RNA polymerase (RdRp) inhibition: an *in-silico* analysis. *J. Biomol. Struct. Dyn.* doi: 10.1080/07391102.2020.1796810. [Epub ahead of print].

Sk, M. F., Jonniya, N. A., and Kar, P. (2020a). Exploring the energetic basis of binding of currently used drugs against HIV-1 subtype CRF01_AE protease via molecular dynamics simulations. *J. Biomol. Struct. Dyn.* doi: 10.1080/07391102.2020.1794965. [Epub ahead of print].

Sk, M. F., Jonniya, N. A., Roy, R., Poddar, S., and Kar, P. (2020b). Computational investigation of structural dynamics of SARS-CoV-2 methyltransferase-stimulatory factor heterodimer nsp16/nsp10 bound to the cofactor SAM. *ChemRxiv [Preprint]*. doi: 10.26434/chemrxiv.12608795.v1

Sk, M. F., Roy, R., Jonniya, N. A., Poddar, S., and Kar, P. (2020c). Elucidating biophysical basis of binding of inhibitors to SARS-CoV-2 main protease by using molecular dynamics simulations and free energy calculations. *J. Biomol. Struct. Dyn.* doi: 10.26434/chemrxiv.12084207. [Epub ahead of print].

Sk, M. F., Roy, R., and Kar, P. (2020d). Exploring the potency of currently used drugs against HIV-1 protease of subtype D variant by using multiscale simulations. *J. Biomol. Struct. Dyn.* doi: 10.1080/07391102.2020.1724196. [Epub ahead of print].

Viswanathan, T., Arya, S., Chan, S. H., Qi, S., Dai, N., Misra, A., et al. (2020). Structural basis of RNA cap modification by SARS-CoV-2. *Nat. Commun.* 11:3718. doi: 10.1038/s41467-020-17496-8

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng. Design Select.* 8, 127–134. doi: 10.1093/protein/8.2.127

Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Model.* 25, 247–260. doi: 10.1016/j.jmgm.2005.12.005

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J. Comput. Chem.* 25, 1157–1174. doi: 10.1002/jcc.20035

Wang, Y., Sun, Y., Wu, A., Xu, S., Pan, R., Zeng, C., et al. (2015). Coronavirus nsp10/nsp16 methyltransferase can be targeted by nsp10-derived peptide *in vitro* and *in vivo* to reduce replication and pathogenesis. *J. Virol.* 89, 8416–8427. doi: 10.1128/JVI.00948-15

Weiser, J., Shenkin, P. S., and Still, W. C. (1999). Fast, approximate algorithm for detection of solvent-inaccessible atoms. *J. Comput. Chem.* 20, 586–596.

Weiss, S. R., and Leibowitz, J. L. (2011). "Coronavirus pathogenesis," in *Advances in Virus Research*, eds K. Maramorosch, A. J. Shatkin, and F. A. Murphy (Elsevier) 81, 85-164. doi: 10.1016/B978-0-12-385885-6.00009-2

Weiss, S. R., and Navas-Martin, S. (2005). Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiol. Mol. Biol. Rev.* 69, 635–664. doi: 10.1128/MMBR.69.4.635-664.2005

Woo, P. C., Huang, Y., Lau, S. K., and Yuen, K.-Y. (2010). Coronavirus genomics and bioinformatics analysis. *Viruses* 2, 1804–1820. doi: 10.3390/v2081803

Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., et al. (2020). Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sinica B* 10, 766–788. doi: 10.1016/j.apsb.2020.02.008

Xu, B., Shen, H., Zhu, X., and Li, G. (2011). Fast and accurate computation schemes for evaluating vibrational entropy of proteins. *J. Comput. Chem.* 32, 3188–3193. doi: 10.1002/jcc.21900

Yang, H., Xie, W., Xue, X., Yang, K., Ma, J., Liang, W., et al. (2005). Design of wide-spectrum inhibitors targeting coronavirus main proteases. *PLoS Biol* 3:e324. doi: 10.1371/journal.pbio.0030428

Zeng, Z. Q., Chen, D. H., Tan, W. P., Qiu, S. Y., Xu, D., Liang, H. X., et al. (2018). Epidemiology and clinical characteristics of human coronaviruses OC43, 229E, NL63, and HKU1: a study of hospitalized children with acute respiratory tract infection in Guangzhou, China. *Eur. J. Clin. Microbiol. Infect. Dis.* 37, 363–369. doi: 10.1007/s10096-017-3144-z

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

Zhou, Y., Hou, Y., Shen, J., Huang, Y., Martin, W., and Cheng, F. (2020). Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov.* 6:14. doi: 10.1038/s41421-020-0153-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 27 January 2021
doi: 10.3389/fmolb.2020.598583

SARS-CoV-2 Proteins Exploit Host's Genetic and Epigenetic Mediators for the Annexation of Key Host Signaling Pathways

Md. Abdullah-Al-Kamran Khan[1] *and Abul Bashar Mir Md. Khademul Islam*[2]*

[1]*Department of Mathematics and Natural Sciences, BRAC University, Dhaka, Bangladesh,* [2]*Department of Genetic Engineering and Biotechnology, University of Dhaka, Dhaka, Bangladesh*

The constant rise of the death toll and cases of COVID-19 has made this pandemic a serious threat to human civilization. Understanding of host-SARS-CoV-2 interaction in viral pathogenesis is still in its infancy. In this study, we utilized a blend of computational and knowledgebase approaches to model the putative virus-host interplay in host signaling pathways by integrating the experimentally validated host interactome proteins and differentially expressed host genes in SARS-CoV-2 infection. While searching for the pathways in which viral proteins interact with host proteins, we discovered various antiviral immune response pathways such as hypoxia-inducible factor 1 (HIF-1) signaling, autophagy, retinoic acid-inducible gene I (RIG-I) signaling, Toll-like receptor signaling, fatty acid oxidation/degradation, and IL-17 signaling. All these pathways can be either hijacked or suppressed by the viral proteins, leading to improved viral survival and life cycle. Aberration in pathways such as HIF-1 signaling and relaxin signaling in the lungs suggests the pathogenic lung pathophysiology in COVID-19. From enrichment analysis, it was evident that the deregulated genes in SARS-CoV-2 infection might also be involved in heart development, kidney development, and AGE-RAGE signaling pathway in diabetic complications. Anomalies in these pathways might suggest the increased vulnerability of COVID-19 patients with comorbidities. Moreover, we noticed several presumed infection-induced differentially expressed transcription factors and epigenetic factors, such as miRNAs and several histone modifiers, which can modulate different immune signaling pathways, helping both host and virus. Our modeling suggests that SARS-CoV-2 integrates its proteins in different immune signaling pathways and other cellular signaling pathways for developing efficient immune evasion mechanisms while leading the host to a more complicated disease condition. Our findings would help in designing more targeted therapeutic interventions against SARS-CoV-2.

Keywords: host-virus interactions, COVID-19, SARS-CoV-2, immune evasion, epigenetic regulation, host immune response

OPEN ACCESS

Edited by:
Nadine Kabbani,
George Mason University,
United States

Reviewed by:
Robson Carvalho,
Institute of Biosciences, São Paulo State University, Brazil
Saba Al Heialy,
Mohammed Bin Rashid University of Medicine and Health Sciences, United Arab Emirates

Correspondence:
Abul Bashar Mir Md. Khademul Islam
khademul@du.ac.bd

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *25 August 2020*
Accepted: *23 December 2020*
Published: *27 January 2021*

Citation:
Khan MA-A-K and Islam ABMMK (2021) SARS-CoV-2 Proteins Exploit Host's Genetic and Epigenetic Mediators for the Annexation of Key Host Signaling Pathways. Front. Mol. Biosci. 7:598583. doi: 10.3389/fmolb.2020.598583

INTRODUCTION

Though several human coronavirus outbreaks caused severe public health crises over the past few decades, the recent coronavirus disease (COVID-19) outbreak caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has beaten the records of the previous ones and the case counts are still on the upswing. About 210 countries and territories around the globe have been affected by this outbreak, ~24 million people are already infected with SARS-CoV-2, and the number was steadily rising at the time of writing this article (Worldometer, 2020). Out of the closed cases, almost 7% of the patients have died and about 1% of the active cases are in critical conditions (Worldometer, 2020). Though the death rates from COVID-19 were estimated to be as low as 3.4% (WHO, 2020), at present, the global fatality rate is changing very rapidly; therefore, more comprehensive studies need to be conducted to effectively control and overturn this pandemic.

Coronaviruses are single-stranded positive-sense, enveloped RNA viruses having ~30 Kb genome (Lu et al., 2020). Among the four genera, SARS-CoV-2 (accession no. NC_045512.2) belongs to the *Betacoronavirus* genus and it has ~29.9 Kb genome encoding 11 genes (NCBI-Gene, 2020). The genome sequence of SARS-CoV-2 is about 90% similar to bat-derived SARS-like coronavirus, whereas this novel virus is only ~79 and ~50% similar to severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East Respiratory Syndrome-related Coronavirus (MERS-CoV), respectively (Lu et al., 2020; Ren et al., 2020). A substantial genomic difference can be observed between SARS-CoV and SARS-CoV-2; as in SARS-CoV-2, there have been 380 amino acids substitution, deletion of ORF8a, elongation of ORF8b, and truncation of ORF3b observed (Lu et al., 2020).

Though the overall mortality rate from SARS-CoV is higher than that of SARS-CoV-2, several unique features of SARS-CoV-2, including increased incubation period and dormancy inside the host, enabled this virus to spread more efficiently (Lauer et al., 2020). This suggests that SARS-CoV-2 might be using some immune evasion strategies to maintain its survival and essential functions within the host.

Upon viral infection, the host innate immune system detects the virion particles and elicits the first sets of antiviral responses (Katze et al., 2008) to eliminate the viral threats. However, viruses themselves have generated various modes of action to evade those immune responses by modulating the host's intracellular signaling pathways (Kikkert, 2020). This arm-wrestling between the host and the infecting virus results in immunopathogenesis. Different human coronaviruses also show similar features of host-pathogen interactions, which range from the viral entry, replication, transcription, translation, and assembly to the evasion from host innate immune response (Fung and Liu, 2019). Moreover, different antiviral cellular responses such as autophagy (Ahmad et al., 2018) and apoptosis (Barber, 2001) can also be moderated by the virus to ensure its survival inside the host cells. Apart from these, several other host-virus interactions are also observed, namely, modulation of the activity of host transcription factors (TFs) (Lyles, 2000) and host epigenetic factors (e.g., histone modifications and host miRNAs) (Adhya and Basu, 2010). All of these multifaceted interactions can lead to the ultimate pathogenesis and progression of the disease.

The interplay between different human coronaviruses and host was previously reported (Fung and Liu, 2019); however, SARS-CoV-2 interactions with the host immune response and its outcome in the pathogenesis are yet to be elucidated. Gordon et al. (2020) identified 332 high-confidence interactions between SARS-CoV-2 proteins and human proteins (Gordon et al., 2020); Blanco-Melo et al. (2020) produced transcriptional signatures of SARS-CoV-2 infected cells (Blanco-Melo et al., 2020). In this study, we aimed to model the complex host-SARS-CoV-2 interactions with the associated differentially expressed genes found in the SARS-CoV-2 infection to gain insights into the probable immune escape mechanisms of SARS-CoV-2. Also, we have compared the differential gene expression profiles of SARS-CoV and SARS-CoV-2 infected cells to find out the pathways uniquely targeted by SARS-CoV-2. Moreover, we incorporated other associated host epigenetic factors that might play a role in the pathogenesis by deregulating the signaling pathways.

MATERIALS AND METHODS

Retrieval of the Host Proteins That Interact With SARS-CoV-2

We have obtained 332 human proteins that form high-confidence interactions with SARS-CoV-2 proteins from the study conducted previously by Gordon et al. (2020) and processed their provided protein names into the associated HGNC official gene symbol (**Supplementary Material S1**).

Analysis of Microarray Expression Data

Microarray expression data on SARS-CoV infected 2B4 cells or uninfected controls for 24 h were obtained from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo) (Barrett et al., 2012), accession: GSE17400 (Yoshikawa et al., 2010). Raw Affymetrix CEL files were background-corrected and normalized using Bioconductor package “affy v1.28.1” using “rma” algorithm. Quality of microarray experiment (data not shown) was verified by Bioconductor package “arrayQualityMetrics v3.2.4” (Kauffmann et al., 2009). Differential expression (DE) between two experimental conditions was detected using Bioconductor package Limma (Smyth, 2005). Probe annotations were converted to genes using in-house python script basing the Ensembl gene model (BioMart 99) (Flicek et al., 2007). The highest absolute expression value was considered for the probes that were annotated to the same gene. We have considered the genes to be differentially expressed, which have false discovery rate (FDR) (Benjamini and Hochberg, 1995) *p* value ≤0.05 and Log_2 fold change value ≥0.25. Positive Log_2 fold change values indicate upregulation, whereas negative ones indicate downregulation.

Analysis of RNA-Seq Expression Data

Illumina sequenced RNA-Seq raw FastQ reads were extracted from the GEO database (https://www.ncbi.nlm.nih.gov/geo) (Barrett et al., 2012), accession: GSE147507 (Blanco-Melo et al., 2020). These data included independent biological triplicates of primary human lung epithelium (NHBE) that were mock-treated or infected with SARS-CoV-2 for 24 h. Mapping of reads was done with TopHat v2.1.1 (with Bowtie v2.4.1) (Trapnell et al., 2009). Short reads were uniquely aligned, allowing at best two mismatches to the human reference genome from GRCh38 as downloaded from the USCS database (Lander et al., 2001). Sequences matched exactly at more than one place with equal quality were discarded to avoid bias (Hansen et al., 2010). The reads that were not mapped to the genome were utilized to map against the transcriptome (junction mapping). Ensembl gene model (Hubbard et al., 2007) (version 99, as extracted from UCSC) was used for this process. After mapping, we used SubRead package featureCount v2.21 (Liao et al., 2013) to calculate absolute read abundance (read count, rc) for each transcript/gene associated with the Ensembl genes. For DE analysis, we used DESeq2 v1.26.0 with R v3.6.2 (2019-07-05) (Anders and Huber, 2010) that uses a model based on the negative binomial distribution. To avoid false positives, we considered only those transcripts where at least 10 reads are annotated in at least one of the samples used in this study. We have considered the genes to be differentially expressed when they have FDR (Benjamini and Hochberg, 1995) *p* value ≤0.05 and Log2 fold change value ≥0.25. Positive Log_2 fold change values mean upregulation, whereas negative ones mean downregulation.

Functional Enrichment Analysis

We utilized Gitools v1.8.4 for enrichment analysis and heatmap generation (Perez-Llamas and Lopez-Bigas, 2011). We have utilized the Gene Ontology Biological Processes (GOBP) and Molecular Function (GOMF) modules (Ashburner et al., 2000), KEGG Pathways (Kanehisa and Goto, 2000), BioPlanet pathways (Huang et al., 2019b) modules, and WikiPathways (Slenter et al., 2017) modules for the overrepresentation analysis. Resulting *p*-values were adjusted for multiple testing using the Benjamini–Hochberg method of FDR (Benjamini and Hochberg, 1995). We have also performed the enrichment analysis based on the KEGG pathway module of the STRING database (Szklarczyk et al., 2019) for 332 proteins (**Supplementary Material S1**) retrieved from the analysis of Gordon et al. (2020) (Gordon et al., 2020) along with the deregulated genes analyzed from SARS-CoV-2 infected cell's RNA-Seq expression data. An enrichment is considered significant if its adjusted *p* value is <0.05.

Obtaining the Transcription Factors That Bind Promoter Regions

We have obtained the TFs that bind to the given promoters from the Cistrome Data Browser (Zheng et al., 2018), which provides TFs from experimental ChIP-seq data. We utilized "Toolkit for CistromeDB," uploaded the 5 Kb upstream promoter with 1 Kb downstream from transcription start site (TSS) BED file of the deregulated genes, and fixed the peak number parameter to "all peaks in each sample."

Obtaining Human miRNAs Target Genes

We extracted the experimentally validated target genes of human miRNAs from miRTarBase database (Huang et al., 2019a). We only considered those miRNAs whose target gene was found downregulated and the associated TF upregulated in the differential gene expression analysis.

Extraction of Transcription Factors That Modulate Human miRNA Expression

We have downloaded the experimentally validated TFs from the TransmiR v2.0 database (Tong et al., 2019), which bind to miRNA promoters and modulate them. We have considered those TFs that are expressed themselves and that can "activate" or "regulate" miRNAs.

Identification of the Host Epigenetic Factors Genes

We used EpiFactors database (Medvedeva et al., 2015) to find human genes related to the epigenetic activity. We only considered those epigenetic factors for the final modeling that are differentially expressed in our analyzed transcriptome of SARS-CoV and SARS-CoV-2.

Mapping of the Human Proteins in Cellular Pathways

We have utilized the KEGG mapper tool (Kanehisa and Sato, 2020) for the mapping of deregulated genes SARS-CoV-2 interacting host proteins in different cellular pathways. We then searched and targeted the pathways found to be enriched for SARS-CoV-2 deregulated genes. From this pathway information, we have manually sketched the pathways to provide a brief overview of the interplay between SARS-CoV-2 and host immune response and their outcomes in the viral pathogenesis.

A brief workflow of the whole study is depicted in **Figure 1**.

RESULTS

Differentially Expressed Genes in SARS-CoV-2 Infection Are Involved in Important Cellular Signaling Pathways

We wanted to identify the pathways that might be modulated upon the infection of SARS-CoV-2 and to highlight their uniqueness from SARS-CoV infection. Therefore, we performed the enrichment analysis using the differentially expressed genes of both SARS-CoV and SARS-CoV-2 by

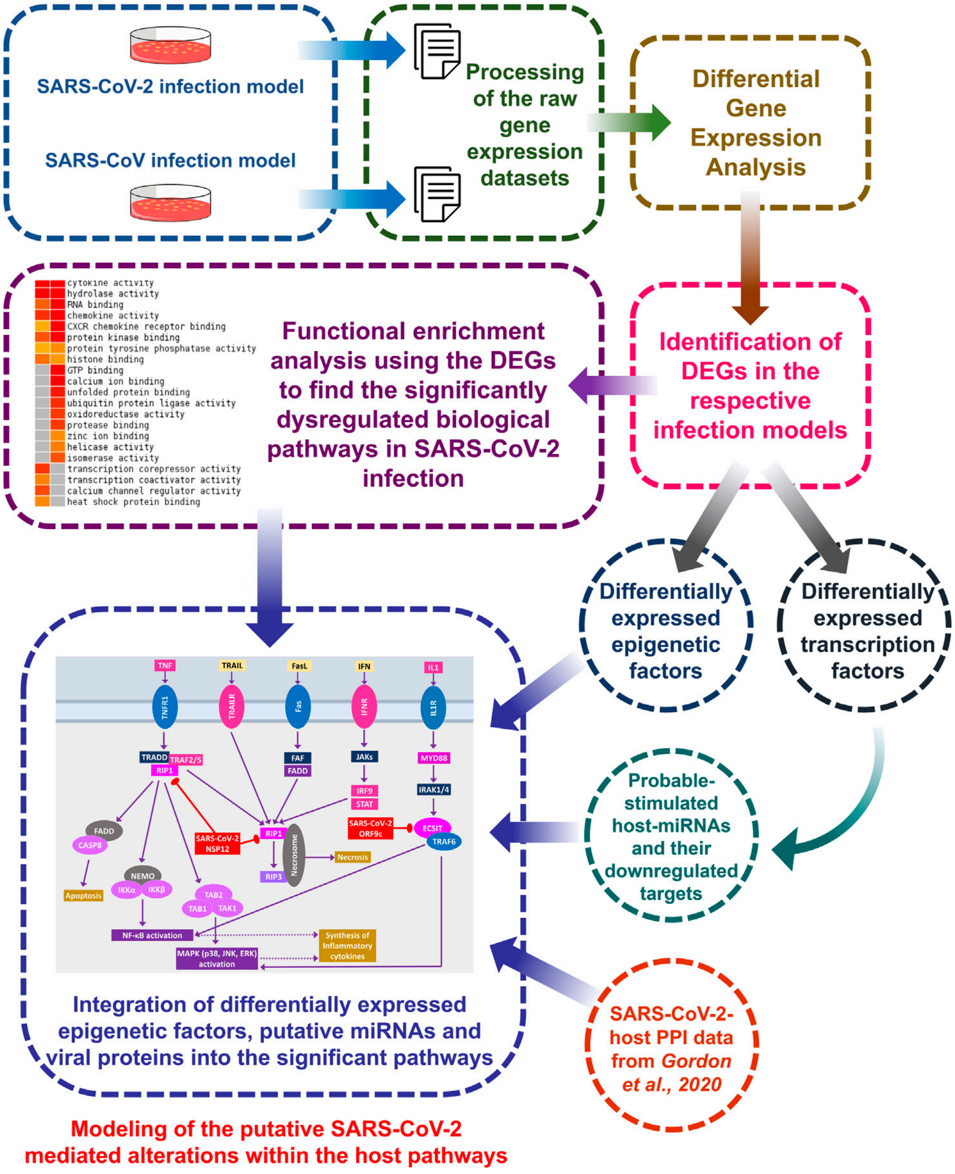

SARS-CoV-2 infection model
SARS-CoV infection model
Processing of the raw gene expression datasets
Differential Gene Expression Analysis
Identification of DEGs in the respective infection models
cytokine activity
hydrolase activity
RNA binding
chemokine activity
CXCR chemokine receptor binding
protein kinase binding
protein tyrosine phosphatase activity
histone binding
GTP binding
calcium ion binding
unfolded protein binding
ubiquitin protein ligase activity
oxidoreductase activity
protease binding
zinc ion binding
helicase activity
isomerase activity
transcription corepressor activity
transcription coactivator activity
calcium channel regulator activity
heat shock protein binding
Functional enrichment analysis using the DEGs to find the significantly dysregulated biological pathways in SARS-CoV-2 infection
Differentially expressed epigenetic factors
Differentially expressed transcription factors
Probable-stimulated host-miRNAs and their downregulated targets
SARS-CoV-2-host PPI data from Gordon et al., 2020
TNF
TRAIL
FasL
IFN
IL1
TNFR1
TRAILR
Fas
IFNR
IL1R
TRADD
TRAF2/5
RIP1
FAF
FADD
JAKs
MYD88
IRF9
STAT
IRAK1/4
SARS-CoV-2 NSP12
SARS-CoV-2 ORF9c
ECSIT
TRAF6
FADD
CASP8
RIP1
Necrosome
Necrosis
RIP3
NEMO
IKKα
IKKβ
Apoptosis
TAB2
TAB1
TAK1
NF-κB activation
MAPK (p38, JNK, ERK) activation
Synthesis of Inflammatory cytokines
Integration of differentially expressed epigenetic factors, putative miRNAs and viral proteins into the significant pathways
Modeling of the putative SARS-CoV-2 mediated alterations within the host pathways

Figure 1 | Overall workflow of the study.

Gitools (Perez-Llamas and Lopez-Bigas, 2011) using GOBP, GOMF, KEGG pathways, BioPlanet pathways, and WikiPathways modules.

We identified 387 upregulated and 61 downregulated genes in SARS-CoV infection (analyzing GSE17400) and 464 upregulated and 222 downregulated genes in SARS-CoV-2 infection (analyzing GSE147507) (**Supplementary Material S2**). Enrichment analysis using these differentially expressed genes exhibited that deregulated genes of SARS-CoV-2 infection can exert biological functions such as regulation of inflammatory response, negative regulation of type I interferon, response to interferon-gamma, interferon-gamma-mediated signaling, NIK/NF-kappaB signaling, regulation of the apoptotic process, cellular response to hypoxia, angiogenesis, negative regulation of inflammatory response, zinc ion binding, and calcium ion binding; all of these were not enriched for SARS-CoV infection (**Figures 2A,B**). Also, different organ-specific functions such as heart development and kidney development were only enriched for differentially expressed genes in SARS-CoV-2 infection (**Figure 2A**).

Deregulated genes of SARS-CoV-2 infection were involved in pathways such as NF-kappaB signaling, Jak-STAT signaling, RIG-I-like receptor signaling, natural killer cell–mediated cytotoxicity, phagosome, HIF-1 signaling, calcium signaling, GnRH signaling, arachidonic acid metabolism, insulin signaling, adrenergic signaling in cardiomyocytes, and PPAR signaling (**Figure 2C**; **Supplementary Figures S1,**

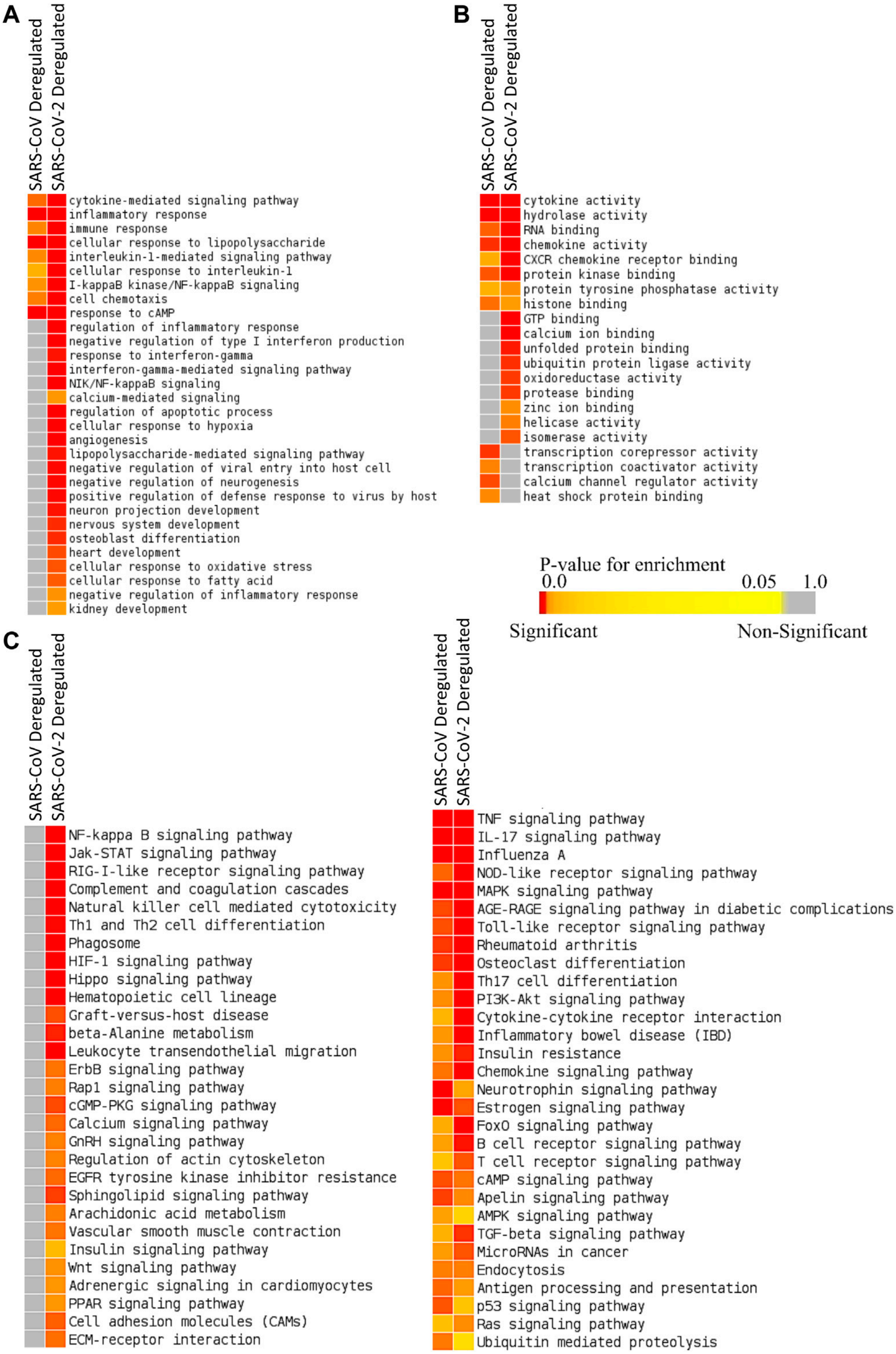


Figure 2 | Enrichment analysis and comparison between deregulated genes in SARS-CoV and SARS-CoV-2 infections using **(A)** GOBP module, **(B)** GOMF module, and **(C)** KEGG pathway module. The significance of enrichment in terms of the adjusted *p* value (<0.05) is represented in color-coded *p* value scale for all heatmaps. Color toward red indicates higher significance and color toward yellow indicates less significance, while gray means nonsignificant.

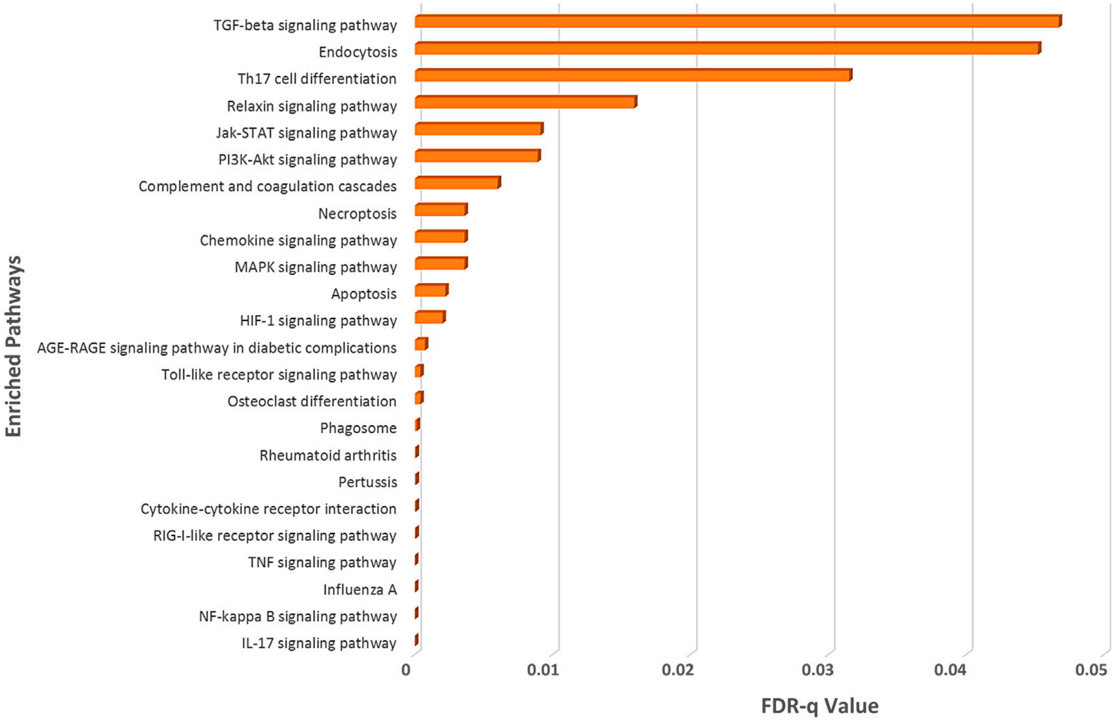


Figure 3 | Enrichment analysis of deregulated genes in SARS-CoV-2 infections and human proteins interacting with SARS-CoV-2 proteins using KEGG pathway module of the STRING database.

S2), which were absent in SARS-CoV infection. These significant differences between these two infections might be crucial in defining the differences in disease pathobiology between these viruses; however, these variations could have been resulted due to the differences in the used infection models.

Both SARS-CoV-2 Interacting Human Proteins and Deregulated Genes from SARS-CoV-2 Infection Have Immunological Roles

To illuminate whether SARS-CoV-2 interacting human proteins and differentially expressed genes in SARS-CoV-2 infection are involved in the same pathways, we sought out an enrichment analysis using STRING (KEGG pathway module) (Szklarczyk et al., 2019), which is a functional protein-protein interaction network database.

From this analysis, we spotted that both the deregulated genes of SARS-CoV-2-infection and the SARS-CoV-2-interacting human proteins are involved in several important immune signaling pathways, namely, IL-17 signaling, NF-kappaB signaling, TNF signaling, Toll-like receptor signaling, phagosome, apoptosis, necroptosis, PI3K-Akt signaling, HIF-1 signaling, and MAPK signaling (**Figure 3**). Moreover, signaling pathways such as relaxin signaling, rheumatoid arthritis, and AGE-RAGE signaling pathway in diabetic complications were also enriched (**Figure 3**).

Differentially Expressed Genes in SARS-CoV-2 Infection Have Different Transcription Factor Binding Preferences Compared to SARS-CoV Infection

We sought to find out the TFs that have promoter binding preferences for differentially expressed genes in SARS-CoV and SARS-CoV-2 infections. To achieve this, we utilized CistromeDB (Zheng et al., 2018). We identified 18 and 29 such TFs overrepresented around differentially expressed genes of SARS-CoV and SARS-CoV-2, respectively (**Figures 4A,B, Supplementary Figure S3**). Among those TFs, only 3 (NFKB1A, TNFAIP3, and BCL3) were common for deregulated genes of both infections. Nineteen of 29 TFs, which were overrepresented for SARS-CoV-2 deregulated genes, were also upregulated upon SARS-CoV-2 infections (data not shown).

Downregulated Genes in SARS-CoV-2 Infection Can Be Targeted by Different Human miRNAs

Viral infection leading to the expression of different host miRNAs is a common phenomenon (Bruscella et al., 2017). To elucidate such roles of the human miRNAs in SARS-CoV and SARS-CoV-2 infections, we considered those miRNAs from miRTarBase database (Huang et al., 2019a) that can particularly target the downregulated genes in these infections. We detected 13 and 389

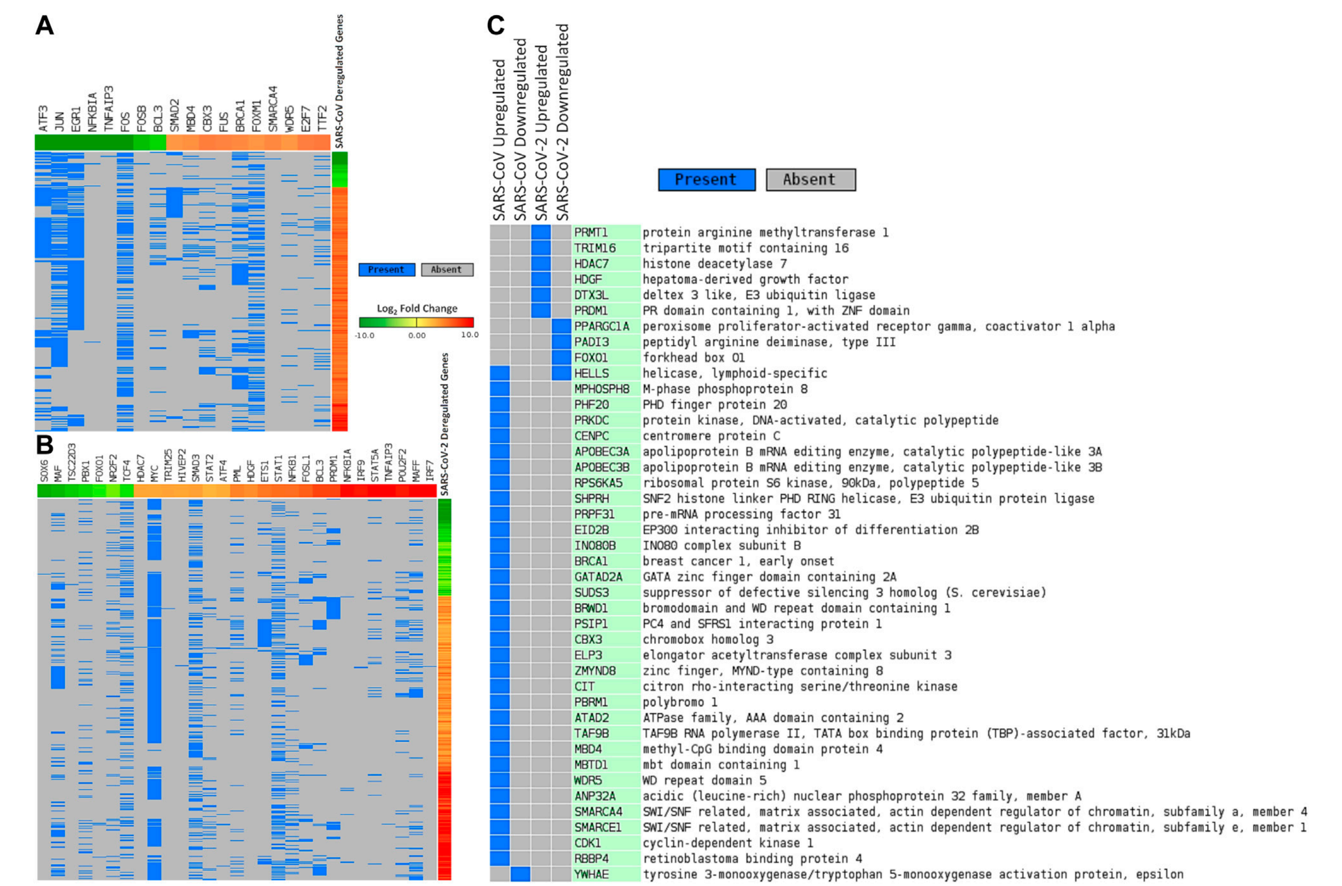


Figure 4 | Differentially expressed genes in **(A)** SARS-CoV and **(B)** SARS-CoV-2 infections and associated deregulated transcription factors that can bind around their promoters. **(C)** Deregulated epigenetic factors in SARS-CoV and SARS-CoV-2 infections. In **Figures 4A,B**, genes are represented vertically, while the associated transcription factors are shown horizontally. Expression values of the TFs/genes are shown in Log_2 fold change scale for both. Color toward red indicates upregulation and color toward green indicates downregulation. Blue color indicates presence and gray color indicates the absence of a term.

candidate miRNAs targeting 17 and 123 downregulated genes in SARS-CoV-infection and SARS-CoV-2-infection, respectively (**Supplementary Material S3**). Among those, only 7 miRNAs (hsa-miR-148a-3p, hsa-miR-146a-5p, hsa-miR-155-5p, hsa-miR-146b-5p, hsa-miR-27a-3p, hsa-miR-146b-3p, and hsa-miR-141-5p) were observed to be targeting the downregulated genes in both infections.

Upregulated Transcription Factors in SARS-CoV-2 Infection Modulate Different Host miRNAs

To detect the upregulated TFs that can preferentially bind around the promoters of the host miRNAs and their associated functional roles, we have utilized the TransmiR v2.0 database (Tong et al., 2019). We obtained 5 and 14 such upregulated TFs that might have a regulatory role in 14 and 90 host miRNAs in SARS-CoV and SARS-CoV-2 infections, respectively (**Supplementary Material S4**). Though these TFs were completly different in both infections, we have found 6 host miRNAs (hsa-miR-146a, hsa-miR-146b, hsa-miR-155, hsa-miR-141, hsa-miR-200a, and hsa-miR-27a) that were commonly modulated by TFs in both infections. Intriguingly in SARS-CoV-2 infection, we noticed 2 host miRNAs (hsa-miR-429 and hsa-miR-1286) that were influenced by the upregulated TF NFKB1 and have associations with several downregulated genes (*BCL2L11*, *FKBP5*, and *TP53INP1* for hsa-miR-429; *CLU* for hsa-miR-1286).

Several Host Epigenetic Factors Can Modulate the Deregulation of Gene Expression in SARS-CoV-2 Infection

Next, we pursued the epigenetic factors that are deregulated and tried to elucidate if their deregulation could play a role in the overall differential gene expression. In this pursuit, we searched the EpiFactors database (Medvedeva et al., 2015) and identified 33 and 10 epigenetic factors that were deregulated in SARS-CoV and SARS-CoV-2 infections, respectively (**Figure 4C**). Among the 10 factors found in SARS-CoV-2 infection, 6 (*PRMT1*, *TRIM16*, *HDAC7*, *HDGF*, *DTX3L*, and *PRDM1*) were upregulated and 4 (*PPARGC1A*, *PADI3*, *FOXO1*, and *HELLS*) were downregulated.

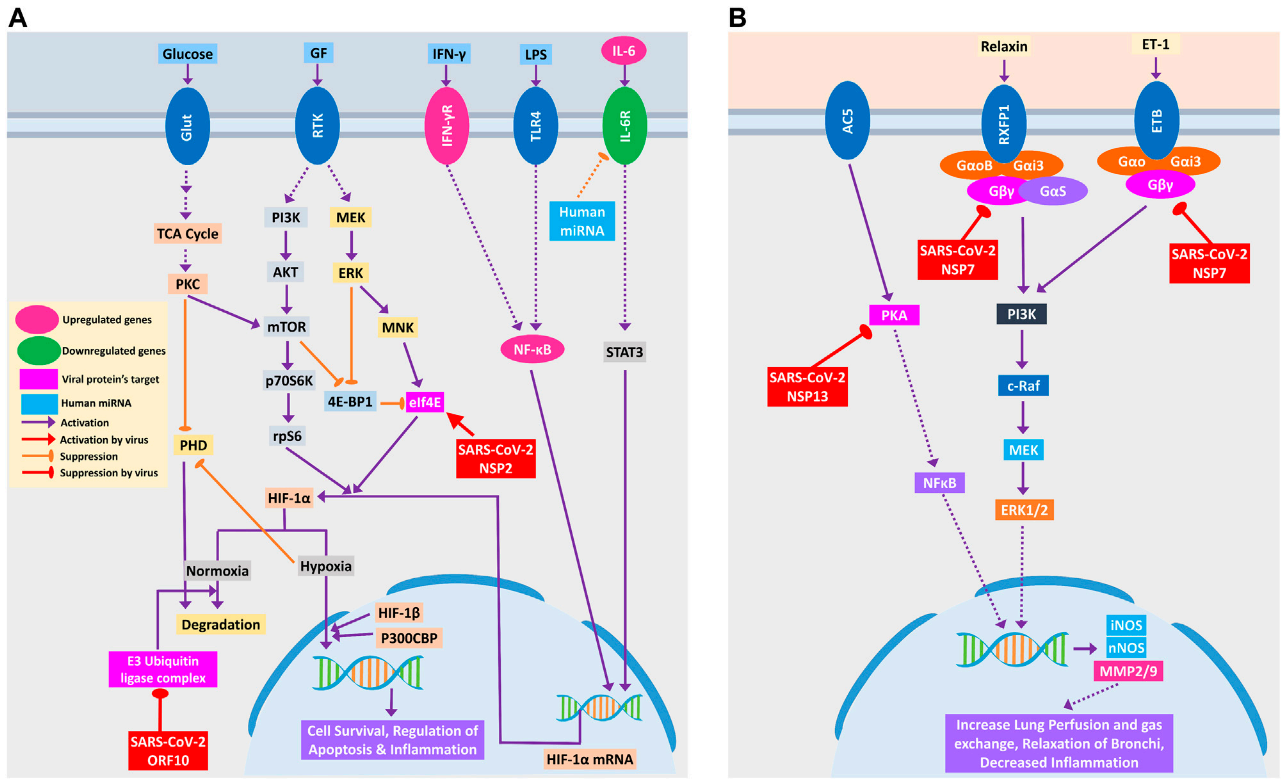


Figure 5 | (A) HIF-1 signaling pathway and **(B)** relaxin signaling pathway in lungs. Upregulated genes are in pink color, while downregulated genes are in green color. Magenta and light blue represent viral protein's target and host miRNAs, respectively. Violet pointed arrows indicate activation, while orange blunt arrows indicate suppression. Red pointed and blunt arrows indicate activation and suppression by virus, respectively.

Putative Roles of Viral Proteins in Immune Evasion and Disease Pathophysiology of COVID-19 Are Evident From Various Signaling Pathways

Although there are some similarities between SARS-CoV and SARS-CoV-2 genetic architecture, it is yet to be known whether they modulate common host pathways or not. Also, it is largely unknown how SARS-CoV-2 exhibits some unique clinical features, despite sharing many similarities, in terms of viral genes, with SARS-CoV.

As now the probable genetic and epigenetic regulators behind the differential gene expression have been identified, we aimed to explore how these deregulated genes are playing a role in the battle between virus and host. To obtain a detailed view of the outcomes resulting from viral-host interactions and how SARS-CoV-2 uses its proteins to evade host innate immune response, we mapped the significantly deregulated genes and host interacting proteins in different overrepresented functional pathways using KEGG mapper (Kanehisa and Sato, 2020). Analyzing the pathways, we modeled several host-virus interactions in signaling pathways leading to the ultimate viral immune escape mechanisms. SARS-CoV-2 can blockade several signaling pathways such as HIF-1 signaling, autophagy, RIG-I signaling, receptor-interacting protein kinase 1- (RIP1-) mediated signaling, beta-adrenergic receptor signaling, insulin signaling, fatty acid oxidation and degradation pathway, IL-17 signaling, Toll-like receptor signaling, phagosome formation, arachidonic acid metabolism, and PVR signaling. Aberration of these pathways might give SARS-CoV-2 a competitive edge over the host immune response. Also, SARS-CoV-2 can prevent the relaxin downstream signaling that plays a crucial role in the lung's overall functionality and its abnormal regulation might result in the respiratory complications found in COVID-19.

From previous studies, we compiled the information on deregulated genes (Blanco-Melo et al., 2020) and virus-host interactome (Gordon et al., 2020) in SARS-CoV-2 infection to get detailed pictures of the affected pathways, which are still obscure. Furthermore, we investigated how our identified host genetic and epigenetic factors are playing a role in these pathways. Taking a closer look, we illustrated some pathways that SARS-CoV-2 might be using but not SARS-CoV.

SARS-CoV-2 Host Interactions Might Lead to Respiratory Complications in COVID-19

COVID-19 patients are reported to have suffered from hypoxic conditions due to breathing complications (Cascella et al., 2020). HIF-1 signaling pathway provides significant support mechanisms during this hypoxic condition by activating a wide range of other stress-coping mechanisms and ultimately

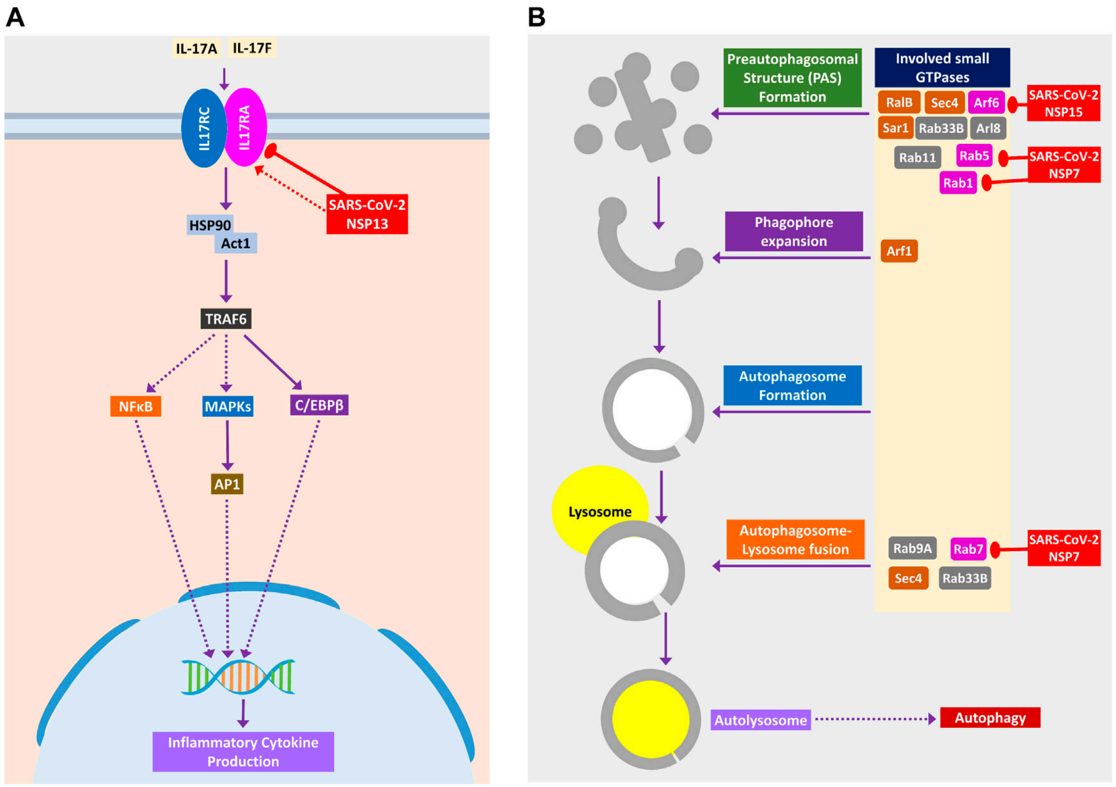


Figure 6 | (A) IL-17 signaling pathway. **(B)** Small GTPases in phagosome formation and maturation. Color codes are as in **Figure 5**.

leads to the survival of the stressed cells (Chen and Sang, 2016). So, if the infected cells utilize this survival mechanism, the viruses propagating within these will also be saved. Thus, SARS-CoV-2 could be supporting these survival mechanisms, as ORF10 protein binds and inhibits the E3 ubiquitin ligase complex, which degrades HIF-1α protein (**Figure 5A**); a similar phenomenon was also evident in other viruses (Mahon et al., 2014). Additionally, it can also stimulate the functions of eIF4E (Eukaryotic Translation Initiation Factor 4E) for the overproduction of HIF-1α (**Figure 5A**); a similar function was recorded for sapovirus protein VPg (Montero et al., 2015).

SARS-CoV-2-mediated overexpression of HIF signaling products might lead to pulmonary hypertension and acute lung injury (Shimoda and Semenza, 2011), which can be correlated with frequent lung failure in critically infected patients. HIF signaling promotes hypoxia-induced endothelial cell proliferation (Krock et al., 2011), which in turn might lead to aberrant clot formation in the presence of amplified inflammation (Yau et al., 2015) that is frequently found in many COVID-19 patients (Rotzinger et al., 2020).

Relaxin signaling plays a significant role in maintaining the lung's overall functionality by maintaining lung perfusion and gas exchange, relaxation of bronchi, and decreased inflammation in the lungs (Alexiou et al., 2013). SARS-CoV-2 NSP7 protein can perturb this signaling by binding and inhibiting the relaxin receptors and can prevent the production of NOS and MMP2/9 through PI3K to ERK1/2 axis (**Figure 5B**). NSP13, another SARS-CoV-2 protein, might bind and block PKA, and its failure to activate NFκB may lead to the blockade of the whole relaxin signaling pathway (**Figure 5B**). Aberration of this signaling pathway by SARS-CoV-2 possibly leads to breathing complications in COVID-19 patients.

Binding of IL-17 receptor by SARS-CoV-2 NSP13 might increase the downstream signaling by activated TRAF6 to NFκB/MAPKs/CEBPB, which might cause some pathogenic inflammatory responses (**Figure 6A**). Though IL-17 signaling is initially helpful in the host defense, still its aberrant expression might lead to pathogenic inflammatory responses leading to lung complications such as chronic obstructive pulmonary disease (COPD), lung fibrosis, pneumonia, and acute lung injury (Gurczynski and Moore, 2018).

SARS-CoV-2 Might Impede Autophagy and Phagocytosis to Ensure its Existence

During viral infection, one important immune response is autophagy that destroys the virus-infected cells, and viruses are often found to modulate it for their survival (Ahmad et al., 2018). SARS-CoV-2 can inhibit the formation of autophagosome (**Figure 6B**) and phagosome (**Figure 7A**) using several of its proteins. SARS-CoV-2 NSP15 and NSP7 might bind and inhibit the small members of Rab, Arf GTPases family (**Figure 6B**) necessary for autophagosome formation (Bento et al., 2013). SARS-CoV-2 NSP7 and ORF8 can inhibit several phagocytoses, promoting receptors such as scavenger receptors

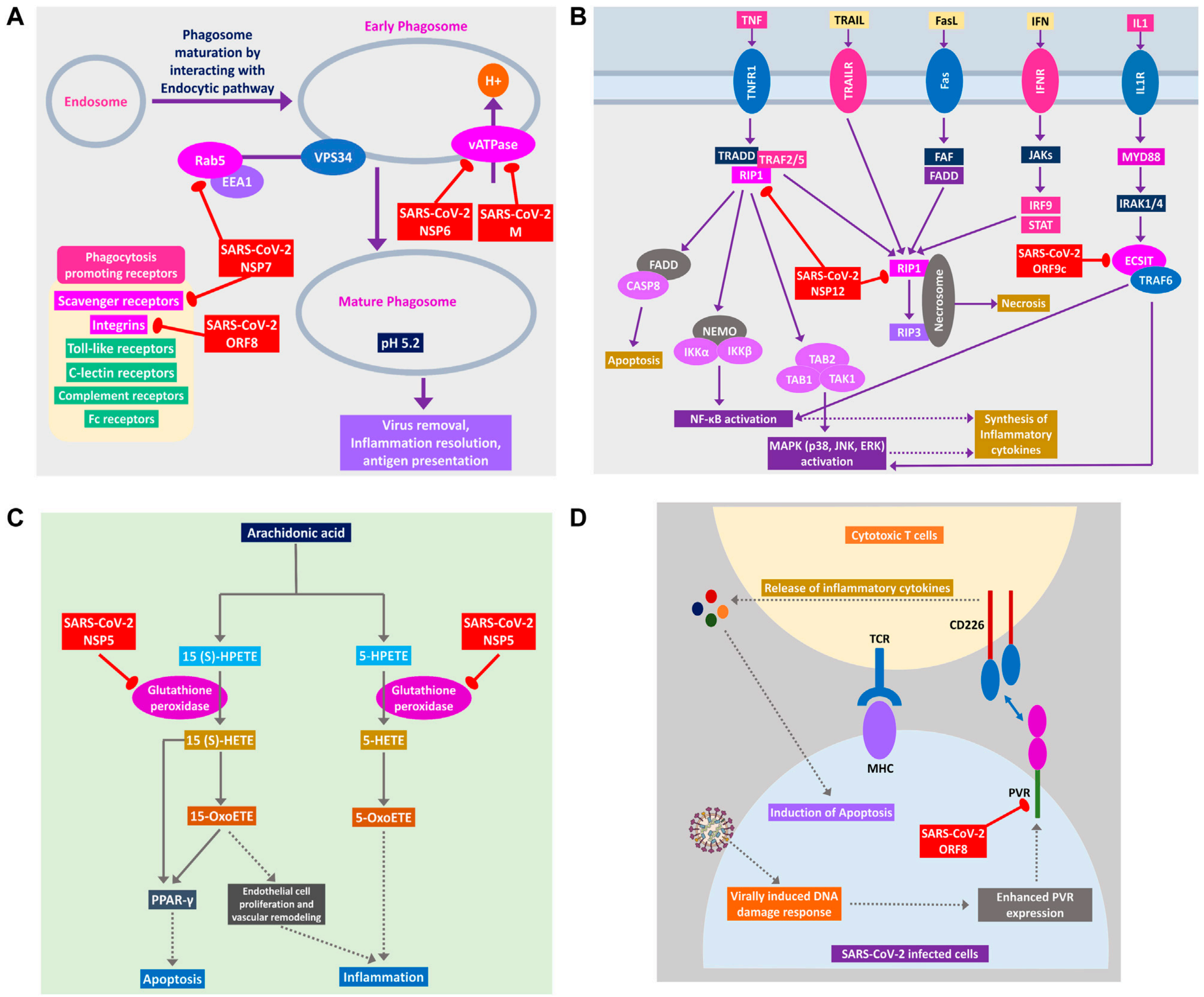


Figure 7 | (A) Phagosome maturation pathway, **(B)** RIP1 and ECSIT signaling pathway, **(C)** arachidonic acid metabolism pathway, and **(D)** cytotoxic T cells signaling pathway using CD226 and PVR axis. Color codes are as in **Figure 5**.

and integrins (**Figure 7A**); SARS-CoV-2 NSP6 and M protein might block v-ATPase, thus preventing the lowering of pH inside phagosome and preventing its maturation (**Figure 7A**); SARS-CoV-2 NSP7 prevents the phagosome-endosome/lysosome fusion by targeting Rab5 GTPase, which might result in viral pneumonia (Jakab et al., 1980) (**Figure 7A**).

SARS-CoV-2 Can Hinder Host Apoptotic Responses Necessary for Its Growth

Apoptosis plays important intracellular host immune response to reduce the further spread of viruses from the infected cells (Barber, 2001). Several signaling pathways are involved in eliciting this apoptotic response inside the infected cells, which can be suppressed by the viral proteins, as follows. NSP12 is found to target RIP1; thus, it might fail to relay signaling to CASP8/FADD-mediated apoptosis and necrosis by RIP1/RIP3 complex (**Figure 7B**). NSP5 might block glutathione peroxidase that is involved in 15(S)-HETE production and ultimately 15-oxoETE production (**Figure 7C**); thus, apoptotic induction by these metabolites through PPARγ signaling axis might not take place (Powell and Rokach, 2015). ORF8 can block PVR-CD226 signaling in cytotoxic T cell–mediated apoptosis (**Figure 7D**), as many viruses were reported to block PVR from expressing in the infected cell's membranes (Cifaldi et al., 2019). Also, infection-induced host miRNAs in β2-adrenergic signaling (**Supplementary Figure S4**) and insulin signaling can block apoptosis of the infected cell.

Host Antiviral Inflammatory Cytokine and Interferon Production Pathways Could Be Perturbed by SARS-CoV-2 Proteins

Cytokine signaling pathways play a major role in suppressing viral infections (Mogensen and Paludan, 2001). Similar inflammatory cytokine production pathways were also

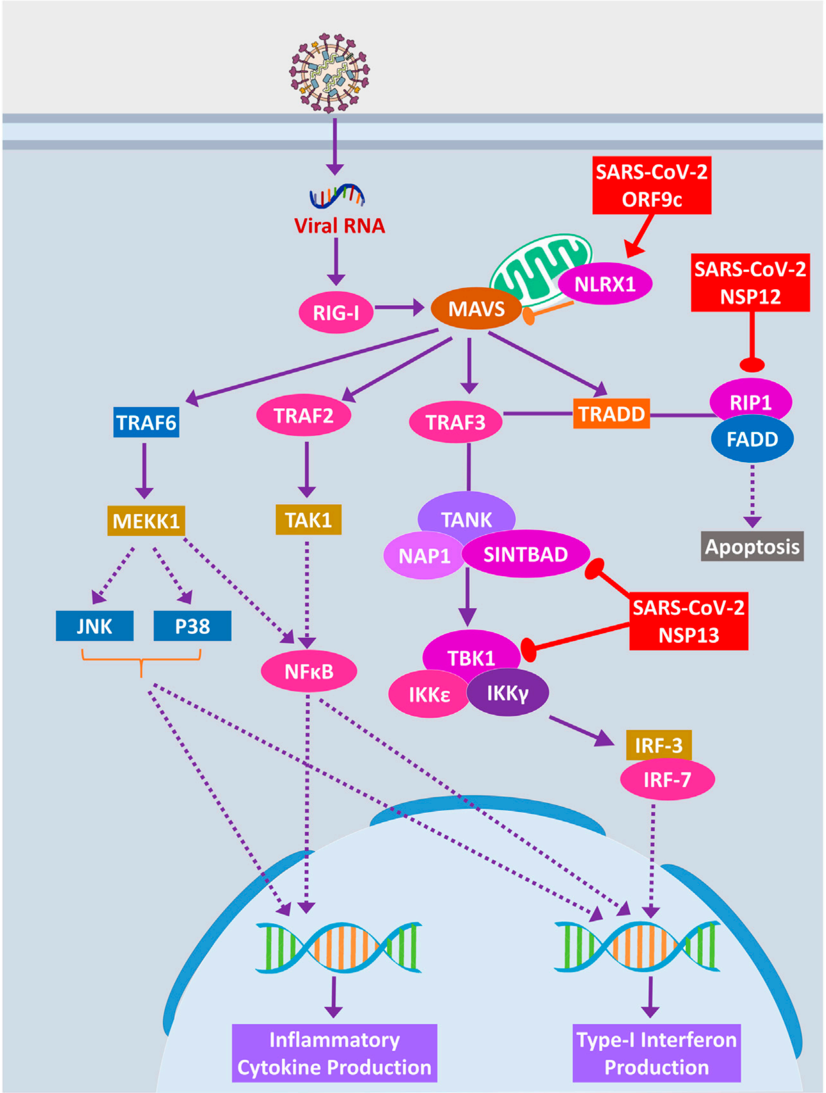


Figure 8 | RIG-I signaling pathway. Color codes are as in **Figure 5**.

reported in human coronaviral infections (Fung and Liu, 2019). We observed that SARS-CoV-2 proteins are interacting with the members of these pathways, and this might alter the signaling outcomes of these pathways to reduce the overall production of virus infection–induced inflammatory cytokines.

RIG-I signaling plays an important role in producing antiviral inflammatory cytokines and interferons and induction of apoptosis (Chan and Gack, 2015). SARS-CoV-2 ORF9c protein can activate NLRX1 to degrade MAVS, which results in failure of inflammatory cytokine production by NFkB via TRAF2/TAK1 or TRAF6/MEKK1 pathways (**Figure 8**). NLRX1 activity was previously found to be upregulated by HCV infection (Qin et al., 2017). By binding TBK1 and SINTBAD, SARS-CoV-2 NSP13 might inhibit the interferon production by IRF3/IRF7 stimulation (**Figure 8**).

A previous study suggested that RIP1 signaling plays a key role in human coronavirus infection (Meessen-Pinard et al., 2017). RIP1 along with TRADD and TRAF2/5 activates NFκB and MAPKs, then inducing the production inflammatory cytokines; this signaling axis might be blocked by SARS-CoV-2 NSP12 protein-RIP1 interaction, therefore inhibiting the antiviral mechanisms exerted by this signaling pathway (**Figure 7B**). SARS-CoV-2 ORF9c can also inhibit the ECSIT/TRAF6 signaling axis (**Figure 7B**), which plays pivotal antiviral roles by activating NFκB and MAPKs signaling (Lei et al., 2015).

β2-Adrenergic signaling plays important antiviral roles in respiratory virus infections (Ağaç et al., 2018). SARS-CoV-2 NSP13 interacts with PKA, which might inhibit the activation of CREB by PKA for producing antiviral inflammatory responses (**Supplementary Figure S4**).

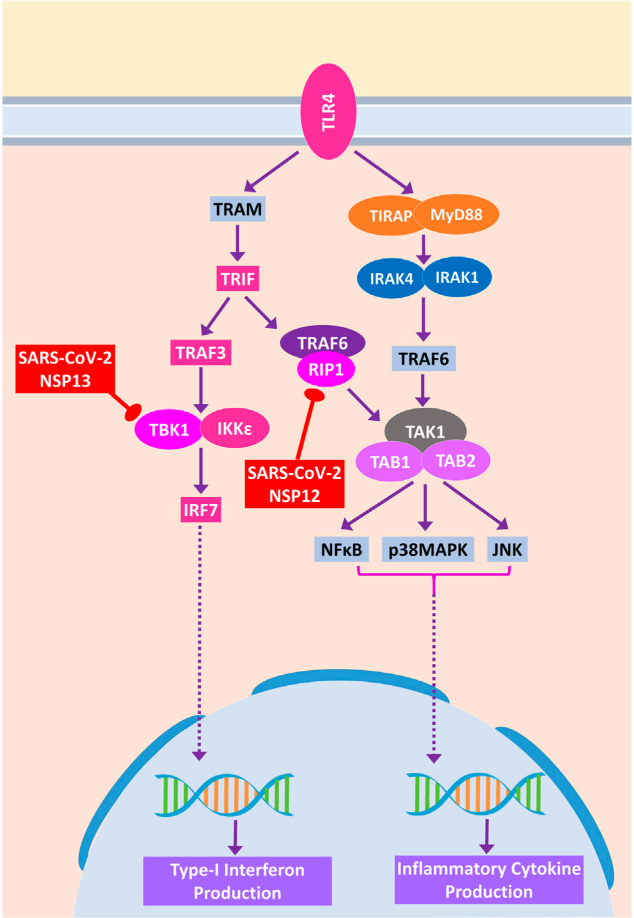


Figure 9 | Toll-like receptor 4 (TLR4) signaling. Color codes are as in **Figure 5**.

Previous studies showed that arachidonic acids suppress the replication of HCoV-229E and MERS-CoV (Yan et al., 2019). 15-OxoETE, an arachidonic acid metabolism product that promotes pulmonary artery endothelial cell proliferation during hypoxia (Ma et al., 2014), in turn results in vascular remodeling and leakage of inflammatory cytokines (Powell and Rokach, 2015). 5-OxoETE, another arachidonic acid metabolism product that can also induce inflammation (Powell and Rokach, 2015), production of both compounds might be hindered by SARS-CoV-2 NSP5 as it interacts with an upstream metabolic enzyme glutathione peroxidase (**Figure 7C**).

Previously, it was reported that IL-17 signaling enhances antiviral immune responses (Ma et al., 2019). SARS-CoV-2 NSP13 can bind IL-17 receptor and inhibit the downstream signaling from IL-17 receptor to TRAF6 for activating NFκB/MAPKs/CEBPB signaling axis, thus decreasing the antiviral inflammatory responses (**Figure 6A**).

During acute viral infections, Toll-like receptor 4 (TLR4) signaling plays an important role in eliciting inflammatory responses (Olejnik et al., 2018). SARS-CoV-2 protein NSP13 interacts with TBK1, which might reduce the signaling from IRF7, resulting in less IFN-I productions, whereas NSP12 interacts with RIP1; as a result, the activation of downstream NFκB and MAPKs (p38 and JNK) pathways and induction of inflammatory responses from these pathways might be stalled (**Figure 9**).

SARS-CoV-2 Infection Can Negatively Regulate Fatty Acid Metabolism for Its Proliferation

Host lipid and fatty acid metabolism play a crucial role in maintaining the viral life cycle and propagation inside the infected cells as viruses tend to utilize host metabolic pathways to their aid (Martín-Acebes et al., 2012). Cui et al. (2019) showed that impairment of fatty acid oxidation could lead to acute lung injury (Cui et al., 2019). SARS-CoV-2 NSP2 can interact with FATP receptor of fatty acid oxidation pathway, whereas M protein can interact and destabilize MCAD of fatty acid oxidation pathway; moreover, two other members can be the targets of human miRNA (**Figure 10**). MCAD deficiency leads to pulmonary hemorrhage and cardiac dysfunction in neonates (Maclean et al., 2005). So, this destabilization might lead to acute lung injury during COVID-19. Increased fatty acid biosynthetic pathways are found in several viral infections for their efficient multiplications (Martín-Acebes et al., 2012), so it is logical for SARS-CoV-2 to inhibit the fatty acid degrading pathways. SARS-CoV-2 NSP13 protein was found to interact with insulin signaling–mediated antilipolysis by PKA-HSL signaling axis; SARS-CoV-2 M and several host miRNAs can inhibit fatty acid degradation to CoA through CPT1-CPT2 metabolic axis (**Figure 10**) so that more fatty acids can be produced.

DISCUSSION

The tug-of-war between the viral pathogens and the infected host's response upon the infection is a critical and complex relationship deciding the ultimate fate of infection. Although most of the time, successful removal of the virus is achieved through host immune response, viruses have also evolved some immune evasion mechanisms to escape the immune surveillance of the host, thus making the outcomes of the disease more complicated (Fung et al., 2020). Similar interactions were also found in other human coronaviruses that modulate the host immune responses (Fung and Liu, 2019; Fung et al., 2020). In this study, we have depicted how SARS-CoV-2 and host protein interactome lead to the probable immune escape mechanisms of this novel virus, along with the functional roles of other host epigenetic factors in this interaction as host epigenetic factors serve important roles in viral infections (Bussfeld et al., 1997; Paschos and Allday, 2010; Girardi et al., 2018).

Our analysis showed several TFs are capable of binding around the promoters of deregulated genes found in SARS-CoV-2 infection, which were absent in SARS-CoV infection. Some of these downregulated TFs in SARS-CoV-2 infection, such as MAF and FOXO1, can elicit proviral responses (Kim and Seed, 2010; Lei et al., 2013). Also, some upregulated TFs in SARS-CoV-2 infection, such as HDAC7 (Herbein and Wendling,

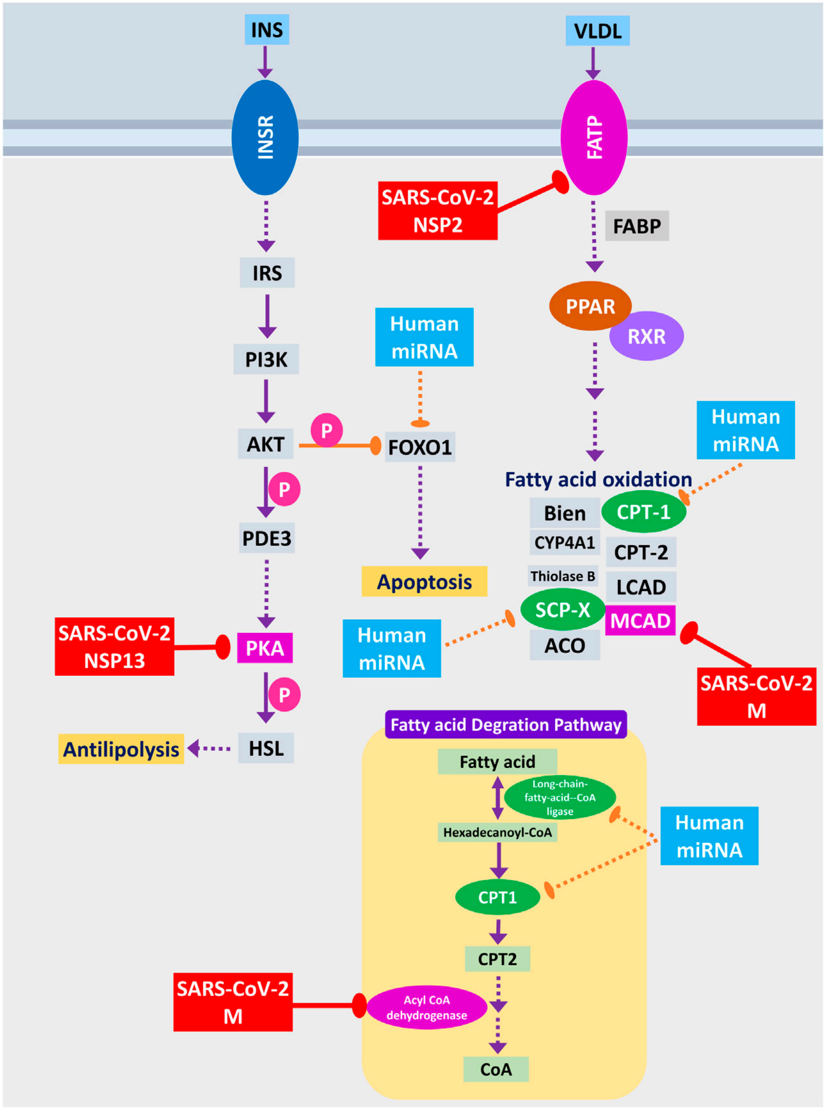


Figure 10 | Fatty acid oxidation, fatty acid degradation, and antilipolysis through insulin signaling. Color codes are as in **Figure 5**.

2010), STAT2 (Le-Trilling et al., 2018), ATF4 (Caselli et al., 2012), and FOSL1 (Cai et al., 2017), can facilitate the progression of the viral life cycle and immune evasion in the host. However, other upregulated TFs in SARS-CoV-2 infection TRIM25 (Martín-Vicente et al., 2017), SMAD3 (Qing et al., 2004), STAT1, IRF7, and IRF9 (Chiang and Liu, 2019) might play a role in the antiviral immunity upon infection.

Interestingly, we detected 2 miRNAs, namely, hsa-miR-429 and hsa-miR-1286, whose associated TFs were upregulated and their target genes were downregulated, suggesting they might have some roles in this host-virus interaction. In RSV infection, hsa-miR-429 was found to be upregulated in severe disease conditions (Inchley et al., 2015). Other studies showed that this miRNA plays an important role in promoting viral replication and reactivation from latency (Ellis-Connell et al., 2010; Bernier and Sagan, 2018). So, the expression of this miRNA in SARS-CoV-2 infection can lead to similar disease outcomes.

Upregulated epigenetic factors in SARS-CoV-2 infections can be both a boon and a bane for the host, as factors such as TRIM16 (van Tol et al., 2017) and DTX3L (Zhang et al., 2015) can provide antiviral responses, whereas factors such as PRDM1 (also known as BLIMP-1) (Lu et al., 2014) and HDAC7 (Herbein and Wendling, 2010) can act as proviral factors.

Upregulated TFs such as SMAD3, MYC, NFKB1, and STAT1 might be involved in the upregulation of hsa-miR-18a, hsa-miR-155, hsa-miR-210, hsa-miR-429, and hsa-miR-433, which can, in turn, downregulate the HIF-1 production (**Supplementary Figure S5**) (Serocki et al., 2018). Also, epigenetic factors like PRMT1 can downregulate HIF-1 expression, whereas factors like HDAC7 can increase the transcription of HIF-1 (Luo and Wang, 2018). We found that when MYC, SMAD3, and TNF TFs are upregulated, they can activate autophagy and apoptosis promoting miRNAs such as hsa-miR-17, hsa-miR-20, and hsa-miR-106 (**Supplementary Figure S5**) (Xu et al., 2012). RIG-I signaling can induce miRNAs such as hsa-miR-24, hsa-miR-32, hsa-miR-125, and hsa-miR-150 to neutralize viral threats (Li and Shi, 2013), and we also found that these miRNAs can be transcribed by the upregulated TFs (**Supplementary Figure S5**). RIP1 can be targeted by induced hsa-miR-24 and hsa-miR-155 (Liu et al., 2011; Tan et al., 2018) in SARS-CoV-2 infection (**Supplementary Figure S5**). IL-17F can be targeted by the upregulated hsa-miR-106a, hsa-miR-17, and hsa-miR-20a, whereas IL-17A expression can be modulated by the upregulated hsa-miR-146a and hsa-miR-30c (**Supplementary Figure S5**) (Mai et al., 2012). We also found that different upregulated TFs induced miRNAs such as hsa-miR-146a and hsa-miR-155 can downregulate TLR4 signaling (**Supplementary Figure S5**) (Yang et al., 2011).

From the enrichment analysis, we found that deregulated genes were also involved in processes and functions such as heart development, kidney development, AGE-RAGE signaling pathway in diabetic complications, zinc ion binding, and calcium ion binding (**Figures 2A,B**). Impairment of these organ-specific functions might suggest the increased susceptibility to COVID-19 in patients having comorbidities. Zinc and calcium ions play significant roles in activating different immune responses; aberrant regulation of these might be lethal for COVID-19 patients (Verma et al., 2011; Read et al., 2019).

Host antiviral immune responses encompassing different mechanisms such as autophagy, apoptosis, interferon signaling, and inflammation play a fundamental role in neutralizing viral threats found not only in human coronaviruses (Fung and Liu, 2019) but also in other viral infections (Barber, 2001; Mogensen and Paludan, 2001; Ahmad et al., 2018). Viruses have also evolved hijacking mechanisms to bypass all those mechanisms for their survival (Fung et al., 2020). From our analyses, we have observed that SARS-CoV-2 might have similar modes of action for its successful immune escape from the host immune surveillance. While all these mechanisms are supporting viral propagation, the host suffers their adverse effects, resulting in severe complications in COVID-19 patients.

All of these findings suggest that a very complex host-virus interaction takes place during the SARS-CoV-2 infections. During the infections, while some host responses have a significant impact on eradicating the viruses from the body, the virus modulates some critical host proteins and epigenetic machinery for its successful replication and evasion of host immune responses. Due to these complex cross-talk between the host and virus, the disease complications of COVID-19 might arise.

In this present study, our goal was to model putative pathogenic alterations within the host biological pathways, and for simplicity, we used only one infection model each for SARS-CoV and SARS-CoV-2. We used several transcriptomes and interactome data from various infection systems to formulize the modeling. Though this modeling could be more precisely done if there were data on uniform infection system, such approach is frequently used in computational systems biology–related modeling approaches when the required data are scarce. We integrated epigenetic factors and miRNAs into our modeling based on the DE patterns observed from the infection models, as the direct evidence of their modulation is yet to be experimentally reported in SARS-CoV-2 infection. Our results would be helpful for further in-depth experimental design to understand the detailed molecular mechanisms of COVID-19 pathogenesis and to develop some potential therapeutic approaches targeting these host-virus interactions.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AI and MK conceived the project, designed the workflow, and performed the analyses. MK and AI wrote the manuscript. Both authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at Biorxiv.org (https://www.biorxiv.org/content/10.1101/2020.05.06.050260v1), (Khan and Islam, 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.598583/full#supplementary-material.

REFERENCES

Adhya, D., and Basu, A. (2010). Epigenetic modulation of host: new insights into immune evasion by viruses. *J. Biosci.* 35 (4), 647–663. doi:10.1007/s12038-010-0072-9

Ahmad, L., Mostowy, S., and Sancho-Shimizu, V. (2018). Autophagy-virus interplay: from cell biology to human disease. *Front. Cell. Dev. Biol.* 6, 155. doi:10.3389/fcell.2018.00155

Alexiou, K., Wilbring, M., Matschke, K., and Dschietzig, T. (2013). Relaxin protects rat lungs from ischemia-reperfusion injury via inducible NO synthase: role of ERK-1/2, PI3K, and forkhead transcription factor FKHRL1. *PLoS One* 8 (9), e75592. doi:10.1371/journal.pone.0075592

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11 (10), R106. doi:10.1186/gb-2010-11-10-r106

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Ağaç, D., Gill, M. A., and Farrar, J. D. (2018). Adrenergic signaling at the interface of allergic asthma and viral infections. *Front. Immunol.* 9, 736. doi:10.3389/fimmu.2018.00736

Barber, G. N. (2001). Host defense, viruses and apoptosis. *Cell Death Differ.* 8 (2), 113–126. doi:10.1038/sj.cdd.4400823

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* 41 (D1), D991–D995. doi:10.1093/nar/gks1193

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* 57 (1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bento, C. F., Puri, C., Moreau, K., and Rubinsztein, D. C. (2013). The role of membrane-trafficking small GTPases in the regulation of autophagy. *J. Cell Sci.* 126 (5), 1059–1069. doi:10.1242/jcs.123075

Bernier, A., and Sagan, S. M. (2018). The diverse roles of microRNAs at the Host⁻Virus interface. *Viruses* 10 (8), 440. doi:10.3390/v10080440

Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., et al. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* 181 (5), 1036–1045.e9. doi:10.1016/j.cell.2020.04.026

Bruscella, P., Bottini, S., Baudesson, C., Pawlotsky, J. M., Feray, C., and Trabucchi, M. (2017). Viruses and miRNAs: more friends than foes. *Front. Microbiol.* 8, 824. doi:10.3389/fmicb.2017.00824

Bussfeld, D., Bacher, M., Moritz, A., Gemsa, D., and Sprenger, H. (1997). Expression of transcription factor genes after influenza A virus infection. *Immunobiology* 198 (1–3), 291–298. doi:10.1016/s0171-2985(97)80049-6

Cai, B., Wu, J., Yu, X., Su, X. Z., and Wang, R. F. (2017). FOSL1 inhibits type I interferon responses to malaria and viral infections by blocking TBK1 and TRAF3/TRIF interactions. *mBio* 8 (1), e02161. doi:10.1128/mBio.02161-16

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., and Di Napoli, R. (2020). *Features, evaluation and treatment coronavirus (COVID-19)*. Treasure Island, FL: StatPearls Publishing.

Caselli, E., Benedetti, S., Grigolato, J., Caruso, A., and Di Luca, D. (2012). Activating transcription factor 4 (ATF4) is upregulated by human herpesvirus 8 infection, increases virus replication and promotes proangiogenic properties. *Arch. Virol.* 157 (1), 63–74. doi:10.1007/s00705-011-1144-3

Chan, Y. K., and Gack, M. U. (2015). RIG-I-like receptor regulation in virus infection and immunity. *Curr. Opin. Virol.* 12, 7–14. doi:10.1016/j.coviro.2015.01.004

Chen, S., and Sang, N. (2016). Hypoxia-Inducible factor-1: a critical player in the survival strategy of stressed cells. *J. Cell. Biochem.* 117 (2), 267–278. doi:10.1002/jcb.25283

Chiang, H. S., and Liu, H. M. (2018). The molecular basis of viral inhibition of IRF- and STAT-dependent immune responses. *Front. Immunol.* 9, 3086. doi:10.3389/fimmu.2018.03086

Cifaldi, L., Doria, M., Cotugno, N., Zicari, S., Cancrini, C., Palma, P., et al. (2019). DNAM-1 activating receptor and its ligands: how do viruses affect the NK cell-mediated immune surveillance during the various phases of infection? *Int. J. Mol. Sci.* 20 (15), 3715. doi:10.3390/ijms20153715

Cui, H., Xie, N., Banerjee, S., Ge, J., Guo, S., and Liu, G. (2019). Impairment of fatty acid oxidation in alveolar epithelial cells mediates acute lung injury, *Am. J. Respir. Cell Mol. Biol.* 60 (2), 167–178. doi:10.1165/rcmb.2018-0152OC

Ellis-Connell, A. L., Iempridee, T., Xu, I., and Mertz, J. E. (2010). Cellular MicroRNAs 200b and 429 regulate the epstein-barr virus switch between latency and lytic replication. *J. Virol.* 84 (19), 10329–10343. doi:10.1128/jvi.00923-10

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Cáccamo, M., Chen, Y., et al. (2007). Ensembl 2008. *Nucleic Acids Res.* 36 (1), D707–D714. doi:10.1093/nar/gkm988

Fung, S.-Y., Yuen, K.-S., Ye, Z.-W., Chan, C.-P., and Jin, D.-Y. (2020). A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses. *Emerg. Microb. Infect.* 9 (1), 558–570. doi:10.1080/22221751.2020.1736644

Fung, T. S., and Liu, D. X. (2019). Human coronavirus: host-pathogen interaction. *Annu. Rev. Microbiol.* 73 (1), 529–557. doi:10.1146/annurev-micro-020518-115759

Girardi, E., López, P., and Pfeffer, S. (2018). On the importance of host MicroRNAs during viral infection. *Front. Genet.* 9, 439. doi:10.3389/fgene.2018.00439

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., O'Meara, M. J., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing.*Nature* 583, 459–468. doi:10.1038/s41586-020-2286-9

Gurczynski, S. J., and Moore, B. B. (2018). IL-17 in the lung: the good, the bad, and the ugly. *Am. J. Physiol. Lung Cell Mol. Physiol.* 314 (1), L6–L16. doi:10.1152/ajplung.00344.2017

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in illumina transcriptase sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38 (12), e131. doi:10.1093/nar/gkq224

Herbein, G., and Wendling, D. (2010). Histone deacetylases in viral infections. *Clin. Epigenet.* 1 (1-2), 13–24. doi:10.1007/s13148-010-0003-5

Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2019a). miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.* 48 (D1), D148–D154. doi:10.1093/nar/gkz896

Huang, R., Grishagin, I., Wang, Y., Zhao, T., Greene, J., Obenauer, J. C., et al. (2019b). The NCATS BioPlanet—an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. *Front. Pharmacol.* 10, 445. doi:10.3389/fphar.2019.00445

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Cáccamo, M., Chen, Y., et al. (2007). Ensembl 2007. *Nucleic Acids Res.* 35 (1), D610–D617. doi:10.1093/nar/gkl996

Inchley, C. S., Sonerud, T., Fjærli, H. O., and Nakstad, B. (2015). Nasal mucosal microRNA expression in children with respiratory syncytial virus infection. *BMC Infect. Dis.* 15, 150. doi:10.1186/s12879-015-0878-z

Jakab, G. J., Warr, G. A., and Sannes, P. L. (1980). Alveolar macrophage ingestion and phagosome-lysosome fusion defect associated with virus pneumonia. *Infect. Immun.* 27 (3), 960–968. doi:10.1128/IAI.27.3.960-968.1980

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Kanehisa, M., and Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci.* 29 (1), 28–35. doi:10.1002/pro.3711

Katze, M. G., Fornek, J. L., Palermo, R. E., Walters, K. A., and Korth, M. J. (2008). Innate immune modulation by RNA viruses: emerging insights from functional genomics. *Nat. Rev. Immunol.* 8 (8), 644–654. doi:10.1038/nri2377

Kauffmann, A., Gentleman, R., and Huber, W. (2009). Array quality metrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25 (3), 415–416. doi:10.1093/bioinformatics/btn647

Khan, M. A.-A.-K., and Islam, A. B. M. M. K. (2020). SARS-CoV-2 proteins exploit host's genetic and epigenetic mediators for the annexation of key host signaling pathways that confers its immune evasion and disease pathophysiology. *bioRxiv*. doi:10.1101/2020.05.06.050260

Kikkert, M. (2020). Innate immune evasion by human respiratory RNA viruses. *J. Innate Immun.* 12 (1), 4–20. doi:10.1159/000503030

Kim, H., and Seed, B. (2010). The transcription factor MafB antagonizes antiviral responses by blocking recruitment of coactivators to the transcription factor IRF3. *Nat. Immunol.* 11 (8), 743–750. doi:10.1038/ni.1897

Krock, B. L., Skuli, N., and Simon, M. C. (2011). Hypoxia-induced angiogenesis: good and evil. *Genes Cancer* 2 (12), 1117–1133. doi:10.1177/1947601911423654

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860–921. doi:10.1038/35057062

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., et al. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Intern. Med.* 172 (9), 577–582. doi:10.7326/m20-0504

Le-Trilling, V. T. K., Wohlgemuth, K., Rückborn, M. U., Jagnjic, A., Maaßen, F., Timmer, L., et al. (2018). STAT2-Dependent immune responses ensure host survival despite the presence of a potent viral antagonist. *J. Virol.* 92 (14). doi:10.1128/jvi.00296-18

Lei, C. Q., Zhang, Y., Li, M., Jiang, L. Q., Zhong, B., Kim, Y. H., et al. (2015). ECSIT bridges RIG-I-like receptors to VISA in signaling events of innate antiviral responses. *J. Innate Immun.* 7 (2), 153–164. doi:10.1159/000365971

Lei, C. Q., Zhang, Y., Xia, T., Jiang, L. Q., Zhong, B., and Shu, H. B. (2013). FoxO1 negatively regulates cellular antiviral response by promoting degradation of IRF3. *J. Biol. Chem.* 288 (18), 12596–12604. doi:10.1074/jbc.M112.444794

Li, Y., and Shi, X. (2013). MicroRNAs in the regulation of TLR and RIG-I pathways. *Cell. Mol. Immunol.* 10 (1), 65–71. doi:10.1038/cmi.2012.55

Liao, Y., Smyth, G. K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41 (10), e108. doi:10.1093/nar/gkt214

Liu, J., van Mil, A., Vrijsen, K., Zhao, J., Gao, L., Metz, C. H., et al. (2011). MicroRNA-155 prevents necrotic cell death in human cardiomyocyte progenitor cells via targeting RIP1. *J. Cell Mol. Med.* 15 (7), 1474–1482. doi:10.1111/j.1582-4934.2010.01104.x

Lu, P., Youngblood, B. A., Austin, J. W., Mohammed, A. U., Butler, R., Ahmed, R., et al. (2014). Blimp-1 represses CD8 T cell expression of PD-1 using a feed-forward transcriptional circuit during acute viral infection. *J. Exp. Med.* 211 (3), 515–527. doi:10.1084/jem.20130208

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395 (10224), 565–574. doi:10.1016/s0140-6736(20)30251-8

Luo, W., and Wang, Y. (2018). Epigenetic regulators: multifunctional proteins modulating hypoxia-inducible factor-α protein stability and activity. *Cell. Mol. Life Sci.* 75 (6), 1043–1056. doi:10.1007/s00018-017-2684-9

Lyles, D. S. (2000). Cytopathogenesis and inhibition of host gene expression by RNA viruses. *Microbiol. Mol. Biol. Rev.* 64 (4), 709–724. doi:10.1128/mmbr.64.4.709-724.2000

Ma, C., Liu, Y., Wang, Y., Zhang, C., Yao, H., Ma, J., et al. (2014). Hypoxia activates 15-PGDH and its metabolite 15-KETE to promote pulmonary artery endothelial cells proliferation via ERK1/2 signalling. *Br. J. Pharmacol.* 171 (14), 3352–3363. doi:10.1111/bph.12594

Ma, W. T., Yao, X. T., Peng, Q., and Chen, D. K. (2019). The protective and pathogenic roles of IL-17 in viral infections: friend or foe? *Open Biol.* 9(7), 190109. doi:10.1098/rsob.190109

Maclean, K., Rasiah, V. S., Kirk, E. P., Carpenter, K., Cooper, S., Lui, K., et al. (2005). Pulmonary haemorrhage and cardiac dysfunction in a neonate with medium-chain acyl-CoA dehydrogenase (MCAD) deficiency. *Acta Paediatr.* 94 (1), 114–116. doi:10.1111/j.1651-2227.2005.tb01797.x

Mahon, C., Krogan, N. J., Craik, C. S., and Pick, E. (2014). Cullin E3 ligases and their rewiring by viral factors. *Biomolecules* 4 (4), 897–930. doi:10.3390/biom4040897

Mai, J., Virtue, A., Maley, E., Tran, T., Yin, Y., Meng, S., et al. (2012). MicroRNAs and other mechanisms regulate interleukin-17 cytokines and receptors. *Front. Biosci.* 4, 1478–1495. doi:10.2741/474

Martín-Acebes, M. A., Vázquez-Calvo, Á., Caridi, F., Saiz, J.-C., and Sobrino, F. (2012). Lipid involvement in viral infections: present and future perspectives for the design of antiviral strategies, lipid metabolism, Rodrigo Valenzuela Baez. *IntechOpen*. doi:10.5772/51068

Martín-Vicente, M., Medrano, L. M., Resino, S., García-Sastre, A., and Martínez, I. (2017). TRIM25 in the regulation of the antiviral innate immunity. *Front. Immunol.* 8, 1187. doi:10.3389/fimmu.2017.01187

Medvedeva, Y. A., Lennartsson, A., Ehsani, R., Kulakovskiy, I. V., Vorontsov, I. E., Panahandeh, P., et al. (2015). EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database* 2015, bav067. doi:10.1093/database/bav067

Meessen-Pinard, M., Le Coupanec, A., Desforges, M., and Talbot, P. J. (2017). Pivotal role of receptor-interacting protein kinase 1 and mixed lineage kinase domain-like in neuronal cell death induced by the human neuroinvasive coronavirus OC43. *J. Virol.* 91 (1). doi:10.1128/jvi.01513-16

Mogensen, T. H., and Paludan, S. R. (2001). Molecular pathways in virus-induced cytokine production. *Microbiol. Mol. Biol. Rev.* 65 (1), 131–150. doi:10.1128/MMBR.65.1.131-150.2001

Montero, H., García-Román, R., and Mora, S. I. (2015). eIF4E as a control target for viruses. *Viruses* 7 (2), 739–750. doi:10.3390/v7020739

Ncbi-Gene (2020). *RE: gene links for nucleotide (select 1798174254)—gene—NCBI*. Bethesda, MD: NCBI.

Olejnik, J., Hume, A. J., and Mühlberger, E. (2018). Toll-like receptor 4 in acute viral infection: too much of a good thing. *PLoS Pathog.* 14 (12), e1007390. doi:10.1371/journal.ppat.1007390

Paschos, K., and Allday, M. J. (2010). Epigenetic reprogramming of host genes in viral and microbial pathogenesis. *Trends Microbiol.* 18 (10), 439–447. doi:10.1016/j.tim.2010.07.003

Perez-Llamas, C., and Lopez-Bigas, N. (2011). Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS One* 6 (5), e19541. doi:10.1371/journal.pone.0019541

Powell, W. S., and Rokach, J. (2015). Biosynthesis, biological effects, and receptors of hydroxyeicosatetraenoic acids (HETEs) and oxoeicosatetraenoic acids (oxo-ETEs) derived from arachidonic acid. *Biochim. Biophys. Acta* 1851 (4), 340–355. doi:10.1016/j.bbalip.2014.10.008

Qin, Y., Xue, B., Liu, C., Wang, X., Tian, R., Xie, Q., et al. (2017). NLRX1 mediates MAVS degradation to attenuate hepatitis C virus-induced innate immune response through PCBP2. *J. Virol.* 91 (23), JVI.01264-17. doi:10.1128/jvi.01264-17

Qing, J., Liu, C., Choy, L., Wu, R. Y., Pagano, J. S., and Derynck, R. (2004). Transforming growth factor beta/Smad3 signaling regulates IRF-7 function and transcriptional activation of the beta interferon promoter. *Mol. Cell Biol.* 24 (3), 1411–1425. doi:10.1128/mcb.24.3.1411-1425.2004

Read, S. A., Obeid, S., Ahlenstiel, C., and Ahlenstiel, G. (2019). The role of zinc in antiviral immunity. *Adv. Nutr.* 10 (4), 696–710. doi:10.1093/advances/nmz013

Ren, L.-L., Wang, Y.-M., Wu, Z.-Q., Xiang, Z.-C., Guo, L., Xu, T., et al. (2020). Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. *Chin. Med. J.* 133 (9), 1015–1024. doi:10.1097/CM9.0000000000000722

Rotzinger, D. C., Beigelman-Aubry, C., von Garnier, C., and Qanadli, S. D. (2020). Pulmonary embolism in patients with COVID-19: time to change the paradigm of computed tomography. *Thromb. Res.* 190, 58–59. doi:10.1016/j.thromres.2020.04.011

Serocki, M., Bartoszewska, S., Janaszak-Jasiecka, A., Ochocka, R. J., Collawn, J. F., and Bartoszewski, R. (2018). miRNAs regulate the HIF switch during hypoxia: a novel therapeutic target. *Angiogenesis* 21 (2), 183–202. doi:10.1007/s10456-018-9600-2

Shimoda, L. A., and Semenza, G. L. (2011). HIF and the lung: role of hypoxia-inducible factors in pulmonary development and disease. *Am. J. Respir. Crit. Care Med.* 183 (2), 152–156. doi:10.1164/rccm.201009-1393PP

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46 (D1), D661–D667. doi:10.1093/nar/gkx1064

Smyth, G. K. (2005). *Limma: linear models for microarray data*. Berlin, Germany: Springer, 397–420.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Tan, H., Qi, J., Fan, B. Y., Zhang, J., Su, F. F., and Wang, H. T. (2018). MicroRNA-24-3p attenuates myocardial ischemia/reperfusion injury by suppressing RIPK1 expression in mice. *Cell. Physiol. Biochem.* 51 (1), 46–62. doi:10.1159/000495161

Tong, Z., Cui, Q., Wang, J., and Zhou, Y. (2019). TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Res.* 47 (D1), D253–D258. doi:10.1093/nar/gky1023

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25 (9), 1105–1111. doi:10.1093/bioinformatics/btp120

van Tol, S., Hage, A., Giraldo, M. I., Bharaj, P., and Rajsbaum, R. (2017). The TRIMendous role of TRIMs in virus-host interactions. *Vaccines* 5 (3), 23. doi:10.3390/vaccines5030023

Verma, S., Hoffmann, F. W., Kumar, M., Huang, Z., Roe, K., Nguyen-Wu, E., et al. (2011). Selenoprotein K knockout mice exhibit deficient calcium flux in immune cells and impaired immune responses. *J. Immunol.* 186 (4), 2127–2137. doi:10.4049/jimmunol.1002878

WHO (2020). *RE: WHO Director-General's opening remarks at the media briefing on COVID-19—3 March 2020.* Geneva, Switzerland: WHO.

Worldometer (2020). Coronavirus Cases. Worldometer, USA. Available at: https://www.worldometers.info/coronavirus/ (Accessed April 4, 2020).

Xu, J., Wang, Y., Tan, X., and Jing, H. (2012). MicroRNAs in autophagy and their emerging roles in crosstalk with apoptosis. *Autophagy* 8 (6), 873–882. doi:10.4161/auto.19629

Yan, B., Chu, H., Yang, D., Sze, K.-H., Lai, P.-M., Yuan, S., et al. (2019). Characterization of the lipidomic profile of human coronavirus-infected cells: implications for lipid metabolism remodeling upon coronavirus replication. *Viruses* 11 (1), 73. doi:10.3390/v11010073

Yang, K., He, Y. S., Wang, X. Q., Lu, L., Chen, Q. J., Liu, J., et al. (2011). MiR-146a inhibits oxidized low-density lipoprotein-induced lipid accumulation and inflammatory response via targeting toll-like receptor 4. *FEBS Lett.* 585 (6), 854–860. doi:10.1016/j.febslet.2011.02.009

Yau, J. W., Teoh, H., and Verma, S. (2015). Endothelial cell control of thrombosis. *BMC Cardiovasc. Disord.* 15, 130. doi:10.1186/s12872-015-0124-z

Yoshikawa, T., Hill, T. E., Yoshikawa, N., Popov, V. L., Galindo, C. L., Garner, H. R., et al. (2010). Dynamic innate immune responses of human bronchial epithelial cells to severe acute respiratory syndrome-associated coronavirus infection. *PLoS One* 5 (1), e8729. doi:10.1371/journal.pone.0008729

Zhang, Y., Mao, D., Roswit, W. T., Jin, X., Patel, A. C., Patel, D. A., et al. (2015). PARP9-DTX3L ubiquitin ligase targets host histone H2BJ and viral 3C protease to enhance interferon signaling and control viral infection. *Nat. Immunol.* 16 (12), 1215–1227. doi:10.1038/ni.3279

Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., et al. (2018). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* 47 (D1), D729–D735. doi:10.1093/nar/gky1094

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 01 April 2021
doi: 10.3389/fmolb.2020.585245

Delayed Modeling Approach to Forecast the Periodic Behavior of SARS-2

Zhenhua Yu[1], *Ayesha Sohail*[2*], *Alessandro Nutini*[3] *and Robia Arif*[2]

[1] Institute of Systems Security and Control, College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, China, [2] Department of Mathematics, Comsats University Islamabad, Lahore, Pakistan, [3] Centro Studi Attività Motorie, Biology and Biomechanics Department, Lucca, Italy

The ongoing threat of Coronavirus is alarming. The key players of this virus are modeled mathematically during this research. The transmission rates are hypothesized, with the aid of epidemiological concepts and recent findings. The model reported is extended, by taking into account the delayed dynamics. Time delay reflects the fact that the dynamic behavior of transmission of the disease, at time t depends not only on the state at time t but also on the state in some period τ before time t. The research presented in this manuscript will not only help in understanding the current threat of pandemic (SARS-2), but will also contribute in making precautionary measures and developing control strategies.

Keywords: SARS-CoV2, dynamical analysis, kinetic modeling, numerical simulations, monoclonal antibody, theoretical analysis

Edited by:
Nadine Kabbani,
George Mason University,
United States

Reviewed by:
Hayat Khan,
Sir Ganga Ram Hospital, Pakistan
Cemil Tunç,
Yüzüncü Yil University, Turkey
Mohsen Motamedi,
University of Shahreza, Iran

****Correspondence:***
Ayesha Sohail
ayeshasohail81@gmail.com;
asohail@cuilahore.edu.pk

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *20 July 2020*
Accepted: *09 November 2020*
Published: *01 April 2021*

Citation:
Yu Z, Sohail A, Nutini A and Arif R (2021) Delayed Modeling Approach to Forecast the Periodic Behavior of SARS-2. Front. Mol. Biosci. 7:585245. doi: 10.3389/fmolb.2020.585245

1. INTRODUCTION

In the field of biological sciences, the delayed processes takes place, not only at macro scale but also at micro scale. The computational framework for such problems can help in understanding the dynamics in a more cost effective manner (Yan, 2007; Fang et al., 2020; Sohail and Nutini, 2020).

Recently, the World Health Organization (WHO) has declared the novel corona virus (2019-nCoV) outbreak a Public Health Emergency of International Concern (PHEIC) It is named as severe acute respiratory syndrome corona virus 2 (SARS-CoV-2). The SARS-CoV-2 has been determined as the seventh member of the corona viruses infected humans (Zhu et al., 2020).

The antiviral immune response, also in the case of SARS-CoV2, behaves according to two characteristics: "lytic" and "non-lytic." These characteristics emerge from the action of two fundamental elements of the immune system in the case of an anti-viral response: antibodies (non-lytic response) and cytotoxic T cells also called CTL—Cytotoxic Lymphocites (lytic response).

At the moment the epitopes that can be identified to analyze these responses, are not clear. There are strong homologies with the SARS-CoV virus (75.5%). Furthermore, there is a strong alteration of antigenicity compared to SARS-CoV2 even if important epitopes in the Spike protein have been identified through the analysis of the localization of the RDB sequences, which can therefore become specific targets for drugs and vaccines (Zheng and Song, 2020).

The non-lytic response occurs against adaptive humeral immunity through the production of antibodies by B cells whose action is to neutralize the virion through direct connection with the same; specific antibodies are important in the defense against viruses during the early stages of infection when the virus is still extra-cellular; neutralizing antibodies facing the virus bind to the capsid or viral pericapside proteins, preventing their adhesion to the cell surface and therefore entry into the cells. Opsonizing antibodies can potentate the elimination of viral particles by phagocytosis.

In the case of SARS-CoV2, virus-specific IgG reached 100% approximately 17–19 days after symptom onset, while the proportion of patients with positive virus-specific IgM reached a peak of 94.1% approximately 20–22 days after symptom onset while during the first 3 weeks after symptom onset, there was increase in virus-specific IgG and IgM antibody titters.

Three types of seroconversion are possible: synchronous seroconversion of IgG and IgM, IgM seroconversion earlier than that of IgG and IgM seroconversion later than that of IgG (Long et al., 2020); specific data on the production of IgG and IgM is crucial to allow the rapid identification of the infection (di Mauro Gabriella et al., 2020). The lytic response depends on the action of the CTL cells which, in order to take place efficiently, must result from a cooperation between $CD4^+$ and $CD8^+$ lymphocytes. $CD8^+$ cells recognize endogenously synthesized viral antigens in association with MHC class I molecules on all cell types; $CD4^+$ cells recognize viral antigens in association with MHC class II antigens and the complete differentiation of $CD8^+$ cells requires the intervention of $CD4^+$ T helper cells. The antiviral effect of CTL is due to the lysis of infected cells, the activation of endocellular enzymes (endonucleases) which cause the degradation of the viral genome and the secretion of cytokines with interferon activity.

CTL epitopes of SARS-CoV-2 have been predicted by several studies, which can be used effectively to understand the pathogenesis (Kumar et al., 2020).

The fundamental mechanism by which the advancement of the virus in an organism is contrasted is given by a balance between the two mechanisms (lytic and non lytic). During multiple stages of the infection, the initial coordinated action of the antibodies join the lytic activity of the CTL cells to eradicate the virus.

The current research is motivated by the possible lack of the "control" and "coordination" between the two types of immune response, such that, a strong CTL response induces a limited and insufficient antibody action (or the reverse); is interlinked with the strong decrease in viral load induced by one of the two types of response.

During this imbalanced process, the coordinative response of the $CD4^+$ helper T cells also varies considerably: in SARS-CoV-2 infected patients, it has been reported that the analysis showed activation and reduction in $CD4^+$ and $CD8^+$ T cell counts (Li et al., 2020).

The purpose of this paper, therefore, is to quantify a possible mathematical model that investigates the possible non-coordination between lytic and non-lytic response and indicate a mathematical quantification of the best antiviral immune response in the case of SARS-CoV2.

2. MATERIALS AND THE METHODS

2.1. Basic Model of Virus Transmission

Consider the basic transmission model, with the coefficients, as listed in **Table 1**.

TABLE 1 | Parameters description.

Symbols	Description	Value	References
ρ	Rate of generation of Susceptible host cells	110	Wodarz, 2005
d	The death rate of suspected host cells	0.01	Wodarz, 2005
μ	The replication rate of the virus	0.01	Wodarz, 2005
a	The death rate of infected cells	0.01	Assumed
p	The strength of the lytic component	1	Wodarz, 2005
κ	Rate of infected cells produced in Free virus	1	Assumed
q	The neutralization death rate by antibodies	1	Wodarz, 2005
ϕ	Decays rate in free virus	1	Assumed
g	Rate of Antibodies develop in response to free virus	20.5	Assumed
h	Rate of decays in Antibodies response	0.1	Wodarz, 2005
c	Rate of CTL response to viral antigen derived from infected cells	20.5	Assumed
b	Rate of decays in the absence of antigenic stimulation	0.1	Assumed

TABLE 2 | Compartments and their description.

Symbols	Description
$U(t)$	Susceptible host cells
$W(t)$	Infected cells
$V(t)$	Free virus
$A(t)$	Antibody response
$C(t)$	CTL response

Model

$$
\begin{aligned}
\frac{dU}{dt} &= \rho - dU - \mu VU,\\
\frac{dW}{dt} &= \mu VU - aW - pWC,\\
\frac{dV}{dt} &= \kappa W - qVA - \phi V,\\
\frac{dA}{dt} &= gVA - hA,\\
\frac{dC}{dt} &= cWC - bC.
\end{aligned}
\tag{1}
$$

The description of compartments $U(t)$, $W(t)$, $V(t)$, $A(t)$, and $C(t)$ are presented in **Table 2**. The basic concept of modeling was adapted from the work of Wodarz (2005).

2.1.1. Positivity of Solution

For the model (1) transmission to be epidemiologically feasible, it is necessary to show that for all time every state variable is non-negative. Thus, the solutions of the model with non-negative initial value for all time $t > 0$ will remain non-negative.

Theorem 2.1. *Suppose that the model (1) consists of all feasible solutions with non-negative initial result then it remains non-negative for all time t.*

Proof: Let the model (1) satisfy the initial non-negative solution, i.e.,
$U(0) \geq 0$, $W(0) \geq 0$, $V(0) \geq 0$, $A(0) \geq 0$ and $C(0) \geq 0$.
It can be deduced from the model (1)

$$\frac{dU}{dt} = \rho - dU - \mu VU, \tag{2}$$

as the solution of variable U can be computed by following result,

$$U = U(0)e^{G} + \int_0^t \pi e^{H(k)} d(k) \geq 0, \tag{3}$$

where $G = -\int_0^t G(k)d(k)$ and $H(k) = -\int_0^t G(l)d(l)$. This gives non-negativity of $U(0) \geq 0$ for all $t \geq 0$. The non-negativity of the rest of the variables in the model (1) is given below.

$$\begin{aligned}
\frac{dW}{dt} &= \mu VU - aW - pWC,\\
\frac{dV}{dt} &= \kappa W - qVA - \phi V,\\
\frac{dA}{dt} &= gVA - hA,\\
\frac{dC}{dt} &= cWC - bC.
\end{aligned} \tag{4}$$

In term of matrix the above can be expressed as,

$$\frac{dF(t)}{dt} = H(t) + MF(t) \tag{5}$$

where

$$F(t) = \begin{pmatrix} W(t) \\ V(t) \\ A(t) \\ C(t) \end{pmatrix}, H(t) = \begin{pmatrix} a \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{6}$$

and M matrix.

$$M = \begin{pmatrix} -a & \frac{\mu\rho}{d} & 0 & 0 \\ \kappa & -\phi & 0 & 0 \\ 0 & 0 & -h & 0 \\ 0 & 0 & 0 & -b \end{pmatrix}. \tag{7}$$

The matrix M is a Matzler matrix so by the result presented in Smith (1996) the model is monotone. The fact thatR^4_+ is invariant with respect to stream of model (1).

2.1.2. Qualitative Analysis

Stability analysis is an essential key to validate the models in the field of science and technology. The remarkable contributions by Tunc (2002), Tunç (2004, 2008, 2013) and other researchers in the cross references can not be denied, since their contribution help the researchers to deal with the highly nonlinear models.

Here we present the stability analysis of given mathematical model (1). The model (1) is locally asymptotically stable at uninfected and infected equilibrium points. For uninfected equilibrium, the model is locally stable, if the value of reproduction number $\mathbb{R}_0 < 1$, whereas for infected equilibrium the model is stable if the value of the basic reproduction number $\mathbb{R}_0 > 1$. Furthermore, we will investigate the model (1) is locally stable at uninfected and infected. The uninfected equilibrium point $E^0 = (U^0, 0, 0, 0, 0)$ and infected equilibrium point $E^* = (U^*, W^*, V^*, A^*, C^*)$ of model (1) are constructed by following theorems.

Theorem 2.2. *The uninfected equilibrium point E^0 of the model (1) is given by*

$$E^0 = (U^0, 0, 0, 0, 0),$$

where:

$$U^0 = \frac{\rho}{d}.$$

Proof: By putting the equations of model (1) equal to zero, the dynamics is determined as follows:

$$\begin{aligned}
\rho - dU - \mu VU &= 0,\\
\mu VU - aW - pWC &= 0,\\
\kappa W - qVA - \phi V &= 0,\\
gVA - hA &= 0,\\
cWC - bC &= 0,
\end{aligned}$$

and after further algebraic manipulation, we got

$$U^0 = \frac{\rho}{d}, \tag{8}$$

which completes the proof.

Theorem 2.3. *The model (1) admits a unique infected equilibrium $E^* = (U^*, W^*, V^*, A^*, C^*)$ if and only if $\mathbb{R}_0 > 1$.*

Proof: The infected equilibrium point is given as:

$$\begin{aligned}
U^* &= \frac{g\rho}{dg + \mu h},\\
W^* &= \frac{b}{c},\\
V^* &= \frac{h}{g},\\
A^* &= \frac{bg\kappa - ch\phi}{chq},\\
C^* &= \frac{-abdg - ab\mu h + \mu ch\rho}{bp(dg + \mu h)}.
\end{aligned} \tag{9}$$

2.1.3. Reproduction Number $\mathbb{R}_0$ and Stability Analysis

The basic reproduction number $\mathbb{R}_0$ is formulated by evaluating infection matrix **F** and transmission matrix **V** constructed from Jacobian matrix J of model (1). The reproduction number $\mathbb{R}_0$ is the spectral radius of matrix $\mathbf{FV}^{-1}$. The Jacobian matrix for J_0 for uninfected equilibrium point is

$$J_0 = \begin{pmatrix} -d & 0 & -\frac{\mu\rho}{d} & 0 & 0 \\ 0 & -a & \frac{\mu\rho}{d} & 0 & 0 \\ 0 & \kappa & -\phi & 0 & 0 \\ 0 & 0 & 0 & -h & 0 \\ 0 & 0 & 0 & 0 & -b \end{pmatrix} \tag{10}$$

The infection matrix F and rest of transmission matrix V are give as follow:

$$\mathbf{V} = \begin{pmatrix} d & 0 & 0 & 0 & 0 \\ 0 & a & 0 & 0 & 0 \\ 0 & -\kappa & \phi & 0 & 0 \\ 0 & 0 & 0 & h & 0 \\ 0 & 0 & 0 & 0 & b \end{pmatrix}, \tag{11}$$

$$F = \begin{pmatrix} 0 & 0 & -\frac{\mu\rho}{d} & 0 & 0 \\ 0 & 0 & \frac{\mu\rho}{d} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{12}$$

The matrix $K = \mathbf{FV}^{-1}$ is constructed as follows:

$$K = \begin{pmatrix} 0 & -\frac{\mu\kappa\rho}{ad\phi} & -\frac{\mu\rho}{d\phi} & 0 & 0 \\ 0 & \frac{\mu\kappa\rho}{ad\phi} & \frac{\mu\rho}{d\phi} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{13}$$

The spectral radius of matrix $K = \mathbf{FV}^{-1}$ give reproductive number of model (1) is

$$\mathbb{R}_0 = \frac{\mu\kappa\rho}{ad\phi}. \tag{14}$$

The stability analysis of system (1) is presented in following Theorem 2.4.

Theorem 2.4. *Uninfected equilibrium point E_0, will be locally asymptotically stable if $\mathbb{R}_0 < 1$ and otherwise it will be unstable.*

Proof: The Jacobian matrix $J(E_0)$ of model (1) for uninfected equilibrium point is formulated as follows:

$$J(E_0) = \begin{pmatrix} -d & 0 & -\frac{\mu\rho}{d} & 0 & 0 \\ 0 & -a & \frac{\mu\rho}{d} & 0 & 0 \\ 0 & \kappa & -\phi & 0 & 0 \\ 0 & 0 & 0 & -h & 0 \\ 0 & 0 & 0 & 0 & -b \end{pmatrix}. \tag{15}$$

The eigenvalues of Jacobian matrix $J(E_0)$ is evaluated by characteristic equation that is $det(J(E_0) - I\lambda_i) = 0$, for $i = 1:5$. The eigenvalues are given as follows:
$\lambda_1 = -b, \lambda_2 = -d, \lambda_3 = -h,$
$\lambda_4 = -\frac{\left(\sqrt{d}\right)(a+\phi)+\sqrt{d(a-\phi)^2+4\mu\kappa\rho}}{2\sqrt{d}}$
and
$\lambda_5 = -\frac{\sqrt{d}(a+\phi)-\sqrt{d(a-\phi)^2+4\mu\kappa\rho}}{2\sqrt{d}}.$
Thus, all eigenvalues λ_i are strictly negative for $i = 1:5$. Hence, model (1) is locally asymptotically stable.

Lemma 2.5. *If $R_0 > 1$, the infected equilibrium point of system is locally asymptotically stable, otherwise it is unstable.*

Proof: The results can by obtained by the same procedure of (2.4).

2.1.4. Sensitivity Analysis

The sensitivity of basis reproductive

$$\mathbb{R}_0 = \frac{\mu\kappa\rho}{ad\phi} \tag{16}$$

is analyze with respect to each parameters is follows:

$$\begin{aligned} \frac{\partial \mathbb{R}_0}{\partial \mu} &= \frac{\kappa\rho}{ad\phi} > 0, \\ \frac{\partial \mathbb{R}_0}{\partial \kappa} &= \frac{\mu\rho}{ad\phi} > 0, \\ \frac{\partial \mathbb{R}_0}{\partial \rho} &= \frac{\mu\kappa}{ad\phi} > 0, \\ \frac{\partial \mathbb{R}_0}{\partial a} &= -\frac{\mu\kappa\rho}{a^2 d\phi} < 0, \\ \frac{\partial \mathbb{R}_0}{\partial d} &= -\frac{\mu\kappa\rho}{ad^2\phi} < 0, \\ \frac{\partial \mathbb{R}_0}{\partial \phi} &= -\frac{\mu\kappa\rho}{ad\phi^2} < 0. \end{aligned} \tag{17}$$

It is clear that reproductive number is directly proportional to the replication rate of the virus, rate of infected cells produced in free virus and rate of generation of susceptible host cells. Decrease with the death rate of infected cells, decays rate in free virus and the death rate of suspected host cells.

The basic reproductive number is $\mathbb{R}_0 = \frac{\mu\kappa\rho}{ad\phi} > 1$ in case of infection. In the absence of an immune responses the model (1) converges to the following equilibrium points $E^{(0)} = (U^{(0)}, W^{(0)}, V^{(0)}, A^{(0)}, C^{(0)})$ which is defined as

$$U^{(0)} = (\frac{\rho}{d}, 0, 0, 0, 0). \tag{18}$$

We assumed that the immune responses is present. this desires following conditions: $cW^{(0)} > b$ and $gV^{(0)} > h$. In this case following can be observed.

1. The anti body response can't established and the CTL response develops. Because the CTL response is strong and decrease virus load to levels which are vary low to stimulate antibody response. It has following equilibrium points,
   $E^{(1)} = (U^{(1)}, W^{(1)}, V^{(1)}, A^{(1)}, C^{(1)})$ where,

$$\begin{aligned} U^{(1)} &= \frac{c\phi\rho}{b\kappa\mu + cd\phi}, \\ W^{(1)} &= \frac{b}{c}, \\ V^{(1)} &= \frac{b\kappa}{c\phi}, \\ A^{(1)} &= 0, \\ C^{(1)} &= \frac{-ab\kappa\mu - acd\phi + c\kappa\mu\rho}{p(b\kappa\mu + cd\phi)}. \end{aligned} \tag{19}$$

This is obtained if $\frac{bg\kappa}{c\phi} < h$ and $\frac{ch\rho\mu}{a(dg+h\mu)} > b$.

2. The sustained CTL response zero and antibody response develop, because the antibody response is strong relative to CTL response and decrease virus load to levels which are vary low to stimulate CTL. It has following equilibrium points.

$$E^{(2)} = (U^{(2)}, W^{(2)}, V^{(2)}, A^{(2)}, C^{(2)}),$$

which is defined as follows:

$$\begin{aligned} U^{(2)} &= \frac{g\rho}{dg + h\mu}, \\ W^{(2)} &= \frac{h\mu\rho}{a(dg + h\mu)}, \\ V^{(2)} &= \frac{h}{g}, \\ A^{(2)} &= \frac{-adg\phi - ah\mu\phi + g\kappa\mu\rho}{aq(dg + h\mu)}, \\ C^{(2)} &= 0. \end{aligned} \tag{20}$$

This is obtained if $\frac{bgk}{c\phi} > h$ and $\frac{ch\rho\mu}{a(dg+h\mu)} < b$.

3. CTL and anti body response develops. This equilibrium points are as follows:

$$E^{(3)} = (U^{(3)}, W^{(3)}, V^{(3)}, A^{(3)}, C^{(3)}) \tag{21}$$

where,

$$\begin{aligned} U^{(3)} &= \frac{g\rho}{dg + h\mu}, \\ W^{(3)} &= \frac{b}{c}, \\ V^{(3)} &= \frac{h}{g}, \\ A^{(3)} &= \frac{bg\kappa - ch\phi}{chq}, \\ C^{(3)} &= \frac{ch\mu\rho - abdg - abh\mu}{bp(dg + h\mu)}. \end{aligned} \tag{22}$$

This is obtained if $\frac{bgk}{c\phi} > h$ and $\frac{ch\rho\mu}{a(dg+h\mu)} > b$.

2.2. Modified Modeling Approach

$$\begin{aligned} \frac{dU}{dt} &= \rho - dU - \mu\frac{VU}{1 + \eta U}, \\ \frac{dW}{dt} &= \mu\frac{VU}{1 + \eta U} - aW - pWC, \\ \frac{dV}{dt} &= \kappa W - qVA - \phi V, \\ \frac{dA}{dt} &= gVA - hA, \\ \frac{dC}{dt} &= cWC - bC, \end{aligned} \tag{23}$$

where the description of compartments $U(t)$, $W(t)$, $V(t)$, $A(t)$, and $C(t)$ are presented in **Table 2**.

2.2.1. Positivity of Solution

For the model (23) transmission to be epidemiologically feasible, it is necessary to show that for all time every state variable is non-negative. Thus, the solutions of the model with non-negative initial value for all time $t > 0$ will remain non-negative.

Theorem 2.6. *Suppose that the model(23) consists of all feasible solutions with non-negative initial result then it remains non-negative for all time t.*

Proof: Let the model (23) satisfy that the initial non-negative solution, i.e.,
$U(0) \geq 0$, $W(0) \geq 0$, $V(0) \geq 0$, $A(0) \geq 0$, and $C(0) \geq 0$.
It can be deduced from model (23) that is,

$$\frac{dU}{dt} = \rho - dU - \mu\frac{VU}{1 + \eta U}. \tag{24}$$

As the solution of variable U can be computed by following result,

$$U = U(0)e^{G} + \int_0^t \pi e^{H(k)} d(k) \geq 0 \tag{25}$$

where $G = -\int_0^t G(k)d(k)$ and $H(k) = -\int_0^t G(l)d(l)$. This gives non-negativity of $U(0) \geq 0$ for all $t \geq 0$. The non-negativity of rest variables in the model (23) is given as follows,

$$\begin{aligned} \frac{dW}{dt} &= \mu\frac{VU}{1 + \eta U} - aW - pWC, \\ \frac{dV}{dt} &= \kappa W - qVA - \phi V, \\ \frac{dA}{dt} &= gVA - hA, \\ \frac{dC}{dt} &= cWC - bC. \end{aligned} \tag{26}$$

In term of matrix the above can be expressed as,

$$\frac{dF(t)}{dt} = H(t) + MF(t) \tag{27}$$

where $F(t)$ and $H(t)$ are as follows:

$$F(t) = \begin{pmatrix} W(t) \\ V(t) \\ A(t) \\ C(t) \end{pmatrix}, H(t) = \begin{pmatrix} a \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{28}$$

and we have matrix M

$$M = \begin{pmatrix} -a & \frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0 \\ \kappa & -\phi & 0 & 0 \\ 0 & 0 & -h & 0 \\ 0 & 0 & 0 & -b \end{pmatrix}. \tag{29}$$

The matrix M is a Matzler matrix so by the result presented in Smith (1996) the model is monotone. the fact thatR_{+}^{4} is invariant with respect to stream of model (23).

2.2.2. Qualitative Analysis

Here we present the stability analysis of (23). The model (23) is locally asymptotically stable at uninfected and infected equilibrium points. We investigate the model (23) is locally stable at uninfected and infected. The uninfected equilibrium point $E^0 = (U^0, 0, 0, 0, 0)$ and infected equilibrium point $E^* = (U^*, W^*, V^*, A^*, C^*)$ of model (23) are constructed by following theorems.

Theorem 2.7. *The uninfected equilibrium point E^0 of the model (23) is given by*

$$E^0 = (U^0, 0, 0, 0, 0),$$

where:

$$U^0 = \frac{\rho}{d}.$$

Proof: By putting the equations in (23) equal to zero, the dynamics is determined as follows,

$$\begin{aligned}
\rho - dU - \mu\frac{VU}{1+\eta U} &= 0,\\
\mu\frac{VU}{1+\eta U} - aW - pWC &= 0,\\
\kappa W - qVA - \phi V &= 0\\
gVA - hA &= 0,\\
cWC - bC &= 0.
\end{aligned}$$

By solving these equations, we got

$$U^0 = \frac{\rho}{d}. \tag{30}$$

This completes the proof.

Theorem 2.8. *The model (23) admits a unique infected equilibrium $E^* = (U^*, W^*, V^*, A^*, C^*)$ if and only if $\mathbb{R}_0 > 1$.*

Proof: Calculating the infected equilibrium point, we obtain

$$\begin{aligned}
U^* &= \frac{-\sqrt{4d\eta g^2\rho + (dg - \eta g\rho + h\mu)^2}}{2dg\eta}\\
&\quad + \frac{\eta g\rho - dg - h\mu}{2dg\eta},\\
W^* &= \frac{b}{c},\\
V^* &= \frac{h}{g},\\
A^* &= \frac{bg\kappa - ch\phi}{chq},\\
C^* &= \frac{-ab - cdU^* + c\rho}{bp}.
\end{aligned} \tag{31}$$

2.2.3. Reproduction Number $\mathbb{R}_0$ and Stability Analysis

The basic reproduction number $\mathbb{R}_0$ is formulated by evaluating infection matrix **F** and transmission matrix **V** constructed from Jacobian matrix J of system (23). The reproduction number $\mathbb{R}_0$ is the spectral radius of matrix $\mathbf{FV}^{-1}$. The Jacobian matrix for J_0 for uninfected equilibrium point is,

$$J_0 = \begin{pmatrix} 0 & 0 & -\frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0\\ 0 & -a & \frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0\\ 0 & \kappa & -\phi & 0 & 0\\ 0 & 0 & 0 & -h & 0\\ 0 & 0 & 0 & 0 & -b \end{pmatrix}. \tag{32}$$

The infection and transmission matrices, F & V are give as follow:

$$\mathbf{V} = \begin{pmatrix} 0 & 0 & \frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0\\ 0 & a & -\frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0\\ 0 & -\kappa & \phi & 0 & 0\\ 0 & 0 & 0 & h & 0\\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{33}$$

$$F = \begin{pmatrix} 0 & -\frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0\\ 0 & \frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{34}$$

The matrix $K = \mathbf{FV}^{-1}$ is constructed as follows:

$$K = \begin{pmatrix} -\frac{\kappa\mu\rho}{ad(\frac{\eta\rho}{d}+1)\phi} & -\frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)\phi} & 0 & 0\\ \frac{\kappa\mu\rho}{ad(\frac{\eta\rho}{d}+1)\phi} & \frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)\phi} & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{35}$$

The spectral radius of matrix $K = \mathbf{FV}^{-1}$ finally provides the reproductive number for given model (23):

$$\mathbb{R}_0 = \frac{\mu\rho(a-\kappa)}{a\phi(d+\eta\rho)}. \tag{36}$$

The stability analysis of system (23) is presented by follows theorem.

Theorem 2.9. *For uninfected equilibrium point E_0 if the real part of eigenvalues of Jacobian matrix J of the system (23) are strictly negative, then the system (23) is locally stable otherwise it is unstable.*

Proof: The Jacobian matrix of model (23) for uninfected equilibrium point $J(E_0)$ is formulated as follows:

$$J(E_0) = \begin{pmatrix} 0 & 0 & -\frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0\\ 0 & -a & \frac{\mu\rho}{d(\frac{\eta\rho}{d}+1)} & 0 & 0\\ 0 & \kappa & -\phi & 0 & 0\\ 0 & 0 & 0 & -h & 0\\ 0 & 0 & 0 & 0 & -b \end{pmatrix}. \tag{37}$$

The eigenvalues of Jacobian matrix $J(E_0)$ is evaluated by characteristic equation $det(J(E_0) - I\lambda_i) = 0$, for $i = 1:5$. The eigenvalues are given as follows:

$\lambda_1 = 0, \lambda_2 = -b, \lambda_3 = -h,$
$\lambda_4 = \frac{1}{2}\left(\sqrt{(a-\phi)^2 + \frac{4\kappa\mu\rho}{d+\eta\rho}} + a + \phi\right)$
and
$\lambda_5 = \frac{1}{2}\left(-\sqrt{(a-\phi)^2 + \frac{4\kappa\mu\rho}{d+\eta\rho}} + a + \phi\right).$

Since the all eigenvalues λ_i, satisfy the criteria (where as $i = 1:5$). Hence, the model (23) is locally stable.

2.2.4. Sensitivity Analysis

The sensitivity of basis reproductive

$$\mathbb{R}_0 = \frac{\mu\rho(a-\kappa)}{a\phi(d+\eta\rho)} \tag{38}$$

is analyze with respect to each parameters is as follows:

$$\begin{aligned}
\frac{\partial \mathbb{R}_0}{\partial a} &= \frac{\kappa\mu\rho}{a^2\phi(d+\eta\rho)} > 0,\\
\frac{\partial \mathbb{R}_0}{\partial \mu} &= \frac{a\rho - \kappa\rho}{ad\phi + a\eta\rho\phi} > 0,\\
\frac{\partial \mathbb{R}_0}{\partial \rho} &= \frac{d\mu(a-\kappa)}{a\phi(d+\eta\rho)^2} > 0,\\
\frac{\partial \mathbb{R}_0}{\partial \kappa} &= -\frac{\mu\rho}{ad\phi + a\eta\rho\phi} < 0\\
,\frac{\partial \mathbb{R}_0}{\partial d} &= -\frac{\mu\rho(a-\kappa)}{a\phi(d+\eta\rho)^2} < 0,\\
\frac{\partial \mathbb{R}_0}{\partial \phi} &= -\frac{\mu\rho(a-\kappa)}{a\phi^2(d+\eta\rho)} < 0,\\
\frac{\partial \mathbb{R}_0}{\partial \eta} &= -\frac{\mu\rho^2(a-\kappa)}{a\phi(d+\eta\rho)^2} < 0.
\end{aligned} \tag{39}$$

Therefore, the reproductive number $\mathbb{R}_0$ is directly proportional to a, μ, and ρ and inversely proportional to κ, d, ϕ and η.
The basic reproductive number is $\mathbb{R}_0 = \frac{\mu\kappa\rho}{ad\phi} > 1$ in case of infection.

In the absence of an immune responses the model (23) converges to the following equilibrium points: $E^{(0)} = (U^{(0)}, W^{(0)}, V^{(0)}, A^{(0)}, C^{(0)})$ which is defined as $U^{(0)} = \frac{\rho}{d}, W^{(0)} = 0, V^{(0)} = 0, A^{(0)} = 0, C^{(0)} = 0$. We assumed that the immune responses is present. This desires following conditions: $cW^{(0)} > b$ and $gV^{(0)} > h$.

In this case, the following can be observed:

1. The anti body response can not be established and the CTL response develops.

   Because of strong CTL response and low virus load to levels, it has the following equilibrium points:

$$\begin{aligned}
E^{(1)} &= (U^{(1)}, W^{(1)}, V^{(1)}A^{(1)}, C^{(1)}),\\
E^{(1)} &= (X, \frac{b}{c}, \frac{b\kappa}{c\phi}, 0, Y),
\end{aligned} \tag{40}$$

where:

$$\begin{aligned}
U^{(1)} &= \frac{-b\kappa\mu - c\phi(d-\eta\rho)}{2cd\eta\phi},\\
&\quad - \frac{\sqrt{(b\kappa\mu + c\phi(d-\eta\rho))^2 + 4c^2d\eta\rho\phi^2}}{2cd\eta\phi},\\
W^{(1)} &= \frac{b}{c},\\
V^{(1)} &= \frac{b\kappa}{c\phi},\\
A^{(1)} &= 0,\\
C^{(1)} &= \frac{-ab - cdU^{(1)} + c\rho}{bp}.
\end{aligned} \tag{41}$$

This equilibrium points is obtain if it follows the following condition:
$\frac{bg\kappa}{c\phi} < h$ and
$\frac{c\sqrt{4d\eta g^2\rho + (-dg+\eta g\rho - h\mu)^2} + cdg + c\eta g\rho + ch\mu}{2ag\eta} > b.$

2. For null sustained CTL response, the antibody response develops. Because of strong antibody response to relative CTL response and low virus load to levels which is too small for stimulate CTL. It has the following equilibrium points:

$$E^{(2)} = (U^{(2)}, W^{(2)}, V^{(2)}, A^{(2)}, C^{(2)}), \tag{42}$$

which is defined as follows:

$$\begin{aligned}
U^{(2)} &= \frac{-dg + \eta g\rho - h\mu}{2dg\eta},\\
&\quad - \frac{\sqrt{4d\eta g^2\rho + (dg - \eta g\rho + h\mu)^2}}{2dg\eta},\\
W^{(2)} &= \frac{g\rho}{a} - U^{(2)},
\end{aligned}$$

$$\begin{aligned}
V^{(2)} &= \frac{h}{g},\\
A^{(2)} &= \frac{g\kappa(\rho - dU^{(2)}) - ah\phi}{ahq},\\
C^{(2)} &= 0.
\end{aligned} \tag{43}$$

The equilibrium point satisfies the following criteria: $\frac{bg\kappa}{c\phi} > h$ and
$\frac{c\sqrt{4d\eta g^2\rho + (-dg+\eta g\rho - h\mu)^2} + cdg + c\eta g\rho + ch\mu}{2ag\eta} < b.$

3. CTL and anti body response develops. The equilibrium point is given as:

$$E^* = (U^*, W^*, V^*, A^*, C^*) \tag{44}$$

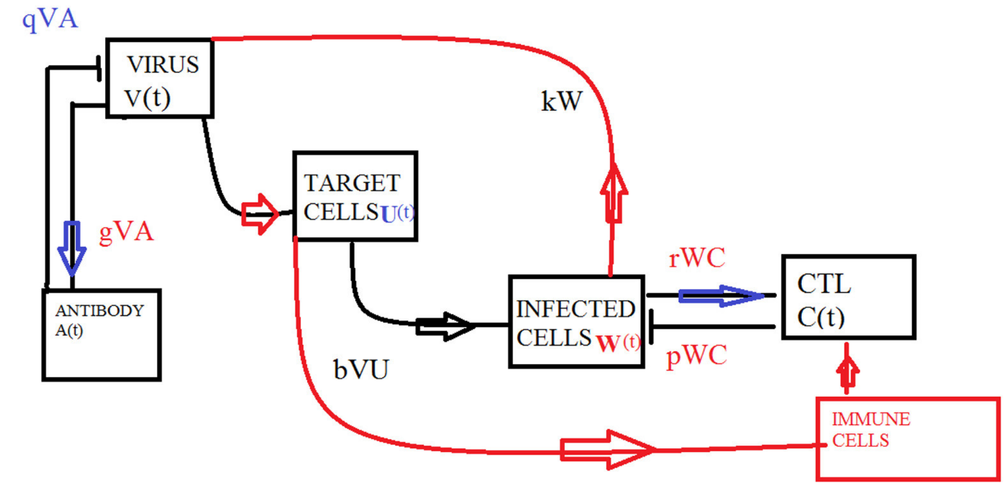


FIGURE 1 | Schematic of the model.

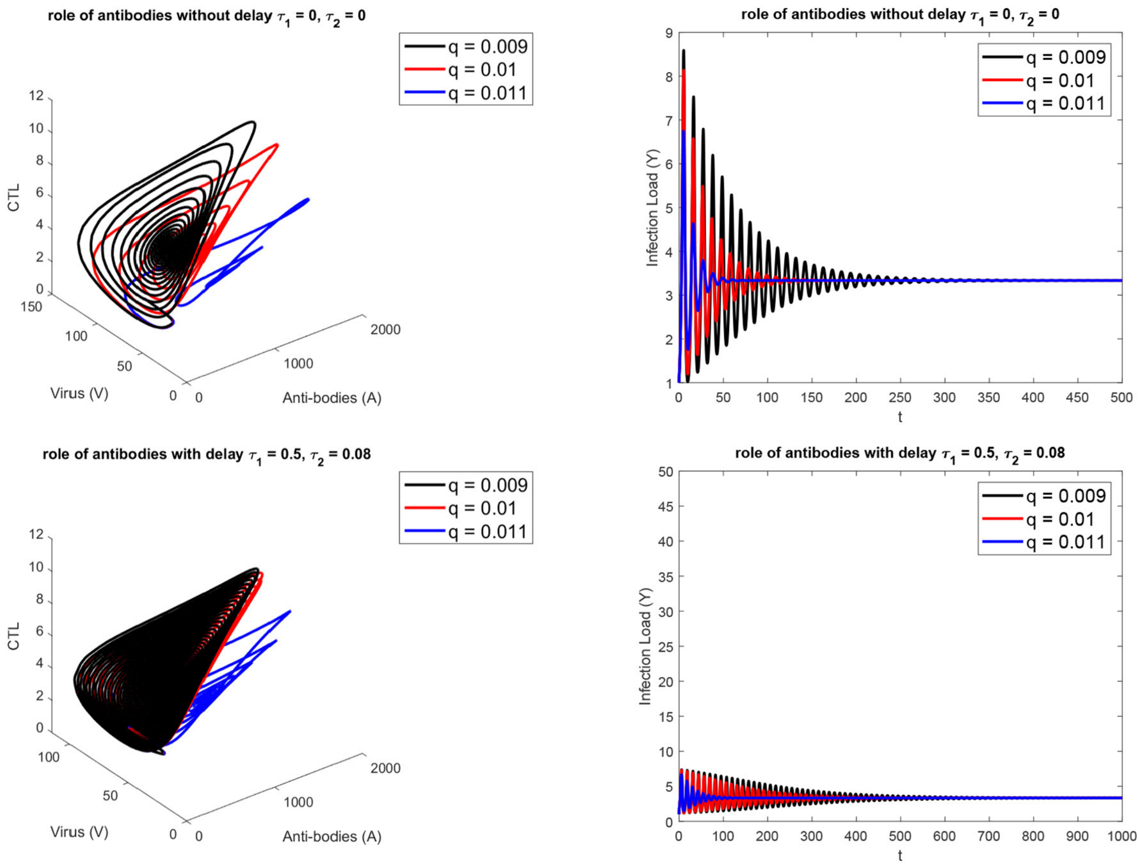


FIGURE 2 | For three different values of antibody-SARS-2 interaction rates (q). **(Top)** without, **(Bottom)** with delay $\tau_1 = 0.5$, $\tau_2 = 0.08$.

where:

$$
\begin{aligned}
U^* &= \frac{-\sqrt{4d\eta g^2\rho + (dg - \eta g\rho + h\mu)^2}}{2dg\eta}, \\
&\quad + \frac{\eta g\rho - dg - h\mu}{2dg\eta}, \\
W^* &= \frac{b}{c}, \\
V^* &= \frac{h}{g}, \\
A^* &= \frac{bg\kappa - ch\phi}{chq}, \\
C^* &= \frac{-ab - cdU^* + c\rho}{bp},
\end{aligned}
\tag{45}
$$

under the criteria: $\frac{bg\kappa}{c\phi} > h$ and $\frac{c\sqrt{4d\eta g^2\rho+(-dg+\eta g\rho-h\mu)^2}+cdg+c\eta g\rho+ch\mu}{2ag\eta} > b.$

2.3. Delayed Model

Different biological models with delayed transmission are available in the literature (Maiti et al., 2008; Lv and Yuan, 2009; Kuniya and Nakata, 2012). The schematic 2 presents the transmission of the virus from one compartment to the other. The delay takes place after the interaction of virus with the target cells at τ_1, it is further assumed that at delay τ_2, the infected cells and the CTLs interact and the antibodies and the virus interact.

$$
\begin{aligned}
\frac{dU(t)}{dt} &= \rho - dU(t) - \mu\frac{V(t)U(t)}{1+\eta U(t)}, \\
\frac{dW(t)}{dt} &= \frac{\mu V(t-\tau_1)U(t-\tau_1)}{1+\eta U(t-\tau_1)} - aW(t), \\
&\quad - pW(t)C(t), \\
\frac{dV(t)}{dt} &= \kappa W(t) - qV(t)A(t) - \phi V(t), \\
\frac{dA(t)}{dt} &= gV(t-\tau_2)A(t-\tau_2) - hA(t), \\
\frac{dC(t)}{dt} &= cW(t-\tau_2)C(t-\tau_2) - bC(t).
\end{aligned}
\tag{46}
$$

2.3.1. Local Stability for Delay Differential Equations

In ODE's, the local stability of consistent condition relies upon the area of underlying foundations of characteristic function, that are polynomial in shape. The unfaltering condition is steady if the majority of the roots having −ve real part. The outstanding Routh-Hurwitz criteria provide exact situation for arbitrary polynomials. For DDE's, nearby stability is likewise controlled by

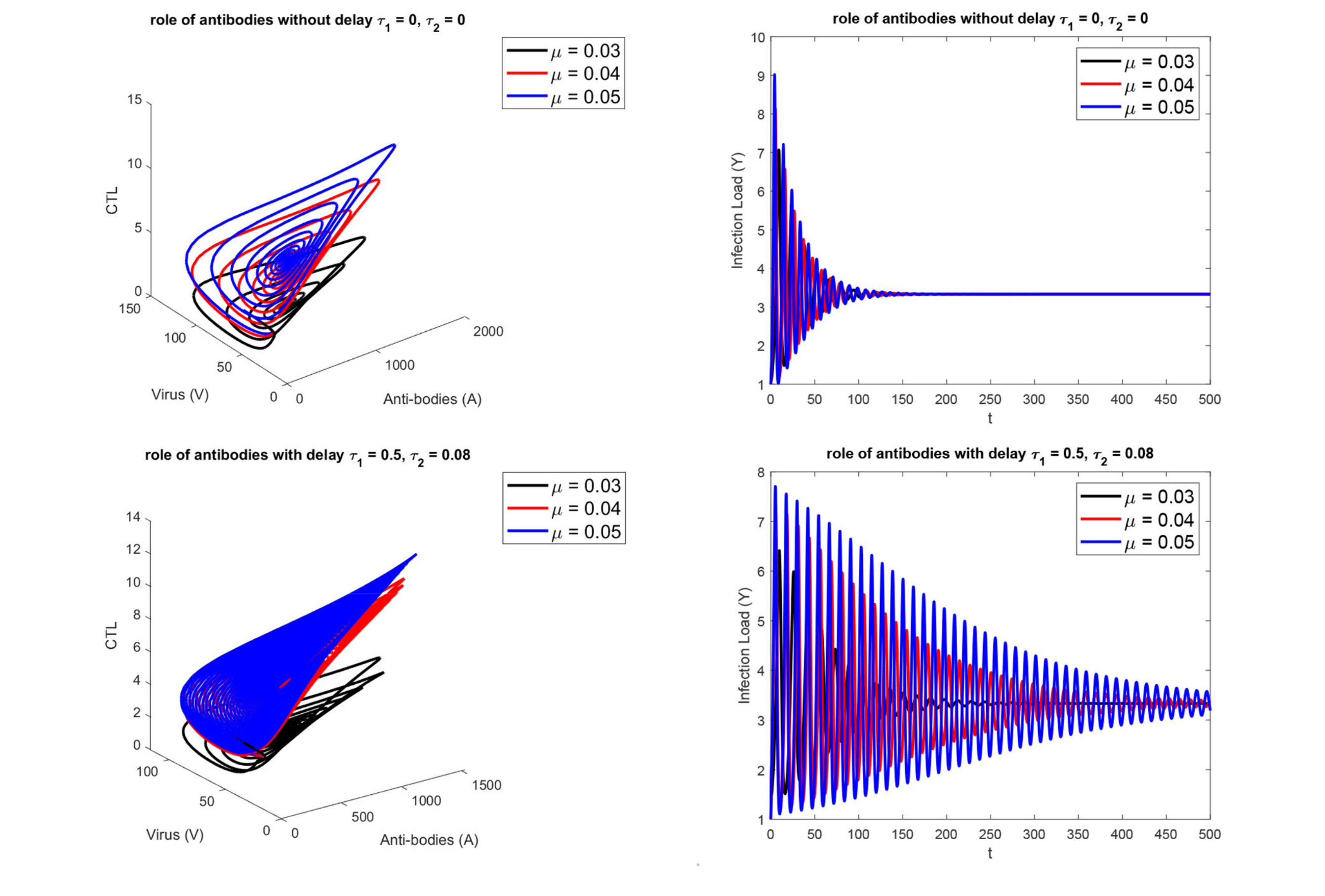


FIGURE 3 | For three different values of the replication rate of virus. **(Top)** without, **(Bottom)** with delay $\tau_1 = 0.5$, $\tau_2 = 0.08$.

the area of trademark work, yet for this situation, this capacity appears as an alleged quasi polynomial, that is supernatural. Hence, there are vastly numerous roots. Moreover, the Routh-Hurwitz criteria are not pertinent. Numerous methodologies decide the stability of steady states delay equations.

2.3.2. Domain Subdivision

D-subdivision or Domain subdivision, utilizes basic details about the actions of the roots of characteristic functions, as a parameter modifies to split parameter space into parts where the no. of roots with +ve real parts is constant. The roots' position depends consistently on the models' parameter and when the parameters are altered, another root rises if imaginary root exists for a set of parameters.

Now we subdivide the parametric domain by hypersurfaces comprising of parameter routines for which at least one simply imaginary roots exist. When the districts are limited by these hypersurfaces, the quantity of roots with $+ve$ real part is constant. Obviously, the locales where the number is zero and their supplements are more interesting. This technique is especially simple to picture when the framework in question relies upon two parameters, so the area is $\mathbb{R}^2$, and, the hypersurfaces are curve.

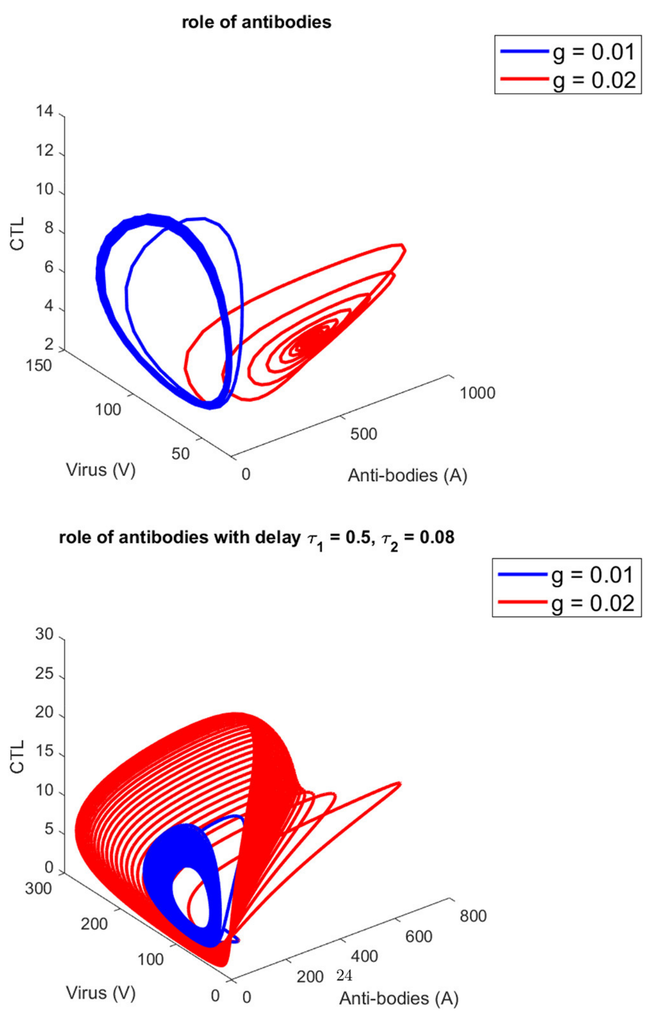


FIGURE 4 | For three different values of *g*. **(Top)** without, **(Bottom)** (with) delay $\tau_1 = 0.5$, $\tau_2 = 0.08$.

3. RESULTS AND DISCUSSION

During this research, we have used the Matlab^{TM} delay differential equations bifurcation analysis tools. For the validity of the computational model, extensive numerical experiments were conducted using simulink toolbox.

3.1. Analysis of SARS-CoV2 and Antibodies Interaction

With the aid of mathematical model we have concluded the following:

1. The analysis of lytic vs. non lytic immune response plays an important role in infection control.
2. The Hill function is important in kinetic modeling and the Hill coefficient is important parameter to forecast a complete cycle of infection.
3. The analytic approach and numerical Matcont bifurcation analysis proved to be efficient in parametric approximation for such complex dynamics.

Figure 1 presents the schematic to understand the interaction of key players. **Figure 2** provides the phase space portraits to explore the interaction between the CTL's, the anti bodies and the virus, for three different values of q. We can see that for increasing values of q, there is reduction in the concentration of CTL, as well as the length of the cycle increases over time (top panel). On the other hand, when delay was considered, the dynamics were different. **Figure 3** provides analysis relative to the replication rate of virus. **Figure 4** presents the dynamics relative to the parameter g. We can see a twist in the phase space portraits when the delay was taken into account. In the supplementary figures, (**Figures S1** and **S2**, we can see the dynamics more clearly relative to change in parametric values). We can see that the dynamics are more visible to witness the rapid action of SARS-2. Thus, a model without delay, and with $\eta = 0$ will not be able to demonstrate the dynamics well. We thus conclude that the disease transmission and the immune response depends on time delay as well as nonlinear Hill formalism.

4. CONCLUSIONS AND FUTURE WORK

The manuscript presents a state of the art model, with delay, from one compartment to the next due to transition. In nature, there is always a delay in the onset of infections. The proposed mathematical model quantifies and analyzes this imbalance and describes the temporal trend of the phenomenon, leaving its application open to possible direct therapies in that sense.

- Mathematical analysis of the non-lytic and lytic action of the immune reaction to SARS-CoV2.
- Construction of a model that describes the balance between antibody reaction and cellular reaction mediated by CTL cells.

- Analysis of the imbalance between non-lytic and lytic action of the immune response.
- Description and quantification of the model related to the infection and functionality of CTL cells over time.
- Evidence of delay in disease transmission.

DATA AVAILABILITY STATEMENT

The original contributions generated for the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ZY and AS did the modeling. AN and RA did the literature review and simulations. All of the authors equally contributed to the manuscript and participated in results and discussion.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.585245/full#supplementary-material

REFERENCES

di Mauro Gabriella, S. C., Concetta, R., Francesco, R., and Annalisa, C. (2020). Sars-cov-2 infection: response of human immune system and possible implications for the rapid test and treatment. *Int. Immunopharmacol*. 84:106519. doi: 10.1016/j.intimp.2020.106519

Fang, J., Liu, C., Simos, T., and Famelis, I. T. (2020). Neural network solution of single-delay differential equations. *Mediterr. J. Math*. 17:30. doi: 10.1007/s00009-019-1452-5

Kumar, S., Maurya, V. K., Prasad, A. K., Bhatt, M. L., and Saxena, S. K. (2020). Structural, glycosylation and antigenic variation between 2019 novel coronavirus (2019-nCov) and SARS coronavirus (SARS-CoV). *VirusDisease* 31, 13–21. doi: 10.1007/s13337-020-00571-5

Kuniya, T., and Nakata, Y. (2012). Permanence and extinction for a nonautonomous seirs epidemic model. *Appl. Math. Comput*. 218, 9321–9331. doi: 10.1016/j.amc.2012.03.011

Li, X., Geng, M., Peng, Y., Meng, L., and Lu, S. (2020). Molecular immune pathogenesis and diagnosis of COVID-19. *J. Pharm. Anal*. 10, 102–108. doi: 10.1016/j.jpha.2020.03.001

Long, Q.-X., Liu, B.-Z., Deng, H.-J., Wu, G.-C., Deng, K., Chen, Y.-K., et al. (2020). Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nat. Med*. 26, 845–848. doi: 10.1038/s41591-020-0897-1

Lv, C., and Yuan, Z. (2009). Stability analysis of delay differential equation models of hiv-1 therapy for fighting a virus with another virus. *J. Math. Anal. Appl*. 352, 672–683. doi: 10.1016/j.jmaa.2008.11.026

Maiti, A., Pal, A., and Samanta, G. (2008). Effect of time-delay on a food chain model. *Appl. Math. Comput*. 200, 189–203. doi: 10.1016/j.amc.2007.11.011

Smith, H. L. (1996). Monotone dynamical systems: an introduction to the theory of competitive and cooperative systems. *Bull. Am. Math. Soc*. 33, 203–209. doi: 10.1090/S0273-0979-96-00642-8

Sohail, A., and Nutini, A. (2020). Forecasting the timeframe of coronavirus and human cells interaction with reverse engineering. *Prog. Biophys. Mol. Biol*. 155, 29–35. doi: 10.1016/j.pbiomolbio.2020.04.002

Tunc, C. (2002). A study of the stability and boundedness of the solutions of nonlinear differential equations of the fifth order. *Indian J. Pure Appl. Math*. 33, 519–529.

Tunç, C. (2004). A note on the stability and boundedness results of solutions of certain fourth order differential equations. *Appl. Math. Comput*. 155, 837–843. doi: 10.1016/S0096-3003(03)00819-1

Tunç, C. (2008). On the stability of solutions to a certain fourth-order delay differential equation. *Nonlin. Dyn*. 51, 71–81. doi: 10.1007/s11071-006-9192-z

Tunç, C. (2013). New results on the stability and boundedness of nonlinear differential equations of fifth order with multiple deviating arguments. *Bull. Malays. Math. Sci. Soc*. 36, 671-682.

Wodarz, D. (2005). Mathematical models of immune effector responses to viral infections: virus control versus the development of pathology. *J. Comput. Appl. Math*. 184, 301–319. doi: 10.1016/j.cam.2004.08.016

Yan, X.-P. (2007). Stability and hopf bifurcation for a delayed prey-predator system with diffusion effects. *Appl. Math. Comput*. 192, 552–566. doi: 10.1016/j.amc.2007.03.033

Zheng, M., and Song, L. (2020). Novel antibody epitopes dominate the antigenicity of spike glycoprotein in SARS-CoV-2 compared to SARS-CoV. *Cell. Mol. Immunol*. 17, 536–538. doi: 10.1038/s41423-020-0385-z

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in china, 2019. *N. Engl. J. Med*. 382, 727–733. doi: 10.1056/NEJMoa2001017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 09 April 2021
doi: 10.3389/fmolb.2021.645216

Structure-Based Virtual Screening to Identify Novel Potential Compound as an Alternative to Remdesivir to Overcome the RdRp Protein Mutations in SARS-CoV-2

Thirumal Kumar D[1†], *Nishaat Shaikh*[2†], *Udhaya Kumar S*[3†], *George Priya Doss C*[2*] and *Hatem Zayed*[4*]

[1] Meenakshi Academy of Higher Education and Research, Chennai, India, [2] School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore, India, [3] Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India, [4] Department of Biomedical Sciences, College of Health and Sciences, Qatar University, QU Health, Doha, Qatar

OPEN ACCESS

Edited by:
Arvind Ramanathan,
Argonne National Laboratory (DOE),
United States

Reviewed by:
Filippo Pullara,
SpIntellx Inc., United States
Debsindhu Bhowmik,
Oak Ridge National Laboratory (DOE),
United States

***Correspondence:**
George Priya Doss C
georgepriyadoss@vit.ac.in
Hatem Zayed
hatem.zayed@qu.edu.qa

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *22 December 2020*
Accepted: *11 March 2021*
Published: *09 April 2021*

Citation:
Kumar D T, Shaikh N, Kumar S U, Doss C GP and Zayed H (2021) Structure-Based Virtual Screening to Identify Novel Potential Compound as an Alternative to Remdesivir to Overcome the RdRp Protein Mutations in SARS-CoV-2. Front. Mol. Biosci. 8:645216. doi: 10.3389/fmolb.2021.645216

The number of confirmed COVID-19 cases is rapidly increasing with no direct treatment for the disease. Few repurposed drugs, such as Remdesivir, Chloroquine, Hydroxychloroquine, Lopinavir, and Ritonavir, are being tested against SARS-CoV-2. Remdesivir is the drug of choice for Ebola virus disease and has been authorized for emergency use. This drug acts against SARS-CoV-2 by inhibiting the RNA-dependent-RNA-polymerase (RdRp) of SARS-CoV-2. RdRp of viruses is prone to mutations that confer drug resistance. A recent study by Pachetti et al. in 2020 identified the P323L mutation in the RdRp protein of SARS-CoV-2. In this study, we aimed to determine the potency of lead compounds similar to Remdesivir, which can be used as an alternative when variants of SARS-CoV-2 develop resistance due to RdRp mutations. The initial screening yielded 704 compounds that were 90% similar to the control drug, Remdesivir. On further evaluation through drugability and antiviral inhibition percentage analyses, we shortlisted 32 and seven compounds, respectively. These seven compounds were further analyzed for their molecular interactions, which revealed that all seven compounds interacted with RdRp with higher affinity than Remdesivir under native conditions. However, three compounds failed to interact with the mutant protein with higher affinity than Remdesivir. Dynamic cross-correlation matrix (DCCM) and vector field collective motions analyses were performed to identify the precise movements of docked complexes' residues. Furthermore, the compound SCHEMBL20144212 showed a high affinity for native and mutant proteins and might provide an alternative against SARS-CoV-2 variants that might confer resistance to Remdesivir. Further validations by *in vitro and in vivo* studies are needed to confirm the efficacy of our lead compounds for their inhibition against SARS-CoV-2.

Keywords: SARS-CoV-2, COVID-19, remdesivir, virtual screening, RdRp, mutations

INTRODUCTION

COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Huang et al., 2020). In December 2019, China reported its first case of SARS-CoV-2 in Wuhan, Hubei province (Zhu et al., 2020). The infection was highly contagious, which led to the global spread of the virus in the following months, thus causing an outbreak (Giovanetti et al., 2020; Phan et al., 2020). It was finally declared as a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO) on January 30, 2020, and ultimately named Coronavirus Disease 2019 (COVID-19) on February 12, 2020. COVID-19 is responsible for high mortality worldwide (WHO, 2020a,b). Currently, there are 172, and 61 vaccines are in pre-clinical and clinical development, respectively. While it takes years to develop a new effective drug against a disease, the COVID-19 outbreak has created a global emergency. Hence, drug repositioning and repurposing are given more attention as it is a rapid solution to identify a drug to combat the disease (Li and Clercq, 2020; Morse et al., 2020). Coronaviruses are enveloped viruses about 80–120 nm in diameter with plus-strand (+) RNA as genetic material (Brian and Baric, 2005). Studies have shown that SARS-CoV-2 shares ∼80% genetic similarity with SARS-CoV, while RNA-dependent RNA polymerase (RdRp) has a 96% similarity (Xu et al., 2020). RdRp and protease are encoded by the viral RNA and play an essential role in replicating and assembling the virus, and hence preexisting drugs that target these proteins are preferred. Antiviral agents that have been developed to combat coronaviruses are used here as potential drugs for the treatment of COVID-19.

A 20-kb gene encodes the replicase complex. It encompasses several viral genes encoding RdRp, RNA helicase, proteases, and some RNA processing enzymes and cellular proteins that aid in replicating and transcribing the genome of coronavirus occurring in the cytoplasm of the host cell (Koonin and Dolja, 1993; Cheng et al., 2005). RNA synthesis takes place in both a continuous and a discontinuous fashion (Gallagher, 1996). Coronaviruses are plus-strand (+) RNA viruses with a tremendously diverse genome, resulting in a variation in their RNA synthesis machinery (Goldbach, 1987). RdRps exhibit a very high transcription error rate leading to variation in the genome of the virus. During the replication process, RdRps adopts the switching mechanism resulting in the recombination of RNA. This process is also responsible for repairing deleterious mutations in the genome of the viruses, leading to gene rearrangements and new gene acquisition (Strauss and Strauss, 1988; Xu et al., 2003).

Once the virus penetrates the host cell by binding to the angiotensin-converting enzyme 2 (ACE2) receptor, the process of uncoating is initiated that releases the plus-strand (+) RNA into the cytoplasm. The ribosomes present in the host cell's cytoplasm proceed to translate the genomic RNA into a polyprotein. This polyprotein undergoes proteolytic cleavage by the protease enzyme that results in the replicase enzyme production and various viral structural proteins. Approximately 67% of the genomic RNA encodes RdRp that arbitrates the genome's synthesis (Thiel et al., 2001). The plus-strand (+) RNA can only produce viral protein products and not genetic material by replication. Thus, to achieve genome replication, RdRp first transcribes and replicates the plus-strand (+) RNA, which acts as a template to generate the minus-strand (–) RNA, and then serves as a template for the RdRPs to produce several plus-strand (+) RNAs. Some of this plus-strand (+) RNA is again translated into proteins.

In contrast, the others are enclosed into the capsid to regenerate complete virions released from the cell by exocytosis. RdRP is a popular target for a selective antiviral strategy against coronaviruses because RNA synthesis by RdRP does not occur in mammalian cells (Casais et al., 2001). **Figure 1** illustrates the role of RdRp in viral replication and effect of RdRp inhibitors (**Figure 1**).

Current potential treatments against COVID-19 that have shown promising results are Lopinavir, a known protease inhibitor of coronavirus (Yao T.-T. et al., 2020), the guanosine analog, ribavirin that targets RdRp and was designed to combat the Ebola outbreak (Falzarano et al., 2013), and chloroquine, an antimalarial drug, that has shown antiviral effects by blocking viral fusion with the cell due to increased endosomal pH (Vincent et al., 2005). A derivative of chloroquine, hydroxychloroquine, is more potent than chloroquine, as reported by several studies (Yao X. et al., 2020). Several corticosteroids have also been studied for their coronavirus effects (Russell et al., 2020). Convalescent plasma transfusion is also currently being administered and has been shown to reduce mortality rate (Mair-Jenkins et al., 2015). Among the treatments, Remdesivir has shown promising results in *in vitro* and animal studies and is currently in phase III trials (Martinez, 2020).

Remdesivir, also known as GS-5734, is a nucleotide prodrug that is metabolized inside the cell into GS-441524 and an adenosine nucleotide analog, phosphorylated into cell-permeable di-and tri-phosphates (Varga et al., 2016; Warren et al., 2016) Viral RdRps can utilize these adenosine triphosphates during genome replication (Sheahan et al., 2017). Remdesivir causes the nascent RNA transcript to terminate prematurely and incorporates mutations into it (Agostini et al., 2018). Studies have shown that remdesivir causes RNA levels to decrease in a dose-dependent manner, and it is 4.5-fold more potent in cells lacking the ExoN proofreading activity. It has also been observed that GS-5734 is more active (∼30 times) than its metabolized form, GS-441524, against the coronaviruses tested (Sheahan et al., 2017). It has been shown that in the presence of GS-441524, there was 5.6-fold resistance in coronavirus mouse hepatitis virus, which was attributed to two mutations, F476L and V553L, in the coding region of nsp12 core polymerase (Agostini et al., 2018). Similar findings were observed in SARS-CoV carrying the mutations F480L and V557L, which are not in the vicinity of the binding site of RdRp; thus, the exact mechanism behind the resistance remains undetermined (Agostini et al., 2018). This raises the concern that resistance may cause the virus to be active and functional, leading to severe disease transmission. Because of its decreased proofreading activity, RdRps have an error rate in the range of $10^{-4}-10^{-6}$ errors/nucleotide/replication, which is very high compared to that of DNA polymerases (Eckerle et al., 2010). This high error rate is responsible for the rapid rate of virus evolution, contributing to its adaptability to the new

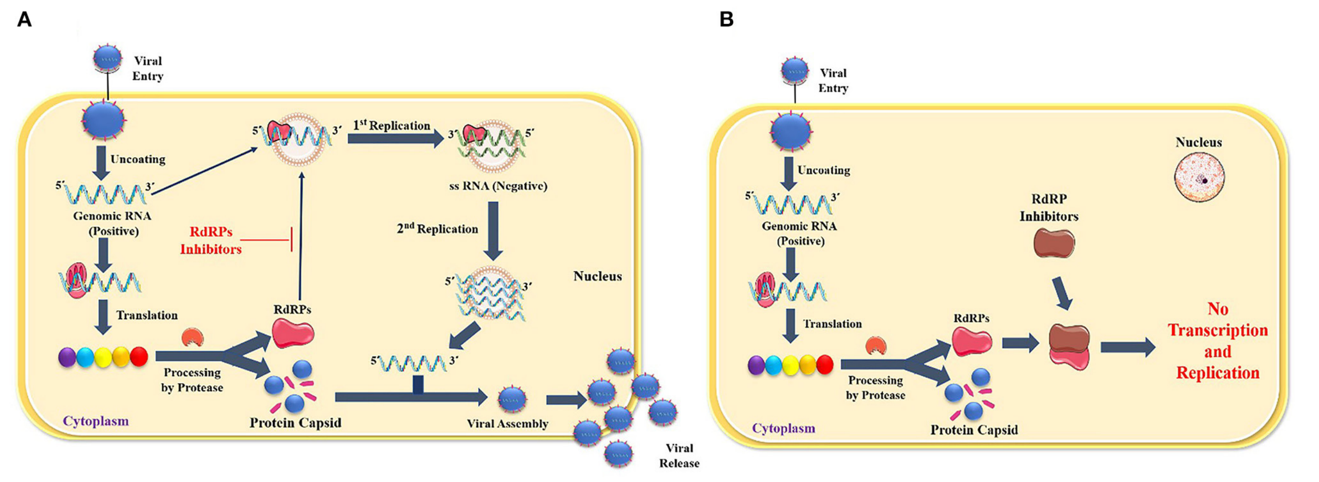


FIGURE 1 | (A) Role of RdRp in viral replication, **(B)** Effect of RdRp inhibitors.

surroundings and expediting its ability to jump between species (Pillay and Zambon, 1998). A recent study by Pachetti et al. in 2020 identified the P323L mutation in the RdRp protein of SARS-CoV-2, which may confer resistance to the drug. In this study, we utilized a virtual screening approach to identify the potential repurposed lead compound that might be efficient in treating COVID-19 due to SARS-CoV-2 variants resistant to Remdesivir.

MATERIALS AND METHODS

Remdesivir and COVID19-RdRp Structure Analysis

The 2D and 3D structures of Remdesvir in canonical SMILES and SDF formats were obtained from the PubChem database with CID 121304016. These formats were also converted to PDB using the OpenBabel software (O'Boyle et al., 2011). The 3D structure of the complexes between the non-structural proteins 12, 7, and 8 and Remdesivir was obtained from the Protein Data Bank with PDB ID—7BV2, and the missing amino acids were theoretically built using the Swiss Model Server (Biasini et al., 2014; Wanchao et al., 2020). Further, the P323L mutation reported by Pachetti et al. in 2020 was introduced in the refined protein structure using the SwissPDB Viewer, and energy was minimized using the same method (Guex and Peitsch, 1997; Pachetti et al., 2020).

Virtual Screening

The inbuilt module of PubChem was used to identify compounds with 90% similarity to the control Remdesivir structure. The identified compounds were screened for absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties using the SwissADME server (Daina et al., 2017). The 90% similar compounds that satisfy the Lipinski drugability properties were further used in an *in silico* antiviral inhibition percentage study using the AntiViral Compound Prediction (AVCpred) server (Qureshi et al., 2017). These filtered compounds were finally subjected to molecular interaction studies against the native and mutant RdRp proteins using the AutoDock standalone package (Morris et al., 2009).

Molecular Interaction Studies

Molecular interaction studies were initiated to understand the differences between the control remdesivir and the virtually similar compounds with RdRp proteins. Hydrogen and charges were added to the RdRp protein, and the torsions were set to Remdesivir and nearly identical compounds. The grid box was placed around the active site (LYS546, SER683, ARG556, THR688, ASP624, SER760, ASN692, ASP761, and ASP762) retrieved from the published crystal structure (Wanchao et al., 2020). The Lamarckian Genetic Algorithm was used to generate the binding pockets and binding affinity of RdRp for Remdesivir or the virtually similar compounds. One hundred different pockets were developed, and those with the best binding affinity were used in the structural visualization.

Dynamic Cross-Correlation Matrix (DCCM) and Collective Motion

The docked complexes were further subjected to the web-server DynOmics ENM to identify the residue cross-correlation matrix. The tool developed with the Elastic network models (ENM) that integrate Anisotropic Network Model (ANM) and GaussianNetwork Model (GNM) (Li et al., 2017). Time-correlated data was represented as a matrix between the protein atoms i and j (cij) in DCCM. Within the form of a map, typical fluctuations, and standardized correlations among residues are typically shown and represented with the following equation:

$$\mathrm{Cij} = <\Delta\mathrm{Ri}.\ \Delta\mathrm{Rj}>$$
$$\mathrm{Cij}\ (n) = <\Delta\mathrm{Ri}.\ \Delta\mathrm{Rj}>/[<\Delta\mathrm{Ri}.\ \Delta\mathrm{Ri}><\Delta\mathrm{Rj}.\ \Delta\mathrm{Rj}>]1/2$$

The range of Cij(n) varies in terms (−1, 1) and analyzes information on the cross-correlation between residue movements i and j (Rader et al., 2005). To determine how mutations affect the internal mechanics of protein

TABLE 1 | Antiviral inhibition percentage prediction of the compounds satisfying Lipinski's drugability properties.

S.No	SMILES	Ligand	PubChem ID	General	HBV	HCV	HHV	HIV
1	CCC(CC)COC(=O)C(C)NP(=O)(OCC1C(C(C(O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O)OC4=CC=CC=C4	Remdesvir	121304016	51.563	21.339	54.616	32.307	65.861
2	CCOC(=O)[C@H](C)NP1(=O)OC[C@@H]2[C@@H](O1)[C@@]([C@](O2)(C)C3=CC=C4N3N=CN=C4N)(C)O	44468492	44468492	57.651	21.148	48.128	33.76	62.468
3	C[C@]1([C@@H]([C@H](O[C@@]1(C#N)C2=CC=C3N2N=CN=C3N)COP(=O)(O)OC4=CC=CC=C4)O)F	51041120	51041120	29.001	18.05	45.644	0.293	73.241
4	CCOC(=O)[C@H](C)NP1(=O)OC[C@@H]2[C@@H](O1)[C@@]([C@](O2)(C)C3=CC=C4N3N=CN=C4C)(C)O	58527341	58527341	59.052	20.445	40.745	13.976	61.67
5	CCOC(=O)[C@H](C)NP1(=O)OC[C@@H]2[C@@H](O1)[C@@]3(C[C@]3(O2)C4=CC=C5N4N=CN=C5N)O	68270743	68270743	43.53	19.756	53.964	40.627	70.965
6	CC(C)C(=O)O[C@@H]1[C@H](O[C@H]([C@]1(C)O)C2=CC=C3N2N=CN=C3N)CO[P+](=O)OC4=CC=CC=C4	70649275	70649275	57.297	20.377	33.6	23.985	61.671
7	C[C@]1([C@@H]([C@H](O[C@@]1(C#N)C2=CC=C3N2N=CN=C3N)CO[P+](=O)OC4=CC=CC=C4)O)O	90048786	90048786	39.527	17.712	55.339	17.268	61.24
8	C[C@@]1([C@@H]([C@@H]([C@H](O1)COP(=O)(N)OC2=CC=CC=C2)O)O)C3=CC=C4N3N=CN=C4N	117913880	117913880	49.781	18.567	24.086	51.373	63.478
9	C[C@@]1([C@@H]([C@@H]([C@H](O1)COP(=O)(N)OC2=CC=CC=C2)O)O)C3=CC=C4N3N=CN=C4N=NN	117913912	117913912	51.369	17.632	55.849	32.075	74.323
10	C[C@@H](C(=O)OC1CCC1)NP(=O)(C)OC[C@@H]2[C@H]([C@H]([C@](O2)(C#N)C3=CC=C4N3N=CN=C4N)O)O	121310126	121310126	43.53	19.756	53.964	40.627	70.965
11	C1=CC=C(C=C1)OP(=O)(O)OC[C@@H]2[C@H]([C@H]([C@](O2)(C#N)C3=CC=C4N3N=CN=C4N)O)O	121313145	121313145	43.449	20.677	51.671	41.029	73.631
12	CCC(CC)COC(=O)[C@H](C)N[P+](=O)OC[C@@H]1CC[C@](O1)(C#N)C2=CC=C3N2N=CN=C3N	126693021	126693021	30.141	11.579	53.901	30.889	66.354
13	C[C@@]1([C@@H]([C@@H]([C@H](O1)CO[P+](=O)OC2=CC=CC=C2)O)O)C3=CC=C4N3N=CN=C4N	126719084	126719084	45.799	21.359	48.486	20.775	66.338
14	C[C@@H](C(=O)OCC(C)(C)C)N[P+](=O)OC[C@@H]1[C@H]([C@H]([C@](O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O	126719091	126719091	29.001	18.05	45.644	0.293	73.241
15	C[C@@H](C(=O)OC)N[P+](=O)OC[C@@H]1[C@H]([C@H]([C@](O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O	126719092	126719092	35.15	19.336	67.523	42.404	62.287
16	C[C@@]1([C@@H]([C@@H]([C@H](O1)COP(=O)(O)OC2=CC=CC=C2)O)O)C3=CC=C4N3N=CN=C4N	130312728	130312728	23.45	23.058	21.02	29.107	63.229
17	CCC(CC)COC(=O)CN[P+](=O)OC[C@@H]1[C@H]([C@H]([C@](O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O	130312749	130312749	37.635	20.432	52.161	59.727	60.1
18	CCC(CC)COC(=O)C(C)NP(OCC1CCC(O1)C2=CC=C3N2N=CN=C3N)OC4=CC=CC=C4	132046034	132046034	45.799	21.359	48.486	20.775	66.338
19	CCC(CC)COC(=O)[C@H](C)N[P+](=O)OC[C@@H]1C[C@H]([C@](O1)(C#N)C2=CC=C3N2N=CN=C3N)O	134443511	134443511	30.21	13.353	53.898	30.99	66.363
20	C[C@@]1([C@@H]([C@@H](C(O1)CO[P+](=O)OC2=CC=CC=C2)O)O)C3=CC=C4N3N=CN=C4N	134502618	134502618	26.991	20.168	68.022	44.334	72.874
21	C1=CC=C(C=C1)O[P+](=O)OC[C@@H]2[C@H]([C@H]([C@](O2)(C#N)C3=CC=C4N3N=CN=C4N)O)O	134502627	134502627	30.818	20.362	67.286	38.195	65.452
22	CC(C(=O)O)NP(=O)(O)OC[C@@H]1C[C@H]([C@](O1)(C#N)C2=CC=C3N2N=CN=C3N)O	134502628	134502628	54.648	17.777	54.79	40.625	62.179
23	CCOC(=O)[C@H](C)NP(=O)(O)OC[C@@H]1[C@H]([C@@]([C@](O1)(C#N)C2=CC=C3N2N=CN=C3N)(C)O)O	137648734	137648734	62.457	20.55	48.165	33.256	63.195
24	C[C@@H](C(=O)OC(C)C)NP(=O)(O)OC[C@@H]1[C@H]([C@@]([C@](O1)(C#N)C2=CC=C3N2N=CN=C3N)(C)O)O	137660077	137660077	40.854	20.469	13.066	18.678	67.667
25	CCC(CC)COC(=O)[C@H](C)N[P+](=O)OC[C@@H]1[C@H](C[C@](O1)(C#N)C2=CC=C3N2N=CN=C3N)O	138525664	138525664	18.701	20.275	56.884	28.924	73.168
26	C1=CC=C(C=C1)OP(=O)(O)OCC2[C@H]([C@H]([C@](O2)(C#N)C3=CC=C4N3N=CN=C4N)O)O	138525701	138525701	26.651	20.42	26.507	38.09	68.021
27	CC(C)OC(=O)CNP(OC[C@@H]1C[C@@H]([C@@H](O1)C2=CC=C3N2N=CN=C3N)C#N)OC4=CC=CC=C4	138529102	138529102	39.45	19.357	62.986	54.581	79.639
28	CCOC(=O)CNP(OC[C@@H]1C[C@@H]([C@@H](O1)C2=CC=C3N2N=CN=C3N)C#N)OC4=CC=CC=C4	138529141	138529141	35.048	20.109	62.385	38.05	69.065

(Continued)

TABLE 1 | Continued

S.No	SMILES	Ligand	PubChem ID	General	HBV	HCV	HHV	HIV
29	C[C@]1([C@H]([C@H]([C@@H](O1)C2=CC=C3N2N=CN=C3N)O)O)CO[P+](=O)OC4=CC=CC=C4	138598986	138598986	47.712	20.613	58.371	33.193	64.964
30	CN=C[C@@]1([C@@H]([C@@H](C(O1)COP(=O)(O)OC2=CC=CC=C2)O)O)C3=CC=C4N3N=CN=C4N	139476292	139476292	22.199	23.047	21.248	13.219	61.262
31	CCC(CC)COC(=O)CN[P@](OCC1CCC(O1)C2=CC=C3N2N=CN=C3N)OC4=CC=CC=C4	145074438	145074438	55.189	20.709	49.992	33.319	63.255
32	CCC(CC)COC(=O)C(C)N[P@](OCC1CCC(O1)C2=CC=C3N2N=CN=C3N)OC4=CC=CC=C4	145074498	145074498	39.424	19.374	62.987	54.504	79.64
33	CC(C)OC(=O)CNP(OCC1CCC(O1)(C#N)C2=CC=C3N2N=CN=C3N)OC4=CC=CC=C4	145074552	145074552	62.213	21.225	48.134	33.256	66.604

conformations, the Bio3D module incorporated with the R studio was used to quantify residue-residue dynamic cross-correlation networks. The normal mode, network analysis, and correlation analysis were called with the function “nma(),” “can(),”and “dccm().” The role from these features is usually a cross-correlation matrix of residue-residue. The results were plotted by calling the functions “plot.dccm()” and “pymol.dccm()” (Grant et al., 2006; Scarabelli and Grant, 2013).

RESULTS

Virtual Screening

As an initiative of virtual screening, the compounds with 90% structural similarity to the control drug, Remdesivir, were screened using the inbuilt PubChem module. The screening identified 704 compounds possessing 90% similarity with Remdesivir. These 704 compounds were subjected to *in silico* ADME analysis using the SwissADME server (**Supplementary Table 1**). Based on Lipinski’s drugability properties, 32 compounds were further filtered from the pool of 704 compounds. Finally, these 32 compounds were subjected to an antiviral inhibition percentage study calculated using the AVCpred server (**Table 1**). This analysis identified the compounds with PubChem IDs 137648734, 145074552, 58527341, 44468492, 70649275, 145074438 134502628 that showed higher antiviral inhibition percentages than Remdesivir. The inhibition percentage of Remdesivir was 51.563%. However, the inhibition percentages of compounds, 137648734, 145074552, 58527341, 44468492, 70649275, 145074438, and 134502628 were found to be 62.457, 62.213, 59.052, 57.651, 57.297, 55.189, and 54.648%, respectively. These seven compounds and Remdesivir were subjected to molecular interaction studies against the native and mutant RdRp molecules.

Molecular Interaction Studies

Molecular interaction studies were performed using AutoDock to understand the interaction pattern of Remdesivir and its similar compounds with the native and mutant RdRp proteins. It was observed that all seven compounds had a higher binding affinity than Remdesivir with native RdRp. For RdRp protein carrying the P323L mutation, we observed that the compounds 134502628, 70649275, 137648734, and 145074552 possessed higher binding affinity than Remdesivir. The binding affinity of Remdesivir for native and mutant RdRp was −4.5 and −4.2 kcal/mol, respectively. However, the compound with PubChem ID 134502628 showed the highest binding affinity of −7.5 and −7.4 kcal/mol for the native and mutant RdRp proteins, respectively (**Tables 2A,B**). The detailed interaction comparison between Remdesivir and top-ranked compound, 134502628 (SCHEMBL20144212), against the native and mutant RdRp is shown in **Table 3**, **Figures 2A–D**.

TABLE 2 | The binding affinity of the ligands against native and mutant SARS-CoV-2 RdRp structures.

S.No	Protein	Ligands	Binding Affinity (kcal/mol)
(A)			
1	Native RdRp	134502628	−7.5
2	Native RdRp	58527341	−6.3
3	Native RdRp	145074552	−6
4	Native RdRp	145074438	−6
5	Native RdRp	137648734	−5.5
6	Native RdRp	70649275	−5.3
7	Native RdRp	44468492	−5.1
8	Native RdRp	Remdesvir	−4.5
(B)			
1	RdRp with P323L	134502628	−7.4
2	RdRp with P323L	70649275	−6.6
3	RdRp with P323L	137648734	−6.3
4	RdRp with P323L	145074552	−6.1
5	RdRp with P323L	Remdesvir	−5.6
6	RdRp with P323L	145074438	−5.5
7	RdRp with P323L	44468492	−5.3
8	RdRp with P323L	58527341	−4.2

***(A)** Against native RdRp, **(B)** against mutant RdRp.*

DCCM and Collective Motion

Firstly, we investigated the residue cross-correlation networks and vector field collective motions with the help of the Bio3D module that was integrated within R studio. All four complexes were subjected to the Bio3D module to retrieve the residue cross-correlation networks and collective motions. As a result,

TABLE 3 | Differences in the binding affinity (kcal/mol), number of hydrogen bonds, number of interacting amino acids, and Interacting amino acids between native and mutant RdRp against Remdesivir and the novel lead compound.

S.No	Protein	Ligands	Binding affinity (kcal/mol)	Number of hydrogen bonds	Number of interacting amino acids	Interacting amino acids
1	Native RdRp	Remdesivir	−4.5	6	14	SER815, TRP618, ASP762, ASP619, ASP761, TYR620, PRO621, LYS622, LYS623, ASP624, ARG625, ARG554, ARG556, LYS546
2	Native RdRp	SCHEMBL20144212	−7.5	7	11	GLU812, SER815, ARG837, ALA551, SER550, LYS552, ALA548, ARG554, ARG556, LYS546, SER683
3	RdRp with P323L mutation	Remdesivir	−4.2	2	4	SER550, ILE549, ARG554, ARG556
4	RdRp with P323L mutation	SCHEMBL20144212	−7.4	6	14	LYS594, GLN816, SER815, CYS814, ARG837, ARG859, ALA841, HIS440, ALA551, SER550, ILE549, LYS552, ARG554, ARG556

the residues from 5 to 120 (α1, α2, α3, β1, β2, and β-hairpin) and 830–933 (α40–45, β22, and β23) were obtained with large cross-correlation networks from RdRp-Remdesivir and SCHEMBL20144212 complexes (**Figures 3A,C**), whereas the P323L-Remdesivir obtained less cross-correlation networks especially at the region from 5 to 120 residues (**Figure 3E**). The residues from 830 to 933 exhibited similar networks for all the docked complexes. However, the differences can be seen with the P323L-Remdesivir and P323L-SCHEMBL20144212 complexes, especially at the region from 5 to 120 residues. The P323L-Remdesivir complex exhibited the least residue cross-correlation; the possible reason for this could be mutation at 323rd position in RdRp. Our results exhibited similar network cross-correlation for P323L-SCHEMBL20144212 (**Figure 3G**) when compared to the RdRp-SCHEMBL20144212 complex. The atomic movements of these docked complexes were investigated by representing the vector field collective motions. The collective motions were largely observed in the 5–120 and 830–933 regions, as seen in **Figures 3B,D,F,H**. Here, we observed the differences in the residues from 5 to 120 from the P323L-Remdesivir complex (**Figure 3F**) compared to the other three docked complexes. The arrows on the docked complex of proteins determine the magnitude and direction of the collective motion. Furthermore, we investigated the DCCM of RdRp (with or without mutation) with remdesivir and SCHEMBL20144212 complexes by utilizing the DynOmics tool. DCCM explores the precise movements of residues of proteins, demonstrates strongly correlated (in red) atomic movements between residues, and highlights atomic movements that are strongly anticorrelated (in blue) (**Figure 4**). The RdRp-SCHEMBL20144212 and P323L-SCHEMBL20144212 complexes (**Figures 4B,C**) illustrate the strongly correlated motions and less anticorrelation as compared to that of RdRp-Remdesivir and P323L-Remdesivir complexes (**Figures 4A,C**).

DISCUSSION

COVID-19, caused by SARS-CoV-2, has become a global pandemic and is responsible for about 2,71,954 deaths worldwide, with more than 3.95 million positive cases. COVID-19 has been transmitted to almost 120 countries, and discovering drugs to fight this disease is a great challenge (Worldometer, 2021). The genome of the SARS-CoV-2 virus was immediately sequenced to improve our understanding of the virus, formulate appropriate diagnostic and preventive techniques, and develop therapeutic strategies (Wang et al., 2020). The high transmission rate of the virus is responsible for the difficulty of implementing efficient preventive measures. Hence, there is an immediate requirement for a drug to inhibit the virus and eliminate the disease (Li et al., 2020). One of the emerging drugs with promising results is Remdesivir. It is a nucleotide analog, initially developed for the Ebola virus, that targets the virus's RdRp (Gordon et al., 2020). The virus's high mutation rate may introduce changes to RdRp, rendering the drug ineffective (Eckerle et al., 2010). A recent study by Pachetti et al. in 2020 identified the P323L mutation in the RdRp of SARS-CoV-2 (Pachetti et al., 2020). The possible increase in mutations highlights the need to develop a series of drugs whose efficiency is not affected by these mutations.

Thus, there is an urgent need to develop novel drugs. The conventional method for drug discovery is expensive and time-consuming, and therefore the *in silico* approach is a means to move forward. Our study formulated a computational pipeline to identify a series of compounds similar to Remdesivir, which could be used as an alternative if SARS-CoV-2 develops resistance to Remdesivir due to mutations. Out of 704 compounds, which were 90% similar to Remdesivir, 32 compounds were found to satisfy the rule of 5 (**Supplementary Table 1**). To assess drugability, these 32 compounds were subjected to *in silico* antiviral inhibition percentage analysis. We observed that the seven virtually similar compounds possessed a higher antiviral inhibition percentage when compared to Remdesivir. Finally, molecular interaction studies against the native and mutant RdRp were carried out with docking analysis. The interactions revealed that all the identified seven compounds could interact with higher affinity with the native protein than Remdesivir.

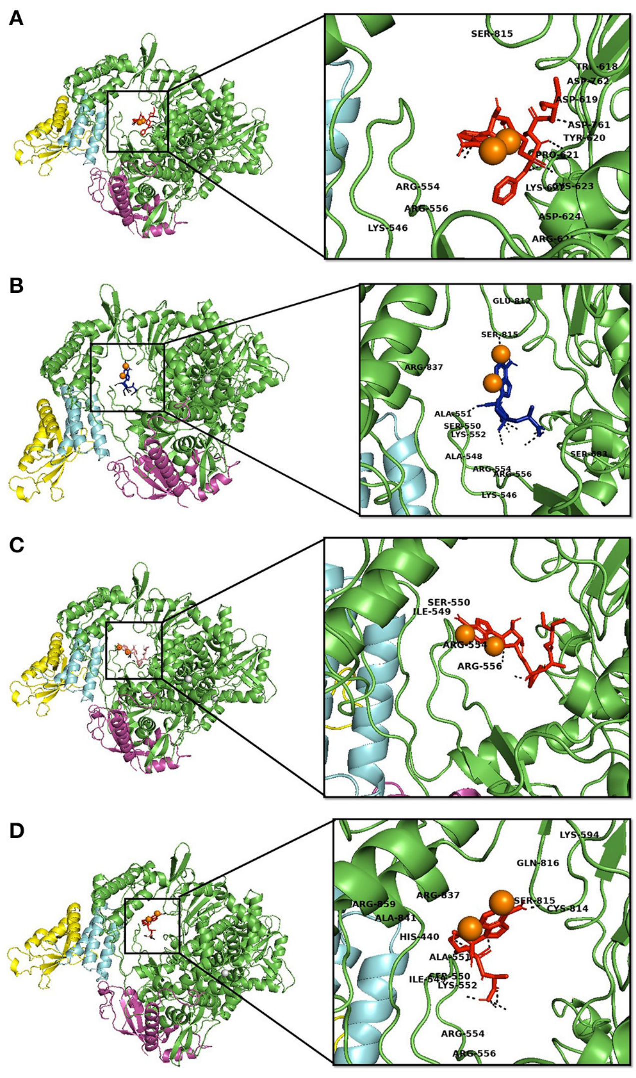


FIGURE 2 | Interaction of Remdesvir and the novel lead compound with the native RdRp and RdRp with the mutation. **(A)** Native RdRp with Remdesvir, **(B)** Native RdRp with novel lead compound (SCHEMBL20144212), **(C)** Mutant RdRp with Remdesvir, **(D)** Mutant RdRp with novel lead compound (SCHEMBL20144212).

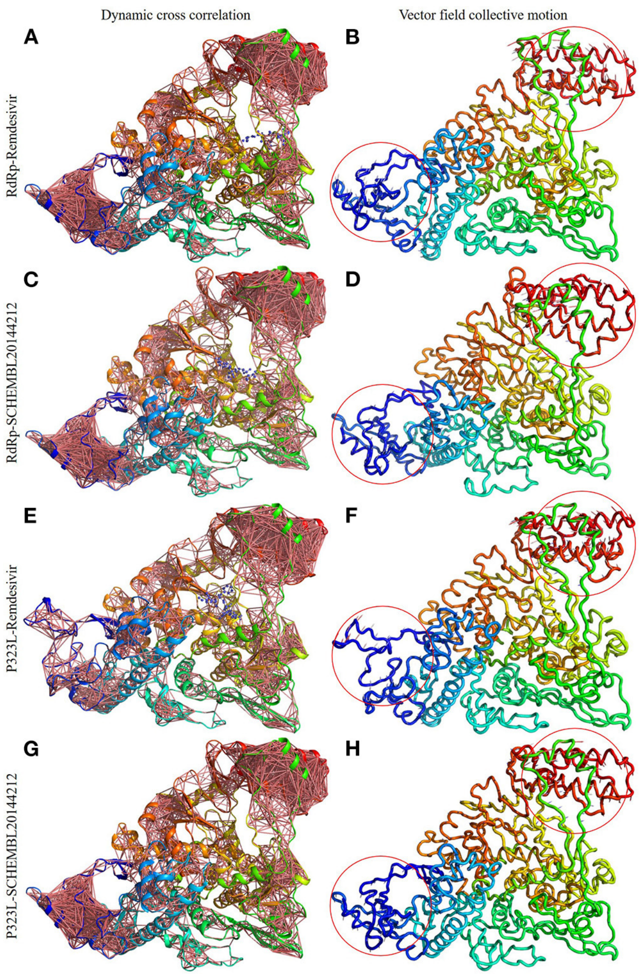


FIGURE 3 | Dynamic cross-correlation networks and vector field collective motions of docked complexes. Depiction of cross-correlation networks: **(A)** RdRp-Remdesivir; **(C)** RdRp-SCHEMBL20144212; **(E)** P323L-Remdesivir; **(G)** P323L-SCHEMBL20144212. Depiction of vector field collective motions: **(B)** RdRp-Remdesivir; **(D)** RdRp-SCHEMBL20144212; **(F)** P323L-Remdesivir; **(H)** P323L-SCHEMBL20144212. Red circles indicate differences in the direction of collective motions by arrows.

However, upon introducing the P323L mutation, only four compounds were found to bind with a higher binding affinity than Remdesivir. The binding efficacy of Remdesivir was also found to decrease from −4.5 kcal/mol in the native to −4.2 kcal/mol in the mutant protein. Interestingly, the compound, SCHEMBL20144212, was topmost in the binding affinity for native and mutant proteins (**Table 2**). Remdesivir was found to interact with the native RdRp through 14 amino acids and 6 hydrogen bonds. However, the identified novel lead compound, SCHEMBL20144212, was found to interact with native RdRp through 11 amino acids and 7 hydrogen bonds.

Similarly, in interaction with the mutant protein, Remdesivir was found to interact through four amino acids and two hydrogen bonds. However, the novel lead compound, SCHEMBL20144212, was found to interact through 14 amino acids and 6 hydrogen bonds (**Table 3**, **Figures 2A–D**). The role of hydrogen bonds and interacting amino acids is crucial for a drug to exhibit its efficacy. We observed higher binding affinity, hydrogen bond interactions, and amino acid interactions with the identified novel lead compound compared to Remdesivir. We have also depicted the detailed pharmacokinetic properties obtained from the pkCSM server in **Table 4**. These results suggest that SCHEMBL20144212 is more effective against the original or the mutated virus compared with Remdesivir.

It is vital to characterize the protein folding via identifying conformational variances, change in conformational motions owing to mutations, mechanism of ion channel's opening/closing, and protein-ligand binding, as these factors

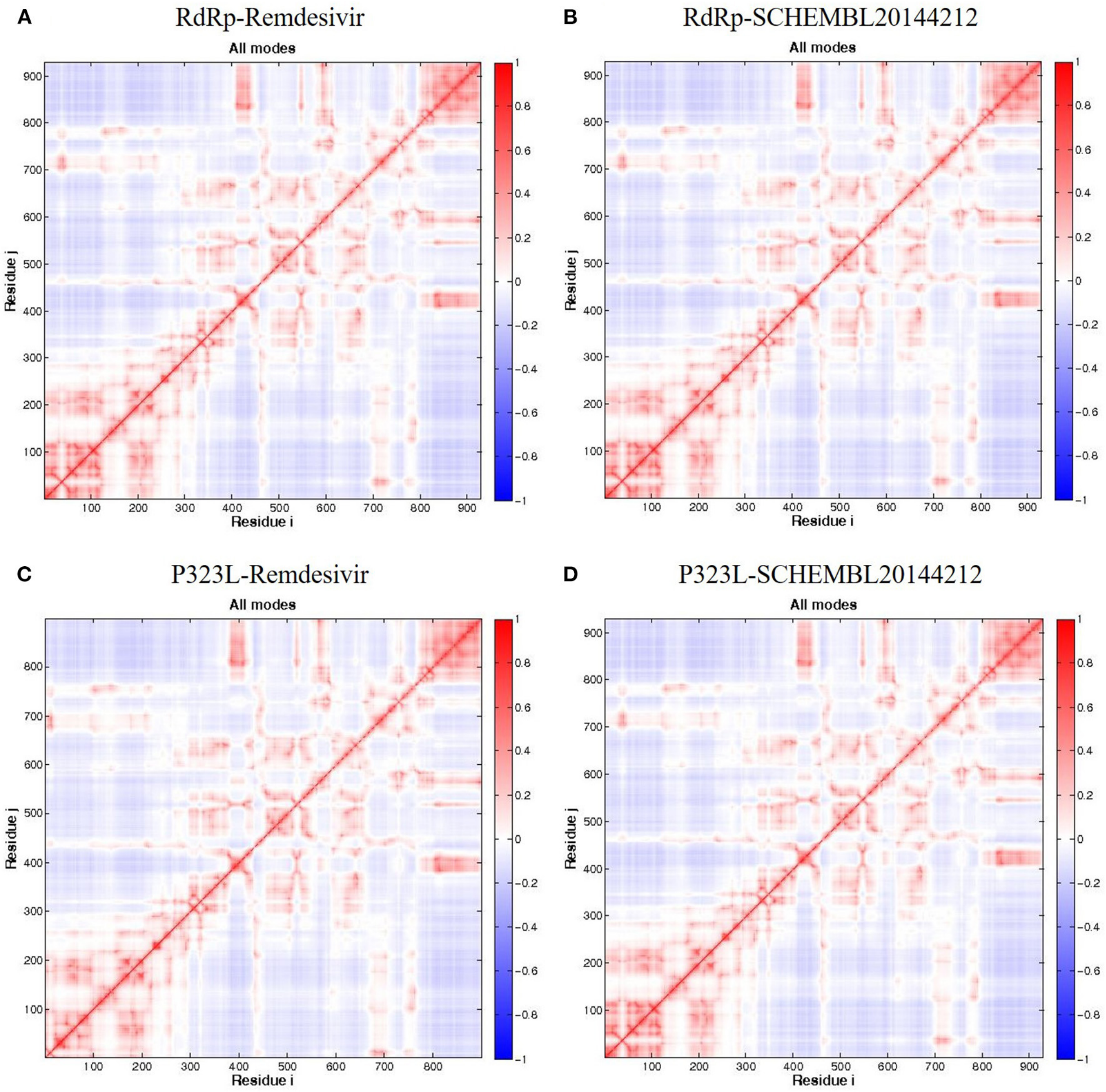


FIGURE 4 | DCCM of the residues around their mean position for all docked RdRp complexes. **(A)** RdRp-Remdesivir; **(B)** RdRp-SCHEMBL20144212; **(C)** P323L-Remdesivir; **(D)** P323L-SCHEMBL20144212. Map size = 100, min i = 5, and min j = 5. The color ranges from −1 (blue) to +1 (red), whereas 0 remains in white color.

directly related to the protein's stability and function (Grottesi et al., 2005; Yang et al., 2008; Kim et al., 2010; S et al., 2020; Udhaya Kumar et al., 2020). The residue cross-correlation networks were obtained for all the docked complexes, and among them, the P323L-Remdesivir complex obtained less correlative network at the region from 5 to 120 residues (**Figure 3E**). A possible reason for this could be a mutation that occurred at position 323; thus, complex resulted in fewer correlative networks. Also, in comparison to our identified lead compound (SCHEMBL20144212) with mutant RdRp (P323L) exhibited large correlative networks and could act as a sturdy inhibitor against mutant RdRp (**Figure 3G**). The vector field collective motions were identified for all the docked complexes to explore each complex's dynamic motion and magnitude (**Figures 3B,D,F,H**). The RdRp and mutant RdRp (P323L) with SCHEMBL20144212 lead compound exhibited a rigid complex as seen in **Figures 3D,H**. This clearly states that the SCHEMBL20144212 lead compound inhibits the RdRp from binding to other molecules and for the catalysis process, whereas the remdesivir complexes exhibited more flexible regions (**Figures 3B,F**) (Huber, 1987; Karshikoff et al., 2015). Furthermore, our results from DCCM showed the RdRp-SCHEMBL20144212 and P323L-SCHEMBL20144212 complexes exhibited strongly correlated motions (**Figures 4B,D**). From the above observations, we strongly believe that the molecular docking of SCHEMBL20144212 lead compound reduces the effect of P323L mutation and could act as an effective inhibitor against SARS-nCoV-2's RdRp.

TABLE 4 | ADMET properties of the novel lead compound obtained from the pkCSM server.

Property	Model name	Predicted value	Unit
Absorption	Water solubility	−2.347	Numeric (log mol/L)
Absorption	Caco2 permeability	−0.496	Numeric (log Papp in 10–6 cm/s)
Absorption	Intestinal absorption (human)	28.391	Numeric (% Absorbed)
Absorption	Skin Permeability	−2.735	Numeric (log Kp)
Absorption	P–glycoprotein substrate	No	Categorical (Yes/No)
Absorption	P-glycoprotein I inhibitor	No	Categorical (Yes/No)
Absorption	P-glycoprotein II inhibitor	No	Categorical (Yes/No)
Distribution	VDss (human)	−0.036	Numeric (log L/kg)
Distribution	Fraction unbound (human)	0.546	Numeric (Fu)
Distribution	BBB permeability	−1.906	Numeric (log BB)
Distribution	CNS permeability	−4.432	Numeric (log PS)
Metabolism	CYP2D6 substrate	No	Categorical (Yes/No)
Metabolism	CYP3A4 substrate	No	Categorical (Yes/No)
Metabolism	CYP1A2 inhibitor	No	Categorical (Yes/No)
Metabolism	CYP2C19 inhibitor	No	Categorical (Yes/No)
Metabolism	CYP2C9 inhibitor	No	Categorical (Yes/No)
Metabolism	CYP2D6 inhibitor	No	Categorical (Yes/No)
Metabolism	CYP3A4 inhibitor	No	Categorical (Yes/No)
Excretion	Total Clearance	0.179	Numeric (log ml/min/kg)
Excretion	Renal OCT2 substrate	No	Categorical (Yes/No)
Toxicity	AMES toxicity	No	Categorical (Yes/No)
Toxicity	Max. tolerated dose (human)	0.643	Numeric (log mg/kg/day)
Toxicity	hERG I inhibitor	No	Categorical (Yes/No)
Toxicity	hERG II inhibitor	No	Categorical (Yes/No)
Toxicity	Oral Rat Acute Toxicity (LD50)	2.376	Numeric (mol/kg)
Toxicity	Oral Rat Chronic Toxicity (LOAEL)	2.307	Numeric (log mg/kg_bw/day)
Toxicity	Hepatotoxicity	Yes	Categorical (Yes/No)
Toxicity	Skin Sensitisation	No	Categorical (Yes/No)
Toxicity	T.Pyriformis toxicity	0.285	Numeric (log ug/L)
Toxicity	Minnow toxicity	3.311	Numeric (log mM)

CONCLUSION

This study was initiated to identify the potential repurposed lead compound to treat COVID-19 in case of possible SARS-CoV-2-virus resistance to Remdesivir due to mutations. The drug inhibits the activity of RdRp protein to a small extent. Identifying the mutation in RdRp of SARS-CoV-2 highlights the possible appearance of mutated viruses that might be resistant to current and future treatments. Therefore, it is necessary to establish a platform to identify compounds similar to but more useful than Remdesivir. Our study identified a total of 704 compounds from the PubChem database, which are 90% similar to Remdesivir. Of these 704 compounds, 32 compounds were found to possess druggability properties, out of these, seven compounds were found to possess higher antiviral inhibition percentages when compared to Remdesivir. Out of these seven compounds analyzed for molecular interaction. The SCHEMBL20144212 compound was found to possess the highest interaction affinity for both the native and mutant RdRp protein. From the DCCM and vector field collective motion analysis, we demonstrated that our lead compound's molecular docking (SCHEMBL20144212) tolerates the effect of the P323L mutation and could act as an effective inhibitor against SARS-nCoV-2's RdRp. Thus, this compound should be further validated through *in vitro* experiments as an alternative to Remdesivir to bypass the resistant effect of the P323L mutation. Our study offers a platform for drug repurposing during the time of viral pandemics.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

TK, UK, HZ, and GD contributed to designing the study and data acquisition, analysis, and interpretation. TK, NS, and UK involved in the acquisition, analysis, and interpreting of the results. TK, UK, and GD contributed to data interpretation, conducted, and drafting the manuscript. GD and HZ supervised the entire study and were involved in study design, acquiring, analyzing, understanding the data, and drafting the manuscript. The manuscript was reviewed and approved by all the authors.

FUNDING

UK, one of the authors, gratefully acknowledges the Indian Council of Medical Research (ICMR), India, for providing him a Senior Research Fellowship [ISRM/11(93)/2019].

ACKNOWLEDGMENTS

The authors would like to thank the Vellore Institute of Technology, India, and Qatar University, Qatar, for providing the necessary research facilities and encouragement to carry out this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.645216/full#supplementary-material

Supplementary Table 1 | The list of 90% similar compounds to the Remdesivir obtained from the PubChem database.

REFERENCES

Agostini, M. L., Andres, E. L., Sims, A. C., Graham, R. L., Sheahan, T. P., Lu, X., et al. (2018). Coronavirus susceptibility to the antiviral remdesivir (GS-5734) is mediated by the viral polymerase and the proofreading exoribonuclease. *MBio* 9, 1–15. doi: 10.1128/mBio.00221-18

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42, W252–W258. doi: 10.1093/nar/gku340

Brian, D. A., and Baric, R. S. (2005). Coronavirus genome structure and replication. *Curr. Top. Microbiol. Immunol.* 287, 1–30. doi: 10.1007/3-540-26765-4_1

Casais, R., Thiel, V., Siddell, S. G., Cavanagh, D., and Britton, P. (2001). Reverse genetics system for the avian coronavirus infectious bronchitis virus. *J. Virol.* 75, 12359–12369. doi: 10.1128/JVI.75.24.12359-12369.2001

Cheng, A., Zhang, W., Xie, Y., Jiang, W., Arnold, E., Sarafianos, S. G., et al. (2005). Expression, purification, and characterization of SARS coronavirus RNA polymerase. *Virology* 335, 165–176. doi: 10.1016/j.virol.2005.02.017

Daina, A., Michielin, O., and Zoete, V. (2017). SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* 7:42717. doi: 10.1038/srep42717

Eckerle, L. D., Becker, M. M., Halpin, R. A., Li, K., Venter, E., Lu, X., et al. (2010). Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog.* 6, 1–15. doi: 10.1371/journal.ppat.1000896

Falzarano, D., de Wit, E., Rasmussen, A. L., Feldmann, F., Okumura, A., Scott, D. P., et al. (2013). Treatment with interferon-α2b and ribavirin improves outcome in MERS-CoV-infected rhesus macaques. *Nat. Med.* 19, 1313–1317. doi: 10.1038/nm.3362

Gallagher, T. M. (1996). Murine coronavirus membrane fusion is blocked by modification of thiols buried within the spike protein. *J. Virol.* 70, 4683–4690. doi: 10.1128/JVI.70.7.4683-4690.1996

Giovanetti, M., Benvenuto, D., Angeletti, S., and Ciccozzi, M. (2020). The first two cases of 2019-nCoV in Italy: where they come from? *J. Med. Virol.* 92, 518–521. doi: 10.1002/jmv.25699

Goldbach, R. (1987). Genome similarities between plant and animal RNA viruses. *Microbiol. Sci.* 4, 197–202.

Gordon, C. J., Tchesnokov, E. P., Feng, J. Y., Porter, D. P., and Götte, M. (2020). The antiviral compound remdesivir potently inhibits RNA-dependent RNA polymerase from Middle East respiratory syndrome coronavirus. *J. Biol. Chem.* 295, 4773–4779. doi: 10.1074/jbc.AC120.013056

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., and Caves, L. S. D. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22, 2695–2696. doi: 10.1093/bioinformatics/btl461

Grottesi, A., Domene, C., Hall, B., and Sansom, M. S. P. (2005). Conformational dynamics of M2 helices in KirBac channels: helix flexibility in relation to gating via molecular dynamics simulations. *Biochemistry* 44, 14586–14594. doi: 10.1021/bi0510429

Guex, N., and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18, 2714–2723. doi: 10.1002/elps.1150181505

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506. doi: 10.1016/S0140-6736(20)30183-5

Huber, R. (1987). Flexibility and rigidity, requirements for the function of proteins and protein pigment complexes. *Eleventh Keilin Memorial Lecture*. (Biochemical Society transactions) 15:1009–20. doi: 10.1042/bst0151009

Karshikoff, A., Nilsson, L., and Ladenstein, R. (2015). Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *FEBS J.* 282:3899–917. doi: 10.1111/febs.13343

Kim, H. J., Choi, M. Y., Kim, H. J., and Llinás, M. (2010). Conformational dynamics and ligand binding in the multi-domain protein PDC109. *PLoS ONE* 5:e9180. doi: 10.1371/journal.pone.0009180

Koonin, E. V., and Dolja, V. V. (1993). Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit. Rev. Biochem. Mol. Biol.* 28, 375–430. doi: 10.3109/104092393090 78440

Li, G., and Clercq, E. D. (2020). Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nat. Rev. Drug Discov.* 19, 149–150. doi: 10.1038/d41573-020-00016-0

Li, H., Chang, Y.-Y., Lee, J. Y., Bahar, I., and Yang, L.-W. (2017). DynOmics: dynamics of structural proteome and beyond. *Nucleic Acids Res.* 45, W374–W380. doi: 10.1093/nar/gkx385

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., et al. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* 368, 489–493. doi: 10.1126/science.abb3221

Mair-Jenkins, J., Saavedra-Campos, M., Baillie, J. K., Cleary, P., Khaw, F.-M., Lim, W. S., et al. (2015). The effectiveness of convalescent plasma and hyperimmune immunoglobulin for the treatment of severe acute respiratory infections of viral etiology: a systematic review and exploratory meta-analysis. *J. Infect. Dis.* 211, 80–90. doi: 10.1093/infdis/jiu396

Martinez, M. A. (2020). Compounds with therapeutic potential against novel respiratory 2019 Coronavirus. *Antimicrob. Agents Chemotherap.* 64, 1–7. doi: 10.1128/AAC.00399-20

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc.21256

Morse, J. S., Lalonde, T., Xu, S., and Liu, W. R. (2020). Learning from the past: possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-nCoV. *Chembiochem* 21, 730–738. doi: 10.1002/cbic.202000047

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *J. Cheminform.* 3:33. doi: 10.1186/1758-2946-3-33

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* 18, 1–9. doi: 10.1186/s12967-020-02344-6

Phan, L. T., Nguyen, T. V., Luong, Q. C., Nguyen, T. V., Nguyen, H. T., Le, H. Q., et al. (2020). Importation and human-to-human transmission of a novel coronavirus in vietnam. *N. Engl. J. Med.* 382, 872–874. doi: 10.1056/NEJMc2001272

Pillay, D., and Zambon, M. (1998). Antiviral drug resistance. *Br. Med. J.* 317, 660–662. doi: 10.1136/bmj.317.7159.660

Qureshi, A., Kaur, G., and Kumar, M. (2017). AVCpred: an integrated web server for prediction and design of antiviral compounds. *Chem. Biol. Drug Des.* 89, 74–83. doi: 10.1111/cbdd.12834

Rader, A. J., Chennubhotla, C., Yang, L.-W., and Bahar, and, I. (2005). "The gaussian network model: theory and applications," in *Normal Mode Analysis Theory and Applications to Biological and Chemical Systems* eds Cui, Q., and Bahar, I. (New York, NY: Chapman and Hall/CRC), 41–64. doi: 10.1201/9781420035070.ch3

Russell, C. D., Millar, J. E., and Baillie, J. K. (2020). Clinical evidence does not support corticosteroid treatment for 2019-nCoV lung injury. *Lancet* 395, 473–475. doi: 10.1016/S0140-6736(20)30317-2

S, U. K., R, B., D, T. K., Doss, C. G. P., and Zayed, H. (2020). Mutational landscape of K-Ras substitutions at 12th position-a systematic molecular dynamics approach. *J. Biomol. Struct. Dyn.* 120, 1–16. doi: 10.1080/07391102.2020.1830177

Scarabelli, G., and Grant, B. J. (2013). Mapping the structural and dynamical features of kinesin motor domains. *PLoS Comput. Biol.* 9:e1003329. doi: 10.1371/journal.pcbi.1003329

Sheahan, T. P., Sims, A. C., Graham, R. L., Menachery, V. D., Gralinski, L. E., Case, J. B., et al. (2017). Broad-spectrum antiviral GS-5734 inhibits both epidemic and zoonotic coronaviruses. *Sci. Transl. Med.* 9:eaal3653. doi: 10.1126/scitranslmed.aal3653

Strauss, J. H., and Strauss, E. G. (1988). Evolution of RNA viruses. *Annu. Rev. Microbiol.* 42, 657–683. doi: 10.1146/annurev.mi.42.100188.0 03301

Thiel, V., Herold, J., Schelle, B., and Siddell, S. G. (2001). Viral replicase gene products suffice for coronavirus discontinuous transcription. *J. Virol.* 75, 6676–6681. doi: 10.1128/JVI.75.14.6676-6681.2001

Udhaya Kumar, S., Thirumal Kumar, D., Mandal, P. D., Sankar, S., Haldar, R., Kamaraj, B., et al. (2020). "Chapter Eight - Comprehensive in silico screening and molecular dynamics studies of missense mutations in Sjogren-Larsson syndrome associated with the ALDH3A2 gene," in *Advances in Protein Chemistry and Structural Biology Inflammatory Disorders - Part B*, ed. R. Donev (Academic Press), 349–377. doi: 10.1016/bs.apcsb.2019.11.004

Varga, A., Lionne, C., and Roy, B. (2016). Intracellular metabolism of nucleoside/nucleotide analogues: a bottleneck to reach active drugs on HIV reverse transcriptase. *Curr. Drug Metab.* 17, 237–252. doi: 10.2174/1389200217666151210141903

Vincent, M. J., Bergeron, E., Benjannet, S., Erickson, B. R., Rollin, P. E., Ksiazek, T. G., et al. (2005). Chloroquine is a potent inhibitor of SARS coronavirus infection and spread. *Virol. J.* 2:69. doi: 10.1186/1743-422X-2-69

Wanchao, Y., Chunyou, M., Xiaodong, L., Dan-Dan, S., Qingya, S., Haixia, S., et al. (2020). Structural basis for the inhibition of the RNA-Dependent RNA polymerase from SARS- CoV-2 by Remdesivir. *bioRxiv* 1560, 1–30. doi: 10.1101/2020.04.08.032763

Wang, P., Anderson, N., Pan, Y., Poon, L., Charlton, C., Zelyas, N., et al. (2020). The SARS-CoV-2 outbreak: diagnosis, infection prevention, and public perception. *Clin. Chem.* 66:hvaa080. doi: 10.1093/clinchem/hvaa080

Warren, T. K., Jordan, R., Lo, M. K., Ray, A. S., Mackman, R. L., Soloveva, V., et al. (2016). Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature* 531, 381–385. doi: 10.1038/nature17180

WHO (2020a). Statement on the second meeting of the International Health Regulations (2005) *Emergency Committee Regarding the Outbreak of Novel Coronavirus (2019-nCoV)*. Available online at: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) (accessed November 15, 2020).

WHO (2020b). WHO *Director-General's Remarks at the Media Briefing on 2019-nCoV on 11 February 2020*. Available online at: https://www.who.int/director-general/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020 (accessed November 15, 2020).

Worldometer (2021). *Coronavirus Update (Live)*. Available online at: https://www.worldometers.info/coronavirus/ (accessed March 13, 2021).

Xu, J., Zhao, S., Teng, T., Abdalla, A. E., Zhu, W., Xie, L., et al. (2020). Systematic comparison of two animal-to-human transmitted human coronaviruses: SARS-CoV-2 and SARS-CoV. *Viruses* 12:244. doi: 10.3390/v12020244

Xu, X., Liu, Y., Weiss, S., Arnold, E., Sarafianos, S. G., and Ding, J. (2003). Molecular model of SARS coronavirus polymerase: implications for biochemical functions and drug design. *Nucleic Acids Res.* 31, 7117–7130. doi: 10.1093/nar/gkg916

Yang, L., Song, G., Carriquiry, A., and Jernigan, R. L. (2008). Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* 16, 321–330. doi: 10.1016/j.str.2007.12.011

Yao, T.-T., Qian, J.-D., Zhu, W.-Y., Wang, Y., and Wang, G.-Q. (2020). A systematic review of lopinavir therapy for SARS coronavirus and MERS coronavirus-A possible reference for coronavirus disease-19 treatment option. *J. Med. Virol.* 92, 556–563. doi: 10.1002/jmv.25729

Yao, X., Ye, F., Zhang, M., Cui, C., Huang, B., Niu, P., et al. (2020). *In vitro* antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Clin. Infect. Dis.* 71, 732–739. doi: 10.1093/cid/ciaa237

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi: 10.1056/NEJMoa2001017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 20 July 2021
doi: 10.3389/fmolb.2021.639423

A Validated Mathematical Model of the Cytokine Release Syndrome in Severe COVID-19

Ruy Freitas Reis[1,2]*, *Alexandre Bittencourt Pigozzo*[3], *Carla Rezende Barbosa Bonin*[4], *Barbara de Melo Quintela*[1,2], *Lara Turetta Pompei*[2], *Ana Carolina Vieira*[5], *Larissa de Lima e Silva*[2], *Maicom Peters Xavier*[6], *Rodrigo Weber dos Santos*[1,2,6] and *Marcelo Lobosco*[1,2,6]

[1]Institute of Exact Sciences, Department of Computing, Federal University of Juiz de Fora, Juiz de Fora, Brazil, [2]FISIOCOMP - Laboratory of Computational Fisiology and High-Performance Computing, Federal University of Juiz de Fora, Juiz de Fora, Brazil, [3]Computer Science Department, Federal University of São João Del-Rei, São João Del-Rei, Brazil, [4]Institute of Education, Science and Technology of Southeast of Minas Gerais - Cataguases Advanced Campus, Cataguases, Brazil, [5]GET-EngComp, Grupo de Educação Tutorial Engenharia Computacional, Federal University of Juiz de Fora, Juiz de Fora, Brazil, [6]Graduate Program on Computational Modeling, Federal University of Juiz de Fora, Juiz de Fora, Brazil

OPEN ACCESS

Edited by:
James Leland Olds,
George Mason University,
United States

Reviewed by:
Carlo Graziani,
Argonne National Laboratory (DOE),
United States
Ruy Ribeiro,
University of Lisbon, Portugal

***Correspondence:**
Ruy Freitas Reis
ruy.reis@ufjf.br

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *08 December 2020*
Accepted: *30 June 2021*
Published: *20 July 2021*

Citation:
Reis RF, Pigozzo AB, Bonin CRB, Quintela BM, Pompei LT, Vieira AC, Silva LL, Xavier MP, Weber dos Santos R and Lobosco M (2021) A Validated Mathematical Model of the Cytokine Release Syndrome in Severe COVID-19. Front. Mol. Biosci. 8:639423. doi: 10.3389/fmolb.2021.639423

By June 2021, a new contagious disease, the Coronavirus disease 2019 (COVID-19), has infected more than 172 million people worldwide, causing more than 3.7 million deaths. Many aspects related to the interactions of the disease's causative agent, SAR2-CoV-2, and the immune response are not well understood: the multiscale interactions among the various components of the human immune system and the pathogen are very complex. Mathematical and computational tools can help researchers to answer these open questions about the disease. In this work, we present a system of fifteen ordinary differential equations that models the immune response to SARS-CoV-2. The model is used to investigate the hypothesis that the SARS-CoV-2 infects immune cells and, for this reason, induces high-level productions of inflammatory cytokines. Simulation results support this hypothesis and further explain why survivors have lower levels of cytokines levels than non-survivors.

Keywords: computational immunology, SARS-CoV-2, COVID-19, mathematical modelling, sensitivity analysis, citokine storm

1 INTRODUCTION

By the beginning of June 2021, the number of confirmed deaths caused by the novel coronavirus pneumonia, COVID-19, surpassed 3.7 millions, while more than 172 millions worldwide were infected. The causative virus was first identified as SARS-CoV-2, also referred to as HCoV-19. It has been shown that SARS-CoV-2 infects the alveolar epithelial cells, mainly type 2 alveolar epithelial cells (AEC2), via the angiotensin-converting enzyme receptor 2 (ACE2) (Catanzaro et al., 2020; Hoffmann et al., 2020; Shang et al., 2020). The ensuing destruction of the epithelial cells and the increase in cell permeability lead to the release of the virus. During the fight against the virus, the innate immune system cells release a large number of extracellular molecular regulators, like cytokines and chemokines, that will induce the adaptive response to recruit more cells from the innate system (Coperchini et al., 2020; Prompetchara et al., 2020). In most individuals, recruited cells (mainly CD8$^+$ T cells) are sufficient to clear the infection. However, in some patients, a dysfunctional immune response occurs, which triggers a "cytokine storm" that mediates widespread inflammation and damage (Tay et al., 2020), mainly in the lung.

The Cytokine Release Syndrome (CRS) or cytokine storm has been associated with a wide variety of infectious and non-infectious diseases for the past decades, including influenza and SARS-CoV (Tisoncik et al., 2012; Shimabukuro-Vornhagen et al., 2018). Nevertheless, the exact signalling pathways that lead to the CRS are yet to be determined (Lukan, 2020). Several mechanisms have been proposed to explain the CRS and the differences between survivors and non-survivors concerning viral dynamics and the immune response to SARS-CoV-2. One hypothesis is that SARS-CoV-2 infects macrophages, CD4^{+} T cells, and CD8^{+} T cells in addition to alveolar epithelial cells (Davanzo et al., 2020; Grant et al., 2021), thus causing the production of several pro-inflammatory cytokines (mainly IL-6) and also impairments in the immune response mediated by macrophages and T cells. The infection of CD8^{+} T cells, for example, prevents those cells from killing other infected cells and induces high-level production of some inflammatory cytokines, including IL-6 (Li H. et al., 2020; Blanco-Melo et al., 2020; Chen et al., 2020).

In this work, the hypothesis above is tested by a mathematical model of the immune response to SARS-CoV-2. The model developed is an extension of a prior model of the adaptive immune response, which was validated using experimental data obtained from vaccination against yellow fever (Bonin et al., 2019). The original model considers the main cells and molecules present in the immune response, such as, for example, antigen-presenting cells, CD4^{+} and CD8^{+} T cells, B cells, and IgM and IgG antibodies. We further extended the model in this work, including pro-inflammatory cytokines and infected immune system cells. Also, the model is validated with experimental data of the viremia, antibodies (IgM and IgG), and cytokines obtained from patients with COVID-19 (Long et al., 2020; To et al., 2020; Zhou et al., 2020). Moreover, a sensitivity analysis (SA) revealed important characteristics of the immune response to SARS-CoV-2.

2 RELATED WORKS

A large number of works use mathematical and computational tools to model the Human Immune System (HIS) using distinct techniques, such as ordinary differential equations (ODEs) (Perelson, 1989; Baker et al., 1997; Chang et al., 2005; Vodovotz et al., 2006; Jarrett et al., 2015; Bonin et al., 2016), partial differential equations (PDEs) (Pettet et al., 1996; Su et al., 2009; Flegg et al., 2012; Pigozzo et al., 2013; Quintela et al., 2014), stochastic methods (Chao et al., 2004; Xavier et al., 2017), cellular automaton and agents (Celada and Seiden, 1992; Morpurgo et al., 1995; Kohler et al., 2000; Bernaschi and Castiglione, 2001; Pappalardo et al., 2018). On the other hand, very few papers were published describing the dynamics of SARS-CoV-2 (Almocera et al., 2020; Du and Yuan, 2020; Hernandez-Vargas and Velasco-Hernandez, 2020; Xavier et al., 2020), although some additional non-peer-reviewed papers can also be found on the internet. Three out four published works (Almocera et al., 2020; Du and Yuan, 2020; Hernandez-Vargas and Velasco-Hernandez, 2020) are based on the target cell-limited model proposed by Perelson (2002): a system of three ODEs to model target cells, infected cells, and viruses. In this work, we use a set of fifteen ODEs to model not only the virus but also immune cells, antibodies and cytokines.

Du and Yuan (2020) developed a mathematical model to investigate the dynamics of the immune response to influenza and the SARS-CoV-2 virus, including in their analysis the effects of a hypothetical antiviral drug on the SARS-Cov-2 infection. The model uses constants to represent the adaptive system CD8^{+} cells, IgM and IgG antibodies. The authors argue, based on their numerical results, that the innate immune system is the main responsible for clearing the influenza virus, while the adaptive system is the main responsible for controlling the SARS-CoV-2 virus. Their numerical results also suggest that the peak concentration of the adaptive immune cells for patients with COVID-19 is more likely to occur before the number of infected cells by SAR-CoV-2 reaches its peak (Du and Yuan, 2020). However, these results have not been validated: the viral loads found were not compared to experimental results.

A model similar to the one of Du and Yuan (2020) was proposed by Hernandez-Vargas and Velasco-Hernandez (2020), but including latent cells. The idea is that newly infected cells spend time in a latent phase, a concept similar to the "Eclipse phase" (Beauchemin et al., 2008). Another difference is that instead of using a constant to represent T cells (Du and Yuan, 2020), Hernandez-Vargas and Velasco-Hernandez (2020) used an equation to represent them. The viral loads obtained by numerical experiments were compared to values found in the literature (Wölfel et al., 2020), with good fitness between numerical and experimental results. The authors then present the Stability Analysis of their model (Almocera et al., 2020), which suggests that the SARS-CoV-2 virus replicates fast enough to overcome T cell response and cause infection.

Xavier et al. (2020) proposed a model based on five Ordinary Differential Equations to model the immune response to SARS-CoV-2. The model parameters and initial conditions were adjusted to cohort studies that collected viremia and antibody data. The results have shown that the model was able to reproduce both viremia and antibodies dynamics successfully. However, the model does not take into account the CRS. The mathematical model proposed in this paper can describe the immune response to SARS-CoV-2 for survivors and non-survivors. The difference between the two scenarios is the cytokine storm, which leads to a deregulated immune response in the second group.

A nonlinear differential equation model was proposed by Waito et al. (2016) to represent the dynamics of the CRS. The authors consider in their model that the rate of production of cytokines is dependent on interactions with other cytokines. The model was adjusted using type I interferon (IFN) receptor (IFNAR)-knockout mice data since mice lacking the IFNAR on their leukocytes experienced a profound cytokine storm (Waito et al., 2016). Although thirteen cytokines were considered in their model, the computational results have

TABLE 1 | Major differences between the models proposed in a previous work (Bonin et al., 2019) and in this work.

	Bonin et al. (2019)	This work
Number of equations	12	15
Number of parameters	32	38
Number of compartments	1	1
Number of considered populations	9	12

shown that TNF-α, IL-10, IL-6, and MIP-1β, have the largest effects on the dynamics of the cytokine storm. In this work, we do not distinguish the cytokines types nor consider the interactions among cytokines in their production rate.

The mathematical model proposed in this paper is an extension of a prior model of the adaptive immune response (Bonin et al., 2019). The original model (Bonin et al., 2019) had their parameters adjusted to reproduce the immune response against the Yellow Fever vaccine. The model of Bonin et al. (2019) considers the main cells and molecules present in the immune response against a virus such as antigen-presenting cells, CD4$^+$ and CD8$^+$ T cells, B cells and antibodies.

In order to represent the immune response to SARS-CoV-2, the model presented in this paper has slight differences from the model shown in our previous work (Bonin et al., 2019), and these differences are summarised in **Table 1**. Some changes were necessary to represent the hypothesis that some immune system cells can be infected by the SARS-CoV-2 virus (Davanzo et al., 2020; Grant et al., 2021). To implement these changes, a new population was included in the model (I, to represent immune defence cells infected by SARS-CoV-2) as well as new terms were included in the defence cells to represent their infection. Furthermore, the production of pro-inflammatory citokines was introduced in order to represent the dynamics of the CRS. To achieve this purpose, a new population was included (C, to represent the cytokine), as well as new terms were included to represent their production by the immune cells. Finally, we differentiate the production of antibodies into I_gM and I_gG. These changes will be presented in the next section.

3 MATERIALS AND METHODS

3.1 Mathematical Model

The model proposed in this work consists of a set of 15 Ordinary Differential Equations (ODEs), one to represent the behaviour of each population: virus (V), naive (A_p) and mature (A_{pm}) antigen-presenting cells (APCs), immune cells infected by the SARS-CoV-2 virus (I), naive (T_{hn}) and effector (T_{he}) T helper (CD4$^+$) cells, naive (T_{kn}) and effector (T_{ke}) T Killer (CD8$^+$) cells, B cells (B), short- (P_s) and long-lived (P_l) plasma cells, B memory cells (B_m), IgM (Ig_M) and IgG (Ig_G) antibodies and cytokines (C) (**Figure 1**).

The first equation describes the virus (V) behaviour. SARS-CoV-2 uses the cell surface receptor ACE2 to infect healthy cells (Catanzaro et al., 2020; Hoffmann et al., 2020; Shang et al., 2020), using its machinery to replicate itself. The replication is

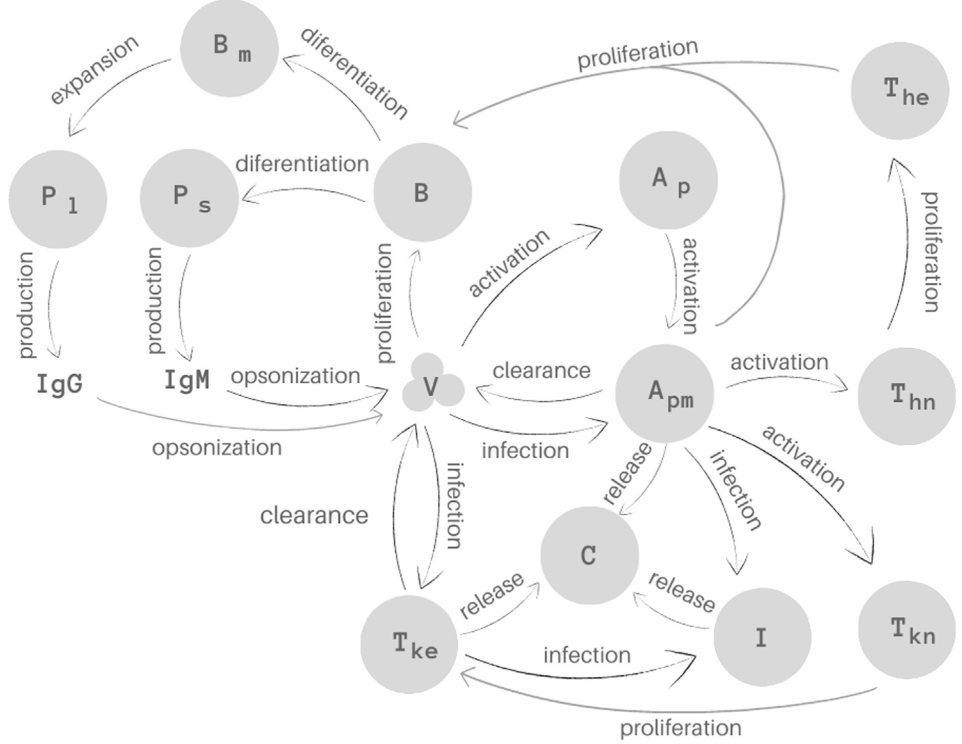


FIGURE 1 | Representation of the 15 biological species and their interactions as considered in the mathematical model. Homeostasis is not illustrated in this Figure.

represented implicitly by the first term of **Eq. 1**, $\pi_v V$, where π_v represents virus increase rate. The remaining terms of **Eq. 1** represent the elimination of the virus by the immune system. The virus can be opsonized by antibodies, which facilitate its binding to receptor molecules present on the phagocytes (Paul, 2008). This is illustrated by the second and third term of **Eq. 1**. Effector CD8$^+$ T-cells kills cells that are infected with viruses (Sompayrac, 2012). The term $k_{v1} VI_{gG}$ represents the elimination of virus due to the opsonization by I_{gG} and the term $k_{v1} VI_{gM}$ represents the elimination of virus due to the opsonization by I_{gM}, where k_{v1} is the opsonization rate. The term $k_{v2} VT_{ke}$ denotes specific viral clearance due to the induction of apoptosis of cells infected by the SARS-CoV-2 virus, where k_{v2} is the clearance rate. Finally, $k_{v3} VA_{pm}$ denotes specific viral clearance due to phagocytosis by mature APCs, such as macrophages, where k_{v3} is the clearance rate.

$$\frac{d}{dt}V = \pi_v V - k_{v1} VI_{gG} - k_{v1} VI_{gM} - k_{v2} VT_{ke} - k_{v3} VA_{pm}. \tag{1}$$

APCs are found in two stages, naive and mature (Murphy and Weaver, 2008). The second and third equations represent these two stages of the APCs, naive (A_p) and mature (A_{pm}). In this work, we consider that macrophages are the main APCs. In **Eq. 2**, the naive APCs homeostasis and activation are described by the first and second terms, respectively. The first term, $\alpha_{ap}(C+1)(A_{p0} - A_p)$, α_{ap} represents the homeostasis rate. Pro-inflammatory cytokines influence the homeostatic balance of the APCs (Murphy and Weaver, 2008). The term $\beta_{ap} A_p \frac{c_{ap1} V}{c_{ap2}+V}$ denotes the conversion of immature APCs into mature ones and, for this reason, the same term appears in **Eq. 3** with positive sign. This function models growth combined with the saturation phenomenon (Goutelle et al., 2008).

$$\frac{d}{dt}A_p = \alpha_{ap}(C+1)\left(A_{p0} - A_p\right) - \beta_{ap} A_p \frac{c_{ap1} V}{c_{ap2} + V}. \tag{2}$$

In **Eq. 3**, which represents mature APCs, $\beta_{apm} A_{pm} V$ denotes A_{pm} infection by the SARS-CoV-2 virus where β_{apm} is the infection rate. The third term, $\delta_{apm} A_{pm}$, means the natural decay of the mature APCs, where δ_{apm} is the decay rate.

$$\frac{d}{dt}A_{pm} = \beta_{ap} A_p \frac{c_{ap1} V}{c_{ap2} + V} - \beta_{apm} A_{pm} V - \delta_{apm} A_{pm}. \tag{3}$$

The dynamics of the infected immune system cells is represented by **Eq. 4**. The first term, $\beta_{apm} A_{pm} V$, represents A_{pm} infection and the second term, $\beta_{tke} T_{ke} V$, represents CD8$^+$ T cells infection. The infection rates are, respectively, β_{apm} and β_{tk}. Infected cells die with a rate δ_{apm}.

$$\frac{d}{dt}I = \beta_{apm} A_{pm} V + \beta_{tke} T_{ke} V - \delta_{apm} I. \tag{4}$$

Equation 5 represents the population of naive CD4$^+$ T cells (T_{hn}). The term $\alpha_{th}(T_{hn0} - T_{hn})$ represents the homeostasis of CD4$^+$ T cells, where α_{th} is the homeostasis rate. APCs are responsible for activating naive CD4$^+$ T cells (Murphy and Weaver, 2008). The term $\beta_{th} A_{pm} T_{hn}$ denotes the activation of naive CD4$^+$ T cells, where β_{th} is the activation rate.

$$\frac{d}{dt}T_{hn} = \alpha_{th}(T_{hn0} - T_{hn}) - \beta_{th} A_{pm} T_{hn}. \tag{5}$$

Equation 6 represents effector CD4$^+$ T cell population (T_{he}). The term $\pi_{th} A_{pm} T_{he}$ represents the proliferation of effector CD4$^+$ T cells, where π_{th} is the proliferation rate. The term $\delta_{th} T_{he}$ represents the natural death of these cells, with δ_{th} representing its death rate.

$$\frac{d}{dt}T_{he} = \beta_{th} A_{pm} T_{hn} + \pi_{th} A_{pm} T_{he} - \delta_{th} T_{he}. \tag{6}$$

Equations 7, 8 represent the population of naive (T_{kn}) and effector (T_{ke}) CD8$^+$ T cells, respectively. In **Eq. 7**, the naive CD8$^+$ T cells homeostasis and activation are described by the first and second terms, respectively. In the first term, $\alpha_{tk}(C+1)(T_{kn0} - T_{kn})$, α_{tk} represents the homeostasis rate. The term $\beta_{tk}(C+1)A_{pm} T_{kn}$ denotes the activation of naive CD8$^+$ T cells, where β_{tk} is the activation rate. Pro-inflammatory cytokines have an influence on both the homeostatic balance and activation of naive CD8$^+$ T cells.

$$\frac{d}{dt}T_{kn} = \alpha_{tk}(C+1)(T_{kn0} - T_{kn}) - \beta_{tk}(C+1)A_{pm} T_{kn}. \tag{7}$$

In **Eq. 8**, the term $\pi_{tk} A_{pm} T_{ke}$ represents the proliferation of effector CD8$^+$ T cells. The terms $\beta_{tke} T_{ke} V$ and $\delta_{tk} T_{ke}$ represent the infection and death of effector CD8$^+$ T cells, respectively.

$$\frac{d}{dt}T_{ke} = \beta_{tk}(C+1)A_{pm} T_{kn} + \pi_{tk} A_{pm} T_{ke} - \beta_{tke} T_{ke} V - \delta_{tk} T_{ke}. \tag{8}$$

Equation 9 represents both naive and effector B cells (B). These populations were not considered separately in order to simplify the model. The term $\alpha_b(B_0 - B)$ represents the B cells homeostasis, where α_b is the homeostasis rate. The terms $\pi_{b1} VB$ and $\pi_{b2} T_{he} B$ represent the proliferation of B cells activated by the T-cell independent and T-cell dependent mechanisms (Sompayrac, 2012), respectively. The terms $\beta_{ps} A_{pm} B$, $\beta_{pl} T_{he} B$ and $\beta_{bm} T_{he} B$ denote the differentiation of active B cells into short-lived plasma cells, long-lived plasma cells and memory B cells, respectively. The activation rates are respectively given by β_{ps}, β_{pl} and β_{bm}.

$$\begin{aligned}\frac{d}{dt}B = {} & \alpha_b(B_0 - B) + \pi_{b1} VB + \pi_{b2} T_{he} B - \beta_{ps} A_{pm} B \\ & -\beta_{pl} T_{he} B - \beta_{bm} T_{he} B.\end{aligned} \tag{9}$$

Equation 10 represents the short-lived plasma cells (P_s) (Sompayrac, 2012). The term $\delta_{ps} P_s$ denotes the natural decay of short-lived plasma cells, where δ_{ps} is the decay rate.

$$\frac{d}{dt}P_s = \beta_{ps} A_{pm} B - \delta_{ps} P_s. \tag{10}$$

Equation 11 represents the long-lived plasma cells (P_l). The term $\delta_{pl} P_l$ denotes the natural decay of long-lived plasma cells, with δ_{pl} representing the decay rate. The term $\gamma_{bm} B_m$ represents the resupply of these cells by memory B cells, where γ_{bm} is the production rate.

$$\frac{d}{dt}P_l = \beta_{pl} T_{he} B - \delta_{pl} P_l + \gamma_{bm} B_m. \tag{11}$$

Memory B cells (B_m) dynamics is represented by **Eq. 12**. The term $\pi_{bm1} B_m \left(1 - \frac{B_m}{\pi_{bm2}}\right)$ represents the logistic growth of memory B cells, *i.e.*, there is a limit to this growth (Bonin et al., 2019). π_{bm1} represents the growth rate, and π_{bm2} limits the growth.

$$\frac{d}{dt}B_m = \beta_{bm} T_{he} B + \pi_{bm1} B_m \left(1 - \frac{B_m}{\pi_{bm2}}\right) - \gamma_{bm} B_m. \tag{12}$$

Equations 13, **14** represents the production of antibodies. The terms $\pi_{ps} P_s$ and $\pi_{pl} P_l$ are the production of the antibodies by the short-lived and long-lived plasma cells, respectively. The production rates are given by π_{ps} and π_{pl}, respectively. The terms $\delta_{am} I_{gM}$ and $\delta_{ag} I_{gG}$ denotes the natural decay of IgM and IgG antibodies, respectively, where δ_{am} and δ_{ag} are the decay rate.

$$\frac{d}{dt}I_{gM} = \pi_{ps} P_s - \delta_{am} I_{gM}. \tag{13}$$

$$\frac{d}{dt}I_{gG} = \pi_{pl} P_l - \delta_{ag} I_{gG}. \tag{14}$$

Finally, **Eq. 15** describes the pro-inflammatory cytokine dynamics. In this equation, the first term, $\pi_{c_{apm}} A_{pm}$, represents the production of cytokines by A_{pm}, where $\pi_{c_{apm}}$ is the production rate. The second term represents the production of cytokines by infected cells, where π_{ci} is the production rate. The third term represents the production of cytokines by T_{ke} cells, where $\pi_{c_{tke}}$ is the production rate. Finally, the last term, $\delta_c C$, represents the cytokine natural decay, where δ_c is the decay rate.

$$\frac{d}{dt}C = \pi_{c_{apm}} A_{pm} + \pi_{ci} I + \pi_{c_{tke}} T_{ke} - \delta_c C. \tag{15}$$

We did not include in the model infected and non-infected epithelial cells because the model would be more complex, with more constants to adjust and, the worst, without data available to validate these cell populations along time. The use of implicit antigen replication does not affect the quality of the result, more specifically those related to the virus population, as our previous works have demonstrated (Bonin et al., 2018; Pigozzo et al., 2018; Bonin et al., 2019; Reis et al., 2019).

We are assuming that only mature cells are infected by the SARS-CoV-2 virus. In our simplification, we assume that the virus is located in the tissue, and that most of the naive cells are activated either in (or just after leaving) the bloodstream (APCs) or in the lymph nodes (CD4 and CD8) (Murphy and Weaver, 2008; Paul, 2008; Sompayrac, 2012). Another simplification adopted in this paper is that infected cells do not produce new virions: we assume that virus is mainly produced by the epithelial tissue despite the evidence that infected alveolar macrophages may support viral replication (Grant et al., 2021). We also assume that the phagocytic activity by infected cells does not occur. Finally, we assume that infected cells continue to produce pro-inflammatory cytokines and we implicitly consider the effects of different pro-inflammatory cytokines, the main effects of which are related to the IL-6 cytokine (Tanaka et al., 2014; Narazaki and Kishimoto, 2018; Zhang et al., 2020).

3.2 Experimental Data

Cohort studies available in the literature with people infected with SARS-CoV-2 were used to evaluate the mathematical model's performance. In particular, viremia, antibodies (IgG and IgM), and cytokine levels for patients that survived and died due to COVID-19 were collected from three different papers.

The first paper presents a temporal profile of serial viral load from a set of 23 patients admitted at two hospitals in Hong Kong, all of them with laboratory-confirmed COVID-19 cases (To et al., 2020). Most of the viral load data reported in the paper were collected daily for 29°days from posterior oropharyngeal saliva samples. For this reason, we have used only this subset from this first paper. The number of patients who provided a sample on each day varies from one to ten. Data were extracted from the paper using the WebPlotDigitalizer tool (Rohatgi, 2020). WebPlotDigitizer is an on-line tool that can extract data in a semi-automatic way from graphics uploaded to the website.

The second paper presents antibody responses to SARS-CoV-2 (Long et al., 2020). A cohort composed of 285 Chinese patients confirmed to be infected with SARS-CoV-2 by RT-PCR assays were enrolled in this study from three hospitals. To measure the level of IgG and IgM against SARS-CoV-2, serum samples were collected at four different time intervals after the reported symptoms onset (Long et al., 2020). Antibody levels were measured using magnetic chemiluminescence, which provides values divided by the cutoff (S/CO) (Long et al., 2020), and were presented as $log_2(S/CO+1)$. The number of patients who provided a sample on each time interval varies from seven to one hundred and thirty. The dataset is available for download.

The third paper presents a retrospective cohort study with 191 adult inpatients from two Chinese hospitals (Zhou et al., 2020) diagnosed with COVID-19 according to WHO interim guidance and a clear outcome (dead or discharged) at the early stage of the outbreak. According to Zhou et al. (2020), 54 patients died during the stay in the hospital (non-survivors) while 137 were discharged (survivors). IL-6 plays a central role in the cytokine storm. For this reason, IL-6 data from this paper were collected using the WebPlotDigitalizer tool and used to adjust the cytokine levels of our mathematical model.

In all these three papers, the reported temporal profile of viremia, antibodies, and cytokines starts after symptom onset (Long et al., 2020; To et al., 2020; Zhou et al., 2020). This imposes an additional challenge because the exact day each patient has been infected is not clear. Epidemiological studies carried with 425 laboratory-confirmed COVID-19 cases in Wuhan, China, have estimated that the mean incubation period is about 5.2°days (Li Q. et al., 2020). Based on these results, we adjusted the cohort data accordingly to reflect the incubation period, *i.e.*, we have added 5°days at the beginning of each dataset to represent the time between the estimated infection until the symptom onset.

3.3 Parameter Estimation

One of the challenges related to the set of equations proposed to describe the immune response against the SARS-CoV-2 virus is

TABLE 2 | Bounds used for parameters calibration of the COVID-19 model for survivors. V_0 is the initial condition of the virus.

Parameter	Interval
V_0	$[1.0, 1.0 \times 10^2]$
π_v	$[8.0 \times 10^{-1}, 1.5 \times 10^0]$
k_{v1}	$[1.0 \times 10^{-5}, 1.0 \times 10^{-2}]$
k_{v2}	$[1.0 \times 10^{-7}, 1.0 \times 10^{-4}]$
k_{v3}	$[1.0 \times 10^{-4}, 5.0 \times 10^{-1}]$
β_{ap}	$[1.0 \times 10^{-5}, 1.0 \times 10^0]$
β_{apm}	$[1.0 \times 10^{-3}, 1.0 \times 10^0]$
β_{tke}	$[1.0 \times 10^{-6}, 1.0 \times 10^{-4}]$
$\pi_{c_{apm}}$	$[1.0, 5.0 \times 10^2]$
π_{ci}	$[5.0 \times 10^{-3}, 1.0 \times 10^{-1}]$
$\pi_{c_{tke}}$	$[1.0 \times 10^{-5}, 1.0 \times 10^{-1}]$
δ_c	$[1.0 \times 10^1, 1.0 \times 10^3]$

TABLE 3 | Bounds used for parameters calibration of the COVID-19 model for non-survivors.

Parameter	Interval
β_{apm}	$[1.0 \times 10^{-3}, 1.0 \times 10^0]$
β_{tke}	$[1.0 \times 10^{-6}, 1.0 \times 10^{-4}]$
π_{ci}	$[5.0 \times 10^{-3}, 1.0 \times 10^{-1}]$

TABLE 4 | Model variables and their initial values.

Variable	Description	Initial value	Unit
V	Virus	61	Copies/mL
A_p	Immature APCs	10^6	Cells/mL
A_{pm}	Mature APCs	0	Cells/mL
I	Infected cells	0	Cells/mL
T_{hn}	Naive CD4$^+$ T cells	10^6	Cells/mL
T_{he}	Effector CD4$^+$ T cells	0	Cells/mL
T_{kn}	Naive CD8$^+$ T cells	5×10^5	Cells/mL
T_{ke}	Effector CD8$^+$ T cells	0	Cells/mL
B	B Cells	2.5×10^5	Cells/mL
P_s	Short-lived plasma cells	0	Cells/mL
P_l	Long-lived plasma cells	0	Cells/mL
B_m	Memory B cells	0	Cells/mL
I_{gM}	Antibodies I_{gM}	0	S/CO
I_{gG}	Antibodies I_{gG}	0	S/CO
C	Cytokines	0	pg/mL

TABLE 5 | Model parameters values that fit survivors' data. The parameter presented in **Table 2** were estimated using DE. The other values were based on the work of Bonin et al. (2019).

Parameter	Unit	Value
π_v	(day^{-1})	1.47×10^0
k_{v1}	(day^{-1} (mIU/ml)$^{-1}$)	9.82×10^{-3}
k_{v2}	(day^{-1} (cells/mL)$^{-1}$)	6.10×10^{-5}
k_{v3}	(day^{-1} (cells/mL)$^{-1}$)	6.45×10^{-2}
α_{ap}	(day^{-1}(pg/mL)$^{-1}$)	1.0×10^0
β_{ap}	(day^{-1}(copies/mL)$^{-1}$)	1.79×10^{-1}
c_{ap1}	(Copies/mL)	8.0×10^0
c_{ap2}	(Copies/mL)	8.08×10^6
δ_{apm}	(day^{-1})	4.0×10^{-2}
β_{apm}	(day^{-1}(copies/mL)$^{-1}$)	1.33×10^{-2}
β_{tke}	(day^{-1}(copies/mL)$^{-1}$)	3.5×10^{-6}
α_{th}	(day^{-1})	2.17×10^{-4}
β_{th}	(day^{-1} (cells/mL)$^{-1}$)	1.8×10^{-5}
π_{th}	(day^{-1} (cells/mL)$^{-1}$)	1.0×10^{-8}
δ_{th}	(day^{-1})	3.0×10^{-1}
α_{tk}	(day^{-1}(pg/mL)$^{-1}$)	1.0×10^0
β_{tk}	(day^{-1}(pg/mL) $^{-1}$(cell/mL)$^{-1}$)	1.43×10^{-5}
π_{tk}	(day^{-1}(cells/mL)$^{-1}$)	1.0×10^{-8}
δ_{tk}	(day^{-1})	3×10^{-1}
α_b	(day^{-1})	3.58×10^2
π_{b1}	(day^{-1}(copies/mL)$^{-1}$)	8.98×10^{-5}
π_{b2}	(day^{-1}(cells/mL)$^{-1}$)	1.27×10^{-8}
β_{ps}	(day $^{-1}$(cells/mL)$^{-1}$)	6.0×10^{-6}
β_{pl}	(day^{-1}(cells/mL)$^{-1}$)	5.0×10^{-6}
β_{bm}	(day^{-1}(cells/mL)$^{-1}$)	1.0×10^{-6}
δ_{ps}	(day^{-1})	2.5×10^0
δ_{pl}	(day^{-1})	3.5×10^{-1}
γ_{bm}	(day^{-1})	9.75×10^{-4}
π_{bm1}	(day^{-1})	1.0×10^{-5}
π_{bm2}	(Cells/mL)	2.5×10^3
π_{ps}	(day^{-1}(cells/mL)$^{-1}$(S/CO))	8.7×10^{-2}
π_{pl}	(day^{-1}(cells/mL)$^{-1}$(S/CO))	1.0×10^{-3}
δ_{am}	(day^{-1})	7.0×10^{-2}
δ_{ag}	(day^{-1})	7.0×10^{-2}
$\pi_{c_{apm}}$	(day^{-1}(cells/mL)$^{-1}$(pg/ml))	3.28×10^2
π_{ci}	(day^{-1}(cells/mL)$^{-1}$(pg/ml))	6.44×10^{-3}
$\pi_{c_{tke}}$	(day^{-1}(cells/mL)$^{-1}$(pg/ml))	1.78×10^{-2}
δ_c	(day^{-1})	7.04×10^2

the estimation of their parameters, i.e., find the values of the parameters which give the best fit to the set of the cohort studies. Unfortunately most of the parameters cannot be measured directly by experiments, and for this reason their biological ranges are unknown. Many distinct methods for non-linear systems can be used to estimate their values (Varah, 1982; Rodriguez-Fernandez et al., 2006; Schwaab et al., 2008). In this work we adopt the differential evolution (DE) optimisation method (Storn and Price, 1997).

Both initial conditions for the 15 variables and the 38 model parameters were calibrated to data obtained from cohort studies using DE. The DE was used to estimate each of the model parameters presented in **Table 2**, respecting the limits established for each one of them as presented in **Tables 2** and **3**.

The idea used in the fitting is quite simple: minimise the difference between the model curves (viremia, IgG, IgM, and cytokines) to the given data (relative error) accordingly to **Eq. 16**.

$$\min_p \left(\omega_1 R_E(V, \widehat{V}) + \omega_2 R_E(C, \widehat{C}) + \omega_3 R_E(IgG, I\widehat{g}G) + \omega_4 R_E(IgM, I\widehat{g}M)\right) \tag{16}$$

where $\widehat{V}(t)$ is the mean viral load, $I\widehat{g}G(t)$ is the mean IgG antibody level, $I\widehat{g}M(t)$ is the mean IgM antibody level, and $\widehat{C}(t)$ represents the mean IL-6 level, p is the set of parameters to be estimated and ω_n is a weight. For this work, we used $\omega_1 = \omega_2 = 1.0$ and $\omega_3 = \omega_4 = 0.1$. The weights values were determined by the number of points in each dataset. For datasets with a small number of points, we used a small weight. This is due to the fact that, for small datasets, a small distance between the experimental data and the estimated value has a huge impact on error. R_E

TABLE 6 | Model parameters values that fit non-survivors data. All other model parameters values were those from **Table 5**.

Parameter	Unit	Value
β_{apm}	$(\text{day}^{-1}(\text{copies/mL})^{-1})$	1.51×10^{-2}
β_{tke}	$(\text{day}^{-1}(\text{copies/mL})^{-1})$	1.0×10^{-6}
π_{ci}	$(\text{day}^{-1}(\text{cells/mL})^{-1}(\text{pg/ml}))$	9.96×10^{-2}

represents the relative error between cohort data and the numerical result ($R_E(\lambda, \hat{\lambda})$), and is given by the two-norm of λ and $\hat{\lambda}$ divided by $\hat{\lambda}$.

The initial conditions for each variable are presented in **Table 4**. **Table 5** presents the complete set of values used to calibrate our model to survivors' data, including those found by the DE for the parameters listed in **Table 2**. The values that fit the model to non-survivors data are the same, except for those presented in **Table 6**, which were also obtained by the DE using the bounds presented in **Table 3**.

3.4 Sensitivity Analysis

A sensitivity analysis (SA) was performed via main Sobol indices (Sobol, 2001). The SA is used to quantify the contribution of each uncertain model input p_i. Thus, Sobol indices support the process of identifying the parameters of the model that most affect the outputs, Y, predicted by the model. The main indices S_m^i shows the portion of the total variance in Y that could be reduced if the exact value of p_i is known, and it is computed as follows:

$$S_m^i = \frac{\mathbb{V}[\mathbb{E}[Y|p_i]]}{\mathbb{V}[Y]}, \text{ for } i \in [1, N], \tag{17}$$

where N is the number of parameters, $\mathbb{V}$ is the variance, and $\mathbb{E}$ the expected value. Therefore, a high value of S_m^i indicates that the outputs of the models are more sensitive to p_i.

The sensitivity analysis was performed considering all parameters of the model. So, the main Sobol indices were evaluated considering all model parameter as uniform distributions, considering perturbations of 10% around the adjusted value, *i.e.* for a given model parameter value v_i was built a uniform distribution ranging from $[0.9v_i, 1.1v_i]$.

3.5 Implementation

The model was implemented in the Python programming language. Numerical solution of the system of ODEs performed by the solve_ivp function, a member of the integrate package in the scipy library (Scipy, 2021). Among the integrate methods offered by this function it was used the Radau option, *i.e.* the fifth-order implicit Runge-Kutta method of the Radau IIA family. DE was implemented using the differential_evolution method available in the package optimize from the scipy. The SA was executed aided by SALib library (Herman and Usher, 2017).

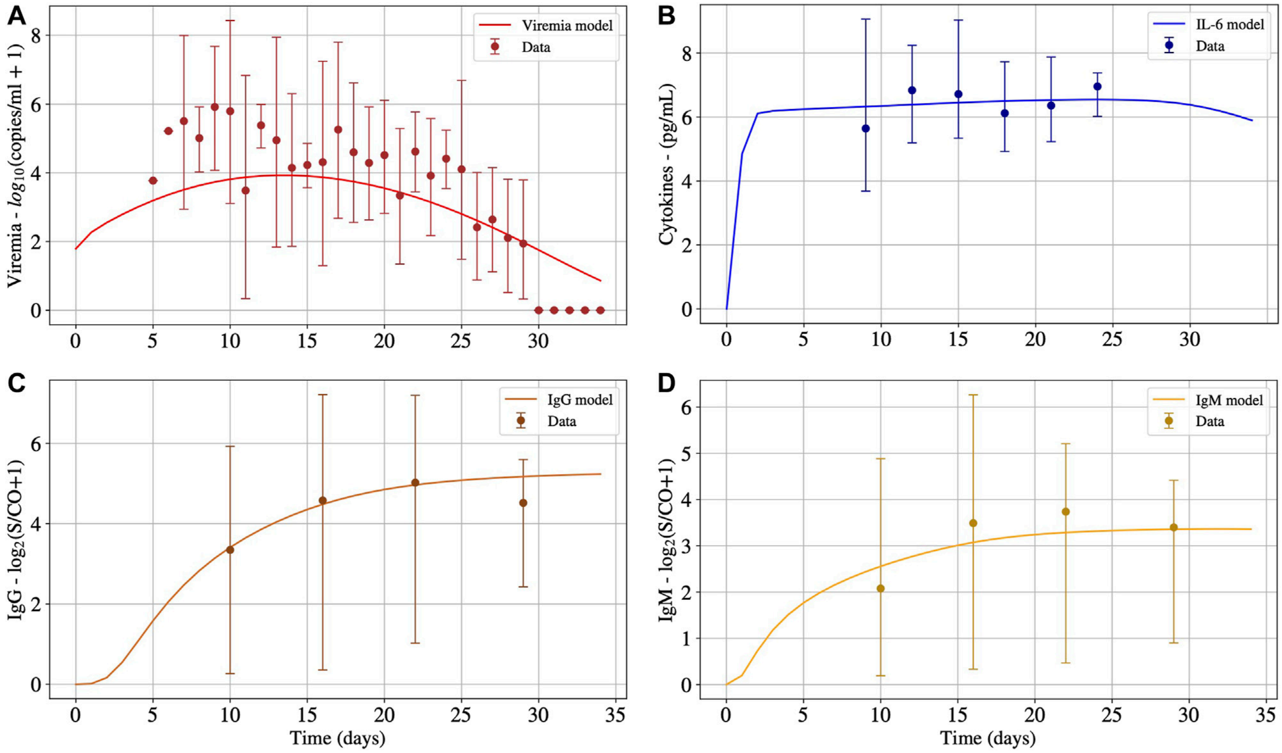


FIGURE 2 | The numerical results presented in Panels A, B, C and D represent the solution of **Eqs 1**, **15**, **14** and **13**, respectively. These numerical solutions were fitted for patients without CRS. Interval used for initial conditions and parameters are given in **Table 2**.

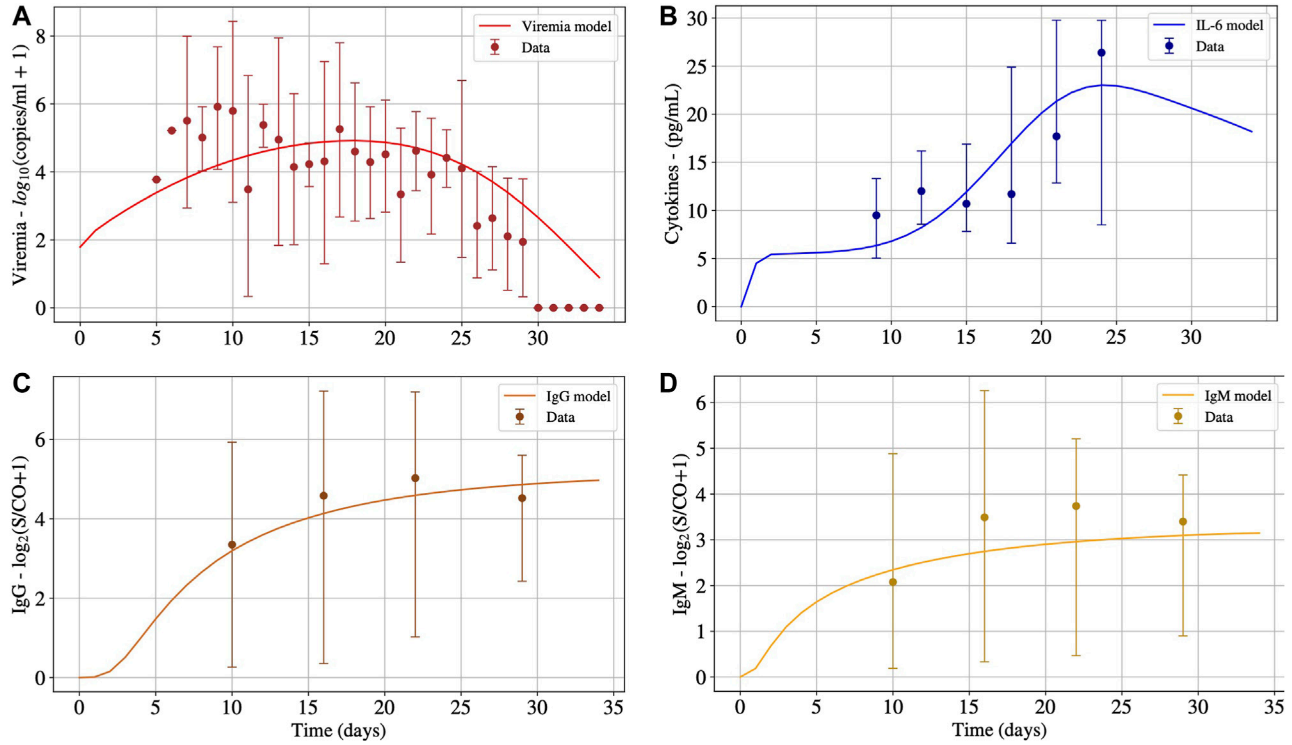


FIGURE 3 | The numerical results presented in Panels A, B, C and D represent the solution of **Eqs 1**, **15**, **14** and **13**, respectively. These numerical solutions were fitted for patients with CRS adjusting the values of β_{apm}, β_{tke} and π_{ci}, and keeping all other parameters and initial conditions values found in the previous adjust. Interval of the parameters are given in **Table 3**.

4 RESULTS

This section presents the predictions of the mathematical model presented in Section Materials and methods, comparing them with experimental data (Long et al., 2020; To et al., 2020; Zhou et al., 2020). Since one of the papers present cytokines levels for patients that survived and died due to COVID-19 (Zhou et al., 2020), we decided to divide our numerical simulations into two distinct scenarios: survivors and non-survivors.

This section also presents the results of the SA, identifying the ten parameters that most affect the outputs of the model.

First Scenario: Survivors

Initially, as a validation step, we calibrated the model using data from patients that survived COVID-19 to check if the proposed model can fit the available cohort data. The model was able to represent viremia, cytokines, IgG and IgM from the patient data without CRS (**Figure 2**).

Second Scenario: Non-survivors

The second scenario simulates the immune response of non-survivors patients. **Figure 3** presents the results. For this scenario, most of the parameters obtained in the first calibration were kept, except for π_{ci}, β_{apm} and β_{tke}. We choose to modify only these three parameters because they are directly related to the hypothesis that SARS-CoV-2 infects effector APCs and T cells, causing the production of pro-inflammatory cytokines in a distinct rate of non-infected cells. More specifically, β_{apm} and β_{tke} represent the rate in which the SARS-CoV-2 infects effector APCs and CD8$^+$ T cells, respectively, and π_{ci} represents the rate in with infected immune cells produce pro-inflammatory cytokines. The new values for these three parameters were found after a new calibration using the DE optimisation method and are presented in **Table 6**.

Figure 4 presents the impacts of β_{apm} and β_{tke} in the cytokine levels. The idea is to evaluate each one separately: first, we consider that β_{apm} is equal to zero, *i.e.*, the SARS-CoV-2 is not able to infect APCs. In this case, only CD8$^+$ T cell can be infected and induce the production of large amounts of cytokines that can start the CRS. Then we do the opposite: we consider that β_{tke} is equal to zero, *i.e.*, the SARS-CoV-2 is not able to infect CD8$^+$ T cells. In this case, only APCs can be infected. Therefore, the model was re-fitted twice: in the first case, all model parameters were adjusted again, except by β_{apm}, which was set to zero. In the second case we did the same, but this time we considered that β_{tke} is equal to zero. The results of both evaluations are then compared to the results obtained when both cells can be infected. **Figure 4** presents the impacts of these changes for viremia, cytokines, IgG and IgM. **Table 7** presents the new values of some model parameters to fit non-survivors data when β_{apm} or β_{tke} are equal to zero.

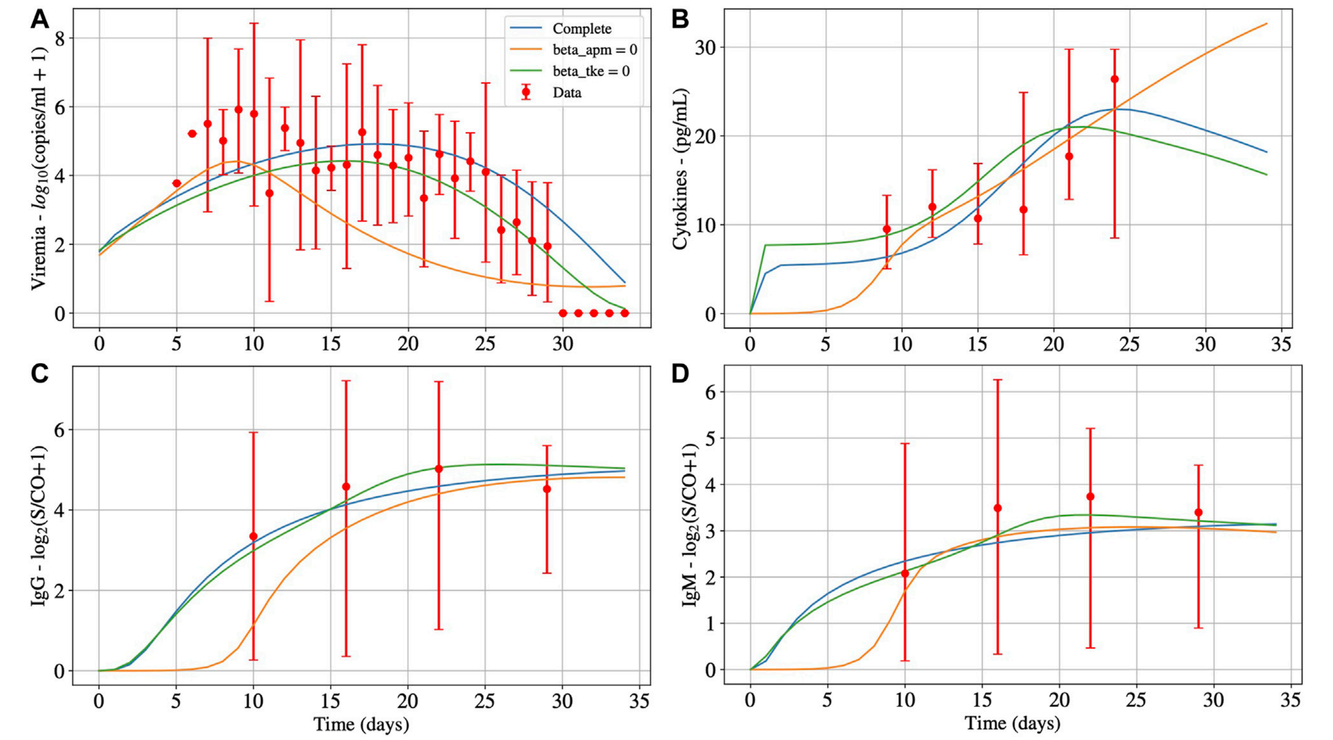


FIGURE 4 | Deterministic numerical solution of **Eqs 1**, **15**, **14** and **13** for the non-survivors case considering three hypothesis: 1) Infection of both effector APCs and CD8$^+$ T cells (complete, with $\beta_{apm} \neq 0$ and $\beta_{tke} \neq 0$); 2) Infection of CD8$^+$ T cells only (β_{apm} = 0); and 3) Infection of effector APCs only (β_{tke} = 0). Panels A, B, C and D present the results for viremia, cytokines, IgG and IgM, respectively.

TABLE 7 | Model parameters values that fit non-survivors' data, when considering that only APCs (β_{tke} = 0) or only T CD8$^+$ cells (β_{apm} = 0) are infected by SARS-CoV-2. The parameter presented in **Table 2** were re-estimated using DE. All other model parameters values were those from **Table 5**.

Parameter	$\beta_{\text{tke}} = 0$	$\beta_{\text{apm}} = 0$
V	67	48
β_{apm}	9.24×10^{-2}	0.0
β_{tke}	0.0	8.0×10^{-5}
β_{ap}	8.15×10^{-1}	1.52×10^{-4}
$\pi_{c_{apm}}$	3.67×10^{2}	4.81×10^{1}
π_{ci}	3.53×10^{-2}	2.03×10^{-2}
$\pi_{c_{tke}}$	1.55×10^{-2}	8.88×10^{-2}
π_{v}	9.70×10^{-1}	8.80×10^{-1}
k_{v1}	3.61×10^{-4}	3.04×10^{-3}
k_{v2}	8.00×10^{-5}	2.23×10^{-7}
k_{v3}	3.18×10^{-2}	1.03×10^{-1}
δ_c	4.15×10^{2}	8.25×10^{1}

Sensitivity Analysis

Figure 5 shows the main Sobol index (S_m^i) over time for **Eqs 1, 13–15**, considering that $\beta_{apm} \neq 0$ and $\beta_{tke} \neq 0$, *i.e.* both APCs and T CD8$^+$ can be infected by SARS-CoV-2. Although the SA was performed for all 40 model parameters, some of then have small influence in the output produced for these equations and, for this reason, we decided to present in **Figure 5** only the 10 parameters that have more impact in the results produced by the model.

5 DISCUSSION

It can be seen in Panel A of **Figure 2** that after 30°days, viruses were almost eliminated, but the concentrations of IL-6 (Panel B), IgG (Panel C), and IgM (Panel D) antibodies remain at high levels. Cytokines start to decrease after 27°days approximately. It is expected that IgG remains at high levels after the resolution of the inflammation. For how long these levels remain high is still an open question in the case of COVID-19. Moreover, after 30°days, the immune response has not ended, and the inflammation has not been regulated. In the presence of inflammation and viruses, the cells of the immune system continue to migrate, causing more inflammation. It is important to highlight that, after the complete elimination of the virus, the inflammation will decrease until it ceases, and the concentration of cells and molecules of the immune system will return to a homeostatic level. However, this process will occur slowly because the model does not consider an anti-inflammatory response.

As one can observe in **Figure 3**, after 30°days, the virus concentration (Panel A) tends to zero. On the other hand, the IL-6 concentration (Panel B) starts to increase considerably after ten days, achieving its peak around the 21st°day. After the peak, it decreases slowly. It can also be observed that the IL-6 concentrations are much higher in this scenario than in the previous one presented in Panel B of **Figure 2**, *i.e.*, the non-survivors peak is about three times higher than the survivors one. We believe that this huge increase in IL-6 levels was due to the

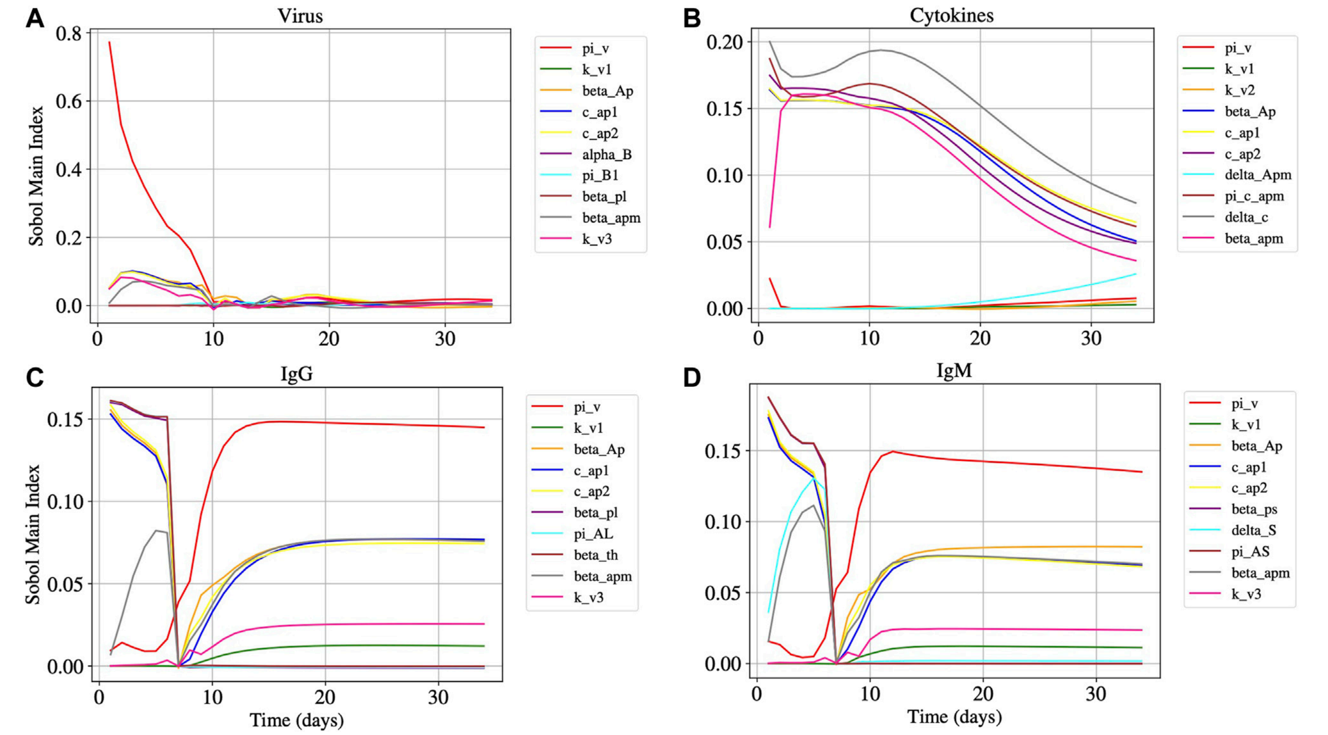


FIGURE 5 | Panels A, B, C, and D represents the Sobol main index over time of **Eqs 1**, **15**, **14** and **13**, respectively. The main Sobol indexes were evaluated for all model parameters, but Panels A, B, C, and D presents the 10 parameters with largest influence in the results.

CRS. Numerical results reveal that the CRS was caused mainly by effector APCs that are producing a considerable amount of pro-inflammatory cytokines. The infection of APCs was responsible for deregulating the innate immune response as well as for impairing the activation of the adaptive immune system. It is noteworthy that several other factors may contribute to CRS, among them a deficiency of the immune system in building an effective response against the virus at the beginning of the infection and a deficiency in controlling exacerbated inflammation.

In **Figure 4**, we observe that the infection of mature APCs, CD8$^+$ T cells, or both cells simultaneously by SARS-CoV-2 can be determinant to the outcome of the disease. In other words, all hypotheses are plausible from a numerical perspective. The main differences in results are that the production of cytokines, IgG, and IgM (Panels B, C and D, respectively) starts later when β_{apm} is equal to zero, *i.e.*, when only CD8$^+$ T cells are infected by SARS-CoV-2. Also, infected CD8$^+$ T cells impact both the day of the peak as well as its value (Panel A). We also observed that it is possible to obtain a similar result to the new fit of infected APC cells if we make $\beta_{apm} = 0$ in the original adjustment, *i.e.*, those obtained when both mature APCs and CD8$^+$ T cells are infected simultaneously by SARS-CoV-2. This result was not presented in this paper because it is close to the one obtained by the new fit, indicating that infected APCs influence the joint fitting more than infected CD8$^+$ T cells.

From the results of the sensitivity analysis (**Figure 5**), we can observe that the most influential parameters in respect to the dynamics of the virus (Panel A), cytokines (Panel B), IgG (Panel C), and IgM (Panel D) populations are related to the APCs. We observe that β_{ap}, c_{ap1} and c_{ap2} are among the ten most influential parameters for the virus, cytokines, IgG and IgM populations. The parameter β_{apm} is among the most influential for the virus, cytokines, IgG, and IgM populations. The virus replication rate π_v has a great influence on the virus, IgG, and IgM dynamics.

It is worth mentioning that the experimental data presents a huge variation (observe that the scale adopted is $\log_{10}$), challenging the calibration process reported in **Section 4**. The severity of the disease could explain this. The literature reports that severe/critical patients tend to peak in the second week of illness, with values ranging from 5.57 to 9.66 $\log_{10}$ copies/mL, whereas mild/moderate patients tend to peak in the first week of illness, with values ranging from 3.25 to 6.40 $\log_{10}$ copies/mL (Lui et al., 2020). Thus, one possible explanation for the considerable variation observed in the viremia is that the dataset mixes patients with distinct disease severity degrees. In such a case, the numerical results represent an average value for distinct severity degrees. Nevertheless, some factors suggest a prevalence of severe/critical patients in the dataset: 1) the need for hospital admission; 2) the fact that the viremia peak observed in the numerical result occurs around the 15th°day, with a value about 6.0 $\log_{10}$ copies/mL, which is compatible with the one reported for severe/critical patients (Lui et al., 2020).

COVID-19 is a new disease, and for this reason, some studies and datasets seem to contradict each other. Different studies in the literature adopt distinct methods and metrics, so it is not easy to gather their data to create a much larger dataset. In addition, the experimental data used in this paper to calibrate the model comprises a reduced number of patients and measurements. In this sense, the results presented in this work have limitations due to the restrictions imposed by data availability.

This paper has analysed the hypothesis that the infection of immune defence cells causes the CRS in patients with COVID-19, mainly macrophages and CD8$^+$ T cells, by the SARS-CoV-2. Although at first our numerical results suggest that this hypothesis may be correct, we must stress that the limitations described in this section could lead us to obtain an adjustment that supports, falsely, the hypothesis. Another limitation that can weaken our conclusions is that the combinations of values associated with other parameters except β_{apm}, β_{tke}, and π_{ci}, were not explored. We did not include parameters that are not related to the hypothesis evaluated in this paper and that can be associated to other theories found in the literature. For example, a paper from Garvin et al. (2020) indicates that the pathology of COVID-19 is likely the result of Bradykinin (BK) Storms rather than CRS. However, given the induction of IL-2 by BK, the two may be intricately linked. These theories could eventually lead to results similar or better to those obtained in our work.

Other techniques could be considered to distinguish survivors and non-survivors groups. The use of ODEs was motivated by our prior experience in using this tool to describe the dynamics of the immune response. Although we are not specialists in other techniques, such as deep learning, we believe that present limitations in the available data set can make it difficult or even impossible to use data driven models.

In the near future, we intend to explicitly represent the dynamics of different pro-inflammatory cytokines such as GM-CSF, TNF-α, IL-6, IL-8, among others. In addition, the model can be extended by considering anti-inflammatory responses through the incorporation of regulatory T cells (Treg cells), macrophages in a regulatory phenotype (Mreg phenotype), and anti-inflammatory cytokines such as IL-10. Finally, we also plan to use this model to reproduce the immune response against the SARS-CoV-2 vaccines.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conception and design of the mathematical model: RR, CB, AP, RW, and ML. Computational implementation of the mathematical model: RR, CB, AP, BMQ, LP, AV, and ML have also been involved in drafting the manuscript. LS, MX, and RR ajusted the numerical results to experimental data. LP, LS, and AV were responsible for extracting experimental data from the literature. RR, AP, BMQ, RW, and ML revised the final manuscript. All authors have read and approved the final manuscript.

FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)-Brasil-Finance Code 001 (scholarship), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-Brasil-PIBIC/CNPq/UFJF and Produtividade em Pesquisa (scholarships), Universidade Federal de Juiz de Fora-Brasil-BIC/UFJF (scholarship), and Fundação de Amparo á Pesquisa do Estado de Minas Gerais (FAPEMIG)-Brasil-Bolsa de Doutorado (scholarship) and CEX APQ 02830/17 (equipments).

ACKNOWLEDGMENTS

The authors would like to thank the infrastructure provided by Universidade Federal de Juiz de Fora (UFJF), Universidade Federal de São João del-Rey (UFSJ), Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais-Campus Avançado de Cataguases (IF Sudeste MG).

REFERENCES

Almocera, A. E. S., Quiroz, G., and Hernandez-Vargas, E. A. (2021). Stability Analysis in COVID-19 Within-Host Model with Immune Response. *Commun. Nonlinear Sci. Numer. Simul* 95, 105584. doi:10.1016/j.cnsns.2020.105584

Baker, C. T. H., Bocharov, G. A., and Paul, C. A. H. (1997). Mathematical Modelling of the Interleukin-2 T-Cell System: A Comparative Study of Approaches Based on Ordinary and Delay Differential Equation. *J. Theor. Med.* 1, 117–128. doi:10.1080/10273669708833012

Beauchemin, C. A. A., McSharry, J. J., Drusano, G. L., Nguyen, J. T., Went, G. T., Ribeiro, R. M., et al. (2008). Modeling Amantadine Treatment of Influenza a Virus *In Vitro*. *J. Theor. Biol.* 254, 439–451. doi:10.1016/j.jtbi.2008.05.031

Bernaschi, M., and Castiglione, F. (2001). Design and Implementation of an Immune System Simulator. *Comput. Biol. Med.* 31, 303–331. doi:10.1016/s0010-4825(01)00011-7

Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., et al. (2020). Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* 181, 1036–1045. doi:10.1016/j.cell.2020.04.026

Bonin, C. R. B., Fernandes, G. C., Menezes, R. M., Camacho, L. A. B., da Mota, L. M. H., de Lima, S. M. B., et al. (2019). "Quantitative Validation of a Yellow Fever Vaccine Model," in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18-21 Nov. 2019 (. IEEE), 2113–2120.

Bonin, C. R. B., dos Santos, R. W., Fernandes, G. C., and Lobosco, M. (2016). Computational Modeling of the Immune Response to Yellow Fever. *J. Comput. Appl. Math.* 295, 127–138. doi:10.1016/j.cam.2015.01.020

Bonin, C. R. B., Fernandes, G. C., Dos Santos, R. W., and Lobosco, M. (2018). A Qualitatively Validated Mathematical-Computational Model of the Immune Response to the Yellow Fever Vaccine. *BMC Immunol.* 19, 15. doi:10.1186/s12865-018-0252-1

Catanzaro, M., Fagiani, F., Racchi, M., Corsini, E., Govoni, S., and Lanni, C. (2020). Immune Response in COVID-19: Addressing a Pharmacological challenge by Targeting Pathways Triggered by SARS-CoV-2. *Sig Transduct Target. Ther.* 5, 1–10. doi:10.1038/s41392-020-0191-1

Celada, F., and Seiden, P. E. (1992). A Computer Model of Cellular Interactions in the Immune System. *Immunol. Today* 13, 56–62. doi:10.1016/0167-5699(92)90135-t

Chang, S. T., Linderman, J. J., and Kirschner, D. E. (2005). Multiple Mechanisms Allow mycobacterium Tuberculosis to Continuously Inhibit Mhc Class Ii-Mediated Antigen Presentation by Macrophages. *Proc. Natl. Acad. Sci.* 102, 4530–4535. doi:10.1073/pnas.0500362102

Chao, D. L., Davenport, M. P., Forrest, S., and Perelson, A. S. (2004). A Stochastic Model of Cytotoxic T Cell Responses. *J. Theor. Biol.* 228, 227–240. doi:10.1016/j.jtbi.2003.12.011

Chen, X., Zhao, B., Qu, Y., Chen, Y., Xiong, J., Feng, Y., et al. (2020). Detectable Serum Severe Acute Respiratory Syndrome Coronavirus 2 Viral Load (RNAemia) Is Closely Correlated with Drastically Elevated Interleukin 6 Level in Critically Ill Patients with Coronavirus Disease 2019. *Clin. Infect. Dis.* 71, 1937–1942. doi:10.1093/cid/ciaa449

Coperchini, F., Chiovato, L., Croce, L., Magri, F., and Rotondi, M. (2020). The Cytokine Storm in COVID-19: An Overview of the Involvement of the Chemokine/chemokine-Receptor System. *Cytokine Growth Factor. Rev.* 53, 25–32. doi:10.1016/j.cytogfr.2020.05.003

Davanzo, G. G., Codo, A. C., Brunetti, N. S., Boldrini, V., Knittel, T. L., Monterio, L. B., et al. (2020). *Sars-cov-2 Uses Cd4 to Infect T Helper Lymphocytes*. New York, NY: EUA. doi:10.1101/2020.09.25.20200329

Du, S. Q., and Yuan, W. (2020). Mathematical Modeling of Interaction between Innate and Adaptive Immune Responses in COVID-19 and Implications for Viral Pathogenesis. *J. Med. Virol.* 92, 1615–1628. doi:10.1002/jmv.25866

Flegg, J. A., Byrne, H. M., Flegg, M. B., and Sean McElwain, D. L. (2012). Wound Healing Angiogenesis: the Clinical Implications of a Simple Mathematical Model. *J. Theor. Biol.* 300, 309–316. doi:10.1016/j.jtbi.2012.01.043

Garvin, M. R., Alvarez, C., Miller, J. I., Prates, E. T., Walker, A. M., Amos, B. K., et al. (2020). A Mechanistic Model and Therapeutic Interventions for COVID-19 Involving a RAS-Mediated Bradykinin Storm. *eLife* 9, e59177. doi:10.7554/eLife.59177

Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., et al. (2008). The Hill Equation: a Review of its Capabilities in Pharmacological Modelling. *Fundam. Clin. Pharmacol.* 22, 633–648. doi:10.1111/j.1472-8206.2008.00633.x

Grant, R. A., Morales-Nebreda, L., Morales-Nebreda, L., Markov, N. S., Swaminathan, S., Querrey, M., et al. (2021). Circuits between Infected Macrophages and T Cells in Sars-Cov-2 Pneumonia. *Nature* 590, 635–641. doi:10.1038/s41586-020-03148-w

Herman, J., and Usher, W. (2017). SALib: an Open-Source python Library for Sensitivity Analysis. *Joss* 2, 97. doi:10.21105/joss.00097

Hernandez-Vargas, E. A., and Velasco-Hernandez, J. X. (2020). In-host Mathematical Modelling of COVID-19 in Humans. *Annu. Rev. Control* 50, 448–456. doi:10.1016/j.arcontrol.2020.09.006

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181, 271–280. doi:10.1016/j.cell.2020.02.052

Jarrett, A. M., Cogan, N. G., and Shirtliff, M. E. (2015). Modelling the Interaction between the Host Immune Response, Bacterial Dynamics and Inflammatory Damage in Comparison with Immunomodulation and Vaccination Experiments. *Math. Med. Biol.* 32, 285–306. doi:10.1093/imammb/dqu008

Kohler, B., Puzone, R., Seiden, P. E., and Celada, F. (2000). A Systematic Approach to Vaccine Complexity Using an Automaton Model of the Cellular and Humoral Immune System. *Vaccine* 19, 862–876. doi:10.1016/s0264-410x(00)00225-5

Li, H., Liu, L., Zhang, D., Xu, J., Dai, H., Tang, N., et al. (2020a). SARS-CoV-2 and Viral Sepsis: Observations and Hypotheses. *The Lancet* 395, 1517–1520. doi:10.1016/S0140-6736(20)30920-X

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., et al. (2020b). Early Transmission Dynamics in Wuhan, china, of Novel Coronavirus–Infected Pneumonia. *New Engl. J. Med.* 382, 1199–1207. doi:10.1056/nejmoa2001316

Long, Q. X., Liu, B. Z., Deng, H. J., Wu, G. C., Deng, K., Chen, Y. K., et al. (2020). Antibody Responses to SARS-CoV-2 in Patients with COVID-19. *Nat. Med.* 26, 845–848. doi:10.1038/s41591-020-0897-1

Lui, G., Ling, L., Lai, C. K., Tso, E. Y., Fung, K. S., Chan, V., et al. (2020). Viral Dynamics of SARS-CoV-2 across a Spectrum of Disease Severity in COVID-19. *J. Infect.* 81, 318–356. doi:10.1016/j.jinf.2020.04.014

Lukan, N. (2020). "Cytokine Storm", Not Only in COVID-19 Patients. Mini-Review. *Immunol. Lett.* 228, 38–44. doi:10.1016/j.imlet.2020.09.007

Morpurgo, D., Serenthà, R., Seiden, P. E., and Celada, F. (1995). Modelling Thymic Functions in a Cellular Automaton. *Int. Immunol.* 7, 505–516. doi:10.1093/intimm/7.4.505

Murphy, K., and Weaver, C. (2008). *Janeway's Immunobiology*. Garland Science. New York, NY: EUA.

Narazaki, M., and Kishimoto, T. (2018). The Two-Faced Cytokine Il-6 in Host Defense and Diseases. *Ijms* 19, 3528. doi:10.3390/ijms19113528

Pappalardo, F., Russo, G., Pennisi, M., Sgroi, G., Palumbo, G. A. P., Motta, S., et al. (2018). "Agent Based Modeling of Relapsing Multiple Sclerosis: A Possible Approach to Predict Treatment Outcome," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), Madrid, Spain, 3-6 Dec. 2018 (IEEE), 1380–1385. doi:10.1109/BIBM.2018.8621109

Paul, W. (2008). Fundamental Immunology. *Fundamental Immunology*. Wolters Kluwer/Lippincott Williams & Wilkins. Philadelphia, PA: EUA.

Perelson, A. S. (1989). "Modeling the Interaction of the Immune System with Hiv," in *Mathematical And Statistical Approaches to AIDS Epidemiology* (Springer), 350–370. doi:10.1007/978-3-642-93454-4_17

Perelson, A. S. (2002). Modelling Viral and Immune System Dynamics. *Nat. Rev. Immunol.* 2, 28–36. doi:10.1038/nri700

Pettet, G. J., Byrne, H. M., McElwain, D. L. S., and Norbury, J. (1996). A Model of Wound-Healing Angiogenesis in Soft Tissue. *Math. biosciences* 136, 35–63. doi:10.1016/0025-5564(96)00044-2

Pigozzo, A. B., Macedo, G. C., Dos Santos, R. W., and Lobosco, M. (2013). On the Computational Modeling of the Innate Immune System. *BMC bioinformatics* 14, S7. doi:10.1186/1471-2105-14-s6-s7

Pigozzo, A. B., Missiakas, D., Alonso, S., Dos Santos, R. W., and Lobosco, M. (2018). Development of a Computational Model of Abscess Formation. *Front. Microbiol.* 9, 1355. doi:10.3389/fmicb.2018.01355

Prompetchara, E., Ketloy, C., and Palaga, T. (2020). Immune Responses in COVID-19 and Potential Vaccines: Lessons Learned from SARS and MERS Epidemic. *Asian Pac. J. Allergy Immunol.* 38, 1–9. doi:10.12932/AP-200220-0772

Quintela, Bd. M., dos Santos, R. W., and Lobosco, M. (2014). On the Coupling of Two Models of the Human Immune Response to an Antigen. *Biomed. Research International* 2014, 410457. doi:10.1155/2014/410457

Reis, R. F., Fernandes, J. L., Schmal, T. R., Rocha, B. M., Dos Santos, R. W., and Lobosco, M. (2019). A Personalized Computational Model of Edema Formation in Myocarditis Based on Long-axis Biventricular MRI Images. *BMC Bioinformatics* 20, 532. doi:10.1186/s12859-019-3139-0

Rodriguez-Fernandez, M., Egea, J. A., and Banga, J. R. (2006). Novel Metaheuristic for Parameter Estimation in Nonlinear Dynamic Biological Systems. *BMC Bioinformatics* 7, 483. doi:10.1186/1471-2105-7-483

Rohatgi, A. (2020). Webplotdigitizer: Version 4.3.

Schwaab, M., Biscaia, Jr., E. C., Jr, Monteiro, J. L., and Pinto, J. C. (2008). Nonlinear Parameter Estimation through Particle Swarm Optimization. *Chem. Eng. Sci.* 63, 1542–1552. doi:10.1016/j.ces.2007.11.024

Scipy (2021). Scipy's Homepage, accessed on april, 2021. (2021).

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., et al. (2020). Cell Entry Mechanisms of SARS-CoV-2, *Proc. Natl. Acad. Sci. USA* 117. 11727–11734. doi:10.1073/pnas.2003138117

Shimabukuro-Vornhagen, A., Gödel, P., Subklewe, M., Stemmler, H. J., Schlößer, H. A., Schlaak, M., et al. (2018). Cytokine Release Syndrome. *J. Immunotherapy Cancer* 6. doi:10.1186/s40425-018-0343-9

Sobol′, I. M. (2001). Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates. *Mathematics Comput. simulation* 55, 271–280. doi:10.1016/S0378-4754(00)00270-6

Sompayrac, L. (2012). *How the Immune System Works*. Oxford: Wiley-Blackwell.

Storn, R., and Price, K. (1997). Differential Evolution–A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. optimization* 11, 341–359. doi:10.1023/A:1008202821328

Su, B., Zhou, W., Dorman, K. S., and Jones, D. E. (2009). Mathematical Modelling of Immune Response in Tissues. *Comput. Math. Methods Med.* 10, 9–38. doi:10.1080/17486700801982713

Tanaka, T., Narazaki, M., and Kishimoto, T. (2014). Il-6 in Inflammation, Immunity, and Disease. *Cold Spring Harbor Perspect. Biol.* 6, a016295. doi:10.1101/cshperspect.a016295

Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A., and Ng, L. F. P. (2020). The trinity of COVID-19: Immunity, Inflammation and Intervention. *Nat. Rev. Immunol.* 20, 363–374. doi:10.1038/s41577-020-0311-8

Tisoncik, J. R., Korth, M. J., Simmons, C. P., Farrar, J., Martin, T. R., and Katze, M. G. (2012). Into the Eye of the Cytokine Storm, *Microbiol. Mol. Biol. Rev.*, 76, 16–32. doi:10.1128/MMBR.05015-11

To, K. K.-W., Tsang, O. T.-Y., Leung, W.-S., Tam, A. R., Wu, T.-C., Lung, D. C., et al. (2020). Temporal Profiles of Viral Load in Posterior Oropharyngeal Saliva Samples and Serum Antibody Responses during Infection by SARS-CoV-2: an Observational Cohort Study. *Lancet Infect. Dis.* 20, 565–574. doi:10.1016/S1473-3099(20)30196-1

Varah, J. M. (1982). A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations. *SIAM J. Sci. Stat. Comput.* 3, 28–46. doi:10.1137/0903003

Vodovotz, Y., Chow, C. C., Bartels, J., Lagoa, C., Prince, J. M., Levy, R. M., et al. (2006). In Silico models of Acute Inflammation in Animals. *Shock* 26, 235–244. doi:10.1097/01.shk.0000225413.13866.fo

Waito, M., Walsh, S. R., Rasiuk, A., Bridle, B. W., and Willms, A. R. (2016). "A Mathematical Model of Cytokine Dynamics during a Cytokine Storm," in *Mathematical and Computational Approaches in Advancing Modern Science and Engineering* (Springer), 331–339. doi:10.1007/978-3-319-30379-6_31

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., et al. (2020). Virological Assessment of Hospitalized Patients with COVID-2019. *Nature* 581, 465–469. doi:10.1038/s41586-020-2196-x

Xavier, M. P., Bonin, C. R. B., dos Santos, R. W., and Lobosco, M. (2017). "On the Use of gillespie Stochastic Simulation Algorithm in a Model of the Human Immune System Response to the Yellow Fever Vaccine," in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13-16 Nov. 2017 (. IEEE), 1476–1482. doi:10.1109/BIBM.2017.8217880

Xavier, M. P., Reis, R. F., dos Santos, R. W., and Lobosco, M. (2020). "A Simplified Model of the Human Immune System Response to the COVID-19," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (Los Alamitos, CA, USA: . IEEE Computer Society)), 1311–1317. doi:10.1109/BIBM49941.2020.9313418

Zhang, C., Wu, Z., Li, J.-W., Zhao, H., and Wang, G.-Q. (2020). Cytokine Release Syndrome in Severe COVID-19: Interleukin-6 Receptor Antagonist Tocilizumab May Be the Key to Reduce Mortality. *Int. J. Antimicrob. Agents* 55, 105954. doi:10.1016/j.ijantimicag.2020.105954

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., et al. (2020). Clinical Course and Risk Factors for Mortality of Adult Inpatients with COVID-19 in Wuhan, china: a Retrospective Cohort Study. *The Lancet* 395, 1054–1062. doi:10.1016/S0140-6736(20)30566-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 27 July 2021
doi: 10.3389/fmolb.2021.614443

Non-RBM Mutations Impaired SARS-CoV-2 Spike Protein Regulated to the ACE2 Receptor Based on Molecular Dynamic Simulation

Yaoqiang Du[1†], Hao Wang[2†], Linjie Chen[3†], Quan Fang[3], Biqin Zhang[4], Luxi Jiang[1], Zhaoyu Wu[5], Yexiaoqing Yang[5], Ying Zhou[5], Bingyu Chen[1], Jianxin Lyu[3,5]* and Zhen Wang[1,3]*

[1]Allergy Center, Department of Transfusion Medicine, Ministry of Education Key Laboratory of Laboratory Medicine, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, China, [2]National Clinical Research Center for Child Health, National Children's Regional Medical Center, Department of Clinical Laboratory, The Children's Hospital, Zhejiang University Shcool of Medicine, Hangzhou, China, [3]School of Laboratory Medicine, Hangzhou Medical College, Hangzhou, China, [4]Department of Hematology, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, China, [5]School of Laboratory Medicine and Life Sciences, Wenzhou Medical University, Wenzhou, China

The emergence of novel coronavirus mutants is a main factor behind the deterioration of the epidemic situation. Further studies into the pathogenicity of these mutants are thus urgently needed. Binding of the spinous protein receptor binding domain (RBD) of SARS-CoV-2 to the angiotensin-converting enzyme 2 (ACE2) receptor was shown to initiate coronavirus entry into host cells and lead to their infection. The receptor-binding motif (RBM, 438–506) is a region that directly interacts with ACE2 receptor in the RBD and plays a crucial role in determining affinity. To unravel how mutations in the non-RBM regions impact the interaction between RBD and ACE2, we selected three non-RBM mutant systems (N354D, D364Y, and V367F) from the documented clinical cases, and the Q498A mutant system located in the RBM region served as the control. Molecular dynamics simulation was conducted on the mutant systems and the wild-type (WT) system, and verified experiments also performed. Non-RBM mutations have been shown not only to change conformation of the RBM region but also to significantly influence its hydrogen bonding and hydrophobic interactions. In particular, the D364Y and V367F systems showed a higher affinity for ACE2 owing to their electrostatic interactions and polar solvation energy changes. In addition, although the binding free energy at this point increased after the mutation of N354D, the conformation of the random coil (Pro384-Asp389) was looser than that of other systems, and the combined effect weakened the binding free energy between RBD and ACE2. Interestingly, we also found a random coil (Ala475-Gly485). This random coil is very sensitive to mutations, and both types of mutations increase the binding free energy of residues in this region. We found that the binding loop (Tyr495-Tyr505) in the RBD domain strongly binds to Lys353, an important residue of the ACE2 domain previously identified. The binding free energy of the non-RBM mutant group at the binding loop had positive and negative changes, and these changes were more obvious than that of the Q498A system. The results of this study

Edited by:
*Nadine Kabbani,
George Mason University,
United States*

Reviewed by:
*Pavel Srb,
Academy of Sciences of the Czech
Republic (ASCR), Czechia
Qiwei Zhang,
Jinan University, China*

***Correspondence:**
*Zhen Wang
wangzhen@hmc.edu.cn
Jianxin Lyu
ljx@hmc.edu.cn*

[†]These authors have contributed equally to this work

Specialty section:
*This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences*

Received: *06 October 2020*
Accepted: *16 July 2021*
Published: *27 July 2021*

Citation:
Du Y, Wang H, Chen L, Fang Q, Zhang B, Jiang L, Wu Z, Yang Y, Zhou Y, Chen B, Lyu J and Wang Z (2021) Non-RBM Mutations Impaired SARS-CoV-2 Spike Protein Regulated to the ACE2 Receptor Based on Molecular Dynamic Simulation. Front. Mol. Biosci. 8:614443. doi: 10.3389/fmolb.2021.614443

elucidate the effect of non-RBM mutation on ACE2-RBD binding, and provide new insights for SARS-CoV-2 mutation research.

Keywords: SARS-CoV-2, spike protein, non-RBM mutations, ACE2 receptor, molecular dynamic simulation

INTRODUCTION

Several cases of unexplained pneumonia occurred in Wuhan, China, at the end of 2019 (Wang et al., 2020a). Patients' clinical symptoms were related to infectious atypical pneumonia (SARS-CoV) and Middle East respiratory syndrome (MERS-CoV), such as fever, cough, and difficulty breathing (Huang et al., 2020). After investigation, it was found that the first case of pneumonia originated from a seafood and farmer's market in Wuhan. Whole-genome sequencing revealed that the pathogen causing pneumonia was a new type of coronavirus. The virus was named 2019-nCoV by the World Health Organization and was subsequently renamed SARS-CoV-2 by the International Commission for Classification of Viruses (Tan et al., 2020). The novel virus has become a global public health event, and many countries have been deeply affected. More than 176 million patients have been infected, and there have been about 3.8 million deaths worldwide until June 2021.

Immediately after the virus broke out, it was sequenced on January 9, 2020, and the sequencing data were posted on the Internet afterwards (Gralinski and Menachery, 2020). Coronaviruses are 26,000 to 32,000 bases in length and are single-stranded positive-stranded RNA (Su et al., 2016). They mainly infect the respiratory tract and digestive tract. According to their genome characteristics, they can be divided into α-coronavirus, β-coronavirus, γ-coronavirus, and δ-coronavirus (Li, 2016). There are six types of coronaviruses that can cause human infection, two of which belong to α-coronavirus, and four belong to β-coronavirus (Tang et al., 2015), and the remaining virus types cannot infect humans. The most aggressive coronaviruses are SARS-CoV and MERS-CoV (belonging to β-coronavirus), as both of these viruses have caused global public health events. These two viruses are believed to have been transmitted from bats to civet cats or camels and, eventually, to humans (Guan et al., 2003; Azhar et al., 2014; Cui et al., 2019). SARS-COV-2 has a Spike protein has 93.1% homology with RaTG13 (bat-like coronavirus), but it has less than 80% homology with other SARS-CoV (Zhou et al., 2020).

To infect the human body, the virus must first bind to the corresponding receptor. For example, angiotensin-converting enzyme 2 (ACE2), CD209L, and dipeptidyl peptidase 4 (DPP4) were the main receptors for the SARS-CoV outbreak in 2003 and MERS-CoV in 2012 (Wu et al., 2020). The receptor of SARS-CoV-2 is also ACE2 (Wang et al., 2020b). After the virus enters the human body, the receptor binding domain (RBD) on the virus Spike protein should first bind to the receptor of the cell before it can fuse with the cell membrane and complete the infection. The Spike protein of SARS-CoV-2 is a structural protein encoded by the end of the viral genome (Wu et al., 2020), when it binds to cell-related receptors, it will be broken down into two subunits, S1 and S2, of which the S1 subunit will directly bind to the receptor (Wang et al., 2020b). Structural changes in RBD may lead to enhanced or weakened binding of the virus to the receptor (Lan et al., 2020) and then affect the probability of virus infection. Here we found a mutant of SARS-CoV-2, which will cause changes in the amino acid sequence of the SARS-CoV-2 RBD region, but the consequences of this change are not yet clear. We used molecular dynamics (MD) simulation methods to describe and compare various atomic forces in wild-type (WT) and mutant SARS-CoV-2 RBD and try to have a more detailed understanding of the binding energy changes caused by mutations, which will help promote new targeted therapies against this emerging pathogen.

As for the relationship between ACE2 and the RBD region, the role of the receptor-binding motif (RBM) is self-evident. The amino acid changes in other regions could also have an impact on the overall population. A previous study showed that the D614G mutation in the Spike protein could increase viral infectivity (Hu et al., 2020). The G614 mutation can enhance the ability of protease to cut the Spike protein, thus promoting the virus to be more infectious. We constructed three mutations on the basis of the RBD information found in actual cases (Ou et al., 2021) and selected the important 498 residues in RBM for the alanine mutation as the control group. We aim to explore the impact of mutations in different regions of RBD on its binding effect and pay attention to the changes in the RBM region.

RESULTS AND DISCUSSION

Predicting Mutation Effects on Protein–Protein Interactions

Protein–protein interactions (PPIs) play an important role in various biological processes that include cell regulation and signal transduction. Various studies have shown that many disease-related amino acid mutations are located at the protein–protein interface, thereby affecting PPIs by changing the binding affinity or specificity. SAAMBE-3D is a newly developed machine learning algorithm used to predict the impact of single amino acid mutations on PPIs (Pahari et al., 2020). Using the PDB structure of 6LZG as the input file, the binding free energy change caused by the mutation and the prediction of whether the mutation disturbs PPIs were obtained. According to the analysis of the results in **Table 1**, the binding energy of

TABLE 1 | The predicted effect of single amino acid mutation of PPIs.

System	N354D	D364Y	V367F	Q498A
ΔG Prediction[a]	0.59	0.59	−0.42	2.93
Effect	Disruptive	Disruptive	Nondisruptive	Nondisruptive

[a]*Prediction of changed binding free energy (kJ/mol) caused by the mutation.*

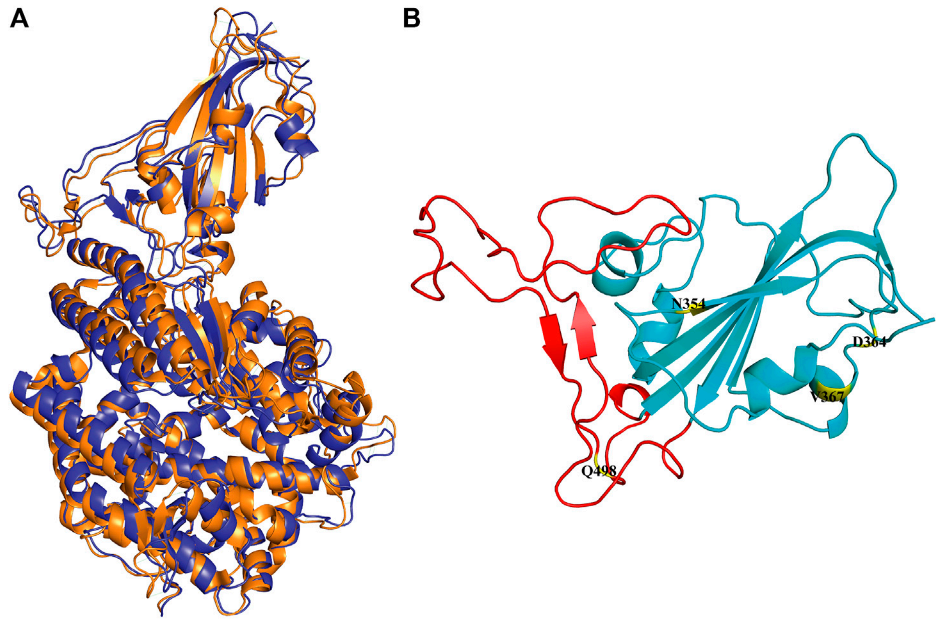


FIGURE 1 | Conformation of WT RBD/hACE2 and position of four mutations in the RBD domain. **(A)** The crystal is illustrated in orange, and the representative structure after simulation is in blue. **(B)** The four mutation sites are shown in yellow, and the range of red fragments represents RBM.

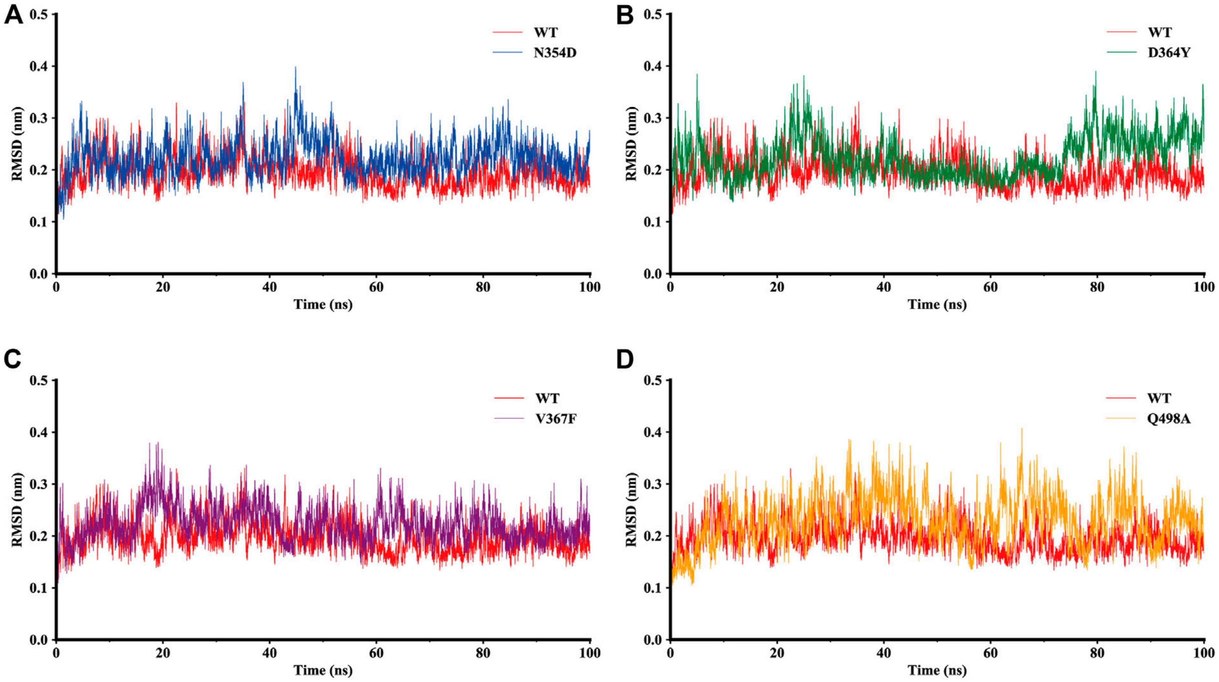


FIGURE 2 | RMSD of backbone atoms relative to the initial structure during the MD trajectories. **(A)** RMSD values in the WT system (red) and N354D system (blue). **(B)** WT system and D364Y system (green). **(C)** WT system and V367F system (purple). **(D)** WT system and Q498A system (orange).

N354D and D364Y were reduced, and the entire PPI networks were destroyed. On the contrary, the binding energy of the Q498A mutation located in the RBM region was also reduced, but the PPI network was not destroyed. Together, these results using a machine learning algorithm revealed that residue mutations that are not present in the RBM region may have a

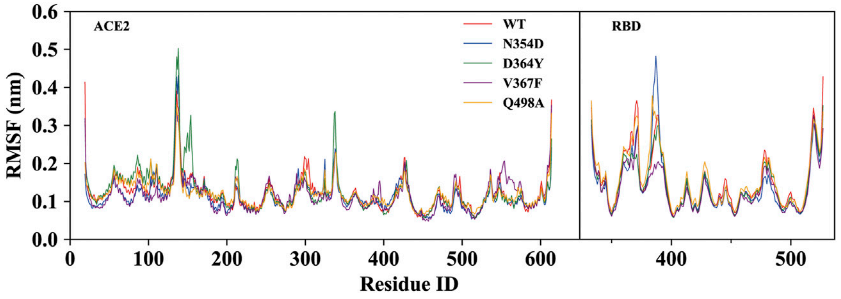


FIGURE 3 | RMSF of backbone atoms relative to the initial structure during the MD trajectories. RMSF values of ACE2 and RBD in the WT system (red), N354D system (blue), D364Y system (green), V367F system (blue) and Q498A system (orange).

greater effect on the complex, a finding that we found particularly interesting. Considering the limited amount of raw data and the particularity of the genome, we then used dynamic simulation to unravel detailed outcomes.

Structural Flexibility and Stability of the Simulation Systems

In the simulation process, the superposition of representative structures with the crystal shows that they are very similar (**Figure 1A**), which indicates that the simulation was performed under ideal conditions. The non-RBM mutations are found structurally far away from the RBM region (**Figure 1B**), conducting kinetic simulation could thus provide insights into their interactions.

The dynamic behavior of the WT system and the mutant system were analyzed by 100-ns MD simulations. The stability of the WT and mutant systems was determined using the root-mean-square deviation (RMSD) of the backbone atoms relative to the initial structure (**Figure 2**). It can be seen from the figure that the conformations of the five complexes reached equilibrium 10 ns after the start of the simulation, with no obvious fluctuations. The RMSD values of N354D, D364Y, and V367F in the non-RBM group all approximated 0.22 nm, which is slightly higher than the 0.19 nm value observed in the WT system (**Figures 2A–C**). The RMSD value of Q498A in the RBM group was 0.23 nm, which was higher than for the other four systems (**Figure 2D**). This indicates that the stability of mutant systems is lower than that of the WT system and that the Q498A system located in the RBM region is the most unstable.

To investigate the detailed residual atomic fluctuations, the Gromacs was applied to compute the root-mean-square fluctuation (RMSF) of the backbone atoms versus residue ID for all systems (**Figure 3**). It is generally believed that the fluctuation of the residue was high if the RMSF value was equal to or greater than 0.23 nm. Comparing the fluctuations of the RMSF value between different systems, we found an area with large fluctuations, which is a random coil (Pro384-Asp389) between β2 and β3 in the RBD domain. The N354D system displayed the highest flexibility in this segment, whereas the V367F system, also from the non-RBM group, had the lowest flexibility. Although this segment was not the key area of the combination, its influence cannot be ignored. Coincidentally, the random coils in the RBM region (Ala475-Gly485) of all mutant systems had high RMSF values, suggesting that this region was sensitive to both types of mutations. In addition, the random coil (Asn134-Glu140) between α1 and α2 in the ACE2 domain of the D364Y system fluctuated greatly, and the RMSF value was 0.5 nm, compared to 0.4 nm for the WT system. Therefore, we speculate that D364Y not only changes the conformation of the RBD domain but also affects the binding state of the complex. In order to understand the specific changes of the random coil described, we further analyzed secondary structures over time.

Secondary Structure Analysis of Fluctuation Regions

The changes with time of the secondary structures were obtained for each amino acid using DSSP plugin analysis in Gromacs. We compared the fluctuations of random coils (Asn134-Glu140) in the ACE2 domain of the D364Y and WT systems. At the beginning of the simulation, the secondary structures of the two systems in this region was found to fluctuate between β-sheets and bends (**Figures 4A,B**). As the simulation progressed, the secondary structure of the D364Y system changed from β-sheet to coil after 47 ns(**Figure 4B**). This change enhanced the flexibility of the region while affecting the upstream binding of α1 to RBM. By comparing the proportion of secondary structures in the two domains of the complex, we found that the helical ratio of the RBD domain was much smaller than that for the ACE2 domain, which indicated that the conformation of the virus was unstable and prone to mutation. The random coil Pro384-Asp389 in the WT system is a mixture of 3-helix, bend, and turn (**Figure 4C**), but the 3-helix in the N354D system is almost entirely replaced by coil (**Figure 4D**). This indicates that the N354D mutation increases the flexibility of this region and thus affects the spatial position of adjacent secondary structures. In contrast, the V367F system was almost entirely 3-helix during the whole simulation (**Figure 4E**), which indicated that the conformation of the random coil Pro384-Asp389 of the V367F system was very stable. In addition, comparison with the WT system revealed that the secondary conformation of both the D364Y and Q498A systems did not

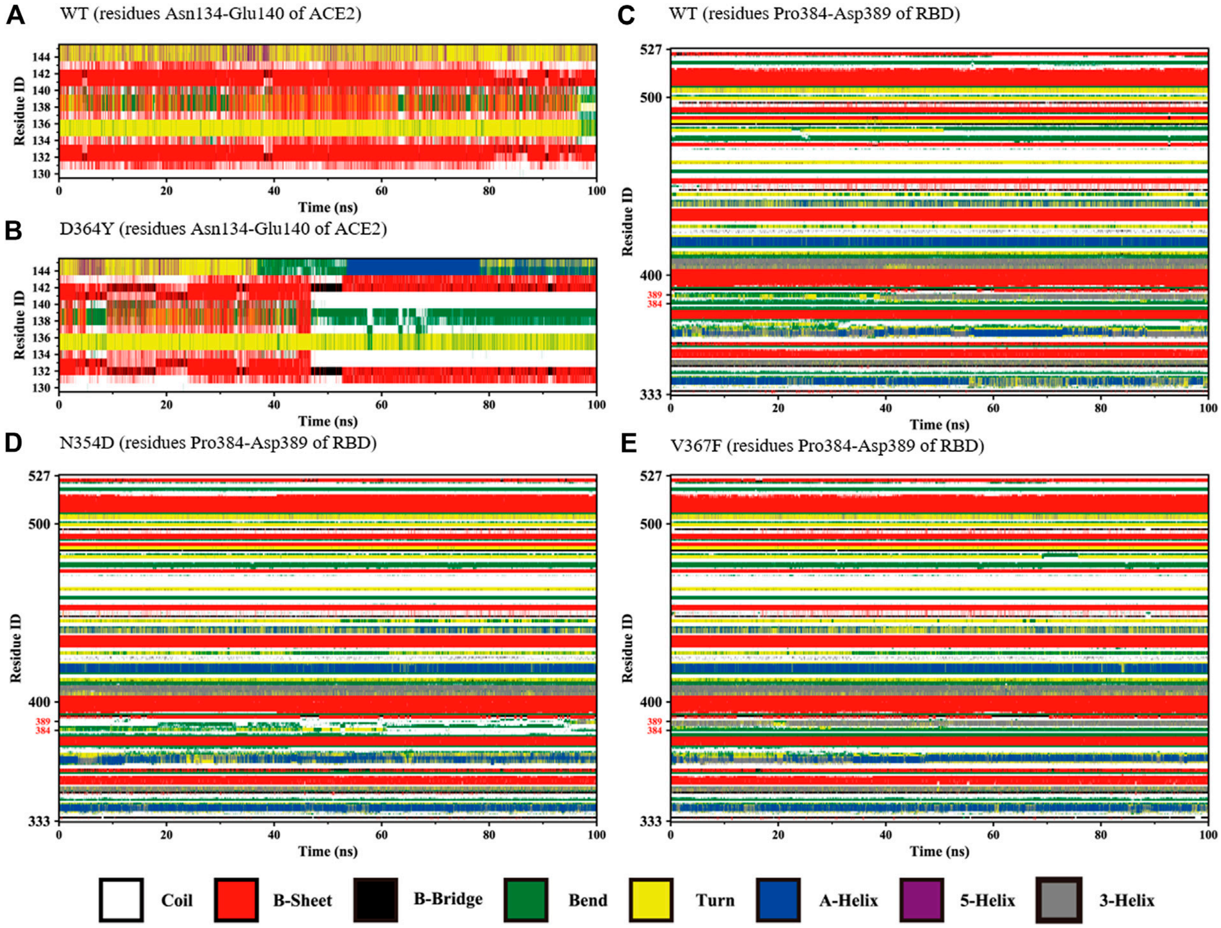


FIGURE 4 | Secondary structures of the regions (residues Asn134-Glu140 of ACE2 and Pro384-Asp389 of RBD) during the simulation. **(A)** Residues Asn134-Glu140 of ACE2 in the WT system. **(B)** Residues Asn134-Glu140 of ACE2 in the D364Y system. **(C)** Residues Pro384-Asp389 of RBD in the WT system. **(D)** Residues Pro384-Asp389 of RBD in the N354D system. **(E)** Residues Pro384-Asp389 of RBD in the V367F system.

change significantly in this region (**Supplementary Figures S1A,B**).

Considering that the RMSF values of the random coil (Ala475-Gly485) were slightly different, we used the DSSP tool to calculate the proportion of secondary structures in this region. All the coils in the non-RBM group became turn (**Figure 5A**), while the Q498A and WT systems showed a similar trend, they were converted from coil to turn in less than half of the simulation frames (**Figure 5D**). These results indicate that this region is more sensitive to non-RBM mutations, which decrease the flexibility of the region.

Binding Free Energy and Decomposition Analyses of Mutant and WT Complexes

The binding free energy was calculated as the sum of gas phase energy ($\Delta E_{vdw} + \Delta E_{ele}$), solvation free energy ($\Delta G_{SA} + \Delta G_{PB}$), and entropy ($-T\Delta S$). Recent studies have shown that the interaction entropy (IE) method provides more reliable predictions for the free energy of protein–ligand and protein–protein binding and entropy contribution of hot residue interactions. The results showed that the binding free energy of the N354D system was −52.085 kJ/mol, lower than that of the WT system −85.611 kJ/mol. The binding free energy of the D364Y, V367F and Q498A was higher than that of the WT system, which was −298.563, −200.852, and −222.705 kJ/mol, respectively (**Table 2**).

By comparing the decomposition energies, we found that the main driving force in the binding process is the polar solvation energy, which is the difference between the electrostatic energy of solvent and vacuum. The contributions of the polar solvation energy of the WT, N354D, D364Y, V367F, and Q498A systems are 1,086.042, 861.460, 920.047, 871.004, and 795.021 kJ/mol, respectively. The polar solvation energy of all mutation systems were decreased, and the Q498A system changed the most, which may be related to the hydrophobicity of alanine. The polar solvation energy of all mutant systems is lower than that of WT system, which indicated that the conformational change reduces the solubility of the RBD domain in the polar solution. In

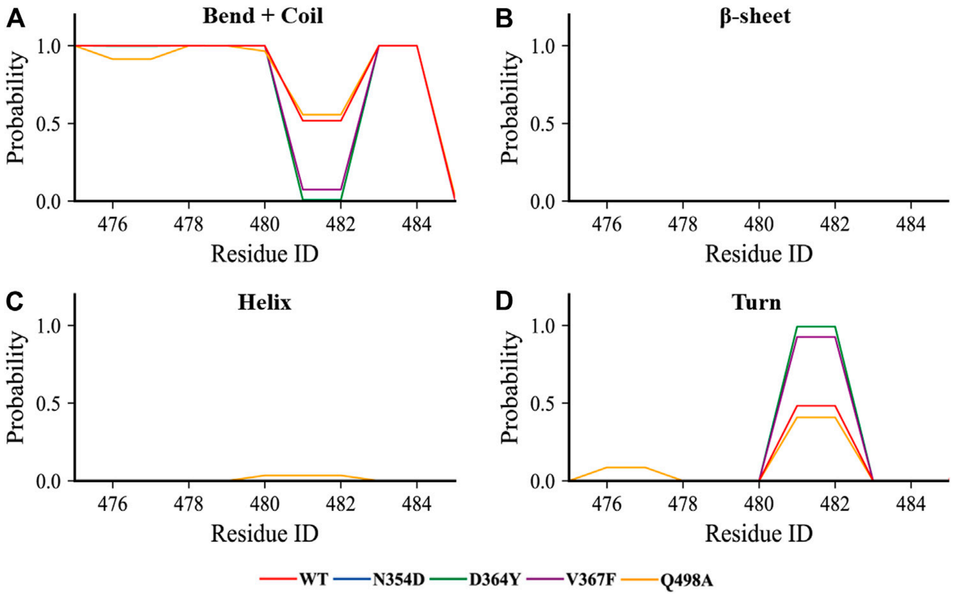


FIGURE 5 | The content of various secondary structures in the region (Ala475-Gly485) in the RBD of each system. **(A–D)** The proportions of the four secondary structures in this region are bend + coil, β-sheet, helix, and turn.

TABLE 2 | Binding free energy components for the ACE2-RBD complex by using the MM-PB/SA method.

Component (kJ/mol)	WT	N354D	D364Y	V367F	Q498A
$\Delta \boldsymbol{E}_{vdw}$	−383.448	−375.295	−392.392	−406.108	−382.819
$\Delta \boldsymbol{E}_{ele}$	−811.617	−587.415	−873.228	−707.774	−686.019
$\Delta \boldsymbol{G}_{polar}$	1086.042	861.460	920.047	871.004	795.021
$\Delta \boldsymbol{G}_{nopolar}$	−47.215	−47.084	−47.618	−48.289	−47.176
$\Delta \boldsymbol{H}$ [a]	−156.238	−148.335	−393.191	−291.167	−320.993
$-\boldsymbol{T}\Delta \boldsymbol{S}$	70.627	96.277	94.628	90.315	98.288
$\Delta \boldsymbol{G}_{bind}$ [b]	−85.611	−52.085	−298.563	−200.852	−222.705

[a] $\Delta H = \Delta E_{vdw} + \Delta E_{ele} + \Delta G_{polar} + \Delta G_{nopolar}$.

[b] $\Delta G_{bind} = \Delta H - T\Delta S$.

these five systems, the relatively small non-polar solvation energy indicated that the packing of the cavity region is quite closed. Similar to the non-polar solvation energy, the van der Waals force in the five systems were not different from each other. The electrostatic interaction of the D364Y system is 62 kJ/mol, which is higher than that of the WT system, whereas the electrostatic interaction of the N354D system is much smaller than that of the other four systems. As a result, the electrostatic interaction was the main factor leading to the change in the binding free energy of the N354D and D364Y systems. For the V367F and Q498A systems, although the electrostatic interaction and polar solvation energy were reduced, the polar solvation energy decreased more dramatically and the binding free energy of these two systems (V367F and Q498A) was higher than that of the WT system. The interaction entropy was determined by the floating of the binding energy in the gas phase, and the WT system has the lowest entropy value, which is similar to the previous RMSD numerical trend. It is important to emphasize here that, in the Amber force field, electrostatic interactions were described by fixed-point charge interactions without considering the electrostatic polarization effect of protein. This may be the reason why the values of our combined free energy are not completely consistent with the experimental results.

In order to clarify the specific impact of mutations on the binding affinity of the system, we decomposed the binding free energy to each residue (**Supplementary Table S1**). In the RBD domains of all systems, the top amino acids that contributed to binding free energy were Lys and Arg, and their electrostatic energy contributions were much higher than those of other amino acids. The electrostatic interaction of these basic amino acids was the main reason for the stable combination of ACE2 and the RBD domain. For the N354D system, the polarity of the mutation point did not change, but an uncharged amino acid changed to a charged amino acid. The most obvious observation is that the electrostatic interaction at the 354 mutation site was significantly reduced, whereas the solvation free energy is only slightly reduced. Because of the mutation at site 354, the binding free energy contribution at this point changed from 0.52 to

80.07 kJ/mol. This should be the main reason for the significant reduction of electrostatic interaction in the N354D system. The mutation of charged Asp to uncharged Tyr in the D364Y system modified the electrostatic interaction at site 364 from 71.99 to 0.01 kJ/mol. In addition, the binding free energy of the seven amino acids contained in the random coil (Asn134-Glu140) of the system was increased, and the total binding energy was reduced to −16.51 kJ/mol (**Supplementary Table S1**), compared with −11.35 kJ/mol for the WT system. The polarity and electrification of the V367F system residue did not change, but the binding free energy was greater than that of the WT system. This may be related to the residue Tyr449 in the RBM region, and its polar solvation energy decreased to 13 kJ/mol. Similar to that of Tyr449, the polar solvation energy of other residues in the RBM region was also improved to different degrees. Because the mutated alanine in the Q498A system is a hydrophobic amino acid, the original strong polar solvation energy of this site is weakened. At the same time, the mutation of the Q498A system caused the loss of glutamine hydrogen donor, which also led to the loss of hydrogen bonds. The superimposed effect of electrostatic interaction and polar solvation energy resulted in an increase in the binding energy contribution of residue 498 by 7 kJ/mol. In addition, the random coil (Ala475-Gly485) mentioned above is an interesting structure considering that its binding free energy increased whether the mutation was located in the RBM region or not.

Hydrogen Bonds and Bonding Interfaces in RBM

The hydrogen bonds between the RBD and ACE2 were extracted using the Gromacs program with default criteria (D–A distance cutoff = 0.35 nm and H–D–A angle cutoff = 30°, where D, A and H are the donor atom, acceptor atom, and hydrogen atom linked to the donor atom, respectively), and we only retained the hydrogen bond pairs whose trajectory ratio were more than 20% (**Supplementary Table S2**). All the mutant systems had fewer hydrogen bonds than had the WT system; in particular, the V367F system only produced 14 hydrogen bonds. Lys417 is the only residue that was not located in the RBM region, interacting instead with Asp30 in the ACE2 domain to form hydrogen bonds and salt bridges. Moreover, the hydrogen bond of 417LYS-30ASP was found to account for more than 70% of all hydrogen bonds in all the investigated systems, indicating that this hydrogen bond is very important for the stability of the complex. Compared with the WT system, both the non-RBM group and the Q498A system lacked six sets of hydrogen bonds [34HIS (HE2)-494SER (O), 487ASN (D21)-24GLN (O), 498GLN (E21)-38ASP (OD1), 500THR (HG1)-355ASP (OD1), 478THR (HG1)-24GLN (OE1), and 24GLN (E21)-487ASN (OD1)]. The breaking of these hydrogen bonds has a certain relationship with the aforementioned changes in the secondary structure of the RBM region. Many studies have shown that hydrogen bonds play a crucial role in the structural stability of proteins, and the disappearance of hydrogen bonds can change the biological activity of protein (Joh et al., 2008). A new 493GLN (E21)-34HIS (O) hydrogen bond was generated in both the N354D and V367F systems after excluding the hydrogen bond contained in the WT system. The D364Y and Q498A systems lost the 353LYS (HZ1)-496GLY (O) hydrogen bond, but formed three new hydrogen bonds: 500THR (HG1)-355ASP (OD2), 353LYS (HZ1)-495TYR (O), and 83TYR (HH)-487ASN (ND2). Although mutations in the Q498A system resulted in the disappearance of all associated hydrogen bonds, the formation of new hydrogen bonds near position 498 kept the entire RBM region stable.

A residue is considered part of the interface if one of its atoms is within 0.4 nm from any atom of the other partner in at least 30% of the 10,000 MD simulation frames (10 ps as an interval). Taking this as the standard, we also analyzed the occupancy rate of the combined interface in the simulation process (**Supplementary Table S3**). We found that there is a binding loop (Tyr495-Tyr505) in the RBM region that is the main component of the binding interface. Then we visualize the binding interface of the binding loop in order to understand the specific details.

Hydrogen Bonds and Hydrophobicity of the Binding Loop

We obtained the representative structure of each system through the clustering method, and then used Ligplot+ to display the information of the complex binding interface, which helps us understand the details of the binding loop (Tyr495-Tyr505) (Laskowski and Swindells, 2011). In all systems, Lys353 of the ACE2 domain has the most hydrogen bond combinations at the binding loop, indicating that it is a key residue in the binding interface. We speculate that this may be related to the strong hydrophobic interaction between Lys353 and Tyr505, which makes the binding loop and Lys353 tightly bound together (**Figure 6A**). This is also consistent with previous findings that mutation in Lys353 caused a significant reduction in binding free energy (Lupala et al., 2020).

The binding loop of the N354D system (**Figure 6B**) of the RBM group generates the most hydrogen bonds and hydrophobic interactions, which is related to the changes in the adjacent random coil (Ala475-Gly485). In the D364Y system, the hydrogen bond between Lys353 and the binding loop of Tyr505 was disrupted, significantly increasing the distance of their hydrophobic interaction (**Supplementary Figure S2A**). Although no hydrogen bond was formed at site 498 in the binding interface of both the V367F (**Supplementary Figure S2B**) and Q498A (**Supplementary Figure S2C**) systems, the hydrophobic interaction at this site was found to increase. The free energies of the binding loop in WT, N354D, D364Y, V367F and Q498A were of 8.30, 5.30, 13.51, 6.38, and 8.11 kJ/mol, respectively. These findings indicate that mutations in the non-RBM group have a greater effect on the binding loop, with these mutations either increasing or decreasing the binding free energy.

Principal Component Analysis of the RBM Region

The first principal component (PC1) could reflect large-amplitude motions of the protein C-alpha conformations, as illustrated in the principal component analysis (PCA) for the

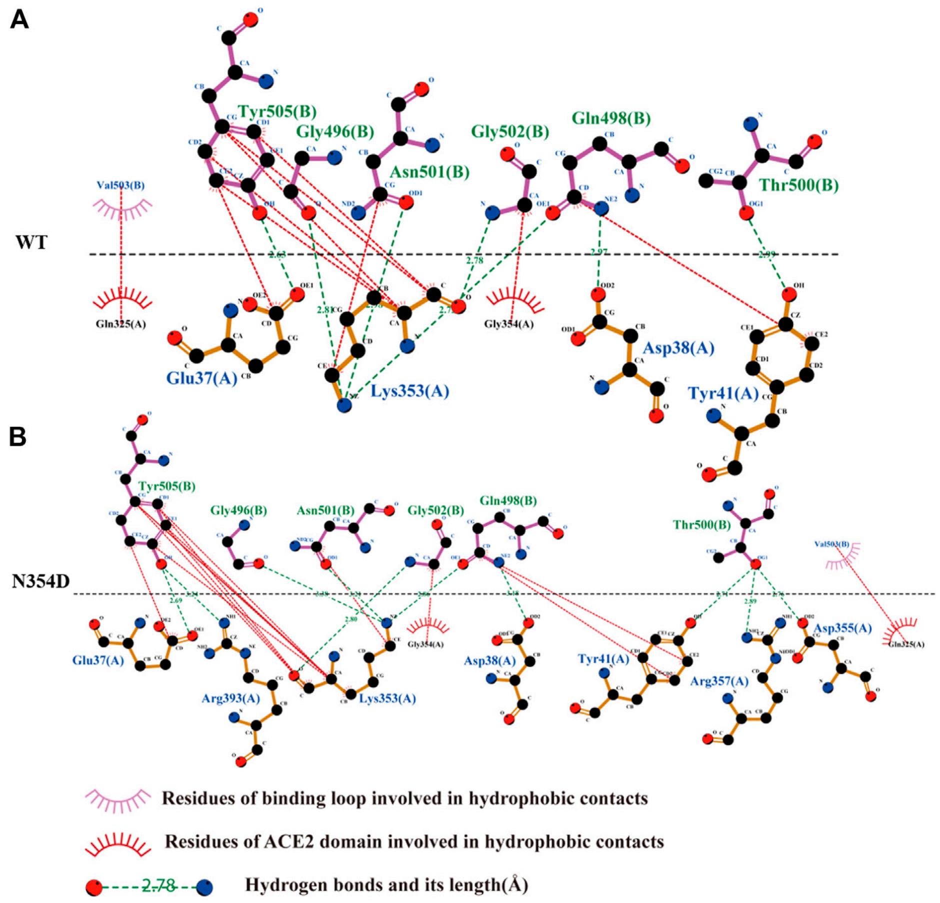


FIGURE 6 | Network of interactions associated with binding loop in RBD derived from Ligplot+ software analysis. **(A)** Analysis of hydrogen bond and hydrophobic interaction of binding interface in the WT system. **(B)** N354D system.

WT and mutant systems (**Figure 7**). The PCA explained how the mutations affect the motions. The direction of the motion is indicated by the direction of the arrow, and the magnitude of the motion is expressed by the length of the arrow. The RBM region (marked blue) in the N354D system has much less motion amplitude than has the WT system (**Figures 7A,B**), and the motion direction is disordered. The motion direction of the ACE2 domain at the interface is also irregular, which is the main reason for the decrease of binding energy of the system. In sharp contrast, the interface of ACE2 and RBD in the D364Y system has the same motion trend (**Figure 7C**), which makes the binding state between them very stable. The movement trend of the RBM region in the V367F system deviated from some angles (**Figure 7D**), which may be related to the overall decrease of the polar solvation energy in the RBM region. In the RBM region of the Q498A system (**Figure 7E**), the movement trend of most residues is similar to that of the WT system, but the movement trend of residues near the alanine mutation point is chaotic, which may be related to the rearrangement of water molecules around the mutation point. In summary, the motion of the RBM region in the complex is closely related to the binding free energy, and the motion direction and amplitude indirectly indicate the binding effect of ACE2 and RBD.

Cell–Cell Fusion Assay Analysis of the Binding Between Mutated Spike and ACE2

The Spike protein mutants N354D, D364Y, and especially V367F, which has been detected at high density (Gong et al., 2020; Song et al., 2020; Zhao et al., 2020), were detected in the early breakout area. We conducted cytological experiments to confirm the functional outcomes of V367F mutation on the binding of Spike protein with ACE2. WT and Q498A mutant were used as controls (**Figure 8**). A cell–cell fusion assay was used to mimic the SARS-CoV-2 virus spreading among ACE2-expressing cells (Kruglova et al., 2021; Zeng et al., 2021). This result showed that compared with the WT (RLU Ratio = 4.27 ± 0.54), the V367F mutation enhanced luciferase activity (RLU Ratio = 8.97 ± 0.91). This indicates that the affinity between Spike protein and human ACE2 is enhanced after V367F mutation. The Q498A mutation also increased luciferase activity (RLU Ratio = 8.36 ± 0.61), but it was slightly lower than the V367F mutation. This is slightly different from the simulation results, which may be due to the

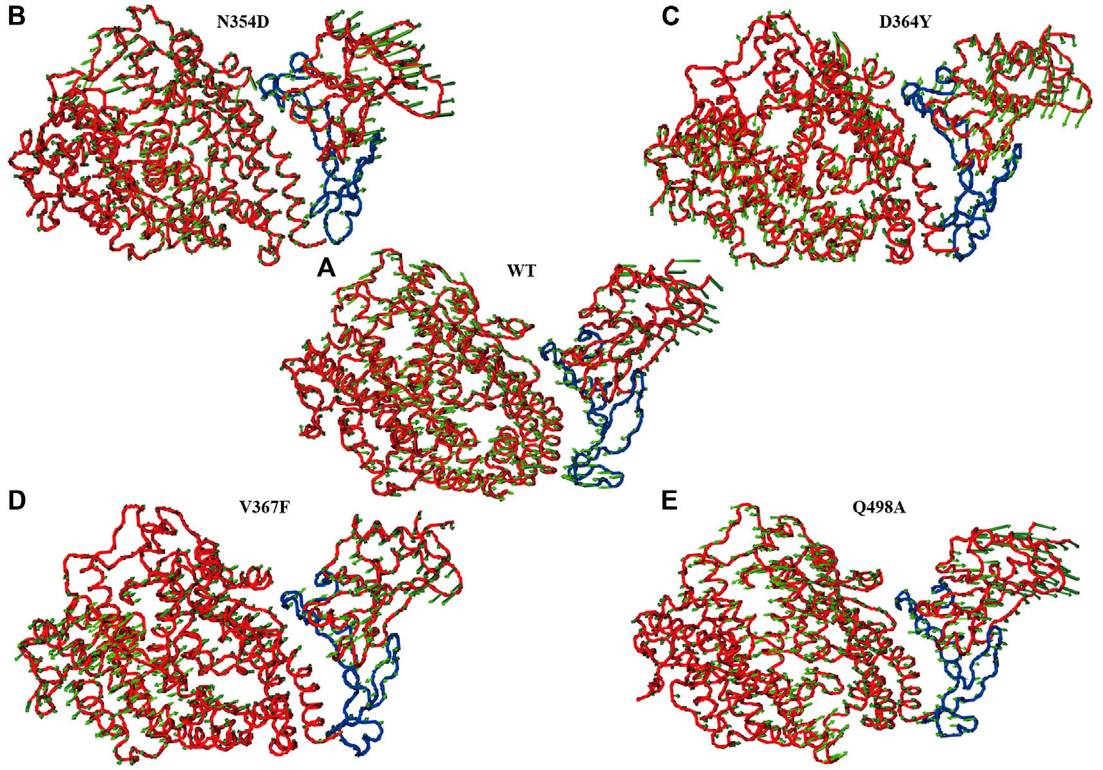


FIGURE 7 | Principal component analysis of each system. The arrow (green) plot of the first (PC1) motion modes for the RBM (blue) region in the WT system **(A)**, N354D system **(B)**, D364Y system **(C)**, V367F system **(D)**, and Q498A system **(E)**.

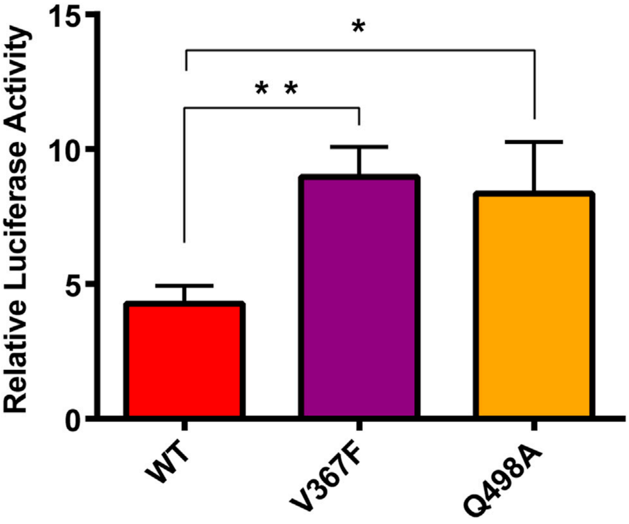


FIGURE 8 | Cell-cell fusion assay analysis of the binding between mutated spike and ACE2. HEK293T cells expressing with pcDNA3.1-spike/ mutant spike were co-incubated with cells expressing ACE2. After 48 h, Cells were lysis and the luciferase activity were detected. One-way ANOVA and Sidak's multiple comparisons test were performed by GraphPad PRISM 6. $^{*}p < 0.05$; $^{**}p < 0.01$.

influence of other sequences in Spike protein that do not include RBD region. Consistent with the result of our cell-cell fusion assay, recently, a study also found the enhanced affinity and infectivity of the V367F Spike mutant based on ELISA, SPR and the pseudovirus entry assay (Ou et al., 2021), implying the important role of V367F mutant in this epidemic strain.

CONCLUSION

Whether the mutation is located in the RBM region or not, it will change the overall stability of the complex. The secondary structure of the random coil Pro384-Asp389 of the N354D system and the V367F system has changed, which may indirectly affect the conformation of the RBM region. We also found a region (Ala475-Gly485) with obvious contrast, the secondary structure of the non-RBM group in this region was mostly changed from bend + coil to turn, whereas the Q498A system remained unchanged. Except for that of the N354D system, the binding free energy of the other three mutation systems is enhanced, and the electrostatic energy provides the greatest contribution. The chargeability of the base acid before and after the mutation in the N354D system and the D364Y system is different, which makes the electrostatic energy of the mutation site greatly changed, so the energy contribution change at this site is the main reason for the overall binding free energy change.

However, the V367F system of the non-RBM group is similar to the Q498A system. The decrease in polar solvation energy of most residues in the RBM region is the main reason for the increase in the overall binding free energy. We found that there are six extremely sensitive hydrogen bonds in the RBM region, which are broken in all mutation systems. However, the key hydrogen bonds that maintain the stability of the system have a high occupancy rate in the simulation process, such as the hydrogen bond 417LYS-30ASP located outside the RBM region. In addition, through the analysis of the binding interface of the binding loop, it can be seen that Lys353 of the ACE2 domain is tightly wrapped inside the binding ring and forms a strong hydrogen bond and hydrophobic effect. Therefore, we believe that Lys353 is an important residue in the binding interface, which is also consistent with the results of previous study. In the three-dimensional motion diagram of PCA, the movement direction of the RMB region in the N354D system is disordered and the motion amplitude is very low. The D364Y, V367F, and Q498A systems show little difference in the movement of the RBM region compared with the WT system. A cell–cell fusion assay was used to mimic the SARS-CoV-2 virus spreading among ACE2-expressing cells, thus the V367F and Q498A mutant displayed a significantly increased luciferase activity. In summary, this study reports the details of the changes in the binding of SARS-CoV-2 Spike protein RBD to the human ACE2 receptor by mutations in the non-RBM region. The non-RBM mutation will affect the secondary structure, hydrogen bonding, hydrophobic interaction and binding free energy, etc., and most of these effects act on the RBM region.

MATERIALS AND METHODS

Preparation of the Structures

The SARS-CoV-2-RBD/hACE2 complex was obtained from the RCSB website named 6LZG (https://www.rcsb.org/) (Wang et al., 2020). The complex structure was solved at 0.25 nm resolution with a SARS-CoV-2-RBD binding to a single hACE2 molecule in the asymmetric unit. For hACE2, clear electron densities could be traced for 596 residues from S19 to A614 of the N-terminal peptidase domain, as well as glycans N-linked to residues 53, 90, and 322. In the complex structure, the SARS-CoV-2-RBD contains 195 consecutive density-traceable residues, spanning T333 to P527, together with N-linked glycosylation at N343.

On the basis of the WT structure, the starting structures of mutants N354D, D364Y, V367F, and Q498A were generated by the PyMOL software (Mooers and Brown, 2021). Five initial structures were prepared for the subsequent study. All missing hydrogen atoms were added using the pdb2gmx module in the Gromacs package (Van Der Spoel et al., 2005). The AMBER99SB-ILDN force field in Gromacs was applied to produce the parameters for the system (Lindorff-Larsen et al., 2010). Sodium ions were added to keep the whole system neutral. The system was solvated with water in a truncated octahedron box with a 1.5 nm distance around the solute, and the TIP3P model was used to describe the water (Van Der Spoel and van Maaren, 2006).

Molecular Dynamics Simulations

MD simulations for all complexes were performed using the Gromacs package (Van Der Spoel et al., 2005). The particle mesh Ewald method was used to treat the long-range electrostatic interactions under periodic boundary conditions (Darden et al., 1993; Essmann et al., 1995). The short-range non-bonded interactions were calculated on the basis of a cutoff of 1 nm. The structures were initially fixed with a 1,000 kcal mol^{-1} nm^{-2} harmonic constraint, and both solvent and ions were energy minimized for 20,000 steps of the steepest descent method for each system. And then the systems were heated to 300 K in the NVT ensemble with velocity of 1 K ps^{-1}. Then the restrain was gradually decreased within 1 ns from 1,000 to 0 kcal mol^{-1} nm^{-1}. Finally, 100 ns MD simulations at a temperature of 300 K and a pressure of 1 atm were carried out without any restrain. The temperature was maintained at 300 K with the collision frequency 0.1 ps^{-1} using the Berendsen thermostat, and a constant isotropic press was maintained at 1 bar using the Berendsen barostat (Berendsen et al., 1984). All bonds involving hydrogen atoms were restricted using the SHAKE algorithm (Coleman et al., 1977; Ryckaert et al., 1977). A time step of 2 fs was used in all MD simulations.

MM/PBSA Calculations

As one of the most widely used methods, MM/PBSA has always played a significant role on calculation of binding free energy. In the MM/PBSA approach, the binding free energyΔG_{bind} was calculated as follows:

$$\Delta G_{bind} = \Delta H - T\Delta S = < E_{pl}^{int} > + \Delta G_{sol} - T\Delta S \tag{1}$$

where ΔH represents enthalpy change and $< E_{pl}^{int} >$ represents the ensemble averaged protein–ligand interaction including electrostatic interaction and van der Waals (vdw) interaction. ΔG_{sol} and $-T\Delta S$ represent the solvation free energy and contribution of entropy change, respectively. ΔG_{sol} can be divided into two parts:

$$\Delta G_{sol} = \Delta G_{pb} + \Delta G_{np} \tag{2}$$

where ΔG_{pb} and ΔG_{np} represent the polar and non-polar solvation free energy, respectively. ΔG_{pb} was calculated using the PB equation. The exterior and interior dielectric constants were set to 80 and 2, respectively. In the meantime, ΔG_{np} was calculated using the following equation:

$$\Delta G_{np} = \gamma \times SASA + \beta \tag{3}$$

where SASA represents solvent-accessible surface area, and it can be calculated using the g_mmpbsa program. The numerical values of γ and β are the standard values of 0.00542 kcal/(mol Å^2) and 0.92 kcal/mol, respectively (Sanner et al., 1996). For each system, the average binding free energy was calculated for 500 snapshots extracted from the last 5 ns of the trajectories at 10-ps interval for the complex structure. The MM/PBSA energy decomposition was performed to address the contributions of each residue to the binding free energy.

Interaction Entropy

In addition to the N mode, a new more rigorous and concise method, that is, the IE method, is employed to calculate entropy change. It can be defined as follows:

$$-T\Delta S = KT\ln < e^{\beta\Delta E_{pl}^{int}} > \tag{4}$$

where ΔE_{pl}^{int} represents the fluctuation of protein–ligand interaction energy (E_{pl}^{int}) around the average energy ($< E_{pl}^{int} >$). It can be calculated as follows:

$$\Delta E_{pl}^{int} = E_{pl}^{int} + < E_{pl}^{int} > \tag{5}$$

The protein–ligand interaction energy (E_{pl}^{int}) consists of electrostatic interaction and vdW interactions. The efficiency of this approach lies in the fact that the two averages $< E_{pl}^{int} >$ and $< e^{\beta\Delta E_{pl}^{int}} >$ can be calculated simply using the following equations:

$$< E_{pl}^{int} > = \frac{1}{N}\sum\nolimits_{i}^{N} E_{pl}^{int}(t_i) \tag{6}$$

and

$$< E_{pl}^{int} > = \frac{1}{N}\sum\nolimits_{i}^{N} e^{\beta\Delta E_{pl}^{int}(t_i)} \tag{7}$$

where β is $1/KT$.

In the IE method, the residue decomposition of entropy change (Wang et al., 2017) is performed using the following equations:

$$-\mathrm{T}\Delta S_{rl} = KT\ln < e^{\beta\Delta E_{rl}^{int}} > \tag{8}$$

where ΔE_{rl}^{int} represents the fluctuation of residue–ligand interaction energy (E_{rl}^{int}) around the average energy ($< E_{rl}^{int} >$).

Clustering and Hydrogen Bonding

The cluster analysis of protein conformations was carried out using Gromos as the clustering algorithm and all-protein atom RMSD as the similarity metric. The cluster analysis performed as follows: count the number of neighbors using a cutoff value, take the structure with the largest number of neighbors with all its neighbors as a cluster, and eliminate it from the pool of clusters. Repeat for the remaining structures in the pool (Daura et al., 1999).

Hydrogen bonds were determined via the distance between the D and A heavy atoms using a cutoff value of 0.35 nm and the angle H–D–A using a cutoff value of 30° (Van Der Spoel et al., 2005).

Principal Component Analysis

PCA is one of the most popular postdynamic techniques (Kurylowicz et al., 2010) that has been widely used to provide a better understanding of the dynamics of a biological system. It defines atomic displacement in a collective manner that transforms the original HD set of (possibly) correlated variables into a reduced set of uncorrelated variables—the principal components (PCs). The most significant fluctuation modes of a protein together with the motion of the system can be identified using PCA in terms of planarity of motion (eigenvectors) and its magnitude (eigenvalues) (Balmith and Soliman, 2017). The eigenvectors, also called PCs, give the direction of the coordinated motion of C-alpha atoms, and the eigenvalues represented the magnitude of the motion with the corresponding eigenvectors. ON the basis of the covariance matrix C_{ij} for coordinates i and j, the principal elements of the protein motion were computed as the eigenvectors and defined by the following ensemble formula:

$$C_{ij} = < (X_i - < X_i >)\big(X_j - < X_j >\big) >,$$

where $X_{i/j}$ are Cartesian atomic coordinates of the C-alpha atom i or j and $<X_i>$ and $<X_j>$ stood for the average coordinates derived from the MD simulation trajectory. The ProDy and VMD software were used to generate the PCA porcupine plot (Bakan et al., 2014).

Cell–Cell Fusion Assay

The Spike protein, hACE2, and TMPRSSR2 proteins were purchased from MiaoLing Plasmid Sharing Platform (Wuhan, China). The V367F and Q498A mutants were constructed using overlapping PCR and cloned into a pcDNA3.1 vector. The primers used are: pc3.1NheI-f ggagacccaagctggctagc; PC3.1XbaI-r gggtttaaacgggccctctaga; V367F-F gactactctttcctgtac aacagcgcctct; V367F-R tgtacaggaaagagtagtcggccacgca; Q498A-F acggcttcgcgcctacaaacggcgtgggc; Q498A-R ttgtaggcgcgaagccgtaag actggag. All the plasmids were sequenced after PCR.

The cell–cell fusion assay was conducted as previously reported (Yi et al., 2020). In brief, HEK293T cells were collected when they were ~90% confluent in 10-cm dishes and then seeded into 24-well plates. The next day, pcDNA3.1-S/mutant S and pcDNA3.1-luc-RE (the plasmid was reconstructed on the pCDNA3.1 backbone and the renilla luciferase-coding region is under the control of the T7 promoter) were cotransfected and transferred in another well for subsequent cotransfection of pcDNA3.1-ACE2, pcDNA3.1-TMPRSSR2 and pCAG-T7pol (addgene# #59926). After 4–6 h of culture at 37°C with 5% CO_2, the culture media were changed to DMEM (10% FBS), and the cells remained in culture for another 48 h. After 48 h of cotransfection, the two groups of HEK293T cells (Spike/mutant Spike and ACE2) were trypsinized and mixed at a 1:1 ratio and then plated on 96-well plates. The cells were further incubated at 37°C for 48 h, lysed with lysis buffer, and tested for luciferase activity (Promega, United States). The binding ability of mutants to ACE2 was analyzed by comparing the luciferase values (N = 3), the binding ability to ACE2 of mutants were analyzed by One-way ANOVA and Sidak's multiple comparisons test (Software: GraphPad PRISM 6), $*p < 0.05$; $**p < 0.01$.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

Conceptualization, YD, JL, and ZWa; Data curation, HW, LC, and ZWu; Formal analysis, YD and HW; Funding acquisition, YD, JL, and ZWa; Investigation, QF and BZ; Methodology, LC and LJ; Project administration, ZWa; Resources, LC and BZ; Software, ZWu and YY; Supervision, BC; Validation, QF and LC; Visualization, YD and YZ; Writing – original draft, YD, HW, and LC; Writing – review & editing, YD and HW. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the Emergency Project of Science and Technology of Zhejiang Province in China (2020C03128), the Medicine and Health Research Foundation of Zhejiang Province in China (2019RC012, 2019RC124, 2019KY017, WKJ-ZJ-2019), the National Natural Science Foundation of China (81772664, 81830071), the Natural Science Foundation of Zhejiang Province in China (LQ21H200007), and the Outstanding Young Scientific Research Funds of Zhejiang Provincial People's Hospital in China (ZRY 2019C008).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.614443/full#supplementary-material

REFERENCES

Azhar, E. I., El-Kafrawy, S. A., Farraj, S. A., Hassan, A. M., Al-Saeed, M. S., Hashem, A. M., et al. (2014). Evidence for Camel-To-Human Transmission of MERS Coronavirus. *N. Engl. J. Med.* 370, 2499–2505. doi:10.1056/NEJMoa1401505

Bakan, A., Dutta, A., Mao, W., Liu, Y., Chennubhotla, C., Lezon, T. R., et al. (2014). Evol and ProDy for Bridging Protein Sequence Evolution and Structural Dynamics. *Bioinformatics* 30, 2681–2683. doi:10.1093/bioinformatics/btu336

Balmith, M., and Soliman, M. E. S. (2017). Non-active Site Mutations Disturb the Loop Dynamics, Dimerization, Viral Budding and Egress of VP40 of the Ebola Virus. *Mol. Biosyst.* 13, 585–597. doi:10.1039/C6MB00803H

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular Dynamics with Coupling to an External bath. *J. Chem. Phys.* 81, 3684–3690. doi:10.1063/1.448118

Coleman, T. G., Mesick, H. C., and Darby, R. L. (1977). Numerical Integration. *Ann. Biomed. Eng.* 5, 322–328. doi:10.1007/BF02367312

Cui, J., Li, F., and Shi, Z.-L. (2019). Origin and Evolution of Pathogenic Coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi:10.1038/s41579-018-0118-9

Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald: AnN·Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397

Daura, X., Gademann, K., Jaun, B., Seebach, D., Van Gunsteren, W. F., and Mark, A. E. (1999). Peptide Folding: when Simulation Meets experiment. *Angew. Chem. Int. Ed.* 38, 236–240. doi:10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.0.CO;2-M

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* 103, 8577–8593. doi:10.1063/1.470117

Gong, Z., Zhu, J. W., Li, C. P., Jiang, S., Ma, L. N., Tang, B. X., et al. (2020). An Online Coronavirus Analysis Platform from the National Genomics Data Center. *Zool. Res.* 41, 705–708. doi:10.24272/j.issn.2095-8137.2020.065

Gralinski, L. E., and Menachery, V. D. (2020). Return of the Coronavirus: 2019-nCoV. *Viruses* 12, 135. doi:10.3390/v12020135

Guan, Y., Zheng, B. J., He, Y. Q., Liu, X. L., Zhuang, Z. X., Cheung, C. L., et al. (2003). Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China. *Science* 302, 276–278. doi:10.1126/science.1087139

Hu, J., He, C.-L., Gao, Q.-Z., Zhang, G.-J., Cao, X.-X., Long, Q.-X., et al. (2020). D614G Mutation of SARS-CoV-2 Spike Protein Enhances Viral Infectivity. *BioRxiv* 06, 161323. doi:10.1101/2020.06.20.161323202020

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *The Lancet* 395, 497–506. doi:10.1016/S0140-6736(20)30183-5

Joh, N. H., Min, A., Faham, S., Whitelegge, J. P., Yang, D., Woods, V. L., et al. (2008). Modest Stabilization by Most Hydrogen-Bonded Side-Chain Interactions in Membrane Proteins. *Nature* 453, 1266–1270. doi:10.1038/nature06977

Kruglova, N., Siniavin, A., Gushchin, V., and Mazurov, D. (2021). SARS-CoV-2 Cell-To-Cell Infection Is Resistant to Neutralizing Antibodies. *BioRxiv* 2021, 442701. doi:10.1101/2021.05.04.442701

Kurylowicz, M., Yu, C.-H., and Pomès, R. (2010). Systematic Study of Anharmonic Features in a Principal Component Analysis of Gramicidin A. *Biophysical J.* 98, 386–395. doi:10.1016/j.bpj.2009.10.034

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., et al. (2020). Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor. *Nature* 581, 215–220. doi:10.1038/s41586-020-2180-5

Laskowski, R. A., and Swindells, M. B. (2011). LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* 51, 2778–2786. doi:10.1021/ci200227u

Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu. Rev. Virol.* 3, 237–261. doi:10.1146/annurev-virology-110615-042301

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* 78, 1950–1958. doi:10.1002/prot.22711

Lupala, C. S., Li, X., Lei, J., Chen, H., Qi, J., Liu, H., et al. (2020). Computational Simulations Reveal the Binding Dynamics between Human ACE2 and the Receptor Binding Domain of SARS-CoV-2 Spike Protein. *BioRxiv* 2020, 005561. doi:10.1101/2020.03.24.005561

Mooers, B. H. M., and Brown, M. E. (2021). Templates for Writing PyMOL Scripts. *Protein Sci.* 30, 262–269. doi:10.1002/pro.3997

Ou, J., Zhou, Z., Dai, R., Zhang, J., Zhao, S., Wu, X., et al. (20212021). V367F Mutation in SARS-CoV-2 Spike RBD Emerging during the Early Transmission Phase Enhances Viral Infectivity through Increased Human ACE2 Receptor Binding Affinity. *J. Virol.*, JVI0061721. doi:10.1128/JVI.00617-21

Pahari, S., Li, G., Murthy, A. K., Liang, S., Fragoza, R., Yu, H., et al. (2020). SAAMBE-3D: Predicting Effect of Mutations on Protein-Protein Interactions. *Ijms* 21, 2563. doi:10.3390/ijms21072563

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* 23, 327–341. doi:10.1016/0021-9991(77)90098-5

Sanner, M. F., Olson, A. J., and Spehner, J. C. (1996). Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers.* 38, 305–320. doi:10.1002/(sici)1097-0282(199603)38:3<305::aid-bip4>3.0.co;2-y

Song, S., Ma, L., Zou, D., Tian, D., Li, C., Zhu, J., et al. (20202020). The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes in 2019nCoVR. *Genomics, Proteomics & Bioinformatics* -0229 (20), S1672–S1675. doi:10.1016/j.gpb.2020.09.00130131

Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C. K., Zhou, J., et al. (2016). Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol.* 24, 490–502. doi:10.1016/j.tim.2016.03.003

Tan, W., Zhao, X., Zhao, X., Ma, X., Wang, W., Niu, P., et al. (2020). A Novel Coronavirus Genome Identified in a Cluster of Pneumonia Cases - Wuhan, China 2019−2020. *China CDC Wkly* 2, 61–62. doi:10.46234/ccdcw2020.017

Tang, Q., Song, Y., Shi, M., Cheng, Y., Zhang, W., and Xia, X.-Q. (2015). Inferring the Hosts of Coronavirus Using Dual Statistical Models Based on Nucleotide Composition. *Sci. Rep.* 5, 17155. doi:10.1038/srep17155

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* 26, 1701–1718. doi:10.1002/jcc.20291

Van Der Spoel, D., and van Maaren, P. J. (2006). The Origin of Layer Structure Artifacts in Simulations of Liquid Water. *J. Chem. Theor. Comput.* 2, 1–11. doi:10.1021/ct0502256

Wang, C., Horby, P. W., Hayden, F. G., and Gao, G. F. (2020a). A Novel Coronavirus Outbreak of Global Health Concern. *The Lancet* 395, 470–473. doi:10.1016/S0140-6736(20)30185-9

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., et al. (2020b). Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* 181, 894–904. doi:10.1016/j.cell.2020.03.045

Wang, Y., Liu, J., Zhang, L., He, X., and Zhang, J. Z. H. (2017). Computational Search for Aflatoxin Binding Proteins. *Chem. Phys. Lett.* 685, 1–8. doi:10.1016/j.cplett.2017.07.024

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., et al. (2020). Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host & Microbe* 27, 325–328. doi:10.1016/j.chom.2020.02.001

Yi, C., Sun, X., Ye, J., Ding, L., Liu, M., Yang, Z., et al. (2020). Key Residues of the Receptor Binding Motif in the Spike Protein of SARS-CoV-2 that Interact with ACE2 and Neutralizing Antibodies. *Cell. Mol. Immunol.* 17, 621–630. doi:10.1038/s41423-020-0458-z

Zeng, C., Evans, J. P., King, T., Zheng, Y.-M., Oltz, E. M., Whelan, S. P. J., et al. (2021). SARS-CoV-2 Spreads through Cell-To-Cell Transmission. *BioRxiv* 2021, 446579. doi:10.1101/2021.06.01.446579

Zhao, W. M., Song, S. H., Chen, M. L., Zou, D., Ma, L. N., Ma, Y. K., et al. (2020). The 2019 Novel Coronavirus Resource. *Yi Chuan* 42, 212–221. doi:10.16288/j.yczz.20-030

Zhou, P., Yang, X.-L., Wang, X.-G., Wang, B., Zhang, L., Zhang, W., et al. (2020). A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature* 579, 270–273. doi:10.1038/s41586-020-2012-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 20 August 2021
doi: 10.3389/fmolb.2021.639614

Molecular Dynamics Simulations Reveal the Interaction Fingerprint of Remdesivir Triphosphate Pivotal in Allosteric Regulation of SARS-CoV-2 RdRp

*Mitul Srivastava†, Lovika Mittal†, Anita Kumari† and Shailendra Asthana**

Translational Health Science and Technology Institute (THSTI), Faridabad, India

The COVID-19 pandemic has now strengthened its hold on human health and coronavirus' lethal existence does not seem to be going away soon. In this regard, the optimization of reported information for understanding the mechanistic insights that facilitate the discovery towards new therapeutics is an unmet need. Remdesivir (RDV) is established to inhibit RNA-dependent RNA polymerase (RdRp) in distinct viral families including Ebola and SARS-CoV-2. Therefore, its derivatives have the potential to become a broad-spectrum antiviral agent effective against many other RNA viruses. In this study, we performed comparative analysis of RDV, RMP (RDV monophosphate), and RTP (RDV triphosphate) to undermine the inhibition mechanism caused by RTP as it is a metabolically active form of RDV. The MD results indicated that RTP rearranges itself from its initial RMP-pose at the catalytic site towards NTP entry site, however, RMP stays at the catalytic site. The thermodynamic profiling and free-energy analysis revealed that a stable pose of RTP at NTP entrance site seems critical to modulate the inhibition as its binding strength improved more than its initial RMP-pose obtained from docking at the catalytic site. We found that RTP not only occupies the residues K545, R553, and R555, essential to escorting NTP towards the catalytic site, but also interacts with other residues D618, P620, K621, R624, K798, and R836 that contribute significantly to its stability. From the interaction fingerprinting it is revealed that the RTP interact with basic and conserved residues that are detrimental for the RdRp activity, therefore it possibly perturbed the catalytic site and blocked the NTP entrance site considerably. Overall, we are highlighting the RTP binding pose and key residues that render the SARS-CoV-2 RdRp inactive, paving crucial insights towards the discovery of potent inhibitors.

Keywords: SARS-CoV-2 RNA-dependent RNA polymerase (RdRp), RDV triphosphate (RTP), molecular dynamics simulation, PCA, FEL, NTP entrance site, allosteric inhibition, remdesivir

Edited by:
S. Chakra Chennubhotla,
University of Pittsburgh, United States

Reviewed by:
Dhruv Kumar,
Amity University, India
Brijesh Rathi,
University of Delhi, India

****Correspondence:***
Shailendra Asthana
sasthana@thsti.res.in

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *09 December 2020*
Accepted: *24 June 2021*
Published: *20 August 2021*

Citation:
Srivastava M, Mittal L, Kumari A and Asthana S (2021) Molecular Dynamics Simulations Reveal the Interaction Fingerprint of Remdesivir Triphosphate Pivotal in Allosteric Regulation of SARS-CoV-2 RdRp. Front. Mol. Biosci. 8:639614. doi: 10.3389/fmolb.2021.639614

INTRODUCTION

SARS-CoV-2 genome consists of a (+) ssRNA virus and its RNA-dependent RNA polymerase (RdRp) which help in viral replication and synthesis of its genome (Eastman et al., 2020; Gao et al., 2020). RdRp has a high sequence and structural conservation that makes it an attractive drug target for various RNA viruses diseases (Mittal et al., 2019, 2020; Gao et al., 2020; Maddipati et al., 2020).

Currently, RDV is one of the most promising drugs for COVID-19 being evaluated in phase III clinical trials in China and the United States (Eastman et al., 2020). It was originally developed to treat Ebola virus and possess a broad-spectrum activity (Babadaei et al., 2020). RDV is a prodrug that consists of nucleoside analogue scaffold and it is metabolised into an alanine metabolite (GS-704277) inside the cells, then gets converted into the monophosphate derivative (RMP) and eventually into the active nucleoside triphosphate derivative (RTP) (Eastman et al., 2020). The nucleotide analog drugs have been a cornerstone of antiviral and anticancer therapy that target viral polymerases that cause chain termination. Nucleotide analogs are not highly permeable to cells, and once within the cell they need di- and then tri-phosphorylation to generate the nucleoside triphosphate (NTP) which could be used as a mimic of natural NTPs (competing the natural substrates) of genome replication by viral RNA-dependent polymerases (Chen et al., 2014; Eastman et al., 2020). Such inhibitors will initially reach the hepatic cells as non-phosphorylated or monophosphate prodrugs, after which cellular kinases turn them into active triphosphate (Fung et al., 2014). The viral RdRp will then mis-incorporate these incoming inhibitors as NTPs into their viral RNA strand, eventually leading to the chain termination process (Eastman et al., 2020). Generally, nucleoside TP analogs lack a 3′OH group that causes complete chain termination; however, RTP has this 3′OH group but also consists of a 1′CN group that delays the mechanism of chain termination (Eastman et al., 2020). The 1′CN group of RTP is reported to sterically clash with residue S861 of RdRp that affects the chain elongation step (Eastman et al., 2020).

Recently, the crystal structure of SARS-CoV-2 RdRp in complex with RMP, two cofactors NSP7 and NSP8, and primer-template strands were released in Protein Data Bank (PDB-ID: 7BV2), showing intermolecular interactions between RMP, primer-template, and RdRp (Kato et al., 2020). It was determined using cryo-electron microscopy which inhabits all 932 amino acids (Kato et al., 2020; Yin et al., 2020). It contains an N-terminal β-hairpin (residues 31–50) and an extended nidovirus RdRp-associated nucleotidyl-transferase domain (NiRAN, residues 115–250), composed of seven folded helices with three β-strands. Additionally, there is an interface domain (residues 251–365), which consists of three helices and five β-strands, that is further connected to the RdRp domain (residues 366–920) (Yin et al., 2020) (**Supplementary Figure S1**). The RdRp is surrounded by three subdomains: a fingers subdomain (residues L366 to A581 and K621 to G679), a palm subdomain (residues T582 to P620 and T680 to Q815), and a thumb subdomain (residues H816 to E920) (**Supplementary Figure S1**). It also contains seven evolutionary conserved motifs, such as in other HCV, DENV, BVDV, and poliovirus (PV) RdRps, which perform specific and essential functions during polymerization (Gong and Peersen, 2010; Appleby et al., 2015; Mittal et al., 2020). These are termed as motifs A–G, out of which the finger domain consists of two important motifs, G (residue 499–511) and F (residue F544-560), whereas the palm domain contains a catalytic core of the RdRp and four conserved motifs A to D, and partially motif E (motif A: residues 611–626; motif B: residue 678–710, motif C: residues 753–767, motif D: residues 771–796, motif E: residues 810–820) (Zhang et al., 2020). The motif A contains the conserved residue D618 which is classic divalent-cation–binding and motif C consists of conserved catalytic residues ^{759}SDD^{761} (Yin et al., 2020). Motifs F (residues K545 and R555) and G (residues K500 and S501) are involved in interaction with RNA template strand and escort it to the active site (catalytic site) (Yin et al., 2020). The residues involved in RNA binding as well as the residues comprising the catalytic active site are highly conserved, highlighting the preserved mechanism of RdRp genome replication in these various RNA viruses, and indicating that wide spectrum antiviral inhibitors could be created.

The arrival of SARS-CoV-2 RdRp crystal structures of bound RMP (PDB ID: 7BV2) and APO (PDB ID: 7BV1) has provided an opportunity to elucidate their complex behavior and to understand the comprehensive inhibition mechanism. The recent studies are based on understanding the dynamicity of the modeled RdRp and its interaction with either RDV (Koulgi et al., 2020) or RTP (Zhang and Zhou, 2020). In the modeled RdRp, RDV and RTP are stated to interact strongly at the catalytic site and NTP entry sites, respectively (Zhang and Zhou, 2020). However, neither RDV nor RMP are the active forms but RTP is known to completely inhibit the polymerization activity of RdRp (Yin et al., 2020). Hence, our aim is to understand the interaction pattern of RTP when it binds to the target in comparison with RDV and RMP as it is the active form and is able to render the RdRp inactive. Therefore, our objective is to elucidate the mechanistic details of RTP. With this objective, we are aiming to provide a comparative insight in terms of interaction fingerprinting to identify the key residues and their pivotal role in the mechanism of inhibition. Their dynamic and energetic contributions were mapped through MD simulations in terms of dynamical and thermodynamical components. Further, we have implemented the PCA and FEL analysis to highlight the major conformational changes and consequently the porcupine plot distinguished the activity of these molecules on the basis of differential atomic motions. The comprehensive analysis offers a rational for RTP being the active state of RDV as it fully overshadows the NTP entry channel and gains favorable contact with key residues which are crucial for NTP entry and other residues, namely K545, R553, R555, D618, P620, K621, R624, K798, and R836. These residues contribute significantly in holding RTP at the NTP entrance site and therefore perturbing the replication process. Apart from these residues the conservation in allosteric sites and SiteMap analysis was also carried out in BVDV, DENV, and HCV RdRps that could further be exploited. Based on these observations and understanding that highlight the mechanism of inhibition as well as the structural level insights, further structural-functional studies are recommended. The study provides scope to target these residues for the discovery of potential non-nucleoside inhibitors (NNIs).

MATERIALS AND METHODS

System Preparation

The protein structure of the SARS-CoV-2 RdRp was retrieved from the RCSB protein data bank (PDB ID: 7BV2, 7BV1). We also included crystal structures of HCV, BVDV, and DENV RdRp (PDB-ID:4NLD (for T1 and P2 site), 3FRZ (for T2 site), 4EAW (P1 site), 5I3Q (Thumb and Palm junction), and 3FRZ, 5IQ6, and 1S48 (for Template entrance site) (Mittal et al., 2019, 2020; Maddipati et al., 2020) to identify the probable allosteric sites in RdRp protein employed on the basis of a literature survey. The protein structure was prepared using the Protein Preparation Wizard module of Maestro (Sastry et al., 2013; Anang et al., 2018) (Schrödinger release 2020–1: Maestro, Schrödinger, LLC, New York, NY, 2020.) (Sarkar et al., 2021). OPLS3 (Jorgensen et al., 1996) force field model was used for preparation (Jorgensen et al., 1996). It has been found that the crystal structure has several breaks from K50-E84, D100-R118 (regions of NiRAN domain), and M906-E919 (region of RdRp domain). To maintain the structural integrity during dynamics, all loop breaks were interpolated using the PLOP algorithm (Srivastava et al., 2018). The loop conformations were optimized and again the structure was minimized. In the crystal structure, RDV monophosphate (RMP) is bound to RdRp rather than the RDV triphosphate (RTP), where the later is known to be the active form. Hence, RDV along with its metabolites RMP and RTP were prepared using LigPrep module for further study (LigPrep, version 4.2; Schrödinger, LLC, New York, NY, 2020–1). The optimization was done using the OPLS3 force field (Jorgensen et al., 1996).

Molecular Docking

To understand the broader spectrum of the catalytic site as well as the NTP entry site, we included RDV as a parent molecule in our study. Since no bound crystal with RDV is yet reported, we performed targeted docking (Asthana et al., 2013; Asthana et al., 2014) of RDV as well as RTP on the grid coordinates of RMP (bound in crystal 7BV2). Afterwards, to identify different binding modes, focused docking was performed by taking the best conformer achieved from targeted docking. We have included targeted docking as initial coordinates of RMP were available and afterwards, to enrich the conformations and to explore any other binding modes, the focused docking was implemented. The docking experiments were performed using AUTODOCK4.2 (Trott and Olson, 2010; Tyagi et al., 2020; Singh et al., 2021).

Molecular Dynamic Simulations

All-atom MD simulations of each system were performed by using AMBER16 software suite (Case et al., 2005; Tyagi et al., 2021). The systems were initially subjected for 100ns each while the docked system of RTP was further extended to 100ns to obtain its convergence. The systems (COM-RMP, COM-RDV, and COM-RTP) were simulated, including catalytic site Mg^{2+} ions. Model systems by Allner et al. and Carlson et al. were used for ions and triphosphate group, respectively (Meagher et al., 2003; Allnér et al., 2012). The parameters for RTP compatible with force fields were derived from the RED server (Vanquelef et al., 2011). The systems were solvated inside an orthorhombic box of TIP3P waters with a 12 Å padding in each direction. In leap, AMBERff14SB force field (Maier et al., 2015) was assigned to proteins, hydrogens were added, and counter ions were added to neutralize the system. The systems were minimized in three steps for system relaxation and removal of bad contacts. In the first step, 2000 steps minimization was implemented comprising 1,000 steps of steepest descent method followed by 1,000 steps of conjugate gradient. The first and second stages involved the position restraints of 10 and 2 kcal mol^{-1} $Å^{-2}$, respectively, on the whole protein systems to relax the solvent molecules. The third stage of minimization was carried out unrestrained. The SHAKE algorithm is used to constrain all bonds involving hydrogen atoms. The systems were heated for 20 ps from 0–300 K and equilibrated for 500 ps at 300 K and 1 atm pressure. Thereafter, the simulations were performed in the NPT ensemble at a time step of 2.0 ps. The coordinates were saved at every 20 ps and are referred to as 'frames' in this study. The details of the simulated systems are mentioned in **Supplementary Table S1**. For post-MD analysis, CPPTRAJ module was used for quantifying the RMSD (root mean square deviation) of backbone atoms, RMSF (root mean square fluctuation) of Cα atoms, SASA (solvent accessible solvent area), and rGyr (radius of gyration). For the calculation of these values, the starting frame, which represent the crystal pose, was chosen as the reference frame. The hydrogen bonds (HBs) calculation was implemented on the stable tarjectory using HBonds Tcl script with the criteria 3.5 Å for donor-acceptor distance and 180° for the angle between acceptor, hydrogen, and donor atoms.

Free Energy Analysis Through MM-GBSA/ PBSA Methods

The average binding energy was calculated for equilibrated MD trajectories for COM systems using MM-GBSA/PBSA approaches in Amber16 (Suri et al., 2014; Guo et al., 2012; Chen et al., 2016). For this, the 500 frames were extracted at equal intervals from the last 25 ns trajectory. The binding free energy (ΔG_{bind}) on each system is evaluated as follows:

$$\Delta G_{bind} = G_{com} - \left(G_{rec} + G_{lig}\right) \quad (1)$$

where G_{com}, G_{rec}, and G_{lig} are the absolute free energies of complex, receptor, and ligand respectively, arranged over the equilibrium trajectory. The free energy, G, for each species can be calculated by using MM-GBSA and MM-PBSA approaches as follows:

$$G = E_{gas} + G_{sol} - TS \quad (2)$$

$$E_{gas} = E_{int} + E_{ele} + E_{vdw} \quad (3)$$

$$G_{polar,PB(GB)} = E_{ele} + G_{sol-polar,PB(GB)} \quad (4)$$

$$G_{non-polar,PB(GB)} = E_{vdw} + G_{sol-np,PB(GB)} \quad (5)$$

$$G_{sol} = G_{PB(GB)} + G_{sol-np} \quad (6)$$

$$G_{sol-np} = \gamma SAS \quad (7)$$

where T and S are the temperature and total solute entropy, respectively. E_{gas} describes the gas phase energy and is the sum of internal energy (E_{int}), van der Waals interaction energy (E_{vdw}), and electrostatic interaction energy (E_{ele}). It is calculated using the parameters described in the AMBERff14SB force field (Maier et al., 2015). G_{sol} is the solvation free energy that can be decomposed into polar and non-polar contributions. $G_{sol\text{-}polar,PB(GB)}$ is the polar solvation contribution calculated by solving the Poisson-Boltzmann (PB) and Generalised-Boltzmann (GB) equations (Genheden and Ryde, 2015). Total polar interaction contributions ($G_{polar,PB(GB)}$) were obtained by sum of the E_{ele} and ($G_{sol\text{-}polar,PB(GB)}$) (Asthana et al., 2014). $G_{sol\text{-}np}$ is the non-polar solvation contribution that was estimated using 0.0072 kcal mol^{-1} $Å^{-2}$ (value of constant γ) and by determining the solvent-accessible surface area (SAS) using a water probe radius of 1.4 Å (Sitkoff et al., 1994). The dielectric constants for solute and solvents were set to 1 and 80 respectively. Total non-polar interaction contributions ($G_{non\text{-}polar,PB(GB)}$) were calculated by E_{vdw} and $G_{sol\text{-}np,PB(GB)}$ (Asthana et al., 2014). The binding free energy contribution for each residue in protein-ligand complex formation was computed using the GB model using the same number of frames (Suri et al., 2015; Srivastava et al., 2018).

Electrostatic Potential Calculations

The calculations were done using the APBS program (Baker et al., 2001) in VMD (Humphrey et al., 1996; Mittal et al., 2019) for final complexes achieved by dynamics. The protonated (.pqr) file for both proteins was generated using pdb2pqr module (Dolinsky et al., 2004) and the iso contour value of (+5 kTe^{-1}) and (−5 kTe^{-1}) was taken for positive and negative potentials, respectively, to generate the iso-surface of the protein.

Principal Component Analysis and Free Energy Landscape

PCA was implemented to undermine the conformational changes in RdRp protein and especially essential motions were captured in APO and complex systems. The PCA analysis was performed to highlight the essential motions in the proteins (Tripathi et al., 2016; Khan et al., 2017; Mittal et al., 2021). The PCA calculations were performed using the CPPTRAJ program in amber tools and performed over Cα atoms for all the equilibrated systems. In our study, PCA analysis was performed on a stable trajectory after 40 ns. The elements of the positional covariance matrix C is calculated which is defined by the equation below:

$$C_i = < \left(q_i - \langle q_i \rangle\right) \cdot \left(q_j - < q_j >\right) > (i, j, = 1, 2, \ldots .3N) \tag{8}$$

where q_i and q_j are the Cartesian coordinates of all Cα atoms and N is the number of Cα atoms used in building the matrix C. The < > sign indicates the ensemble average of the atomic positions in the Cartesian space. The resulting principal components (eigenvectors) were ranked by the corresponding total motion captured by them. In the end, there will be 3N eigenvectors generated, where N is the number of Cα atoms used in the building matrix. Porcupine graphs were generated using the "nmode" files obtained after PCA analysis. It basically provides the residue-based mobility plots for each eigenvector. Further, the free energy landscape was performed to explore the conformation change of proteins based on the PCA by using the 'g_sham' module (Tripathi et al., 2016; Khan et al., 2017). The free energy, ΔG(X), is calculated by Equation;

$$\Delta G(X) = -K_B T \, \text{In} \, P(X) \tag{9}$$

where K_B is the Boltzmann constant and T is absolute temperature, X represents the PCs, and P(X) is the probability distribution of the conformation ensemble along the PCs.

Computational Alanine Scanning

To confirm the *hot-spot* amino acids in the RTP complex, we have performed computational alanine scanning for the residues highlighted in residue-wise energy decomposition results. The calculations were run on 100 frames from the last 100 ns of MD trajectory by using the MM-GBSA approach. In this method, an amino acid of interest is replaced with alanine and relative binding free energy is recalculated. Finally, the difference in the binding free energies of the wild type and mutant, $\Delta\Delta G_{bind}$, was computed as follows:

$$\Delta\Delta G_{bind} = \Delta G_{bind}[\text{Wild Type}] - \Delta G_{bind}[\text{Mutant}] \tag{10}$$

Negative values of ΔΔG bind indicate the favorable contributions of residues in wild type while positive values indicate the unfavourable contributions. The mutant models of all the *hot-spot* residues were generated by using the maestro module.

SiteMap

To map the potential allosteric sites, the SiteMap (Halgren, 2009) module in the Schrodinger suite was used. It identifies putative binding sites by implementing different parameters. The different parameters on the basis of which a potential binding site is considered are: *site score, size, exposure score, enclosure, hydrophobic/hydrophilic character, contact,* and *donor/acceptor character*. As per Halgren's analysis, the average number of sites for sub-micromolar sites is 132, where lower exposure scores of 0.52 and higher exposure scores of 0.76 on average are considered better for sub-micromolar sites. For the donor/acceptor character and site score, the average for the sub-micromolar sites is 0.76 and 1.01, respectively. Druggability of the site is denoted by Dscore. Dscore values provide a rough estimate of whether the site is druggable. These scores were derived by Halgren (2009) by executing the SiteMap program on a number of proteins that have inhibitors bound with potencies in the sub-micromolar range and performing statistical analyses to produce optimized scores. The OPLS-2003 force field (Jorgensen, 2002) was employed, and a standard grid was used with 15 site points per reported site and cropped at 4.0 Å from the nearest site point.

Figures

All the images were generated using VMD (Humphrey et al., 1996) and graphs were plotted using XMGRACE (Turner, 2005. XMGRACE, Version 5.1.19. Center for Coastal and Land-Margin

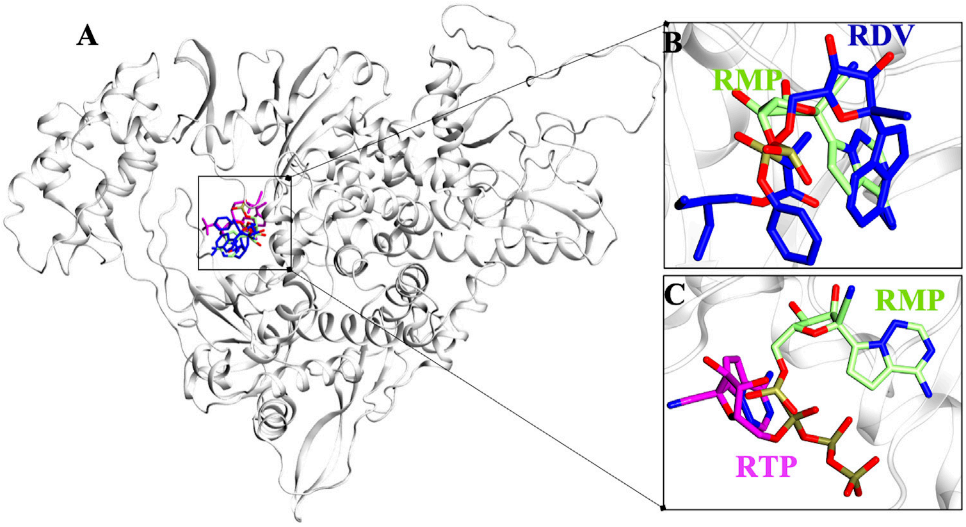


FIGURE 1 | Possible binding pose of RDV and RTP revealed through molecular docking. **(A)** Superimposition of docked pose of RDV and RTP on RMP-crystal pose. The inset shows the docked pose of **(B)** RDV on RMP bound site and **(C)** RTP on RMP bound site. Protein is represented in new cartoon and white in color. RMP, RDV and RTP are displayed in lime, blue and magenta and rendered in licorice atom-wise.

Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR; 2005) (Mittal et al., 2020).

RESULT AND DISCUSSION

Remdesivir and Remdesivir Triphosphate Binding Mode

Molecular docking has revealed the most likely binding pose of RDV and RTP. Targeted docking followed by focused docking has provided their final binding pose at the RMP bound site. The cluster representative of −7.5 kcal/mol and −8.9 kcal/mol docking energy were found for RDV and RTP, respectively (**Figure 1A**). The superimposition of RMP with docked complex of RDV and RTP has revealed that the obtained poses were close to RMP as the RMSD divergence from the common backbone was less (RDV ~0.4 Å and RTP ~1.0 Å) (**Figures 1B,C**).

MD Simulation Reveals Stability of Remdesivir monophosphate, Remdesivir, and Remdesivir Triphosphate in SARS-CoV-2 RdRp

Protein structures must be converged in order to achieve robust results from MD simulations. The quantification of RMSD, RMSF, and Radius of gyrations are good indicators of the overall stability of any protein system. The APO-vs.-COM RMSD comparison reveals that the systems COM-RMP and COM-RDV are achieving stability after ~20 ns of simulation while system COM-RTP was evolving and did not attain stability initially. The extension of COM-RTP simulation reveals that the system was stable in its next production stretch of 200 ns (**Figure 2A**). However, to confirm COM-RTP stability, we further decided to perform a triplicate study of this particular system. We found that all the different runs of COM-RTP have shown the same trend where the trajectory was evolving till 80ns while afterwards the system has attained the convergence till 200 ns (**Figure 2A**). In all systems, the divergence shift from 0 to ~4 ns is mainly due to the flexible nature of loops (residues 35–100), which has been found to be inconsistent during dynamics. These regions correspond to the NiRAN domain that does not exist in crystals and hence were interpolated during protein preparation. However, their dynamicity is frozen at a certain time-step in all systems, which is satisfactory and eventually makes systems stable (**Figure 2A**). The RMSF analysis mainly snapped two highly fluctuating regions, Region1-residues 30 to 130, which majorly belong to the interpolated flexible loops of NiRAN domain and another Region2-residues 875 to 900, which is the C-terminal and was also flexible in nature (**Figure 2B**). In all systems, respective ligands have granted stability to the protein which is reflected in RMSF values, as in many parts COMs are more stable than APO.

Along with these values, Radius of gyration also highlights that the compactness of COM systems is greater than APO systems (**Figure 2C**). Further, ligand independent stability was analysed in all systems. The RMP was bound in crystal and its stability was well comprehended by MD simulations (**Supplementary Figure S2**). We then observed movements of both RDV and RTP. As both were docked poses, their fluctuation was quite justified; however, it was observed that both were trying to gain favorable interactions at respective pockets. We found that RDV achieved complete stability after ~40 ns (**Supplementary Figure S2**). A separate introspective RTP stability in triplicate systems was analyzed. In all systems, it was found that RTP's docked pose was fluctuating initially. However, later it maintained interaction with the Mg^{2+} ion and thus gained stability in the later phase of simulation (**Figure 3A**). We also curated the movement of RTP throughout the simulation timeline and visualised the same in

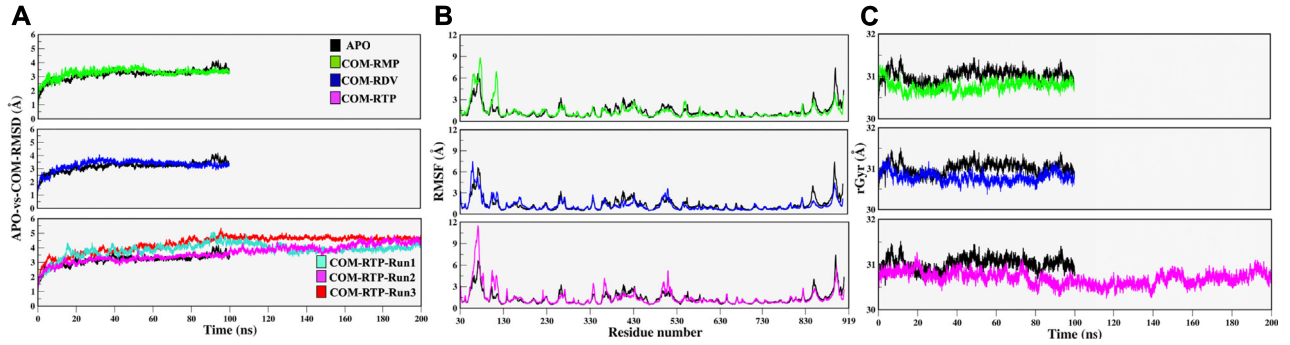


FIGURE 2 | Estimation of systems stability through Molecular Dynamics Simulation over different time scale. **(A)** Root mean square deviation (RMSD) comparison of APO with COM-RMP, COM-RDV and COM-RTP **(B)** Root mean square fluctuations (RMSF) of RdRp residues in different systems, and **(C)** Compactness of all systems assessed through Radius of gyration. All quantitative values are measured in angstrom (Å) unit.

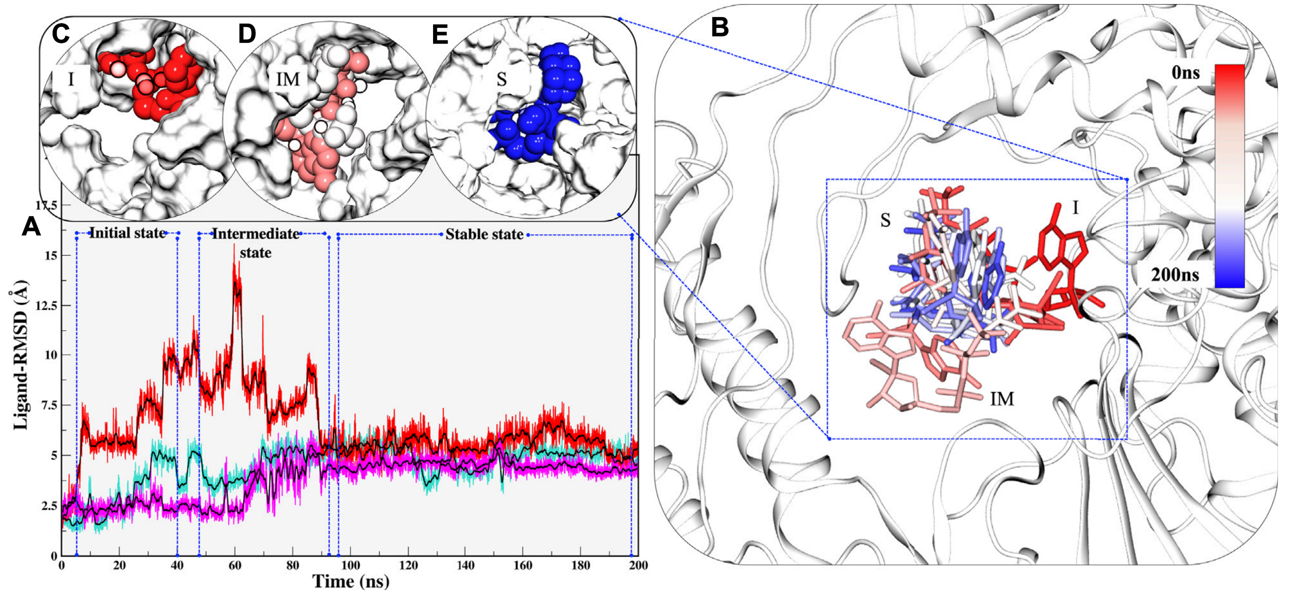


FIGURE 3 | Estimation of different RTP states at MD scale. **(A)** RTP stability in different runs, Run1: Turquoise, Run2: Magenta and Run3: Red. **(B)** Three different states were observed, I: Initial state, IM: Inter-mediate state and S: MD stable state, where RTP is shown in licorice and protein in new cartoon. The conformations of RTP are shown in licorice with time-step coloring method, RTP conformation shown in vdW and protein is represented in surf at **(C)** I state **(D)** IM state and **(E)** S state.

MD timestep (**Figure 3B**). We found that RTP movement provided three different states: initial state (I) was the docked pose which, after some time, starts fluctuating and captures itself at intermediate (IM) state and eventually RTP settles itself at its final MD stable state (S) (**Figure 3B**). We tried to understand the significance of different binding poses at each time step. We found that at I-state RTP is at the catalytic site while at the IM-state RTP is found at the edge of NTP channel (**Figures 3C,D**). However, MD stable state (S) is found at the centre of NTP channel which might occlude its major portion (**Figure 3E**) (complete significance of RTP-MD pose is explained in detail in the mechanism of inhibition section). We also found that the ligand RDV found itself at the junction of NTP entry channel and catalytic site while RTP tail favoured NTP entry channel site (explained later). The SASA distribution plot of the binding site reveals that COM-RTP system is least exposed to solvent as compared to other systems (**Supplementary Figure S3A**). COM-RDV was an exception when compared with APO. This might be attributed to RDV's self-fluctuation which is higher than RMP and RTP movement in the binding pocket. Adding further insights into ligands' conduct in the binding pocket, we measured the binding site RMSD. The binding site residues flexibility is reduced upon ligand binding as compared to APO which is a good starting indication as the study cements upon understanding the mechanistic details of the binding pocket (**Supplementary Figure S3B**). COM-RTP was notably higher

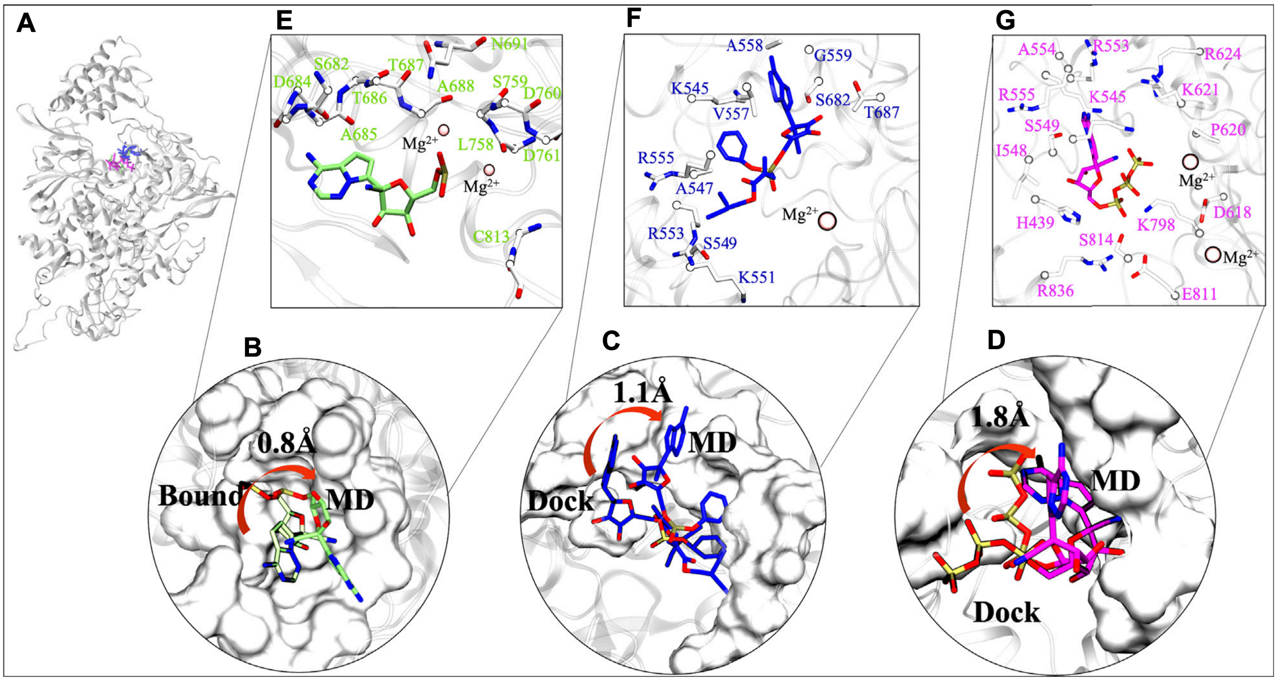


FIGURE 4 | Comparative interaction patterns of RMP, RDV and RTP MD poses. **(A)** MD poses of RMP, RDV and RTP in RdRp. **(B)** RMP's comparison between its bound and MD representative state **(C,D)** RDV and RTP comparison between docked and MD representative state, respectively. **(E–G)** Arrangement of these molecules at their most stable state, RMP (snapshot: 97.51 ns), RDV (snapshot: 96.31 ns) and RTP (snapshot: 198.56 ns) are lined by respective residues. Interacting residues are shown uniformly in white and licorice representation and small molecules RMP, RDV and RTP in lime, blue and magenta rendered in licorice. Catalytic ion Mg2^{+} is represented in Beads:pink.

as it moved from docked to stable MD pose and gained key interactions (discussed in the next section).

Overall, system stability was asserted through MD simulations, where all systems were stable and special focus was to understand the ligand's independent behaviour. In continuation, the focus was to highlight the ligand's integral properties which include key interactions, preferable site, and their respective stability. RDV was fluctuating initially and RMP was most stable. The RTP is given the status of the active form of RDV (Yin et al., 2020) and hence its dynamicity probing is our major concern so that its mechanistic details might be highlighted along with its inhibition mechanism.

Interaction Fingerprinting of Remdesivir monophosphate, Remdesivir, and Remdesivir Triphosphate Revealed the Key Residues

The stability of chosen ligands are thoroughly observed by understanding the interaction with residues at their most stable states obtained from their respective MD trajectories (**Figure 4A**). By comparison from the centre of mass of RMP's bound state to MD state suggests that there is a minor 0.8 Å deviation (**Figure 4B**). This minor transition was observed due to loss of its interaction with residue K545 (motif F), however, it maintains its interactions with residues of motif B, motif C, and Mg^{2+} ion. It also gains interactions with five other residues, four from motif B and C (D684, A685, T686, L758) and one from motif E (C813) (**Supplementary Table S2**). Similarly, RDV does a transition of 1.1 Å from docked to MD state (**Figure 4C**). RDV maintains its interactions with residues of NTP entry channel K545, R553, and R555 and loses its interaction with SDD motif (**Figure 4C** and **Supplementary Table S2**). The RDV gains interactions with other motif F residues A547, S549, K551, V557, A558 and G559, as well as with motif B residues S682 and T687. RTP does a shift from its docked pose (**Figure 4D**) and its RTP-MD state reveals that it majorly forms interactions with all critical residues of motif F and Mg^{2+} ion i.e., K545, R553, and R555. RTP also gains interactions with residues of motif A (D618, P620, K621, and R624) and motif D (K798) which were found missing in its docked state (**Supplementary Table S2**) and in other systems. The interactions of RTP-MD pose is in accordance with the recently published study by Zhang and Zhou (2020). The importance of these residues is discussed in the section of possible mechanism of inhibition.

After exploring the transitions, we further tried to understand the characteristics of the ligand's respective binding pocket. So, RMP was chosen as the preferred active site and was surrounded with twelve residues: four hydrophobic residues (A685, A688, L758, and C813), five polar (S682, T686, T687, S691, and S759) and three acidic residues (D684, D760, and D761) (**Figure 4E**). At its most stable state, RMP is closely surrounded by the SDD motif which forms the base of catalytic activity of RdRp. Further, upon inspection of parent molecules, it has been found that RDV is

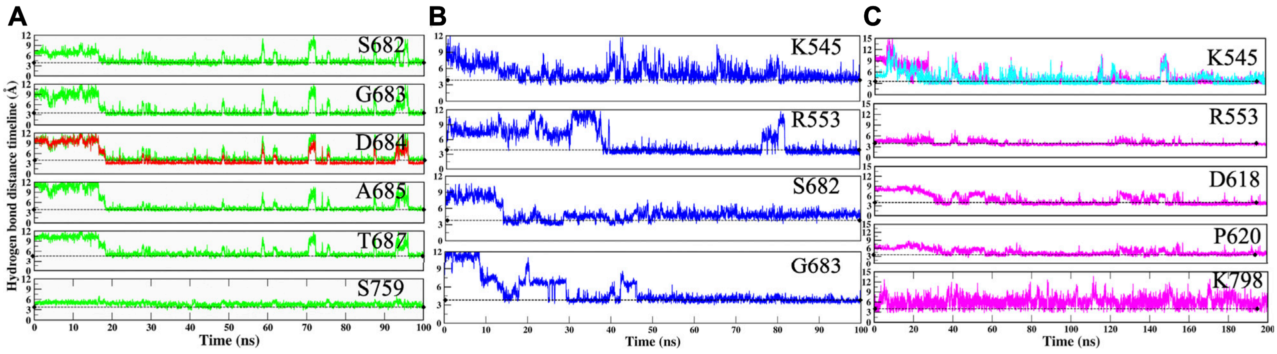


FIGURE 5 | Hydrogen bond analysis throughout MD simulation timeline for **(A)** COM-RMP **(B)** COM-RDV and **(C)** COM-RTP. The dotted lines describe the threshold of 3.5 Å.

TABLE 1 | Hydrogen bond analysis of COM-RMP, COM-RDV and COM-RTP. The hydrogen bonds with occupancy greater than 30% from the equilibrated trajectory are shown.

	Hydrogen bonds of RMP	
Donor	**Acceptor**	**Occupancy (%)**
RMP-N5	S682-Main-O	69.20
RMP-N5	G683-Main-N	48.11
RMP-N5	G683-Main-O	72.22
RMP-N5	D684-Main-N	68.36
RMP-N5	D684-Main-O	73.28
RMP-N5	A685-Main-N	71.88
RMP-N5	T687-Side-OG1	33.76
S759-OG	RMP-Side-O4	70.46
	Hydrogen bonds of RDV	
K545-Side-NZ	RDV-N4	58.20
R553-Side-NH1	RDV-O7	55.52
S682-Side-OG	RDV-O3	59.50
G683-Main-N	RDV-N2	37.14
	Hydrogen bonds of RTP	
R553-Main-O	RTP-N5	66.8
R553-Main-N	RTP-N5	59.3
K545-Side-NZ	RTP-N3	42.7
RTP-O11	D618-Side-OD1	55.3
RTP-O11	P620-Main-N	57.2
K798-Side-NZ	RTP-O8	59.1

surrounded by 11 residues and among them, two residues (S682 and T687) have been found in common with RMP (**Figure 4F**). It is lined with four hydrophobic residues (A547, V557, G559, and V560) as well as two polar (S549 and S682) and has shown interaction with residues K545, K551, R553, and R555. The RDV stable pose reveals that it extends interaction strength at the NTP entry channel. RTP binding pattern reveals that it encompasses residues crucial for NTP entry channel, including three basic residues: K545, R553, and R555 (**Figure 4G**). Some other residues that surround RTP are H439, I548, S549, A554, D618, P620, K621, R624, K798, E811, S814, and R836 and thus the MD state of RTP appears as a "basic rich pocket" due to the presence of seven basic residues (**Figure 4G**). The transition of RTP-docked to RTP-MD pose (RMSD 1.8Å) revealed that it prefers NTP channel residues. Hence, the significance of their contribution are assessed further with hydrogen bond (strength and durability) and thermodynamics analysis.

The RMP and RTP form six hydrogen bonds, while RDV forms four hydrogen bonds. Residues that were crossing 30% cut-off in hydrogen bond occupancy were considered for this study. RMP forms hydrogen bonds with S682, G683, D684, A685, T687, and S759 (**Figure 5A**) and their respective occupancies are mentioned in **Table 1**. Among them, T687 was the lowest contributor with occupancy 33.76% and the rest of the residues formed strong interactions with RMP (Occupancy ~65–75%) (**Table 1**). One of the key residues, S759 from the SDD motif, forms a hydrogen bond with RMP, highlighting that RMP and its direct interaction with catalytic sites. RDV forms four hydrogen bonds, where two residues S682 and G683 were in common with RMP, while the other two residues belong to NTP entry channel K545 and R553 (**Figure 5B** and **Table 1**). The occupancies of these residues are mentioned in **Table 1**. The final analysis on RTP reveals that it forms five hydrogen bonds with K545, R553, D618, P620, and K798 (**Figure 5C** and **Table 1**) and their respective occupancies are described in **Table 1**. This shows that the hydrogen bond forming pattern of RMP and RTP is unrelated as RTP has no direct contact with catalytic site residues, while it gains stability by interacting with residues K545 and R553, which are important for NTP entry. This suggests that RTP could be the active form of RDV, possibly binding at NTP entry site instead of other catalytic sites.

The movements of all ligands can be justified after observing the hydrogen bond distance timeline of each residue (**Figure 5**). Initially, it was observed that ligands were not forming interactions with crucial residues; however, after they gained interactions, the value shrinks down to ~3.0 Å. Thus, it can be deduced that by rearranging themselves, the ligands are forming strong interactions at their preferable site, which will pave the way for the identification of key residues and possible mechanisms of inhibition.

Moreover, further thermodynamics analysis forms a strong base and might complement the binding pattern observations obtained from residue mapping and hydrogen bond analysis.

TABLE 2 | The calculated binding free energies (kcal/mol) by MM-GBSA/PBSA methods.

Contribution	COM-RMP (kcal/mol)	COM-RDV (kcal/mol)	COM-RTP (kcal/mol)
ΔE_{int}	0	0	0
ΔE_{vdW}	−41.83	−52.43	−39.11
ΔE_{ele}	−102.55	−21.05	−110.57
ΔE_{GB}	132.65	53.80	127.86
ΔE_{surf}	−4.86	−5.88	−5.44
ΔG_{gas}	−144.39	−73.48	−149.68
$\Delta G_{solv\ GB}$	127.78	47.92	122.26
ΔG_{GB}	−16.61	−25.55	−27.26
$\Delta G_{solv\ PB}$	121.71	25.85	123.31
ΔE_{PB}	124.73	53.83	127.40
$\Delta E_{n\text{-}polar}$	−3.01	−4.18	−4.08
ΔG_{PB}	−22.67	−23.83	−26.36

Binding Free Energy Analysis Reveals Strong Binding of Remdesivir Triphosphate

The speculation over RTP being the active form of RDV has been completely justified by binding free energy analysis. The result indicates that RTP, despite movement, binds stronger than both RMP and RDV. The RTP binds with ΔG_{bind} = −26.37 kcal/mol, which is higher than RDV ΔG_{bind} = −23.83 kcal/mol and RMP ΔG_{bind} = −22.67 kcal/mol (**Table 2**). This binding pattern was obtained through the PBSA method, while a similar spectrum (RTP > RDV > RMP) was found in the GBSA method as well (**Table 2**). The difference in the binding energies was due to the higher contribution of the polar solvation energy in the PBSA method. As from **Table 2**, the electrostatic contribution is more favorable in COM-RMP and COM-RTP than COM-RDV, however was more favorable in COM-RTP.

Deciphering Residue-Wise Energetic Contributions

The results and observations till now have suggested that findings related to ligand binding have provided a mixed view of the pocket; however, the interaction pattern at stable states explores the key residues. RMP stays at a catalytic site with slight fluctuation, RDV being the bulkiest molecule establishes its interaction from catalytic site to NTP entry channel and RTP, which undergoes rearrangement to establish its prime interaction with residues critically important for NTP entry.

As per cut-off of −0.5 kcal/mol, RMP contributors D623 and D760 have an upper-bound value of −4.0 kcal/mol, S681 has a mid-bound value of −2.9 kcal/mol, and the remaining lower-bound contributors are Y456, K551, V557, C622, T680, and S682 (**Figure 6A**). As we have commented earlier on RDV's fluctuation, we found the same in the energy contribution pattern of residues. Only one residue, V557, has an upper-bound value of −3.8 kcal/mol, while residues K545, R553, R555, and S682 are mid contributors ~2.0 kcal/mol, and residues A558 and G683 snap at lower-bound (**Figure 6B**). Eventually, the RTP pattern was analyzed, in which the highest contributor was R553 with −5.9 kcal/mol and residues K545, R555, D618, and P620 had mid-bound value of ~ −3.0 kcal/mol. At the same time S549, K621, R624, and R798 lie in the lower-bound value (**Figure 6C**). RTP gets the highest value contribution of ~ −6.0 kcal/mol, which is missing in both of the ligands. RTP gets its strong binding affinity at the NTP entry channel and strong interactions with K545, R553, and R555 eventually describing its behaviour at the pocket. The comparison between the RTP-docked pose and MD pose shows the gain in stability at the pocket. The major revelation is that at docked pose catalytic site residues D760 and D761 contribute to RTP while at MD pose, those contributions become weak gradually (**Supplementary Figure S4**). At NTP channel, RTP not only strengthened its interaction with K545 and R553 but also gained interaction with R555 (**Supplementary Figure S4**). RTP gains four additional interactions (D618, P620, K621, and R624) but also loses five interactions (D452, A547, K551, D760, and D761) and hence we did not find any significant energy change from docked to MD pose (Docked pose ΔG_{bind} = −25.42 kcal/mol and MD pose = −26.37 kcal/mol) (**Supplementary Figure S4**). It can be clearly seen that the energy contribution of four gained residues is higher than five lost residues (**Supplementary Figure S4**) and this might be the reason for its stability. Despite the least energy difference, RTP forms strong interaction with NTP channel residues and thus might block the entry of NTP into the active site. As these residues of motif F interact with primer strand RNA and stabilize the incoming nucleotide (Yin et al., 2020), RTP might be destabilizing this interaction.

Additionally, we have characterized the pocket based on contributing residues and its respective chemical nature that might be crucial for designing ligands. Falling from RMP to RTP, it can be observed that RTP site (MD pose) is "basic rich" as residues K545, R553, R555, K621, R624, K798, and R836 are the significant contributors (**Figures 6D–F**) while in the other two systems this uniformity is not found. This makes the RTP site unique and relevant for designing more potential inhibitors.

Charge Complementarity Profile

Electrostatic potential was evaluated for all three systems in order to understand electrostatic complementarity between ligands and protein (**Figures 7A–C**). Complementarity profile is expressed in charged units, indicating a spectrum of optimal complementary charges preferred by protein along its molecular surface and afterwards its complementarity is established with respect to

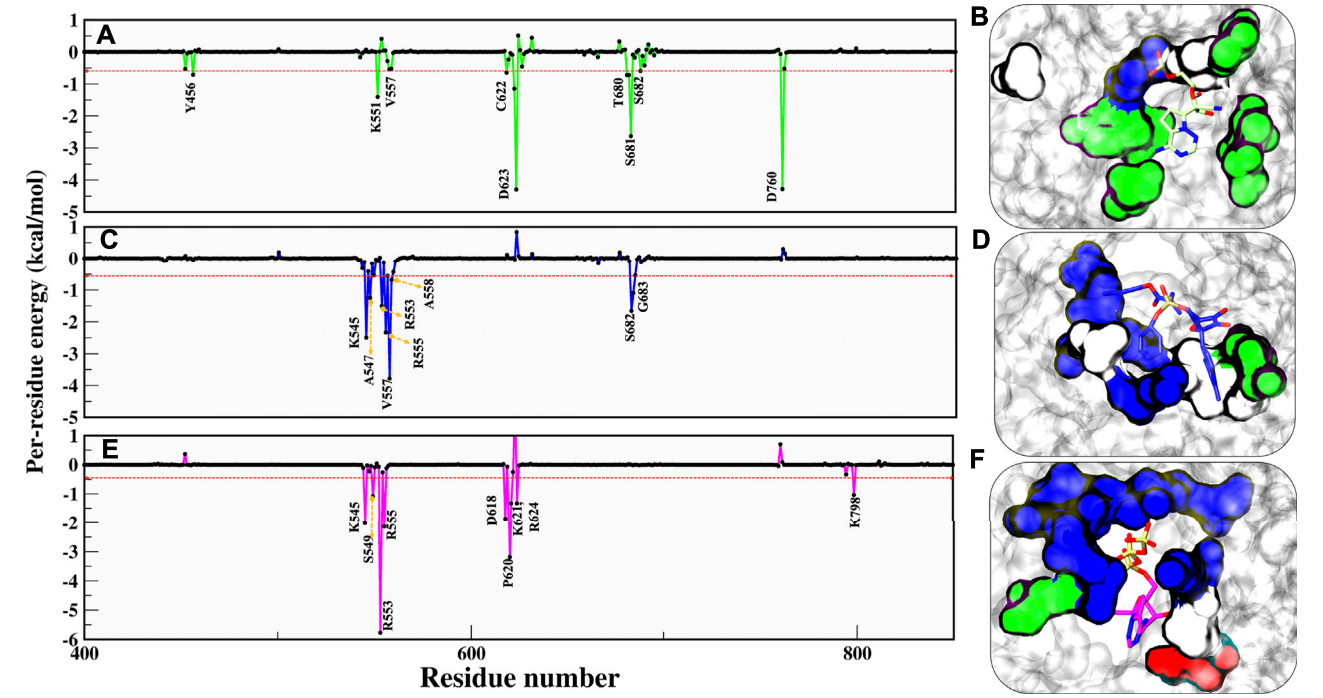


FIGURE 6 | Residue wise decomposition energy (kcal/mol) in all three systems. **(A,C,E)** COM-RMP, COM-RDV and COM-RTP and its respective residue type characteristics is shown panels **(B,D,E)** in surface representation. RMP, RDV and RTP are rendered in licorice and shown in Green, Blue and Magenta respectively. The cut-off energy was -0.5 kcal/mol. Residues chemical properties are shown as: Polar: green, Hydrophobic: white, Basic: blue and Acidic: red.

ligand. RTP gains three slightly favorable patches with its two 'N' atoms and two 'O' atoms (**Figure 7D**), while RDV gains only two small complement patches (**Figure 7E**). Unlike these two, RTP phosphate tail finds a large complementary patch with its respective pocket along with a slightly favorable patch with its 'O' atom (**Figure 7F**).

The final protein-ligand analysis justifies the RTP stability at NTP entry channel as it gains perfect complementarity from protein while the other two clearly miss this gain from their respective preferable site. This analysis also shows that RTP-MD pose is the best pose.

The Free Energy Landscape Undermines the Thermodynamically Most Stable Basin of Remdesivir Triphosphate

To capture the significant conformational changes, PCA was applied to all the simulated systems. The first two PCs were found as important components for capturing the maximum variance in the Cα displacement during the MD simulations. This was found on inspection of the path of the trajectory connecting the minima's and the sub-conformational spaces of different systems. The minima are numbered as per the appearance of the basins w.r.t. time during MD simulations. We found three minimas in APO systems, while only two minimas are observed in all complex systems (**Figure 8**). APO systems show large areas of conformational space that have three minima at ~50 ns, ~60 ns, and ~80 ns and among them, the deepest minima was observed at 80 ns (**Figure 8A**). Upon superimposition of the conformations extracted from these basins, the structural changes were observed, which may be important for RdRp activity. The interpolated loop (residue 51–83) at NiRAN domain and helix of thumb (residue 832–895) have displayed the major fluctuations while other regions are well-aligned (**Figure 8B**). The loops of NiRAN domain have shown bidirectional movement, which has outward (at ~50 ns) to the inward movement (at ~80 ns) as trajectory tends towards stability (**Figure 8B**), while thumb helices show slight dislocation at every minima step (**Figure 8B**). The porcupine plot of PC1 and PC2 shows highest atomic fluctuations in the above-mentioned regions as well (**Figures 8C,D**). The variation captured by PC1 and PC2 is correlated with the RMSF values of these regions (**Figure 2B**). COM-RMP has two minima at ~75 and ~95 ns respectively (**Figure 8E**). The structural changes were the same as in APO (**Figure 8F**) but in this case, the loop of NiRAN domain does not move inwards and in both the basins it has shown outward movement (**Figure 8F**). PC1 and PC2 captured fluctuations in the same regions in COM-RMP. The atomistic fluctuation of NiRAN domain at PC1 shows higher fluctuation than APO, while the rest have shown this with lesser magnitude (**Figures 8G,H**). COM-RDV also displays two minima: one minima at ~50 ns and the other at ~95 ns (**Figure 8I**). The second minima were broader and deeper as compared to minima I (**Figure 8J**). Similar observations were found as of COM-RMP, which depicts the outward movement of NiRAN domain loop along with the changes in helices of thumb subdomain (**Figure 8J**). PC1 and PC2 capture

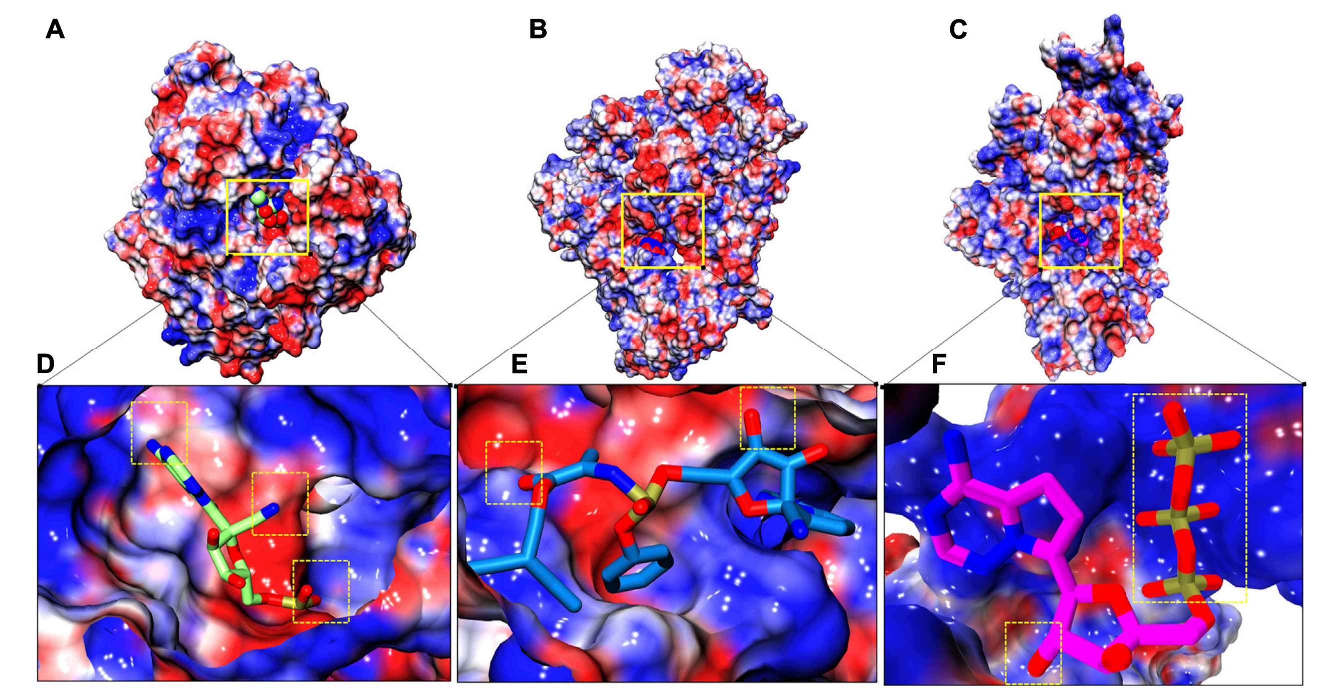


FIGURE 7 | Electrostatic surface complementarity. **(A–C)** Electrostatic potential isosurface of whole complexes: COM-RMP, COM-RDV and COM-RTP respectively and site of embedded small molecules is highlighted in solid yellow box **(D–F)** Respective binding site snapshot of systems in aforementioned order. Positive and negative electrostatic potential is rendered in blue and red, respectively. Dotted yellow boxes show positive complementarity patches observed at their respective stable states.

the highest atomistic fluctuation at similar regions (**Figures 8K,L**). The most different trends were observed in the case of COM-RTP. Upon landscape inspection, we found that COM-RTP surpassed all systems with broader and deepest minima at ~160 ns (**Figure 8M**). As such deep minima were missing in other systems, COM-RTP achieves its lowest free energy state. Another difference is the inward movement of NiRAN domain loop from outward movement (**Figure 8N**). This is one of the states that was snapped in APO trajectory (previously mentioned). The equilibrium shift from minima I (at ~50 ns) to minima II (at ~160 ns) has elucidated the lowest free energy state of RdRp in RTP bound condition (**Figure 8N**). Another feature that has been observed is the minimization of thumb domain fluctuation while in previous systems these fluctuations were notable (**Figure 8O**). Also, the directionality of the projections at PC1 of NiRAN domain changes to inward, while in other systems it was outward (**Figure 8O**). Similarly, projections of the helices of thumb domain were also inward as shown by PC2 (**Figure 8P**). The inward movement COM-RTP shows that RTP binding at palm sub-domain might have an impact on internal wiring as it changes the atomistic fluctuation of distal NiRAN domain. Since NiRAN and palm domain share close proximity and may engage in a functional interaction (Zhang et al., 2020), the strong binding of RTP at NTP entry channel might be justified.

Overall, COM-RTP displays different behavior as compared to RMP/RDV systems. Two major regions have been identified through PCA and FEL analysis. Thumb subdomain is crucial for template entry while NiRAN domain is known for its interaction with other protein partners (Zhang et al., 2020). COM-RTP follows one trend of APO, where the inward movement of the loop was snapped while all other systems follow the either trend of APO i.e. outward movement. One of the major findings observed from PCA and FEL analysis is the deepest basin obtained in system COM-RTP which justifies the impact of RTP binding on protein significantly. The presence of RTP has not only minimized the nearby regions (thumb domain helices) but also led to significant changes in the NiRAN domain which may have caused an allosteric regulatory effect. Thumb domain acts as a door for NTP entry, and our observation does suggest that it was completely minimized in COM-RTP, thus these changes may justify the role of RTP rearrangement at NTP entry site. In continuation, it may not allow the entry of new nucleotides, which might lead to inhibition of replication.

Computational Alanine Scanning

The key residues K545, R553, and R555 that are crucial for NTP entry channel have shown a remarkable drop in their contribution to RTP stability after they were mutated to alanine (**Supplementary Figure S5**). Residue P620 that has shown backbone displacement to rigidify the channel was also considered and it was found that it has a slight contribution. Residues of catalytic site S549, D760, and D761 were also considered to justify the RTP binding at NTP channel, and found that their contribution is negligible. These observations

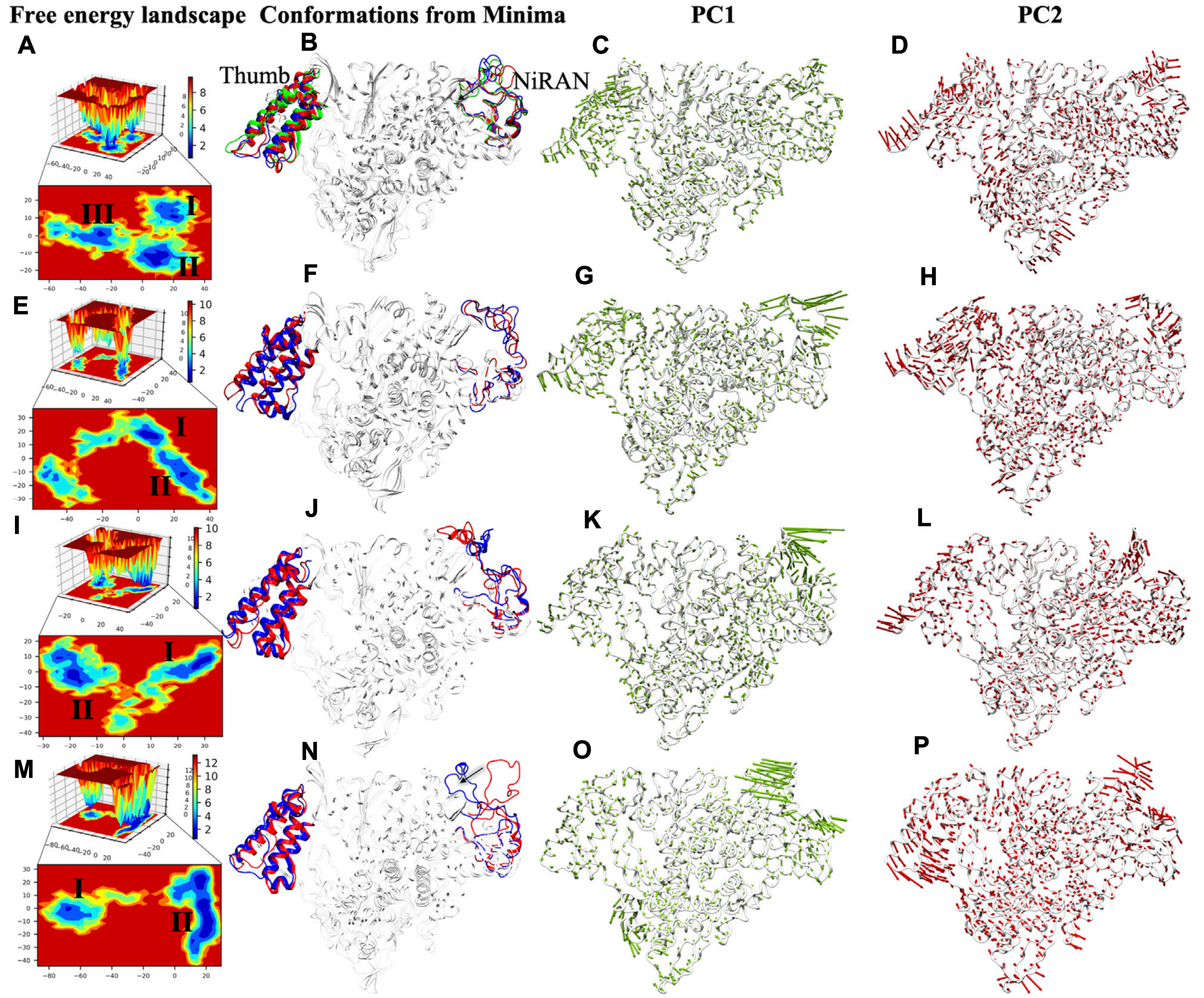


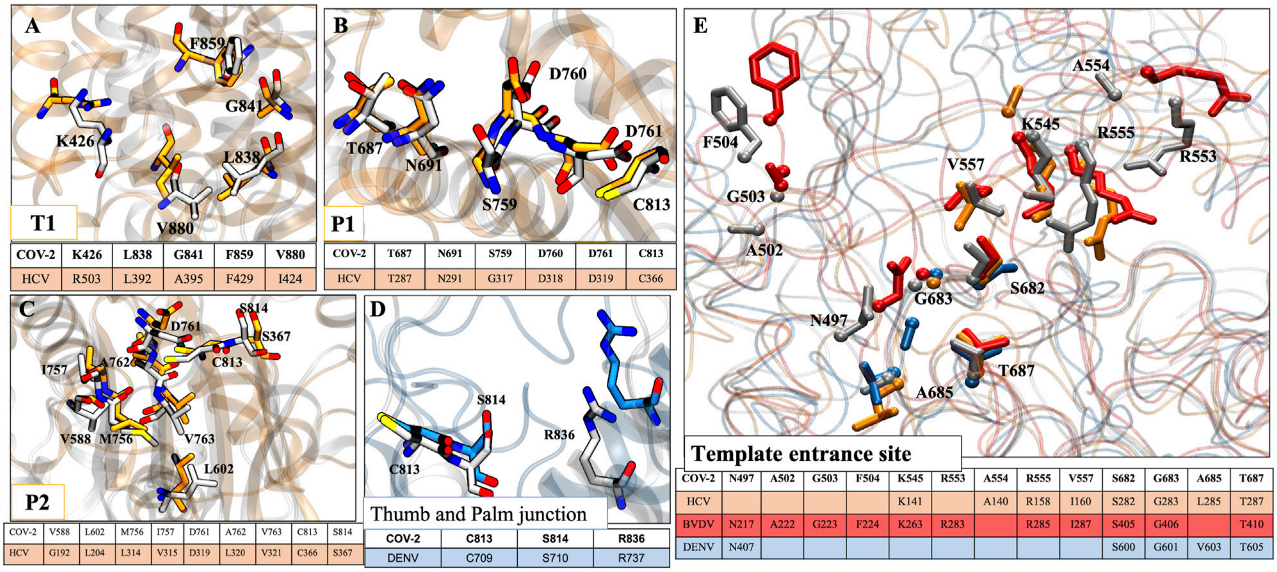
FIGURE 8 | FEL analysis and porcupine plots of SARS-CoV-2 RdRp in bound and unbound states. The panels **(A,E,I,M)** of 2D and 3D free energy landscape (FEL) plots between PC1 (*x*-axis) and PC2 (*y*-axis) for APO, COM-RMP, COM-RDV and COM-RTP respectively. The color bar represents the Gibbs free energies in the plot ranging the lowest energy (blue) to highest energy (red) conformation states. The low energy minima states are shown in the 2D FEL plots. The minima extracted from FEL plots are superimposed for all simulated structures depicting the conformational changes in the panels **(B,F,J,N)**. The red, blue and green conformations represent the minima I-III respectively in the panels **(B,F,J,N)**. Porcupine plots generated using extreme PC1 panels **(C,G,K,O)** and PC2 projections panels **(D,H,L,P)** for all the four simulated systems. The direction of the green and red arrows (in PC1 and PC2 respectively) at each Cα shows its direction of motion and the length of arrow depicts its strength. The protein is represented in tube form and colored in white. The regions that are highly fluctuating are labelled.

further confirm the criticality of NTP facilitators residues (K545, R553 and R555) and also justify the significance of RTP MD pose.

Mapping of Potential Residues of SARS-CoV-2 RdRp Protein With the Allosteric Sites of Other Viruses

Since RdRps are structurally and functionally well conserved, we have mapped the possible allosteric sites for SARS-CoV-2 RdRp based on literature and crystallographic data of RdRp of HCV (Mittal et al., 2020), BVDV (Mittal et al., 2019), and DENV (Maddipati et al., 2020). Allosteric inhibitors customize the internal motions and conformational states of the enzyme and are very specific to a target, thus avoiding off-target effects (Davis et al., 2016; Mittal et al., 2020). Although the sequence and structural similarity among their sequences with CoV-2 RdRp is less than 30%, the secondary structures and functional motifs are well conserved. Therefore, we found four probable allosteric sites based on structural alignment with HCV RdRp: T1, P1, P2, and TES sites on CoV-2 RdRp.

T1 site: The general APO HCV RdRp consists of a Δ1 loop tucked on the thumb domain which is displaced by non-nucleoside inhibitors (NNIs) and forms a hydrophobic T1 pocket (Mittal et al., 2020). In CoV-2 RdRp also, we observed the same Δ1 loop protruding from the finger domain and



COV-2	K426	L838	G841	F859	V880
HCV	R503	L392	A395	F429	I424

COV-2	T687	N691	S759	D760	D761	C813
HCV	T287	N291	G317	D318	D319	C366

COV-2	V588	L602	M756	I757	D761	A762	V763	C813	S814
HCV	G192	L204	L314	V315	D319	L320	V321	C366	S367

COV-2	C813	S814	R836
DENV	C709	S710	R737

COV-2	N497	A502	G503	F504	K545	R553	A554	R555	V557	S682	G683	A685	T687
HCV					K141		A140	R158	I160	S282	G283	L285	T287
BVDV	N217	A222	G223	F224	K263	R283		R285	I287	S405	G406		T410
DENV	N407									S600	G601	V603	T605

FIGURE 9 | Conservity analysis of allosteric sites in different viral RdRp's. The panels **(A–C)** represent three conserved allosteric sites in HCV and CoV2 RdRp. Panel **(D)** represents conserved allosteric sites in DENV and CoV2 RdRp. **(E)** The conservation analysis of template entrance site (TES) in CoV-2, HCV, BVDV and DENV. For figure clarity, only sidechain and Cα atoms of residues are shown in this panel. The respective tables at every panels corresponds to conserved residues obtained after structural alignment. Residues color coding is done as follows: HCV: orange, CoV2: white, DENV: blue, BVDV: red.

extends till thumb domain. Along with this, the residues K426, L838, G841, F849, and V880 of CoV-2 RdRp are structurally aligned with the residues R503, L392, A395, F429, and I424 of HCV RdRp but the proper pocket surface is not present (**Figure 9A**). So, we believe that if NNIs are designed for this site based on the aforementioned residues this may result in a similar mechanism of inhibition by causing displacement of the Δ1 loop, which will disrupt the functionality of the RdRp.

T2 site: As per the T2 site of HCV RdRp (Mittal et al., 2020) and allosteric site 1 of DENV RdRp (Maddipati et al., 2020) below the thumb domain and C-terminal, we did not find any relevant or conserved residues in CoV-2 RdRp.

P1 site: As per the P1 site of HCV RdRp (present adjacent to the active site on palm subdomain) (Mittal et al., 2020), we found residues T687, N691, S759, D760, D761, and C813 of CoV-2 RdRp as structurally well-conserved residues with their corresponding residues T287, N291, G317, D318, D319, and C366 in HCV RdRp (**Figure 9B**). Therefore, the NNIs designed at this site could interfere with the active site mechanism as well as incoming NTPs.

P2 site: The residues V588, L602, M756, I757, D761, A762, V763, C813, and S814 of CoV-2 RdRp are found to be aligned with the residues G192, L204, L314, V315, D319, L320, V321, C366, and S367 of HCV RdRp (**Figure 9C**). Although a similar groove of P2 pocket of HCV is not observed in CoV-2 RdRp, the possibility of pocket formation by NNIs is still there that may interfere with the NTP entry into the RdRp.

As per DENV RdRp's allosteric site 2 (interface/junction between thumb and palm or N-pocket) (Hilgenfeld and Vasudevan, 2018; Maddipati et al., 2020), we found three conserved residues, C813, S814, and R836 in CoV-2 RdRp also (**Figure 9D**). Therefore, the NNIs designed for this site would possibly interrupt the *de novo* initiation activity of the viral RdRp (Hilgenfeld and Vasudevan, 2018).

TES site: Since the RNA template entrance site (TES) exists in all polymerases and has been well explored in DENV, BVDV, and HCV RdRps, this site could also be targeted for designing NNIs. We have aligned all these four RdRps and highlighted the conserved residues accordingly. We found more number of conserved residues in BVDV RdRp than others with CoV-2 RdRp. In CoV-2, the conserved residues are N497, A502, G503, F504, K545, R553, A554, R555, V557, S682, G683, A685, and T687 (**Figure 9E**). Among these residues, K545, R553, and R555 are involved in escorting the RNA template and positioning of NTPs at the catalytic site.

Further, we implemented an unbiased and ligand-independent approach to search for potential druggable binding sites in CoV-2 RdRp by using the SiteMap module of Schrodinger. We identified 4 out of 5 sites that could be targeted for designing small molecules. As per the highest Dscore, the best site is site_3 which corroborates with DENV RdRp N-pocket and HCV P1 site. The site_1 and site_4 correspond to the template entrance channel while site_4 corresponds to the P2 site as HCV RdRp (**Supplementary Figure S6**). The residues lining these sites are mentioned in **Supplementary Table S3**. Therefore, the overall analysis hints for these allosteric sites for rational or targeted designing of NNIs for SARS-CoV-2 RdRp.

The conservity analysis revealed the significance of all sites to develop NNIs, however the importance of TES as allosteric site and its lining residue is well documented. The significance of

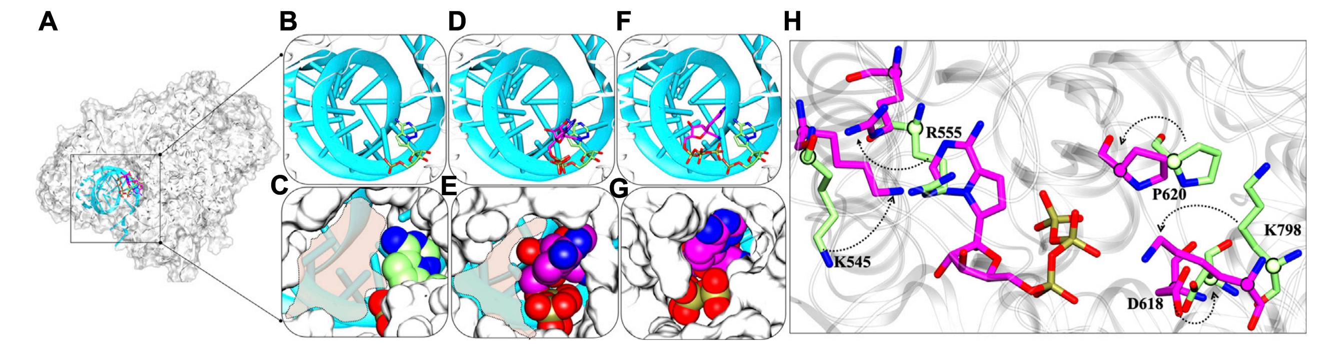


FIGURE 10 | Dynamic insights at NTP entry site **(A)** Superimposed structure of RMP-bound with RTP-docked and RTP-MD pose in the presence of primer-template strand. The inset shows the RMP bound, RTP docked and RTP MD poses in the panels **(B,D,F)** in licorice representation and primer-template strand in cyan New Cartoon representation. The panels **(C,E,G)** show the NTP channel occlusion by these poses with protein rendered in white surface representation and molecules in vdW representation. **(H)** Crucial residues with major backbone displacement are highlighted that rigidifies the channel in the presence of RTP-MD pose. RMP lined residues bound with RMP are shown in Licorice: lime and RTP lined residues are in Licorice: magenta. The transparent shading represents the available space at NTP entry site. The direction of arrows represents the transition of residues from crystal RMP-bound pose to RTP-MD stable pose.

these residues, especially K545, R553, and R555, were well aligned with our simulation studies as well, which justifies the RTP binding and its probable inhibition mechanism. Thus, a comprehensive biochemical and structural-dynamics studies of this enzyme may reveal more insights about other potential druggable allosteric sites (Shi et al., 2020).

Possible Mechanism of Inhibition by Remdesivir Triphosphate

The detailed possible mode of inhibitions in RdRps were explored and based on various analysis, the role of RTP at NTP entry site looks promising for the inhibition of RdRp and thus we proposed the possible inhibition mechanism. We compared the RMP-bound pose with RTP- RTP-MD pose that provided significant information and thus support RTP rearrangement at NTP entrance site (**Figure 10A**). At RMP-bound pose, it can be clearly seen that in this binding mode there is available space for NTP to enter into the cavity (**Figures 10B,C**) whereas a similar trend was found with RTP-docked pose as well (**Figures 10D,E**). The most distinguished feature was observed in RTP-MD pose, extracted from the deepest basin shows that RTP occupies the space completely and therefore occludes the NTP entry channel (**Figures 10F,G**). We obtained a new set of residues when the transition of RTP occurs from dock to MD binding pose (**Figure 4D** and **Supplementary Table S2**). Out of those set of residues, D618, P620, and K798 contributed to RTP stability as observed in thermodynamics analysis (**Figure 6**). Hence, we tried to elucidate their importance and upon inspection, we found that residues K545, D618, P620, and K798 showed backbone displacement towards the NTP entry channel to rigidify the channel and provide tight packing to RTP. Along with these R555 also undergoes backbone displacement which helps to accommodate RTP at this channel. Furthermore, the residues H439, R836, and S814 do not contribute much energetically, however, their presence at the periphery of the channel does provide the additional rigidity to the RTP pocket.

Thus, RTP's pose explored through MD simulations embraces the NTP entry channel dynamicity, by which it occupies a complete area that is left open in case of RMP and RTP-docked pose. It can also be estimated that NTP would be now sterically hindered due to non-availability of space at the channel and the change in conformations of the key residues essential to escort NTPs into the polymerase would be possible to perturb the replication process.

CONCLUSION

In this work, a thorough comparative study between RDV (parent molecule) and its metabolites, RMP (crystal bound), and RTP (active form) was performed to understand the mechanism of RdRp inhibition, the reason behind the active form of RTP only, and its interaction pattern with key residues. Previous studies with RTP or RDV were performed on homology models and after the arrival of crystal structure of SARS-CoV-2 RdRp, we implemented a holistic approach to extract the meaningful insights that lead to inhibition. RMP was bound in crystal, however as stated by Yin et al. in 2020, the metabolite RTP is known to be the active form of RDV (Yin et al., 2020). Therefore, to explore the dynamic spectrum of the binding pocket, we included both the metabolites and parent molecules to compare the binding pattern. Molecular docking of RDV and RTP provided the initial information of RTP with highest docking energy being the best candidate in this competitive analysis. The MD simulations revealed that all systems are stable throughout the trajectory, however, the RTP has shown movement at initial states in the binding pocket, but after gaining the interaction with residues and Mg^{2+} ion, it achieved the most stable state. We observed that the most stable state obtained from MD is not the initial docked-pose. Thermodynamics analysis revealed that RTP specifically interacts with the residues K545, R553, and R555, important for NTP entry (Yin et al., 2020; Zhang and Zhou, 2020). This outcome correlates well with our finding. In RTP interaction map, four other basic residues, K621, R624, K798, and R836, were found trapped in the "basic rich pocket". Apart from these we identified four new residues, S549, D618, P620, and K798, as the key contributors to RTP binding in the terms of energetic analysis. The conformational analysis reveal that their backbone movement facilitates the tight packing of RTP at NTP channel. Overall the combination of motif A, E, and F residues are involved in the interaction with RTP supports the possible inhibition mechanism. PCA analysis elucidates two key regions, i.e. thumb and interpolated loops of NiRAN domains, that have shown high atomistic fluctuations. Atomistic fluctuations were merely observed in thumb regions which might be because of the rearrangement of RTP. As thumb domain is adjacent to template entry site and its structural minimization upon RTP binding lays down the possibility that it possibly hinders the template entry and perturbs the replication process. To speculate the possible inhibition, the conservity analysis of different allosteric sites of RdRps of HCV, BVDV, DENV, and SARS-CoV-2 highlighted the potential allosteric sites apart from RTP's NTP site. Overall, the comparative analysis and computational insights encourage to target these residues at NTP entrance site for the discovery of non-nucleosidic antiviral inhibitors.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SA proposed and designed the study. MS, LM and AK performed experiments and analysis of results. MS, LM, AK, and SA wrote the manuscript. All the authors have read the manuscript.

FUNDING

The work has been supported by the THSTI funding and computational facilities.

ACKNOWLEDGMENTS

The authors acknowledge the reviewer for their valuable suggestions and comments that have improved the content of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.639614/full#supplementary-material

REFERENCES

Allnér, O., Nilsson, L., and Villa, A. (2012). Magnesium Ion-Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theor. Comput.* 8, 1493–1502. doi:10.1021/ct3000734

Anang, S., Kaushik, N., Hingane, S., Kumari, A., Gupta, J., Asthana, S., et al. (2018). Potent Inhibition of Hepatitis E Virus Release by a Cyclic Peptide Inhibitor of the Interaction between Viral Open Reading Frame 3 Protein and Host Tumor Susceptibility Gene 101. *J. Virol.* 92, e00684–18. doi:10.1128/JVI.00684-18

Appleby, T. C., Perry, J. K., Murakami, E., Barauskas, O., Feng, J., Cho, A., et al. (2015). Structural Basis for RNA Replication by the Hepatitis C Virus Polymerase. *Science* 347, 771–775. doi:10.1126/science.1259210

Asthana, S., Shukla, S., Ruggerone, P., and Vargiu, A. V. (2014). Molecular Mechanism of Viral Resistance to a Potent Non-nucleoside Inhibitor Unveiled by Molecular Simulations. *Biochemistry* 53, 6941–6953. doi:10.1021/bi500490z

Asthana, S., Shukla, S., Vargiu, A. V., Ceccarelli, M., Ruggerone, P., Paglietti, G., et al. (2013). Different Molecular Mechanisms of Inhibition of Bovine Viral Diarrhea Virus and Hepatitis C Virus RNA-dependent RNA Polymerases by a Novel Benzimidazole. *Biochemistry* 52, 3752–3764. doi:10.1021/bi400107h

Babadaei, M. M. N., Hasan, A., Vahdani, Y., Bloukh, S. H., Sharifi, M., Kachooei, E., et al. (2020). Development of Remdesivir Repositioning as a Nucleotide Analog against COVID-19 RNA Dependent RNA Polymerase. *J. Biomol. Struct. Dyn.* 39, 3771–3779. doi:10.1080/07391102.2020.1767210

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001). Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci.* 98, 10037–10041. doi:10.1073/pnas.181342398

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., et al. (2005). The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* 26, 1668–1688. doi:10.1002/jcc.20290

Chen, F., Liu, H., Sun, H., Pan, P., Li, Y., Li, D., et al. (2016). Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 6. Capability to Predict Protein-Protein Binding Free Energies and Re-rank Binding Poses Generated by Protein-Protein Docking. *Phys. Chem. Chem. Phys.* 18, 22129–22139. doi:10.1039/c6cp03670h

Chen, H., Beardsley, G. P., and Coen, D. M. (2014). Mechanism of Ganciclovir-Induced Chain Termination Revealed by Resistant Viral Polymerase Mutants with Reduced Exonuclease Activity. *Proc. Natl. Acad. Sci. USA* 111, 17462–17467. doi:10.1073/pnas.1405981111

Davis, B. C., Brown, J. A., and Thorpe, I. F. (2016). Allosteric Inhibitors Have Distinct Effects, but Also Common Modes of Action, in the HCV Polymerase. *Biophys. J.* 111, 2336–2337. doi:10.1016/j.bpj.2016.10.015

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004). PDB2PQR: an Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* 32, W665–W667. doi:10.1093/nar/gkh381

Eastman, R. T., Roth, J. S., Brimacombe, K. R., Simeonov, A., Shen, M., Patnaik, S., et al. (2020). Remdesivir: A Review of its Discovery and Development Leading to Emergency Use Authorization for Treatment of COVID-19. *ACS Cent. Sci.* 6, 672–683. doi:10.1021/acscentsci.0c00489

Fung, A., Jin, Z., Dyatkina, N., Wang, G., Beigelman, L., and Deval, J. (2014). Efficiency of Incorporation and Chain Termination Determines the Inhibition Potency of 2′-Modified Nucleotide Analogs against Hepatitis C Virus Polymerase. *Antimicrob. Agents Chemother.* 58, 3636–3645. doi:10.1128/aac.02666-14

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., et al. (2020). Structure of the RNA-dependent RNA Polymerase from COVID-19 Virus. *Science* 368, 779–782. doi:10.1126/science.abb7498

Genheden, S., and Ryde, U. (2015). The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* 10, 449–461. doi:10.1517/17460441.2015.1032936

Gong, P., and Peersen, O. B. (2010). Structural Basis for Active Site Closure by the Poliovirus RNA-dependent RNA Polymerase. *Proc. Natl. Acad. Sci.* 107, 22505–22510. doi:10.1073/pnas.1007626107

Guo, J., Wang, X., Sun, H., Liu, H., and Yao, X. (2012). The Molecular Basis of IGF-II/IGF2R Recognition: a Combined Molecular Dynamics Simulation, Free-Energy Calculation and Computational Alanine Scanning Study. *J. Mol. Model.* 18, 1421–1430. doi:10.1007/s00894-011-1159-4

Halgren, T. A. (2009). Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* 49, 377–389. doi:10.1021/ci800324m

Hilgenfeld, R., and Vasudevan, S. G. (2018). *Dengue and Zika: Control and Antiviral Treatment Strategies*. Singapore: Springer.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics* 14, 33–38. doi:10.1016/0263-7855(96)00018-5

Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* 118, 11225–11236. doi:10.1021/ja9621760

Jorgensen, W. L. (2002). OPLS Force Fields. *Encyclopedia Comput. Chem.* doi:10.1002/0470845015.coa002s

Kato, K., Honma, T., and Fukuzawa, K. (2020). Intermolecular Interaction Among Remdesivir, RNA and RNA-dependent RNA Polymerase of SARS-CoV-2 Analyzed by Fragment Molecular Orbital Calculation. *J. Mol. Graphics Model.* 100, 107695. doi:10.1016/j.jmgm.2020.107695

Khan, S., Farooq, U., and Kurnikova, M. (2017). Protein Stability and Dynamics Influenced by Ligands in Extremophilic Complexes - a Molecular Dynamics Investigation. *Mol. Biosyst.* 13, 1874–1887. doi:10.1039/c7mb00210f

Koulgi, S., Jani, V., Uppuladinne, M. V. N., Sonavane, U., and Joshi, R. (2020). Remdesivir-bound and Ligand-free Simulations Reveal the Probable Mechanism of Inhibiting the RNA Dependent RNA Polymerase of Severe Acute Respiratory Syndrome Coronavirus 2. *RSC Adv.* 10, 26792–26803. doi:10.1039/d0ra04743k

Maddipati, V. C., Mittal, L., Mantipally, M., Asthana, S., Bhattacharyya, S., and Gundla, R. (2020). A Review on the Progress and Prospects of Dengue Drug Discovery Targeting NS5 RNA- Dependent RNA Polymerase. *Cpd* 26, 4386–4409. doi:10.2174/1381612826666200523174753

Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., Sherman, W., and Sherman, W. (2013). Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided Mol. Des.* 27, 221–234. doi:10.1007/s10822-013-9644-8

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theor. Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Meagher, K. L., Redman, L. T., and Carlson, H. A. (2003). Development of Polyphosphate Parameters for Use with the AMBER Force Field. *J. Comput. Chem.* 24, 1016–1025. doi:10.1002/jcc.10262

Mittal, L., Kumari, A., Suri, C., Bhattacharya, S., and Asthana, S. (2020). Insights into Structural Dynamics of Allosteric Binding Sites in HCV RNA-dependent RNA Polymerase. *J. Biomol. Struct. Dyn.* 38, 1612–1625. doi:10.1080/07391102.2019.1614480

Mittal, L., Srivastava, M., and Asthana, S. (2019). Conformational Characterization of Linker Revealed the Mechanism of Cavity Formation by 227G in BVDV RDRP. *J. Phys. Chem. B* 123, 6150–6160. doi:10.1021/acs.jpcb.9b01859

Mittal, L., Srivastava, M., Kumari, A., Tonk, R. K., Awasthi, A., and Asthana, S. (2021). Interplay Among Structural Stability, Plasticity, and Energetics Determined by Conformational Attuning of Flexible Loops in PD-1. *J. Chem. Inf. Model.* 61, 358–384. doi:10.1021/acs.jcim.0c01080

Sarkar, R., Sharma, K. B., Kumari, A., Asthana, S., and Kalia, M. (2021). Japanese Encephalitis Virus Capsid Protein Interacts with Non-lipidated MAP1LC3 on Replication Membranes and Lipid Droplets. *J. Gen. Virol.* 102. doi:10.1099/jgv.0.001508

Shi, Y., Zhang, X., Mu, K., Peng, C., Zhu, Z., Wang, X., et al. (2020). D3Targets-2019-nCoV: a Webserver for Predicting Drug Targets and for Multi-Target and Multi-Site Based Virtual Screening against COVID-19. *Acta Pharm. Sinica B* 10, 1239–1248. doi:10.1016/j.apsb.2020.04.006

Singh, M., Srivastava, M., Wakode, S. R., and Asthana, S. (2021). *Elucidation of Structural Determinants Delineates the Residues Playing Key Roles in Differential Dynamics and Selective Inhibition of Sirt1-3. J. Chem. Inf. Model.* 61, 1105–1124. doi:10.2210/pdb7bv2/pdb

Sitkoff, D., Sharp, K. A., and Honig, B. (1994). Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* 98, 1978–1988. doi:10.1021/j100058a043

Srivastava, M., Suri, C., Singh, M., Mathur, R., and Asthana, S. (2018). Molecular Dynamics Simulation Reveals the Possible Druggable Hot-Spots of USP7. *Oncotarget* 9, 34289–34305. doi:10.18632/oncotarget.26136

Suri, C., Hendrickson, T. W., Joshi, H. C., and Naik, P. K. (2014). Molecular Insight into γ-γ Tubulin Lateral Interactions within the γ-tubulin Ring Complex (γ-TuRC). *J. Comput. Aided Mol. Des.* 28, 961–972. doi:10.1007/s10822-014-9779-2

Suri, C., Joshi, H. C., and Naik, P. K. (2015). Molecular Modeling Reveals Binding Interface of γ-tubulin with GCP4 and Interactions with Noscapinoids. *Proteins* 83, 827–843. doi:10.1002/prot.24773

Tripathi, S., Srivastava, G., and Sharma, A. (2016). Molecular Dynamics Simulation and Free Energy Landscape Methods in Probing L215H, L217R and L225M βI-tubulin Mutations Causing Paclitaxel Resistance in Cancer Cells. *Biochem. Biophys. Res. Commun.* 476, 273–279. doi:10.1016/j.bbrc.2016.05.112

Trott, O., and Olson, A. J. (2010). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* 31, 455–461. doi:10.1002/jcc.21334

Turner, P. J. (2005). *XMGRACE, Version 5.1.19.* Beaverton, OR: Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology.

Tyagi, R., Srivastava, M., Jain, P., Pandey, R. P., Asthana, S., Kumar, D., et al. (2020). Development of Potential Proteasome Inhibitors against *Mycobacterium tuberculosis. J. Biomol. Struct. Dyn.* 1, 1–15. doi:10.1080/07391102.2020.1835722

Tyagi, R., Srivastava, M., Singh, B., Sharma, S., Pandey, R. P., Asthana, S., et al. (2021). Identification and Validation of Potent Mycobacterial Proteasome Inhibitor from Enamine Library. *J. Biomol. Struct. Dyn.*, 1–11. doi:10.1080/07391102.2021.1914173

Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., et al. (2011). R.E.D. Server: a Web Service for Deriving RESP and ESP Charges and Building Force Field Libraries for New Molecules and Molecular Fragments. *Nucleic Acids Res.* 39, W511–W517. doi:10.1093/nar/gkr288

Yin, W., Mao, C., Luan, X., Shen, D.-D., Shen, Q., Su, H., et al. (2020). Structural Basis for Inhibition of the RNA-dependent RNA Polymerase from SARS-CoV-2 by Remdesivir. *Science* 368, 1499–1504. doi:10.1126/science.abc1560

Zhang, L., and Zhou, R. (2020). Structural Basis of the Potential Binding Mechanism of Remdesivir to SARS-CoV-2 RNA-dependent RNA Polymerase. *J. Phys. Chem. B* 124, 6955–6962. doi:10.1021/acs.jpcb.0c04198

Zhang, W.-F., Stephen, P., Thériault, J.-F., Wang, R., and Lin, S.-X. (2020). Novel Coronavirus Polymerase and Nucleotidyl-Transferase Structures: Potential to Target New Outbreaks. *J. Phys. Chem. Lett.* 11, 4430–4435. doi:10.1021/acs.jpclett.0c00571

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Molecular Biosciences

TECHNOLOGY AND CODE
published: 30 August 2021
doi: 10.3389/fmolb.2021.636077

Models and Processes to Extract Drug-like Molecules From Natural Language Text

Zhi Hong[1]*, *J. Gregory Pauloski*[1], *Logan Ward*[2], *Kyle Chard*[1,2,3], *Ben Blaiszik*[2,3] and *Ian Foster*[1,2,3]

[1]Department of Computer Science, University of Chicago, Chicago, IL, United States, [2]Data Science and Learning Division, Argonne National Laboratory, Lemont, IL, United States, [3]Globus, University of Chicago, Chicago, IL, United States

Researchers worldwide are seeking to repurpose existing drugs or discover new drugs to counter the disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A promising source of candidates for such studies is molecules that have been reported in the scientific literature to be drug-like in the context of viral research. However, this literature is too large for human review and features unusual vocabularies for which existing named entity recognition (NER) models are ineffective. We report here on a project that leverages both human and artificial intelligence to detect references to such molecules in free text. We present 1) a iterative model-in-the-loop method that makes judicious use of scarce human expertise in generating training data for a NER model, and 2) the application and evaluation of this method to the problem of identifying drug-like molecules in the COVID-19 Open Research Dataset Challenge (CORD-19) corpus of 198,875 papers. We show that by repeatedly presenting human labelers only with samples for which an evolving NER model is uncertain, our human-machine hybrid pipeline requires only modest amounts of non-expert human labeling time (tens of hours to label 1778 samples) to generate an NER model with an F-1 score of 80.5%—on par with that of non-expert humans—and when applied to CORD'19, identifies 10,912 putative drug-like molecules. This enriched the computational screening team's targets by 3,591 molecules, of which 18 ranked in the top 0.1% of all 6.6 million molecules screened for docking against the 3CLPro protein.

Keywords: coronavirus disease-19, coronavirus disease-19 open research dataset challenge, named entity recognition, long-short term memory, data mining

Edited by:
James Leland Olds,
George Mason University,
United States

Reviewed by:
Valentina Tozzini,
Consiglio Nazionale delle Ricerche,
Italy
Shang Gao,
Oak Ridge National Laboratory (DOE),
United States

Correspondence:
Zhi Hong
hongzhi@uchicago.edu

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *30 November 2020*
Accepted: *11 August 2021*
Published: *30 August 2021*

Citation:
Hong Z, Pauloski J G, Ward L,
Chard K, Blaiszik B and Foster I (2021)
Models and Processes to Extract
Drug-like Molecules From Natural
Language Text.
Front. Mol. Biosci. 8:636077.
doi: 10.3389/fmolb.2021.636077

1 INTRODUCTION

The Coronavirus Disease (COVID-19) pandemic, caused by transmissible infection of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has resulted in tens of millions of diagnosed cases and over 1,450,000 deaths worldwide (Dong et al., 2020); straining healthcare systems, and disrupting key aspects of society and the wider economy. It is thus important to identify effective treatments rapidly via discovery of new drugs and repurposing of existing drugs. Here, we leverage advances in natural language processing to enable automatic identification of drug candidates being studied in the scientific literature.

The magnitude of the pandemic has resulted in an enormous number of academic publications related to COVID-19 research since early 2020. Many of these articles are collated in the COVID-19

Open Research Dataset Challenge (CORD-19) collection (Allen Institute For AI, 2020; Wang et al., 2020). With 198,875 articles at the time of writing, that collection is far too large for humans to read. Thus, tools are needed to automate the process of extracting relevant data, such as drug names, testing protocols, and protein targets. Such tools can save domain experts significant time and effort.

Extracting named entities from scientific texts has been studied for more than 2 decades. Prior work relied on matching tokens (words) in text to entries in existing databases or ontologies (Rindflesch et al., 1999; Furrer et al., 2019). However, for our task, existing drug databases like DrugBank cover both *too much*, in that they include entities of many types (for example, “rabbit” is in DrugBank as an allergen), and *too little*: because the creation of such databases is time-consuming, they cannot keep up with the new entities in the latest COVID’19-related publications (~200 k papers in the CORD’19 corpus), which is exactly what we want to extract here.

Machine learning (ML) and deep learning (DL) methods have also been used for identifying named entities from free text. While less dependent on a comprehensive ontology, they need labeled data for training. Traditional training data collection employs a brute-force approach, collecting a large corpus and labeling each and every word indiscriminatingly (Bada et al., 2012). The time and human resources involved is enormous. We, in contrast, had access to just a few assistants who worked on labeling only during free time in their day job. Thus we needed a more careful approach of selective labeling to maximize their effectiveness.

Towards this goal, we describe here how we have tackled two important problems: creating labelled training data via judicious use of scarce human expertise, and applying a named entity recognition (NER) model to automatically identify drug-like molecules in text. We are looking not only for novel drugs under development, but also any small molecule drug that has been used to treat patients with COVID’19 or similar infectious diseases like SARS and MERS. In the absence of expert-labeled data for the growing COVID literature, we employ an iterative model-in-the-loop collection process inspired by our previous work (Tchoua et al., 2019a,b) and demonstrate that it can build a high quality training set without input from domain scientists. We first assemble a small bootstrap set of human-verified examples to train a model for identifying similar examples. We then iteratively apply the model, use human reviewers to verify the predictions for which the model is least confident, and retrain the model until the improvement in performance is less than a threshold. (The human reviewers were administrative staff without scientific backgrounds, with time available for this task due to the pandemic.)

Having collected adequate training data via this model-guided human annotation process, we then use the resulting labeled data to re-train a NER model originally developed to identify polymer names in materials science publications (Hong et al., 2020b) and apply this trained model to CORD-19. We show that the labeled data produced by our approach are of sufficiently high quality than when used to train NER models, which achieves a best F-1 score of 80.5%—roughly equivalent to that achieved by non-expert humans.

The labeled data, model, and model results are all available online, as described in **Section 5**.

2 MATERIALS AND METHODS

We aim to develop and apply new computational methods to mine the scientific literature to identify small molecules that have been investigated or found useful as antiviral therapeutics. For example, processing the following sentence should allow us to determine that the drug *sofosbuvir* has been found effective against the Zika virus: “Sofosbuvir, an FDA-approved nucleotide polymerase inhibitor, can efficiently inhibit replication and infection of several ZIKV strains, including African and American isolates.” (Bullard-Feibelman et al., 2017).

This problem of identifying drug-like molecules in text can be divided into two linked problems: 1) identifying references to small therapeutic molecules (“drugs”) and 2) determining what the text says about those molecules. In this work, we consider potential solutions to the first problem.

A simple way to identify entities in text that belong to a specialized class (e.g., drug-like molecules) is to refer to a curated list of valid names, if such is available. In the case of drugs, we might think to use DrugBank (Wishart et al., 2018) or the FDA Drug Database (Center for Drug Evaluation and Research, 2020), both of which in fact list *sofosbuvir*. However, such databases are not in themselves an adequate solution to our problem, for at least two reasons. First, they are rarely complete. The tens of thousands of entity names in DrugBank and the FDA Drug Database together are just a tiny fraction of the billions of molecules that could potentially be used as drugs. Second, such databases may be overly general: DrugBank, for example, includes the terms “rabbit” and “calcium,” neither of which have value as antiviral therapeutics. In general, the use of any such list to identify entities will lead to both false negatives and false positives. We need instead to employ the approach that a human reader might follow in this situation, namely to scan text for words that appear in contexts in which a drug name is likely to appear. In the following, we explain how we combine human and artificial intelligence for this purpose.

2.1 Automated Drug Entity Extraction From Literature

Finding strings in text that refer to drug-like molecules is an example of Named Entity Recognition (NER) (Nadeau and Sekine, 2007), an important NLP task. Both grammatical and statistical (e.g., neural network-based) methods have been applied to NER; the former can be more accurate, but require much effort from trained linguists to develop. Statistical methods use supervised training on labeled examples to learn the contexts in which entities of interest (e.g., drug-like molecules) are likely to occur, and then classify previously unseen words as such entities if they appear in similar contexts. For instance, a training set may contain the sentence “Ribavirin was administered once daily by the i. p. route” (Oestereich et al., 2014), with *ribavirin* labelled as

Drug. With sufficient training data, the model may learn to assign the label *Drug* to *arbidol* in the sentence "Arbidol was administered once daily per os using a stomach probe" (Oestereich et al., 2014). This learning approach can lead to general models capable of finding previously unseen candidate molecules in natural language text.

The development of effective statistical NER models is complicated by the many contexts in which names can occur. For example, while the contexts just given for *ribavirin* and *arbidol* are similar, both are quite different from that quoted for *sofosbuvir* earlier. Furthermore, authors may use different wordings and sentence structures: e.g., "given by i. p. injection once daily" rather than "administered once daily by the i. p. route." Thus, statistical NER methods need to do more than learn template word sequences: they need to learn more abstract representations of the context(s) in which words appear. Modern NLP and NER systems do just that (Chiu and Nichols, 2016).

2.1.1 SpaCy and Keras-Long-Short Term Memory Models

We consider two NER models in this paper, SpaCy and a Keras long-short term memory (LSTM) model. Both models are publicly available on DLHub (Li et al., 2021) and GitHub, as described in **Section 5**.

SpaCy (Honnibal and Montani, 2020a; Honnibal et al., 2020) is an open source NLP library that provides a pre-trained entity recognizer that can recognize 18 types of entities, including PERSON, ORGANIZATION, LOCATION, and PRODUCT. Its model calculates a probability distribution of a word over the entity types, and outputs the type with the highest probability as the predicted type for that word. When pre-trained on the OntoNotes five dataset of over 1.5 million labeled words (Weischedel et al., 2013), the SpaCy entity recognizer can identify supported entities with 85.85% accuracy. However, it does not include drug names as a supported entity class, and thus we would need to retrain the SpaCy model on a drug-specific training corpus. Unfortunately, there is no publicly available corpus of labeled text for drug-like molecules in context. Thus, we need to use other methods to retrain this model (or other NER models), as we describe in **Section 4**.

While SpaCy is easy to use, it lacks flexibility: its end-to-end encapsulation does not expose many tunable parameters. Thus we also explore the use of a Keras-LSTM model that we developed in previous work for identification of polymers in materials science literature (Hong et al., 2020b). This model is based on the Bidirectional LSTM network with a conditional random field (CRF) layer added on top. It takes training data labeled according to the "IOB" schema. The first word in an entity is given the label "B" (Beginning), the following words in the same entity are labeled "I" (Inside), and non-entity words are labeled "O" (outside). During prediction, the Bi-LSTM network tries to assign one of "IOB" to each word in the input sentence, but it has no awareness of the validity of the label sequence. The CRF layer is used on top of Bi-LSTM to lower the probability of invalid label sequences (e.g., "OIO").

We compare the performance of SpaCy and Keras-LSTM models under various conditions in **Section 2.2**.

2.1.2 Model-In-The-Loop Annotation Workflow

We address the lack of labeled training data by using **Algorithm 1** (and see **Figure 1**) to assemble a set of human- and machine-labeled data from CORD-19 (Wang et al., 2020). In describing this process, we refer to paragraphs labeled automatically via a heuristic or model as *silver* and to silver paragraphs for which labels have been corrected by human reviewers as *gold*. We use the Prodigy machine learning annotation tool to manage the review process: reviewers are presented with a silver paragraph, with putative drug entities highlighted; they click on false negative and false positive words to add or remove the highlights and thus produce a gold paragraph. Prodigy saves the corrected labels in standard NER training data format.

Our algorithm involves three main phases, as follows. In the first **bootstrap** phase, we assemble an initial test set of gold paragraphs for use in subsequent data acquisition. We create a first set of silver paragraphs by using a simple heuristic: we select N_0 paragraphs from CORD-19 that contain one or more words in DrugBank with an Anatomical Therapeutic Chemical Classification System (ATC) code, label those words as drugs, and ask human reviewers to correct both false positives and false negatives in our silver paragraphs, creating gold paragraphs. In the subsequent **build test set** phase, we repeatedly use all gold paragraphs obtained so far to train an NER model; use that model to identify and label additional silver paragraphs, and engage human reviewers to correct false positives and false negatives, creating additional gold paragraphs. We repeat this process until we have N_t initial gold paragraphs.

In the third **build labeled set** phase, we repeatedly use an NER model trained on all human-validated labels obtained to date, with the N_t gold paragraphs from the bootstrap phase used as a test set, to identify and label promising paragraphs in CORD-19 for additional human review. To maximize the utility of this human effort, we present the reviewers only with paragraphs that contain one or more *uncertain words*, i.e., words that the NER model identifies as drug/non-drug with a confidence in the range (min, max). We continue this process of model retraining, paragraph selection and labeling, and human review until the F-1 score improves by less than ϵ.

The behavior of this algorithm is influenced by six parameters: N_0, N, N_t, ϵ min, and max. N_0 and N are the number of paragraphs that are assigned to human reviewers in the first and subsequent steps, respectively. N_t is the number of examples in the test set. ϵ is a threshold that determines when to stop collecting data. The min and max determine the confidence range from which words are selected for human review. In the experimental studies described below, we used $N_0 = 278$, $N = 120$, $N_t = 500$, $\epsilon = 0$, min = 0.45, and max = 0.55.

The NER model used in the model-in-the-loop annotation workflow to score words might also be viewed as a parameter. In the work reported here, we use SpaCy exclusively for that purpose, as it integrates natively with the Prodigy annotation

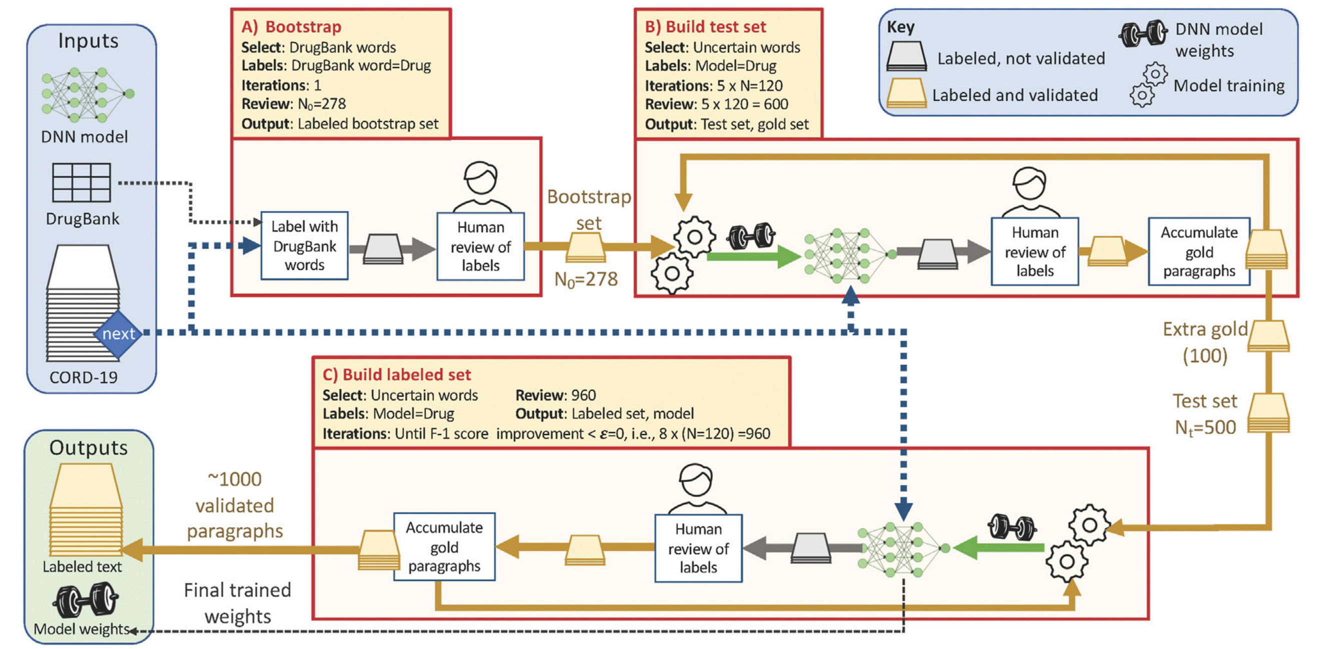


FIGURE 1 | Overview of the training data collection workflow, showing the three phases described in the text and with the parameter values used in this study. Each phase pulls paragraphs from the CORD-19 dataset (blue dashed line) according to the **Select** criteria listed (yellow shaded box). Phases B and C repeatedly update the weights for the NER model (green arrows) that they use to identify and label uncertain paragraphs; human review (yellow and gold arrows) corrects those silver paragraphs to yield gold paragraphs. Total human review work is ~278 + 600+960 = 1838 paragraphs.

tool and trains more rapidly. However, as we show below, the Keras-LSTM model is ultimately somewhat more accurate when trained on all of the labeled data generated, and thus is preferred when processing the entire CORD-19 dataset: see **Section 3.1.1** and **Section 3.2**.

This semi-automated method saves time and effort for human reviewers because they are only asked to verify labels that have already been identified by our model to be uncertain, and thus worth processing. Furthermore, as we show below, we find that we do not need to engage biomedical professionals to label drugs in text: untrained people, armed with contextual information (and online search engines), can spot drug names in text with accuracy comparable to that of experts.

We provide further details on the three phases of the algorithm in the following, with numbers in the list referring to line numbers in **Algorithm 1**.

A) Bootstrap

1. We start with the 2020–03-20 release version of the CORD-19 corpus, which contains 44,220 papers (Wang et al., 2020). We create $\mathcal{C}$, a random permutation of its paragraphs from which we will repeatedly fetch paragraphs via next ($\mathcal{C}$).
2. We bootstrap the labeling process by identifying as $\mathcal{D}$ the 2,675 items in the DrugBank ontology with a Anatomical Therapeutic Chemical Classification System (ATC) code attached (eliminating many, but not all, drug-like molecule entities).
3. We create an initial set of silver paragraphs, $\mathcal{P}_0$, by selecting N_0 paragraphs from $\mathcal{C}$ that include a word from $\mathcal{D}$.
4. We engage human reviewers to remove false positives and label false negatives in $\mathcal{P}_0$, yielding an initial set of gold paragraphs, $\mathcal{B}$.

B) Build test set

5. We expand the test set that we will use to evaluate the model created in the next phase, until we have N_t validated examples.
6. We train the NER model on 60% of the data collected to date and evaluate it on the remaining 40%, to create a new trained model, $\mathcal{M}$, with improved knowledge of the types of entities that we seek.
7. We use the probabilities over entities returned by the model to select, as our N new silver paragraphs, $\mathcal{P}$, paragraphs that contain at least one uncertain word (see above).
8. We engage human reviewers to convert these new silver paragraphs, $\mathcal{P}$, to gold, $\mathcal{V}$.
9. We add the new gold paragraphs, $\mathcal{V}$, to the bootstrap set $\mathcal{B}$.

11–12. Having assembled at least N_t validated examples, we select the first N_t as the test set, $\mathcal{T}$, and use any remaining examples to initialize the new gold set, $\mathcal{G}$.

C) Build labeled set

13. We assemble a training set $\mathcal{G}$, using the test set $\mathcal{T}$ assembled in the previous phases for testing. This process continues until the F-1 score stops improving see **Section 2.2.**

14–17. Same as Steps 6–9, except that we train on $\mathcal{G}$ and test on $\mathcal{T}$. Human reviewers are engaged to review new silver

Algorithm 1 | Model-in-the-loop Annotation Workflow

```
$\mathcal{C} := \textbf{randomized\_paragraphs}(\text{CORD})$ ;            // Prepare CORD-19 paragraphs
/* A) Bootstrap: Identify first silver paragraphs                          */
$\mathcal{D} := \{s : s \in \text{DrugBank} \;\&\; \text{ATC}(s)\}$ ;        // Extract names from DrugBank
$\mathcal{P}_0 := \{p : p = \textbf{next}(\mathcal{C} \;\&\; p \cap \mathcal{D} \neq \varnothing\} \;\&\; |\mathcal{P}_0| = N_0$ ;   // Assemble first silver set
$\mathcal{B} = \textbf{verify}(\mathcal{P}_0)$ ;          // Verify labels; initialize bootstrap set
/* B) Build test set: Expand on bootstrap set from (A)                     */
while $|\mathcal{B}| < N_t$ do
    $\mathcal{M} := \text{NER}(\text{train} = 0.6\mathcal{B}, \text{test} = 0.4\mathcal{B}$ ;     // Train: 60-40 train-test split
    $\mathcal{P} := \{p : p = \textbf{next}(\mathcal{C} \;\&\; \exists w \in p : \mathcal{M}(w) \in [\mathtt{min}, \mathtt{max}])\} \;\&\; |\mathcal{P}| = N$ ;   // Next silver
    $\mathcal{V} := \textbf{verify}(\mathcal{P})$ ;                  // Engage crowd to verify labels
    $\mathcal{B} := \mathcal{B} \cup \mathcal{V}$ ;          // Add corrected paragraphs to bootstrap set
end
$\mathcal{T} = \mathcal{B}[: N_t]$ ;                 // Initialize test set to first $N_t$ examples
$\mathcal{G} = \mathcal{B}[N_t :]$ ;           // Initialize gold set to any remaining examples
/* C) Build labeled set: Use test set from (B) to evolve model             */
while $F\text{-}1$ score improvement $> \epsilon$ do
    $\mathcal{M} := \text{NER}(\text{train} = \mathcal{G}, \text{test} = \mathcal{T}$ ;            // Train on $\mathcal{G}$, test on $\mathcal{T}$
    $\mathcal{P} := \{p : p = \textbf{next}(\mathcal{C} \;\&\; \exists w \in p : \mathcal{M}(w) \in [\mathtt{min}, \mathtt{max}])\} \;\&\; |\mathcal{P}| = N$ ;   // Next silver
    $\mathcal{V} := \textbf{verify}(\mathcal{P})$ ;                  // Engage crowd to verify labels
    $\mathcal{G} := \mathcal{G} \cup \mathcal{V}$ ;              // Add corrected paragraphs to gold set
end
```

paragraphs and produce new gold paragraphs, which are then added to $\mathcal{G}$ instead of $\mathcal{T}$.

2.2 Data-Performance Tradeoffs in Named-Entity Recognition Models

As noted in **Section 2.1.2**, our model-in-the-loop annotation workflow requires repeated retraining of a SpaCy model. Thus we conducted experiments to understand how SpaCy prediction performance is influenced by model size, quantity of training data, and amount of training performed.

As the training data produced by the model-in-the-loop evaluation workflow are to be used to train an NER model that we will apply to the entire CORD-19 dataset, we also evaluate the Keras-LSTM model from the perspectives of big data accuracy and training time.

2.2.1 Model Size

We first need to decide which SpaCy model to use for model-in-the-loop annotation. Model size is a primary factor that affects training time and prediction performance. In general, larger models tend to perform better, but require both more data and more time to train effectively. As our model-in-the-loop annotation strategy requires frequent model retraining, and furthermore will (initially at least) have little data, we hypothesize that a smaller model may be adequate for our purposes.

To explore this hypothesis, we study the performance achieved by the SpaCy medium and large models (Honnibal and Montani, 2020b) on our initial training set of 278 labeled paragraphs. We show in **Figure 2** the performance achieved by the two models as a function of number of training epochs. Focusing on the harmonic mean of precision and recall, the F-1 score (a good measure a model's ability to recognize both true positives and true negatives), we see that the two models achieve similar prediction performance, with the largest difference in F-1 score being around 2%. As the large model takes over eight times longer to train per epoch, we select the medium model for model-in-the-loop data collection.

2.2.2 Word Embedding Models

The Keras LSTM model requires external word vectors since, unlike SpaCy, it does not include a word embedding model. To explore the affect of different word embedding models we trained both BERT (Devlin et al., 2018), a top-performing language model developed by Google, and FastText (Bojanowski et al., 2016), a model shown to have outperformed traditional Word2Vec models such as CBOW and Skipgram in our previous work (Hong et al., 2020b). While Google has released pre-trained BERT models, and researchers often build upon these models by "fine-tuning" them with additional training on small external datasets, it is not suitable to our problem as the vocabulary used in the CORD-19 is very different than the

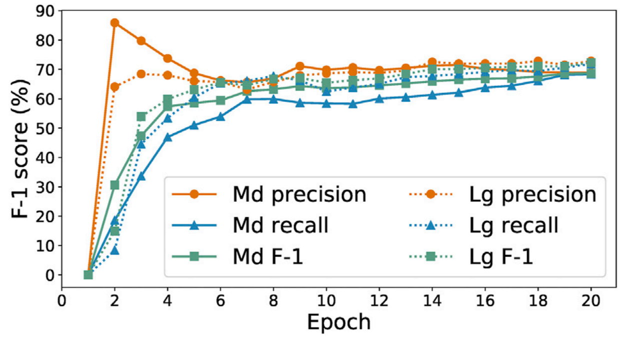


FIGURE 2 | Precision, recall, and F-1 scores of medium and large SpaCy models trained on 278 examples.

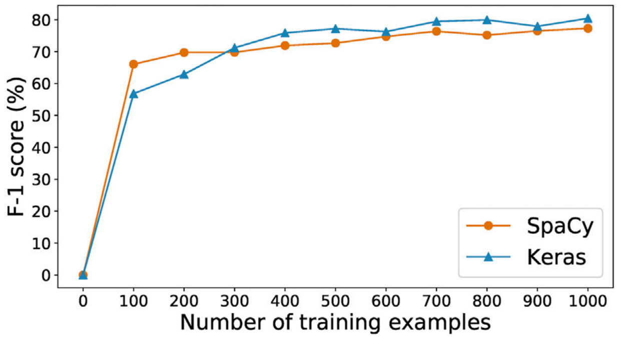


FIGURE 3 | Training curves for the SpaCy and Keras models for different number of examples collected.

datasets used to train these models.[1] Rather than use a pre-trained BERT model, we trained a BERT model on the CORD-19 corpus using a distributed neural network training framework from our previous work (Pauloski et al., 2020). As the CORD-19 corpus is approximately 20% of the size of the training data used by Google used to train BERT, we reduced the word embedding size proportionally from 768 to 128 to avoid over-fitting. For FastText word embeddings we set the size to the default 120.

We used word embeddings derived from both models to train the Keras LSTM model on the same training and testing data collected in **Section 2.1.2**. The model using FastText embeddings achived a slightly higher F-1 score (80.5%) than the model trained with BERT embeddings (78.7%). This result is likely due to the limited training data and embedding size. In short, the humongous BERT model requires an equally humongous amount of data to achieve the best performance, and without such it will not necessarily outperform other much smaller and less computationally intensive word embedding models. In the remainder of the paper, we use the FastText word embedding model.

2.2.3 Amount of Training Data

As data labeling is expensive in both human time and model training time, it is valuable to explore the tradeoff between time spent collecting data and prediction performance. To this end, we manually labeled a set of 500 paragraphs selected at random from CORD-19 (Wang et al., 2020) as a test set. Then, we used that test set to evaluate the results of training the SpaCy and Keras-LSTM models of **Section 2.1.1** on increasing numbers of the paragraphs produced by our human-in-the-loop annotation process. **Figure 3** shows their F-1 score curves as we scale from 0 to 1,000 training samples. With only 100 training examples, SpaCy and Keras-LSTM achieve F-1 scores of 57 and 66%, respectively. SpaCy performs better than Keras-LSTM with fewer training examples (i.e., less than 300), after which Keras-LSTM overtakes it and maintains a steady 2–3% advantage as the number of examples increases. This result motivates our choice of Keras-LSTM for the CORD-19 studies in **Section 3.2**.

We stopped collecting training data after 1,000 examples. We see in **Figure 3** that the performance of the SpaCy and Keras-LSTM models is essentially the same with 1,000 training examples as with 700 examples, with the F-1 score even declining when the number of available examples increases to 800 or 900. At 1,000 examples the F-1 score is greatest for both models. We conclude that the 1,000 training examples, along with the other 500 withheld as the test set, are best-suited to train our models. There are 4,244 and 1861 entities in the training and test set, respectively.

2.2.4 Training Epochs

Prediction performance is also influenced by the number of epochs spent in training. The cost of training is particularly important in a model-in-the-loop setup, as human reviewers cannot work while an model is offline for training.

Figure 4 shows the progression of the loss, precision, recall, and F-1 values of the SpaCy model during 100 epochs of training with the initial 278 examples. We can see that the best F-1 score is achieved within 10–20 epochs. Increasing the number of epochs does not result in any further improvement. Indeed, F-1 score does not tell us all about the model's performance. Sometimes training for more epochs could lead to lower loss values while other metrics (such as precision, recall, or F-1) no longer improve. That would still be desirable because it means the model is now more "confident," in a sense, about its predictions. However, that is not the case here. As shown in **Figure 4**, after around 40 epochs the loss begins to oscillate instead of continuing downwards, suggesting that in this case training for 100 epochs does not result in a better model than only training for 20 epochs.

Figure 5 shows the progression of accuracy and loss value for the Keras-LSTM model with the initial 278 examples. In **Figure 5A**, we see that validation accuracy improves as training accuracy increases during the first 50 epochs. After around epoch 50, the training and validation accuracy curves

[1] A biomedical-focused BERT model PubMdBERT (Gu et al., 2020) was released after the work described in this paper was done. We have fine-tuned the PubMedBERT-base-uncased-abstract-fulltext model on our dataset with the following parameters: batch size: 16, warmup steps: 500, weight decay: 0.01, training epochs: 3. It achieved a F-1 score of 0.885.

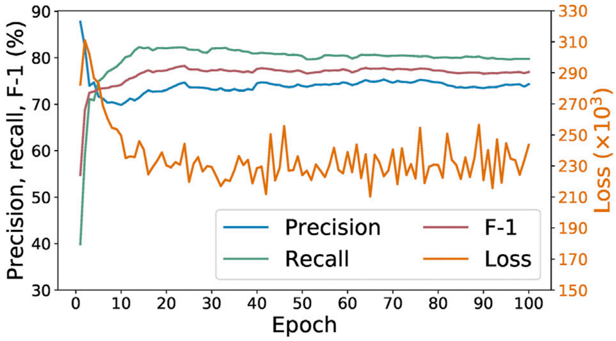


FIGURE 4 | Loss, precision, recall, and F-1 of SpaCy model during training for 100 epochs on 278 paragraphs.

diverge: the training accuracy continues to increase but the validation accuracy plateaus. This trend is suggestive of overfitting, which is corroborated by **Figure 5B**. After about 50 epochs, the validation loss curve turns upwards. Hence we choose to limit the training epochs to 64. After each epoch, if a lower validation loss is achieved, the current model state is saved. After 64 epochs, we test the model with the lowest validation loss on the withheld test set.

3 RESULTS

We present the results of experiments in which we first evaluate the performance of our models from various perspectives and then apply the models to the CORD-19 dataset.

3.1 Evaluating Model and Human Performance

We conducted experiments to compare the performance of the SpaCy and Keras-LSTM NER models; compare the performance of the models against humans; determine how training data influences model performance; and analyze human and model errors.

3.1.1 Performance of SpaCy and Keras Named-Entity Recognition Models

We used the collected data of **Section 2.1.2** to train both the SpaCy and Keras-LSTM NER models of **Section 2.1.1** to recognize and extract drug-like molecules in text. The SpaCy model used is the medium English language model en_core_web_md (Honnibal and Montani, 2020b). For the Keras model, the input embedding size is 120, the LSTM and Fully-connected hidden layers have a size of 32, and the dropout rate is 0.1. The model is trained for 64 epochs with a batch size of 64. We find that the trained SpaCy model achieved a best F-1 score of 77.3%, while the trained Keras-LSTM model achieved a best F-1 score of 80.5%, somewhat outperforming SpaCy.

As shown in **Figure 3**, the SpaCy model performs better than the Keras-LSTM model when trained with small amounts of training data—perhaps because of the different mechanisms employed by the two methods to generate numerical representations for words. SpaCy's built-in language model, pre-trained on a general corpus of blog posts, news, comments, etc., gives it some knowledge about commonly used words in English, which are likely also to appear in a scientific corpus. On the other hand, the Keras-LSTM model uses custom word embeddings trained solely on an input corpus, which provides it with better understanding of multi-sense words, especially those that have quite different meanings in a scientific corpus. However, without enough raw data to draw contextual information from, custom word embeddings can not accurately reflect the meaning of words.

3.1.2 Comparison Against Human Performance

Recognizing drug-like molecules is a difficult task even for humans, especially non-medical professionals (such as our non-expert annotators). To assess the accuracy of the annotators, we asked three people to examine 96 paragraphs, with their associated labels, selected at random from the labeled examples. Two of these reviewers had been involved in creating the labeled dataset; the third had not. For each paragraph, each reviewer decided independently whether each drug molecule entity was labeled correctly (a true positive), was labeled as a

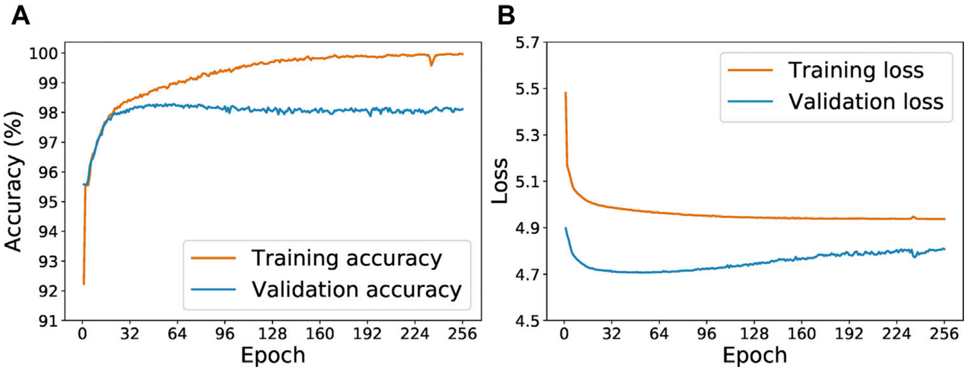


FIGURE 5 | Evolution of the training and validation accuracy **(A)** and loss **(B)** as the Keras-LSTM model is trained for from 1 to 256 epochs.

TABLE 1 | K-fold (*K* = 5) validation of the SpaCy model on 96 paragraphs with original vs. corrected labels. The first five rows are the results of each fold; the last row is the average F-1 score of the five folds.

Original labels			Corrected labels		
Precision	**Recall**	**F-1**	**Precision**	**Recall**	**F-1**
100.0	25.6	40.7	100.0	20.9	34.6
93.8	50.6	65.7	88.9	53.9	67.1
93.0	63.5	75.5	92.0	73.0	81.4
76.5	45.2	57.1	68.6	61.4	64.8
98.9	92.3	95.5	99.2	92.1	95.5
Average	—	66.9	Average	—	68.7

TABLE 2 | K-Fold (K = 5) validation of the Keras-LSTM model on 96 paragraphs with original vs. corrected labels. The first five rows are the results of each fold; the last row is the average F-1 score of the five folds.

Original Labels			Corrected Labels		
Precision	**Recall**	**F-1**	**Precision**	**Recall**	**F-1**
88.5	53.5	66.7	80.6	69.1	74.4
70.6	57.8	63.6	90.2	59.1	71.4
70.0	67.7	68.9	82.9	47.5	60.4
80.8	55.3	65.6	80.8	51.2	62.7
78.1	56.8	65.8	75.0	67.9	71.3
Average	—	66.1	Average		68.0

drug when it was not (a false positive), or was not labeled (a false negative). If all three reviewers agreed in their opinions on a paragraph (the case for 88 of the 96 paragraphs), we accepted their opinions; if they disagreed (the case for eight paragraphs), we engaged an expert.

This process revealed a total of 257 drug molecule entities in the 96 paragraphs, of which the annotators labeled 201 correctly (true positives), labeled 49 incorrectly (false positives), and missed 34 (false negatives). The numbers of true positives and false negatives do not sum up to the total number of drug molecules because in some cases an annotator labeled not to a drug entity but the entity plus extra preceding or succeeding word or punctuation mark (e.g., “sofosbuvir,” instead of “sofosbuvir”) and we count such occurrences as false positives rather than false negatives. In this evaluation, the non-expert annotators achieved an F-1 score of 82.9%, which is comparable to the 80.5% achieved by our automated models, as shown in **Figure 3**. In other words, our models have performance on par with that of non-expert humans.

3.1.3 Effects of Training Data Quality on Model Performance

We described in the previous section how review of 96 paragraphs labeled by the non-expert annotators revealed an error rate of about 20%. This raises the question of whether model performance could be improved with better training data. To examine this question, we compare the performance of our models when trained on original vs. corrected data. As we only have 96 corrected paragraphs, we restrict our training sets to those 96 paragraphs in each case.

We sorted the 96 paragraphs in both datasets so that they are considered in the same order. Then, we split each dataset into five subsets for K-fold cross validation (*K* = 5), with the first four subsets having 19 paragraphs each and the last subset having 20. Since *K* is set to five, the SpaCy and Keras models are trained five times. In the *i*th round, each model is trained on four subsets (excluding the *i*th) of each dataset. The *i*th subset of the corrected dataset is used as the test set. The *i*th subset of the original dataset is not used in the *i*th round.

We present the K-fold cross validation results in **Tables 1**, **2**. The models performed reasonably well when trained on the original dataset, with an average F-1 score only 2% less than that achieved with the corrected labels. Given that the expert input required for validation is hard to come by, we believe that using non-expert reviewers is an acceptable tradeoff and probably the only practical way to gather large amounts of training data.

3.2 Applying the Trained Models

After training the models with the labeled examples, we applied the trained models to the entire CORD-19 corpus (2020–10-04 version with 198,875 articles) to identify potential drug-like molecules. Processing a single article takes only a few seconds; we adapted our models to use data parallelism to enable rapid processing of these many articles.

We ran the SpaCy model on two Intel Skylake 6,148 processors with a total of 40 CPU cores; this run took around 80 core-hours and extracted 38,472 entities. We ran the Keras model on four NVidia Tesla V100 GPUs; this run took around 40 GPU-hours and extracted 121,680 entities. We recorded for each entity the number of the times that it has been recognized by each model, and used those numbers as a voting mechanism to further determine which entities are the most likely to be actual drugs. In our experiments, “balanced” entities (i.e., those whose numbers of detection by the two models are within a factor of 10 of each other) are most likely to appear in the DrugBank list. As shown in **Figure 6**, we sorted all extracted entities in descending order by their total number of detections by both models and compared the top 100 entities to DrugBank. We found that only 77% were exact matches to drug names or aliases, or 86% if we included partial matches (i.e., the extracted entity is a word within a multi-word drug name or alias in DrugBank). In comparison, among the top 100 “balanced” entities, 88% were exact matches to DrugBank, or 91% with partial matches.

Although DrugBank provides a reference metric to evaluate the results, it is not an exhaustive list of known drugs. For instance, remdesivir, a drug that has been proposed as a potential cure for COVID-19, is not in DrugBank. We manually checked via Google searches the top 50 “balanced” and top 50 “imbalanced” entities not matched to DrugBank, and found that 70% in the “balanced” list are actual drugs, but only 26% in the “imbalanced” list. Looking at the false positives in these top 50 lists, the “balanced” false positives are often

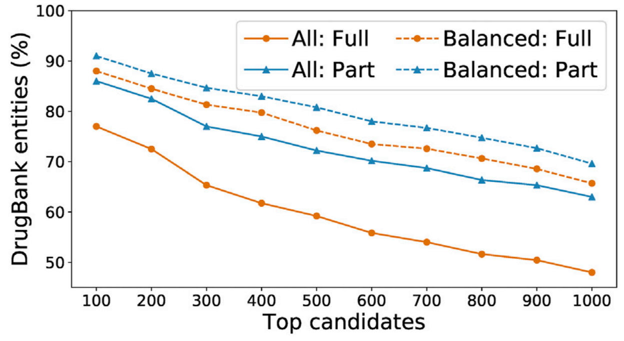


FIGURE 6 | Percentage of detected entities that are also found in DrugBank, when running either on all words found by our model or on just the balanced subset, and with "found" defined as either a full or partial match.

understandable. For example, in the sentence "ELISA plate was coated with … and then treated for 1 h at 37.8°C with dithiothreitol … ", the model mistook the redox reagent *dithiothreitol* for a drug entity, probably due to its context "treated with." On the other hand, we found no such plausible explanations for the false positives in the "imbalanced" list, where most false positives are chemical elements (e.g., silver, sodium), amino acids (e.g., cysteine, glutamine), or proteins (e.g., lactoferrin, cystatin).

Finally, we compared our extraction results to the drugs being used in clinical trials, as listed on the United States National Library of Medicine website (National Institutes of Health (2020). We queried the website with "covid" as the keyword and manually screened the returned drugs in the "Interventions" column to remove stopwords (e.g., tablet, injection, capsule) and dosage information (e.g., 2.5 mg, 2.5%) and only keep the drug names. Then we compared the top 50 most frequently appeared drugs to the automatically extracted drugs from literature. The "balanced" entities extracted by both models matched to 64% of the top 50 drugs in clinical trial, whereas the "imbalanced" entities only matched to 6% in the same list.

The results discussed here are available in the repository described in **Section 5**.

3.3 Validating the Utility of Identified Molecules

We are interested to evaluate the relevance to COVID research of the molecules that we extracted from CORD19. To that end, we compared our extracted molecule list against ZINC and Drugbank sets that a group of scientists at Argonne National Laboratory had used for computational screening. We found that our list contained an additional 3,591 molecules not found in their screening sets (filtered by their canonical SMILE strings). Applying their methods to screen those 3,591 molecules for docking against a main coronavirus protease, 3CLPro, revealed that 18 had docking scores in the top 0.1% of the 6.6 M ZINC molecules that they had screened previously (Saadi et al., 2020)—significantly more than the four that we would expect by chance.

TABLE 3 | The 10 most frequent tokens, excluding stopwords and punctuation marks, within various window sizes around entities incorrectly labeled by human reviewers.

#	Window size = 1		Window size = 3		Window size = 5	
	Token	Count	Token	Count	Token	Count
1	300	4	Mg	19	mg	23
2	oral	3	once/day	7	daily	8
3	dose	3	treatment	6	treatment	8
4	intravenous	2	300	5	once/day	7
5	500	2	treated	4	300	7
6	intravenously	1	Oral	4	oral	6
7	include	1	Once	4	recipients	5
8	Both	1	Dose	4	treated	4
9	resistance	1	Cidofovir	3	twice	4
10	treatment	1	resistance	3	include	4

As reported by Babuji et al. (2020a), those researchers have leveraged the outputs of our models in a computational screening pipeline that leverages HPC resources at scale, coupled with multiple artificial intelligence and simulation-based approaches (including leads from NLP), to identify high-quality therapeutic compounds to screen experimentally. A further manuscript, detailing the end-to-end process from data collection to simulation, incorporation of these results, cellular assays, and identification of high performing therapeutic compounds, is in preparation.

4 DISCUSSION: ANALYSIS OF HUMAN AND MODEL ERRORS

Finally, we explore the contexts in which human reviewers and models make mistakes. Specifically, we study the tokens that appear most frequently near to incorrectly labeled entities. To investigate the effects of immediate and long-distance context, we control, as *window size*, the maximum distance between a token and a entity for that token to be considered as "context" for that entity.

One difficulty with this analysis is that the most frequent tokens identified in this way were mostly stop words or punctuation marks. For instance, when the window size is set to three, the 10 most frequent tokens around mislabeled words are, in descending order, "comma (,)," "and," "mg," "period (.)," "right parenthesis ()),” "with," "of," "left parenthesis ((),” "is," and "or." Only "mg" is neither a stop word nor punctuation mark.

Those tokens provide little insight as to why human reviewers might have made mistakes, and furthermore are unlikely to have influenced reviewer decisions. Thus we exclude stopwords and punctuation marks when providing, in **Table 3**, lists of the 10 most frequent tokens within varying window sizes of words that were *incorrectly* identified as molecules by human reviewers.

TABLE 4 | The 10 most frequent tokens, excluding stop words and punctuation marks, within various window sizes around entities correctly labeled by human reviewers.

#	Window size = 1		Window size = 3		Window size = 5	
	Token	**Count**	**Token**	**Count**	**Token**	**Count**
1	resistance	176	Tetracycline	230	Tetracycline	230
2	treatment	9	resistance	177	resistance	178
3	mM	4	Trimethoprim	118	Trimethoprim	118
4	oral	3	treatment	11	treatment	14
5	after	3	20 ~	7	20 ~	8
6	analogue	3	Figure	5	placebo	7
7	responses	3	concentration	5	effects	6
8	antibiotics	2	compared	4	Figure	6
9	exposure	2	100	4	KLK5	6
10	pharmacokinetics	2	mM	4	Matriptase	6

TABLE 5 | The 20 most frequent tokens, including stop words and punctuation marks, within various window sizes around entities incorrectly labeled by human reviewers. Words that are neither stop words nor punctuation words are in boldface.

#	Window size = 1		Window size = 3		Window size = 5	
	Token	**Count**	**Token**	**Count**	**Token**	**Count**
1	—	27	—	49	—	74
2	—	14	And	21	—	28
3	And	14	**Mg**	19	And	28
4	with	8	—	18	**Mg**	23
5	(	7	)	13	)	21
6	Is	7	With	10	Of	18
7	Of	6	Of	10	(	17
8	Was	6	(	9	with	13
9	Or	4	Is	9	to	12
10	**300**	4	Or	7	the	11
11	**oral**	3	**once/day**	7	a	11
12	Has	3	A	7	in	10
13	To	3	The	6	is	9
14	[	3	Was	6	or	9
15	**dose**	3	To	6	**daily**	8
16	**intravenous**	2	**treatment**	6	was	8
17	In	2	In	5	**treatment**	8
18	may	2	**300**	5	]	8
19	**500**	2	Be	4	**once/day**	8
20	A	2	**Treated**	4	were	7

TABLE 6 | The 20 most frequent tokens, including stop words and punctuation marks, within various window sizes around entities incorrectly labeled by the Keras model. Words that are neither stop words nor punctuation words are in boldface.

#	Window size = 1		Window size = 3		Window size = 5	
	Token	**Count**	**Token**	**Count**	**Token**	**Count**
1	—	166	—	347	—	468
2	(	86	and	126	and	176
3	And	81	(	117	(	162
4	Of	58	)	89	)	143
5	)	30	of	85	of	130
6	To	28	the	73	the	121
7	Or	24	to	60	to	85
8		22		48	in	75
9	**mM**	18	with	44	with	72
10	With	17	a	41		68
11	In	15	in	37	a	62
12	For	15	or	35	was	52
13	Is	14	was	33	or	47
14	As	14	**mM**	32	is	43
15	The	13	is	31	for	42
16	Was	12	as	29	that	37
17	That	11	that	25	**mM**	36
18	[	11	by	22	by	35
19	**treatment**	9	for	22	as	32
20	A	9	[	22	were	31

We see that there are indeed several deceptive contextual words. With a window size of one, the 10 most frequent tokens include "oral," "dose," and "intravenous." It is understandable that an untrained reviewer might label as drugs words that immediately precede or follow such context words. Similar patterns can be seen for window sizes of three and five. Without background knowledge to draw from, non-experts are more likely to rely on their experience gained from labeling previous paragraphs. One may hypothesize that after the reviewers have seen a few dozen to a few hundred paragraphs, those deceptive contextual words must have left a deep impression, so that when those words re-appear they are likely to label the strange unknown word close to them as a drug.

To investigate this hypothesis, we also explored the most frequent words around drug entities that are *correctly* labeled by human reviewers: see **Table 4**. Interestingly, we found overlaps between the lists in **Tables 3**, **4**: in all, three, four, and two overlaps for window sizes of one, three, and five, respectively, when treating all numerical values as identical. This finding supports our hypothesis that those frequent words around real drug entities may confuse human reviewers when they appear around non-drug entities.

We repeat this comparison of context words around human and model errors while considering stopwords and punctuation marks. **Tables 5**, **6** show the 20 most frequent tokens in each case. We see that 20–25% of the tokens in **Table 5**, but only 5–10% of those in **Table 6**, are *not* stop words or punctuation marks. As the model only learns its word embeddings from the input text, if a token often co-occurs with drug entities in the training corpus the model will treat it as an indication of drug entities near its presence, regardless of whether or not it is a stopword. This apparently leads the model to make incorrect inferences. Humans, on the other hand, are unlikely to think that stopword such as "the" is indicative of drug entities, no matter how frequently they appear together.

5 DATA AVAILABILITY AND FORMATS

We have made our annotated training data, trained models, and the results of applying the models to the CORD-19 corpus publicly available online. (Babuji et al., 2020b).

In order to facilitate training of various models, we published the training data in two formats—an unsegmented version in line-delimited JSON (JSONL) format, and a segmented version in Comma Separated Value (CSV) format. The JSONL format contains the most comprehensive information that we have collected on the paragraphs in the dataset. We choose JSONL format rather than a JSON list because it allows for the retrieval of objects without having to parse the entire file. A JSON object in the JSONL file has the following structure:

- text: The original paragraph stored as a string without any modification.
- tokens: The list of tokens from text after tokenization.
- text: The text of the token as a string.
- start: The index of the first character of the token in text.
- end: The index of the first character after the token in text.
- id: Zero-based numbering of the token.
- spans: The list of spans (sequences of tokens) that are labeled as named entities (drugs)
- start: The index of the first character of the span in text.
- end: The index of the first character after the span in text.
- token_start: The index of the first token of the span in text.
- token_end: The index of the last token of the span in text.
- label: The label of the span ("drug")

Another commonly adopted labeling scheme for NER datasets is the "IOB" labeling scheme, in which the original text is first tokenized and each token is assigned a label "I," "O," or "B." The label "B (eginning)" means the corresponding token is the first in a named entity. A label "I (nside)" is given to every token in a named entity except for the first token. All other tokens gets the label "O (utside)" which means they are not part of any named entity. The aforementioned JSONL data are converted according to the IOB scheme and stored in Comma Separated Value (CSV) files with one training example per line. Each line consists of two columns: a first of tokens that made up of the original texts, and a second of the corresponding IOB labels for those tokens. In addition to a different labeling scheme, the samples in the CSV files are segmented, meaning that each sentence is treated as a training sample instead of an entire paragraph. This structure aligns with that used in standard NER training sets such as CoNLL03 (Sang and De Meulder, 2003).

The trained SpaCy and Keras models and the results of applying the models to the 198,875 articles in the CORD-19 corpus are also available in this GitHub repo. Additionally, the pre-trained SpaCy model is provided as a cloud service via DLHub (Hong et al., 2020a; Li et al., 2021). (The Keras model could not be hosted there due to compatibility issues with DLHub.) This cloud service allows researchers to apply the model to any texts they provide with as few as four lines of code.

6 CONCLUSION AND FUTURE DIRECTIONS

We have presented a human-machine hybrid pipeline for collecting training data for named entity recognition (NER) models. We applied this pipeline to create a NER model for identifying drug-like molecules in COVID-19-related research papers. Our pipeline facilitated efficient use of valuable human resources by presenting human labellers only with samples that were most likely to confuse the NER model. We explored various trade-offs, including model size, number of training samples, and training epochs, to find the right balance between model performance and time-to-result. In total, human reviewers working with our pipeline validated labels for 278 bootstrap samples, 1,000 training samples, and 500 test samples. As this work was performed in conjunction with other tasks, we cannot accurately quantify the total effort taken to collect and annotate the above training and test samples, but it was likely around 100 person-hours.

NER models trained with these training data achieved a best F-1 score of 80.5% when evaluated on our collected test set. Our models correctly identified 64% of the top 50 drugs that are in clinical trials for COVID-19, and when applied to 198,875 articles in the CORD-19 collection, identified 10,912 molecules with potential therapeutic effects against the SARS-CoV-2 coronavirus. The code, model, and extraction results are publicly available. Our work provided an additional 3591 SMILES strings to scientists at Argonne National Laboratory to be used in computational screening pipelines, of which 18 ranked in the top 0.1% of the molecules screened. Babuji et al. (2020a) have leveraged the outputs of our models in a computational screening pipeline that leverages HPC resources at scale to identify high-quality therapeutic compounds to screen experimentally. A further manuscript detailing the end-to-end process of identifying high performing therapeutic compounds is in preparation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/globus-labs/covid-nlp/tree/master/drug-ner.

AUTHOR CONTRIBUTIONS

ZH and JGP conducted the computational experiments. BB and LW contributed to discussions of ideas. ZH wrote the first draft. KC and IF revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act. This research used resources of the Argonne Leadership Computing Facility, a DOE Office of Science User Facility supported under Contract DE-AC02–06CH11357. This work was also supported by financial assistance award 70NANB19H005 from United States Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD).

ACKNOWLEDGMENTS

The data generated have been prepared as part of the nCov-Group Collaboration, a group of over 200 researchers working to use computational techniques to address various challenges associated with COVID-19. We are grateful to our human reviewers: India S. Gordon, Linda Novak, Kasia Salim, Susan Sarvey, Julie Smagacz, and Monica Orozco White.

REFERENCES

Allen Institute For AI (2020). [Dataset]: COVID-19 Open Research Dataset challenge. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

Babuji, Y., Blaiszik, B., Brettin, T., Chard, K., Chard, R., Clyde, A., et al. (2020a). *Targeting SARS-CoV-2 with AI-And HPC-Enabled lead Generation: A First Data Release*. arXiv. preprint arXiv:2006.02431.

Babuji, Y., Blaiszik, B., Chard, K., Chard, R., Foster, I., Gordon, I., et al. (2020b). [Dataset]: Lit - A Collection of Literature Extracted Small Molecules to Speed Identification of COVID-19 Therapeutics. Materials Data Facility. https://github.com/globus-labs/covid-nlp/tree/master/drug-ner.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., et al. (2012). Concept Annotation in the CRAFT Corpus. *BMC bioinformatics*. 13, 161. doi:10.1186/1471-2105-13-161

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguistics*. 5, 135–146.

Bullard-Feibelman, K. M., Govero, J., Zhu, Z., Salazar, V., Veselinovic, M., Diamond, M. S., et al. (2017). The FDA-Approved Drug Sofosbuvir Inhibits Zika Virus Infection. *Antiviral Res.* 137, 134–140. doi:10.1016/j.antiviral.2016.11.023

Center for Drug Evaluation and Research (2020). [Dataset]: About Drugs@FDA. https://www.fda.gov/drugs/drug-approvals-and-databases/about-drugsfda.

Chiu, J. P. C., and Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *Tacl* 4, 357–370. doi:10.1162/tacl_a_00104

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. preprint arXiv:1810.04805.

Dong, E., Du, H., and Gardner, L. (2020). An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. *Lancet Infect. Dis.* 20, 533–534. doi:10.1016/s1473-3099(20)30120-1

Furrer, L., Jancso, A., Colic, N., and Rinaldi, F. (2019). OGER++: Hybrid Multi-type Entity Recognition. *J. Cheminform.* 11, 7–10. doi:10.1186/s13321-018-0326-3

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., et al. (2020). *Domain-specific Language Model Pretraining for Biomedical Natural Language Processing*. arXiv. preprint arXiv:2007.15779.

Hong, Z., Pauloski, J., Ward, L., Chard, K., Blaiszik, B., and Foster, I. (2020a). [Dataset]: A Model to Extract Drug-like Molecules from Natural Language Text. Data and Learning Hub for Science. https://10.26311/EV06-C385.

Hong, Z., Tchoua, R., Chard, K., and Foster, I. (2020b). "SciNER: Extracting Named Entities from Scientific Literature," in *International Conference on Computational Science* (Amsterdam: Springer), 308–321. doi:10.1007/978-3-030-50417-5_23

Honnibal, M., and Montani, I. (2020a). [Dataset]: SpaCy Pipeline Design. https://spacy.io/models#design.

Honnibal, M., and Montani, I. (2020b). [Dataset]: SpaCy Pretrained English Models. https://spacy.io/models/en.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). [Dataset]: spaCy: Industrial-Strength Natural Language Processing in Python. https://10.5281/ zenodo.1212303

Li, Z., Chard, R., Ward, L., Chard, K., Skluzacek, T. J., Babuji, Y., et al. (2021). DLHub: Simplifying Publication, Discovery, and Use of Machine Learning Models in Science. *J. Parallel Distributed Comput.* 147, 64–76. doi:10.1016/j.jpdc.2020.08.006

Nadeau, D., and Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Lingvist. Investig.* 30, 3–26. doi:10.1075/li.30.1.03nad

National Institutes of Health (2020). [Dataset]: U.S. COVID-19 Clinical Trials. https://clinicaltrials.gov/ct2/results?cond=covid&cntry=US.

Oestereich, L., Rieger, T., Neumann, M., Bernreuther, C., Lehmann, M., Krasemann, S., et al. (2014). Evaluation of Antiviral Efficacy of Ribavirin, Arbidol, and T-705 (Favipiravir) in a Mouse Model for Crimean-Congo Hemorrhagic Fever. *Plos Negl. Trop. Dis.* 8, e2804. doi:10.1371/journal.pntd.0002804

Pauloski, J., Zhang, Z., Huang, L., Xu, W., and Foster, I. (2020). "Convolutional Neural Network Training with Distributed K-FAC," in 2020 SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (SC) (Atlanta, GA: IEEE Computer Society, 1331–1342. doi:10.1109/sc41405.2020.00098

Rindflesch, T. C., Tanabe, L., Weinstein, J. N., and Hunter, L. (1999). EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. *Biocomputing (World Scientific)* 2000, 517–528.

Saadi, A. A., Alfe, D., Babuji, Y., Bhati, A., Blaiszik, B., Brettin, T., et al. (2020). *IMPECCABLE: Integrated Modeling Pipeline for COVID Cure by Assessing Better Leads*. arXiv. preprint arXiv:2010.06574.

Sang, E. F., and De Meulder, F. (2003). "Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition," in 7th Conference on Natural Language Learning Edmonton, 142–147.

Tchoua, R., Ajith, A., Hong, Z., Ward, L., Chard, K., Audus, D., et al. (2019a). "Active Learning Yields Better Training Data for Scientific Named Entity Recognition," in 15th International Conference on eScience, San Diego, CA (Faro, Algarve: IEEE), 126–135. doi:10.1109/escience.2019.00021

Tchoua, R. B., Ajith, A., Hong, Z., Ward, L. T., Chard, K., Belikov, A., et al. (2019b). "Creating Training Data for Scientific Named Entity Recognition with Minimal Human Effort," in *International Conference on Computational Science*. Springer, 398–411. doi:10.1007/978-3-030-22734-0_29

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., et al. (2020). *CORD-19: The COVID-19 Open Research Dataset*. ArXiv. Preprint 2004.10706.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., et al. (2013). *OntoNotes Release 5.0*. Philadelphia, PA: Linguistic Data Consortium, 23.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi:10.1093/nar/gkx1037

Author Disclaimer: Unless otherwise indicated, this information has been authored by an employee or employees of the UChicago Argonne, LLC, operator of the Argonne National laboratory with the United States Department of Energy. The United States Government has rights to use, reproduce, and distribute this information. The public may copy and use this information without charge, provided that this Notice and any statement of authorship are reproduced on all copies.

While every effort has been made to produce valid data, by using this data, User acknowledges that neither the Government nor UChicago Argonne LLC makes any warranty, express or implied, of either the accuracy or completeness of this information or assumes any liability or responsibility for the use of this information. Additionally, this information is provided solely for research purposes and is not provided for purposes of offering medical advice. Accordingly, the United States Government and UChicago Argonne LLC are not to be liable to any user for any loss or damage, whether in contract, tort (including negligence), breach of statutory duty, or otherwise, even if foreseeable, arising under or in connection with use of or reliance on the content displayed on this site.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 08 November 2021
doi: 10.3389/fmolb.2021.614728

Prediction of Non-canonical Routes for SARS-CoV-2 Infection in Human Placenta Cells

Flávia Bessi Constantino[1†], *Sarah Santiloni Cury*[1†], *Celia Regina Nogueira*[2], *Robson Francisco Carvalho*[1]* *and Luis Antonio Justulin*[1]*

[1]*Department of Structural and Functional Biology, Institute of Biosciences, São Paulo State University (UNESP), Botucatu, Brazil,* [2]*Department of Internal Clinic, Botucatu Medicine School, São Paulo State University (UNESP), Botucatu, Brazil*

The SARS-CoV-2 is the causative agent of the COVID-19 pandemic. The data available about COVID-19 during pregnancy have demonstrated placental infection; however, the mechanisms associated with intrauterine transmission of SARS-CoV-2 is still debated. Intriguingly, while canonical SARS-CoV-2 cell entry mediators are expressed at low levels in placental cells, the receptors for viruses that cause congenital infections such as the cytomegalovirus and Zika virus are highly expressed in these cells. Here we analyzed the transcriptional profile (microarray and single-cell RNA-Seq) of proteins potentially interacting with coronaviruses to identify non- canonical mediators of SARS-CoV-2 infection and replication in the placenta. Despite low levels of the canonical cell entry mediators *ACE2* and *TMPRSS2*, we show that cells of the syncytiotrophoblast, villous cytotrophoblast, and extravillous trophoblast co-express high levels of the potential non-canonical cell-entry mediators *DPP4* and *CTSL*. We also found changes in the expression of *DAAM1* and *PAICS* genes during pregnancy, which are translated into proteins also predicted to interact with coronaviruses proteins. These results provide new insight into the interaction between SARS-CoV-2 and host proteins that may act as non-canonical routes for SARS-CoV-2 infection and replication in the placenta cells.

Keywords: COVID-19, SARS-CoV-2, placenta, virus entry mediator, DPP4, CTSL, DAAM1, PAICS

OPEN ACCESS

Edited by:
James Leland Olds,
George Mason University,
United States

Reviewed by:
Estela Bevilacqua,
University of São Paulo, Brazil
Arvind Ramanathan,
Argonne National Laboratory (DOE),
United States

****Correspondence:***
Luis Antonio Justulin
l.justulin@unesp.br
Robson Francisco Carvalho
robson.carvalho@unesp.br

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *06 October 2020*
Accepted: *15 October 2021*
Published: *08 November 2021*

Citation:
Constantino FB, Cury SS, Nogueira CR, Carvalho RF and Justulin LA (2021) Prediction of Non-canonical Routes for SARS-CoV-2 Infection in Human Placenta Cells. Front. Mol. Biosci. 8:614728. doi: 10.3389/fmolb.2021.614728

INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of the coronavirus disease 2019 (COVID-19) (Wu et al., 2020). It was first notified at the end of 2019, in Wuhan, China, and became a worldwide pandemic (Dong E. et al., 2020). At the beginning of September 2021, COVID-19 infected over 227 million people and is the cause of approximately 4.674.673 deaths worldwide (https://coronavirus.jhu.edu/).

Older age, laboratory abnormalities, and several comorbidities are associated with the more severe cases of COVID-19 (Williamson et al., 2020). For specific groups of COVID-19 patients, for example, pregnant women, the potential impacts of SARS-CoV-2 infection remain mostly unknown, and data are limited. However, considering previous works reporting coronaviruses infections (Schwartz and Graham, 2020), pregnant women are at higher risk of SARS-CoV-2 infection due to physiological changes in the immune, cardiorespiratory, and metabolic systems (Qadri and Mariona, 2020).

Although only a small number of maternal virus infections are transmitted to the fetus, some may cause life-threatening diseases (Pereira, 2018). These viruses use cellular host entry mediators expressed by placenta cells, as described for the cytomegalovirus and the Zika virus (Pierson and

Diamond, 2018; Hashimoto et al., 2020), to infect these cells. The Zika virus (ZIKV) outbreak, associated with fetal brain damage, emphasizes the necessity of further characterization and understanding of placental infection or intrauterine (vertical) transmission of SARS-CoV-2, as well as the possible adverse fetal outcomes. The few studies on the subject have provided contradictory findings, with some reports suggesting no evidence of placental infection or vertical transmission of SARS-CoV-2 (Celik et al., 2020; Chen et al., 2020; Yang and Liu, 2020). Conversely, multiple lines of evidence have shown placental SARS-CoV-2 infection in pregnant women diagnosed with moderate to severe COVID-19 (Schwartz and Morotti, 2020; Shende et al., 2021; Valdespino-Vázquez et al., 2021). Moreover, neonates born from mothers with COVID-19 presented a positive serological test for SARS-CoV-2 immunoglobulin (Ig) M and IgG (Dong L. et al., 2020; Zeng et al., 2020). While the IgG can be transferred from mother to fetus across the placenta, the detection of IgM in newborns suggests a vertical transmission of the virus, since IgM cannot cross the placental barrier due to its high molecular mass (Kimberlin and Stagno, 2020). Accordingly, SARS-CoV-2 RNA transmission was comprehensively confirmed by pathological and virological investigations (Patanè et al., 2020). Also, it was recently shown SARS-CoV-2 particles in syncytiotrophoblast with generalized inflammation, diffuse perivillous fibrin depositions, and tissue damage in an asymptomatic woman (Schoenmakers et al., 2020). Remarkably, these placental alterations due to the SARS-CoV-2 infection lead to fetal distress and neonatal multi-organ failure. These results highlight the importance of exploring the expression profile of potential host mediators of the SARS-CoV-2 that may create a permissive microenvironment to placental infection and enable vertical transmission of the virus.

In fact, like other viruses, SARS-CoV-2 requires diverse host cellular factors for infection and replication. The angiotensin-converting enzyme 2 (ACE2) is the canonical receptor for the SARS-CoV-2 spike protein receptor-binding domain (RBD) for viral attachment (Letko et al., 2020). This process is followed by S protein priming by cellular transmembrane serine protease 2 (TMPRSS2) that allows the fusion of the virus with host cellular membranes (Letko et al., 2020). Single-cell RNA sequencing (scRNA-Seq) has demonstrated that both *ACE2* and *TMPRSS2* are co-expressed in multiple tissues affected by COVID-19, including airway epithelial cells, cornea, digestive and urogenital systems (Sungnak et al., 2020). Few cells express *ACE2* and *TMPRSS2* in the placenta (Pique-Regi et al., 2020; Sungnak et al., 2020), suggesting that SARS-CoV-2 is unlikely to infect the placenta through the canonical cell entry mediators. Therefore, other host interacting proteins may play a role in the biological cycle of the virus and contribute to the pathogenesis of SARS-CoV-2 in the placenta. In this paper, we demonstrate, through transcriptomic (microarray and scRNA-Seq) analysis and *in silico* predictions of virus-host protein-protein interactions, that cells of the syncytiotrophoblast, villous cytotrophoblast, and extravillous trophoblast express high levels of potential non-canonical cell-entry mediators dipeptidyl peptidase 4 (*DPP4*) and cathepsin L (*CTSL*), despite low-levels of *ACE2* and *TMPRSS2*. We also found changes in the expression of *DAAM1* and *PAICS* genes (translated into proteins predicted to interact with coronaviruses proteins) during pregnancy, which co-express with *DPP4* and *CTSL* in placenta single cells (syncytiotrophoblast, villous cytotrophoblast, and extravillous trophoblast). These results open new avenues of investigation of the human placenta infection by the SARS-CoV-2.

RESULTS

Prediction of Non-canonical Routes for SARS-CoV-2 Infection in Human Placenta Cells

We first investigated the gene expression in placental tissues of classical host-virus interacting proteins described in the literature. The canonical entry receptors *ACE2* and *TMPRSS2* were low expressed throughout gestation. *CTSL*, which is translated into a lysosomal cysteine proteinase that plays a role in intracellular protein catabolism - presented the highest level of expression in the placenta during the first, second, and third trimester. Similarly, *DPP4*, which is translated into an intrinsic membrane glycoprotein, was highly expressed throughout gestation (**Figure 1A**).

Next, we analyzed the gene expression profile during gestation in placental tissues. We found 25 differentially expressed genes (DEGs) in the second trimester, and 687 DEGs in the third, when they were independently compared to the first trimester (**Supplementary Table S1**). All DEGs were divided into three clusters by the K-means clustering analysis **(Supplementary Figure S1A)**. Cluster 1 includes genes that increase expression during pregnancy, and these genes enriched terms related to blood vessels morphogenesis, complement cascade, extracellular matrix organization, and cellular response to nitrogen compounds. Cluster 2 encompasses genes that increase expression specifically in the third trimesters, which are related to female pregnancy, growth hormone signaling pathway, homeostasis, and steroid biosynthetic process. Cluster 3 includes genes that decrease expression during pregnancy. These genes enriched terms related to cell division, sulfur compounds biosynthetic process, chromosomal segregation, PID MYC active pathway (**Supplementary Figure S1B**, **Supplementary Table S2**). The gene expression changes in the course of pregnancy revealed that the second and third trimesters enriched genes that are associated with the placental growth, mainly in the third trimester (**Supplementary Figure S1C**, **Supplementary Table S3**). Even with a high discrepancy between the number of DEGs in the second and third trimesters (25 and 687, respectively), we identified similarities in the enriched terms between gene clusters (**Supplementary Figure S1D**) and between the second and third trimesters of gestation (**Supplementary Figure S1E**).

We further asked whether the DEGs in the placenta during gestation are translated into proteins that potentially interact with SARS-CoV proteins. Considering that the current cases of SARS-CoV-2 placental infection are mainly related to the third trimester

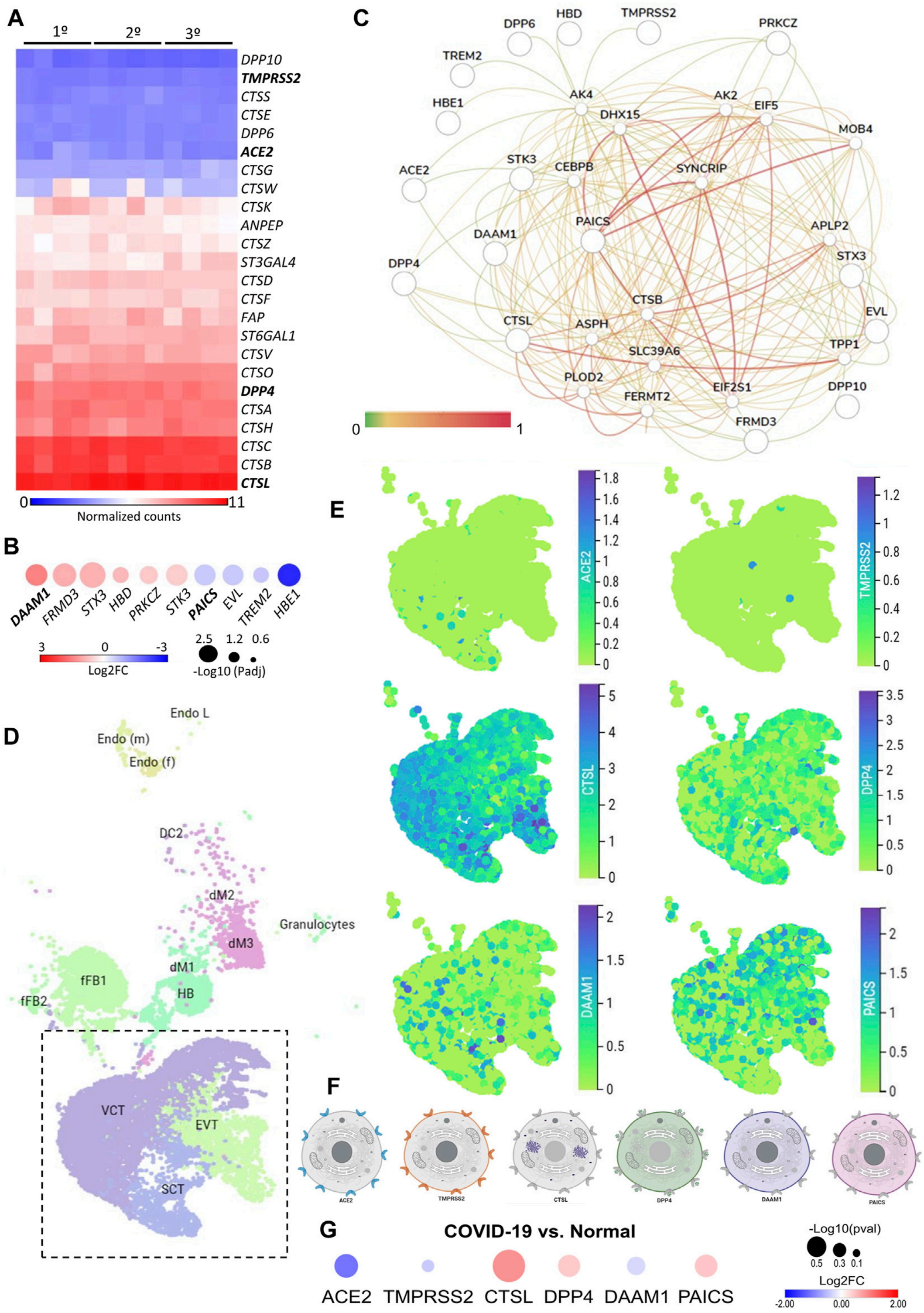


FIGURE 1 | The expression landscape of human placenta proteins potentially interacting with SARS-CoV-2. **(A)**. Heatmap illustrating gene expression (counts RMA normalized) of SARS-CoV-2 interacting proteins from placental samples in the first, second and third trimester of gestation (GSE9984). **(B)**. Heat-scatter plot presenting ten differentially expressed genes in the placenta that encode proteins potentially interacting with SARS-CoV as predicted by the P-HIPSter (http://phipster.org/) database. The color and size of the circles correspond to the log2FC and -log10 transformed FDR adjusted *p*-value, respectively. Fold change represents the gene expression of placenta samples from the third trimester of gestation compared with the first trimester. Bolded genes represent the six candidates selected for the further scRNA-seq evaluations. **(C)**. Tissue-specific gene network of placenta proteins predicted to interact with coronaviruses. The network was generated using the

(Continued)

FIGURE 1 | humanbase (https://hb.flatironinstitute.org/) online tool. **(D)**. Dimensionality reduction demonstrates different cell populations identified in five first-trimester human placental samples in a Uniform Manifold Approximation and Projection (UMAP) plot, by using scRNA-Seq data from Vento-Tormo et al., 2018. The images were generated using the COVID-19 Cell Atlas (https://www.covid19cellatlas.org) functionalities. **(E)**. UMAP plots representing the gene expression (log-transformed normalized counts) of *ACE2*, *TMPRSS2*, *CTSL*, *DPP4*, *DAAM1,* and *PAICS* in human placental cells from the syncytiotrophoblast, villous cytotrophoblast, and extravillous trophoblast. The images were generated using COVID-19 Cell Atlas (https://www.covid19cellatlas.org) functionalities. **(F)**. Cell illustrations depicting the protein localization of ACE2, TMPRSS2, CTSL, DPP4, DAAM1, and PAICS in cell compartments, as described in the Human Protein Atlas database (http://www.proteinatlas.org). Illustrations were created using BioRender App (https://biorender.com/). **(G)**. Heat-scatter plot presenting gene expression of *ACE2*, *TMPRSS2*, *CTSL*, *DPP4*, *DAAM1*, and *PAICS* in the placenta from COVID-19 pregnant women compared to healthy pregnant (GSE17995). The color and size of the circles correspond to the log2FC and -log10 transformed *p*-values, respectively. Fold change represents the gene expression of placenta samples from the third trimester of gestation of COVID-19 women compared with matched third trimester uninfected women. RMA: Robust Multichip Average; FC: Fold Change; FDR: false discovery rate; Padj: adjusted *p*-value; SCT, syncytiotrophoblast; VCT, villous cytotrophoblast; EVT, extravillous trophoblast; HB, Hofbauer cells; FB: fibroblasts; dM, decidual macrophages; Endo, endothelial cells; l, lymphatic; m, maternal; f, fetal.

of pregnancy (Patanè et al., 2020; Shanes et al., 2020), we selected the 687 DEGs in the placenta in the third trimester compared to the first trimester and, among them, 474 and 213 were up- and down-regulated, respectively (**Supplementary Table S1**). Next, we selected the human proteins potentially interacting with SARS-CoV using the P-HIPSter database. We found 32 virus-host interacting proteins of SARS-CoV (**Supplementary Table S4**), and nine virus-host interacting proteins of the ZIKV (**Supplementary Table S5**). We also found that, from this list of virus-host interacting proteins, 10 DEGs (*DAAM1*, *FRMD3*, *STX*3, *HBD*, *PRKCZ*, *CTK3*, *PAICS*, *EVL*, *TREM2*, and *HBE1*) are translated into proteins that interact with SARS-CoV proteins, and one with the ZIKV (**Supplementary Figure S1F,G**). Among these 10 DEGs, six were up-regulated (*DAAM1*, *FRMD3*, *STX*3, *HBD*, *PRKCZ*, and *CTK3*) and four (*PAICS*, *EVL*, *TREM2*, and *HBE1*) were down-regulated in the third trimester (**Figure 1B**). The gene *DAAM1,* which is translated into an intrinsic membrane glycoprotein implicated in cell motility, adhesion, cytokinesis, and cell polarity—showed the highest level of fold change in the third trimester of gestation compared to the first (logFC = 1.43; **Figure 1B**). The host-virus protein-protein interactions (PPI) predicted for these 10 DEGs presented a Likelihood Ratio (LR) > 100 (**Supplementary Table S6**), according to P-HIPSTer (Lasso et al., 2019). Noteworthy, *PAICS* transcript, which is translated into an enzyme that catalyzes the sixth and seventh steps of the novo purine biosynthesis, were predicted to interact with both SARS-CoV and ZIKV (**Figure 1B**, **Supplementary Table S7**). We observed a significant interaction of placental proteins predicted to interact with SARS-CoV-2 (based on the DEGs or not), and PAICS was also predicted to interact with other proteins in the PPI network with the highest degree (**Figure 1C**, **Supplementary Table S8**).

We next used single-cell RNA sequencing to analyze the expression of the 10 DEGs translated into proteins that potentially interact with SARS-CoV in human placental cells (**Supplementary Figure S1H,I**). We selected the genes *ACE2*, *TMPRSS2*, *CTSL*, *DPP4, PAICS,* and *DAAM1* for further investigations using scRNA-seq transcriptome data from cells of the syncytiotrophoblast (*n* = 1,144), villous cytotrophoblast (*n* = 8,244), and extravillous trophoblast (*n* = 2,170) of non-disease human placental tissues, considering the potential relevance of these genes for SARS-CoV infection and replication in the organ (**Figure 1D**). We noticed that the expression of the classical SARS-CoV-2 entry receptor genes (*ACE2* and *TMPRSS2*) was minimally expressed in these cells. In contrast, potential non-canonical cell entry mediator genes *DPP4* and *CTSL*, as well as the genes for the predicted virus-host interaction proteins *DAAM1,* and *PAICS* were expressed at higher levels (**Figure 1E**, **Supplementary Figure S2A**). We also looked for the co-expression of *ACE2* and *TMPRSS2* with *DPP4*, *CTSL*, *DAAM1,* and *PAICS* by using scRNA-Seq in these same cells (**Figure 2**). This analysis revealed that *DPP4, CTSL, DAAM1*, and *PAICS* are co-expressed at high levels, while these genes are co-express at low levels with *ACE2* and *TMPRSS2* (**Figure 2**). Remarkably, we found only three cells co-expressing *ACE2* and *TMPRSS2*. For this reason, we consider that *DPP4, CTSL, DAAM1*, and *PAICS* may represent candidates of an alternative route for SARS-CoV-2 infection in human placentas. We used The *Human Protein Atlas* to predict the subcellular location of proteins encoded by our candidates. DPP4 is predicted as intracellular, membrane, and secreted protein; CTSL is a protein located in the Golgi apparatus and additionally in vesicles; PAICS is a protein found in the cytosol while DAAM1 is in the plasma membrane and cytosol (**Figure 1F**). A recent published study investigated the gene expression alteration in the term placentas from COVID-19 pregnant women compared to uninfected (Lu-Culligan et al., 2021). We have accessed the DEGs identified by the authors, and we found that our candidate genes are not differentially expressed (**Figure 1G**). Moreover, *DAAM1* and *PAICS* are expressed in the opposite direction to the one found during gestational ages (**Figure 1B**). Finally, we analyzed the gene expression profile of *ACE2*, *TMPRSS2*, *CTSL*, *DPP4*, *DAAM1,* and *PAICS* genes on publicly available scRNA-Seq datasets of lung, liver, and thymus fetal tissues. *CTSL*, *DPP4,* and *DAAM1* were found as highly expressed in fetal cells compared to *ACE2* and *TMPRSS2* in all tissues analyzed (**Supplementary Figure S2**).

DISCUSSION

Recent literature has shown placenta cells infected with SARS-CoV-2 in pregnant women diagnosed with moderate to severe COVID-19 (Patanè et al., 2020; Penfield et al., 2020), with findings supporting the vertical transmission of SARS-CoV-2 in early and late pregnancy (Hosier et al., 2020; Patanè et al., 2020; Schwartz and Morotti, 2020; Shende et al., 2021; Valdespino-Vázquez et al., 2021). However, few placenta cells express *ACE2*

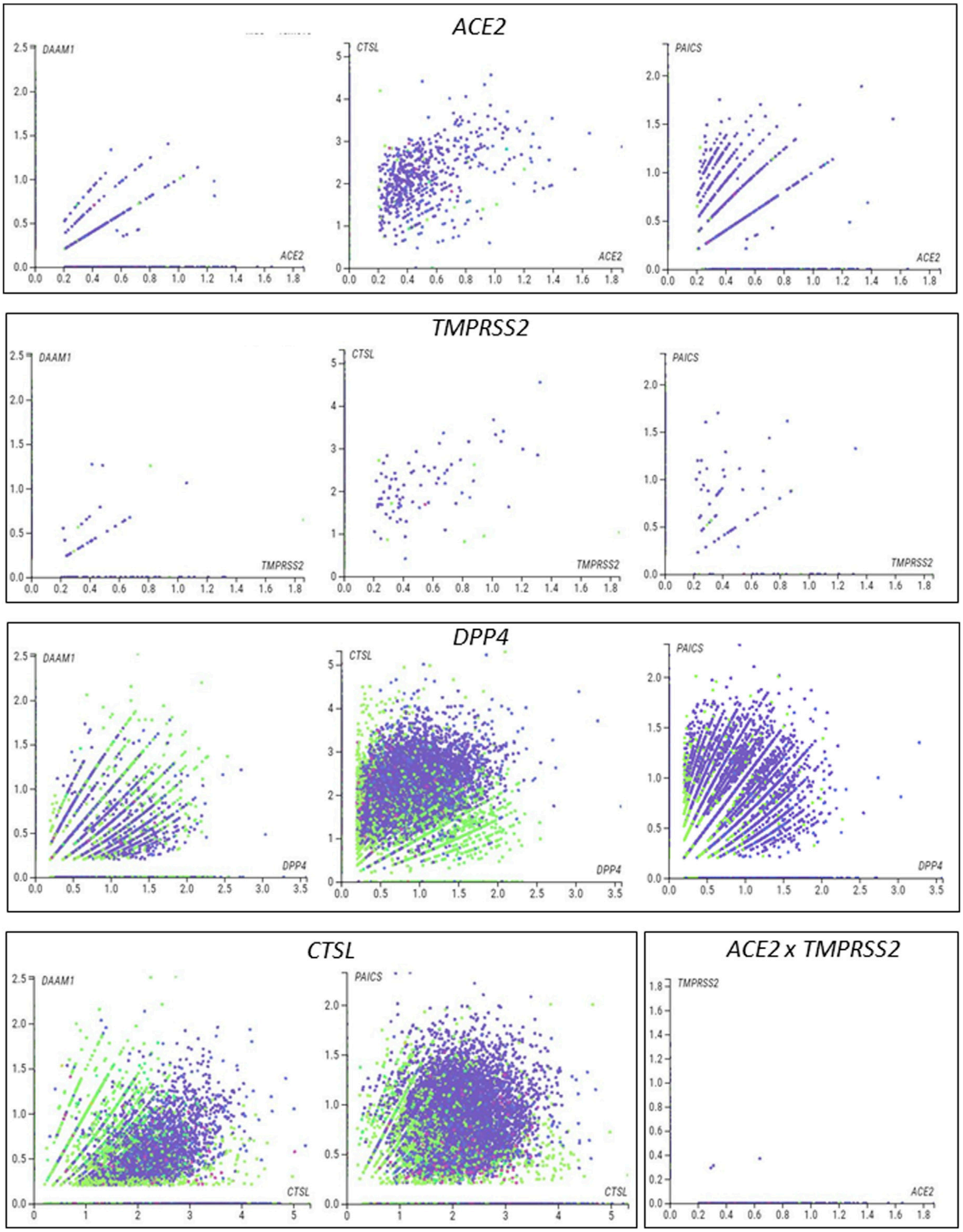


FIGURE 2 | Single-cell analysis of the human placenta indicates potential non-canonical routes of SARS-CoV-2. **(A)**. Co-expression analysis of *DAAM1*, *CTSL*, DPP4, and *PAICS* with the canonical virus entry-associated genes *ACE2* and *TMPRSS2*. **(B)**. Co-expression analysis of *DAAM1* and *PAICS* with the non-canonical virus entry genes *DPP4* and *CTLS*. The images were generated using COVID-19 Cell Atlas (https://www.covid19cellatlas.org) functionalities.

and *TMPRSS2* (Pique-Regi et al., 2020; Sungnak et al., 2020) required for virus entry, raising the question of whether potential non-canonical molecular mechanisms may be involved in the biological cycle of the virus in these cells. Here, we demonstrated by transcriptomic analyses (microarray and scRNA-Seq) and *in silico* predictions of virus-host protein-protein interactions that, despite low-levels of *ACE2* and *TMPRSS2*, villous trophoblast cells express high levels of the potential non-canonical cell-entry mediators *DPP4* and *CTSL*. We also found changes in the expression of *DAAM1* and *PAICS* genes that code for proteins predicted to interact with SARS-CoV proteins during pregnancy. These results provide evidence of potential host-virus PPI that may lead to SARS-CoV-2 infection and replication in the human placenta.

Firstly, we characterized the expression of SARS-CoV-2/ SARS-CoV entry-associated receptors and proteases genes, which revealed that *DPP4* and *CTSL* are highly expressed in villous trophoblast cells during pregnancy. Similar to our findings, *DPP4* and *CTSL* expression was already identified in first and second trimester human placenta single cells (Ashary et al., 2020). DPP4 has a high affinity to the SARS-CoV-2 spike receptor-binding domain, (Li et al., 2020), and critical DPP4 residues share the SARS-CoV-2-S/DPP4 binding as found for MERS-CoV-S/DPP4 (Hoffmann et al., 2020). Although it was demonstrated that DPP4 is not a SARS-CoV-2 entry receptors in BHK-21 cells (Hoffmann et al., 2020), more studies are necessary to address the role of *DPP4* in placental cells. In a pioneer study based on SARS-CoV cell entry mechanisms, (Simmons et al., 2005) described a three-step method for host-virus membrane fusion, involving receptor-binding and induced conformational changes in S glycoprotein with subsequent CTSL proteolysis and activation of membrane fusion within endosomes. During the COVID-19 pandemic, it has been further demonstrated that SARS-CoV-2 can use CTSB and CTSL as well as TMPRSS2 for priming host cells (Hoffmann et al., 2020; Ou et al., 2020). These data provide evidence that CTSL is a potentially promising treatment for COVID-19 by blocking coronavirus host cell entry and intracellular replication (Liu et al., 2020). Secondly, we analyzed whether placental development and growth are associated with transcriptional changes in proteins that potentially interact with SARS-CoV. Among the transcripts translated into proteins predicted as interacting with SARS-CoV, *DAAM1* increased with placental development and growth. DAAM1 was previously identified as a regulator of bacteria phagocytosis by regulating filopodia formation and phagocytic uptake in primary human macrophages (Hoffmann et al., 2014). We predicted that DAAM1 potentially interacts with the viral protein encoded from open reading frame 8 (ORF8) and the hypothetical protein sars7a of SARS-CoV. The SARS-CoV ORF8 shares the lowest homology with all SARS-CoV-2 proteins; however, it was previously demonstrated that SARS-CoV-2 ORF8 expression was able to selectively target MHC-I for lysosomal degradation by an autophagy-dependent mechanism in different cell types (Zhang et al., 2020). We found that the *PAICS* transcript, which is also translated into a protein predicted as interacting with the Nsp3 of SARS-CoV, decreased its levels with placental development and growth. PAICS is an enzyme of the *de novo* purine biosynthesis pathway, which was previously predicted as having a putative interaction with the human influenza A virus (IAV) nucleoprotein and was up-regulated during IAV infection (Generous et al., 2014). The putative interaction of Nsp3 with PAICS is relevant. Nsp3 is one of the 16 non-structural proteins in the replicase ORF1ab gene of SARS-CoV and SARS-CoV-2 (Enjuanes, 2005; Wu et al., 2020), and binds to virus RNA and proteins, including the nucleocapsid protein (Lei et al., 2018). Considering the low levels of *ACE2* and *TMPRSS2* that we found in villous trophoblast cells, our results provide evidence of alternative cell-entry mediators in the placenta. Moreover, the description of the potential interaction between host and SARS-CoV-2 may provide insights into the mechanisms of placental infection in women with COVID-19.

Cellular and molecular composition play a role in placenta infection and intrauterine transmission of viruses (Koi et al., 2001; Barbeito-Andrés et al., 2020). Thus, we finally used single-cell analysis of the syncytiotrophoblast, villous cytotrophoblast, and extravillous trophoblasts—the fetal component of the placenta—that confirmed the low levels of expression of the canonical entry receptor *ACE2* and *TMPRSS2* genes (Pique-Regi et al., 2020; Sungnak et al., 2020). Conversely, we found that these cells express high-levels of *DPP4* and *CTSL*, two potential mediators of SARS-Cov-2 host cell entry (Hoffmann et al., 2020; Li et al., 2020) that may contribute to the SARS-CoV-2 infection (Hosier et al., 2020; Patanè et al., 2020; Penfield et al., 2020). It is essential to highlight that our scRNA-Seq analysis showed that the non-canonical *DPP4* and *CTSL* entry genes (Hoffmann et al., 2020; Li et al., 2020) are highly co-expressed in syncytiotrophoblast, villous cytotrophoblast, and extravillous trophoblasts cells. *DPP4* and *CTSL* expression landscape in different tissues were extensively investigated in scRNA-seq study (Singh et al., 2020). However, there was no evidence of *DAAM1* and *PAICS* role. We found that the transcripts *DAAM1* and *PAICS*, which are translated into proteins predicted as potentially interacting with SARS-CoV-2, were also highly co-expressed with *DPP4* and *CTSL*. These results demonstrate that, although *ACE2* and *TMPRSS2* are poorly expressed in the placenta, other mediators that potentially interact with the virus are highly co-expressed in villous trophoblast cells and, therefore, may represent a valuable alternative route for infection and viral replication.

Although our *in-silico* analyses are a starting point, they have limitations. We reanalyzed published placenta datasets with a limited number of samples (microarray = 12 individuals, and scRNA-seq = 5 individuals) and, consequently, validations in a large cohort of samples must be performed. We understand that under sampling in transcriptomic data are a limitation and statistical analysis as imbalanced learn and ablation tests would benefit our methodology, however, our main results are supported by the relevant literature. Moreover, the predicted interactions presented here should also be experimentally validated in infected cells or placenta to circumvent these limitations. Moreover, the analyses of single-cell data should be conducted for the entire period of gestation, considering that we only evaluated single-cell data from the first trimester of pregnancy. The detection of SARS-CoV-2 in the placenta needs to be performed in a large cohort of pregnant women with COVID-19, including asymptomatics. Considering the vertical transmission of SARS-CoV-2, the follow-up of newborns from mothers with COVID-19 during pregnancy should be necessary since, if it occurs even in the non-asymptomatic, the long-term consequences are mostly unknown.

In conclusion, despite low levels of *ACE2* and *TMPRSS2*, our analyses demonstrate that villous trophoblast cells express high levels of potential non-canonical cell-entry mediators *DDP4* and *CTSL.* We also found changes in the expression of the *DAAM1* and *PAICS* genes coding for proteins predicted to interact with SARS-CoV proteins during pregnancy. These results provide new insight into the interaction between SARS-CoV-2 and the host proteins and indicate that coronaviruses may use multiple mediators for virus infection and replication.

MATERIALS AND METHODS

Differential Expression Analysis of the Placenta During Pregnancy Trimesters

We reanalyzed microarray data from villous trophoblast tissues of first (45–59 days, $n = 4$), second trimester (109–115 days, $n = 4$), and third trimesters ($n = 4$) of uncomplicated pregnancies, available in Gene Expression Omnibus (GEO), under the accession number GSE9984 (Mikheev et al., 2008). Next, we identified the DEGs during the pregnancy, by comparing the second trimester and third trimester with the first trimester using the GEO2R tool (https://www.ncbi.nlm.nih.gov/geo/geo2r). We applied the default settings of GEO2R to perform transcriptome analysis on original submitter-supplied processed data tables (RMA normalized signal log2) using the GEOquery and limma R packages from the Bioconductor project. Genes with Log2 of Fold Change ≥ |0.5| and False Discovery Rate (FDR) < 0.05 were considered as DEGs. We further grouped the gene expression profiles during pregnancy by using the K-means clustering analysis based on One Minus Pearson Correlation (Robust Multi-array Average, RMA, normalization log2, K-means = 3).

Enrichment Analysis

We used the Metascape tool (https://metascape.org) (Zhou et al., 2019) to perform functional enrichment analysis of the genes lists generated in the clustering and differential expression analyses.

Placenta Proteins That Potentially Interact With Coronaviruses and Zika Virus

SARS-CoV-2 cell entry mediators were selected for first screening using the human proteins already described in the COVID-19 Cell Atlas (https://www.covid19cellatlas.org/) and the literature (Hoffmann et al., 2020; Li et al., 2020; Sungnak et al., 2020; Zhou et al., 2020). Next, we used gene expression data to generated a list of PPIs that potentially interact with human coronaviruses and the ZIKV from the Pathogen–Host Interactome Prediction using Structure Similarity (P-HIPSTer, http://phipster.org/) database (Hosier et al., 2020) (**Supplementary Tables S4,S5**, respectively). SARS-CoV was included in the analysis considering the evolutionary relationship between the novel SARS-CoV-2 (Zhou et al., 2020), and ZIKV considering its impact on placental infection and fetal microcephaly (Pierson and Diamond, 2018; Barbeito-Andrés et al., 2020).

Tissue-specific Gene Networks of Potential Virus-Host Placenta Interactome

The humanbase webtool (https://hb.flatironinstitute.org) (Greene et al., 2015) was used to generate the tissue-specific gene network of placental genes coding for proteins that potentially interact with SARS-CoV-2 based on P-HIPSTER or that we identified as either associated with COVID-19 or that we hypothesized may be associated with the disease, based on the literature (**Supplementary Table S8**). We considered co-expression, interaction, transcriptional factor binding, Gene Set Enrichment Analysis (GSEA) perturbations as active interaction sources, applying minimum interaction confidence of 0.07, and a maximum number of 15 genes.

Single-Cell Transcriptome Analysis of Human Placentas and Fetal Tissues

We used the functionalities of the COVID-19 Cell Atlas (https://www.covid19cellatlas.org) to explore single-cell RNA sequencing (scRNA-Seq) data from five samples from first-trimester placentas, previously described by Vento-Tormo et al. (2018). The processed scRNA-seq data (log-transformed normalized counts) of placental genes potentially interacting with SARS-CoVs was visualized in placental cell types using the "interactive viewer" developed by cellxgene v0.15.0 (https://cellxgene.cziscience.com) by the Chan Zuckerberg Initiative available at https://cellxgene.cziscience.com/d/fetal_maternal_interface_10x-22.cxg/. We also characterized the gene expression profile of canonical and non-canonical placental mediators of viruses interaction per cell using cellxgene v0.15.0 bar plots. The placental single-cell transcriptome is available for re-analysis at Human Cell Atlas Data Portal (https://data.humancellatlas.org) or E-MTAB-6701 (EMBL-EBI, ArrayExpress; https://www.ebi.ac.uk/arrayexpress). We applied this same strategy to investigate the gene expression of our candidates using scRNA-seq in human fetal tissues: lung, liver, and thymus (ArrayExpress; E-MTAB-8221, E-MTAB-7407, and E-MTAB-8581, respectively. Last accessed June 2020.

Placenta RNA Sequencing From Pregnant Women With COVID-19

We downloaded the list of differentially expressed genes from 3rd trimester placental villi of pregnant women with COVID-19 ($n = 5$) compared with uninfected control individuals matched for maternal age, gestational age, maternal comorbidities, and mode of delivery (n = 3), publicly available at GSE17995 (Lu-Culligan et al., 2021). The authors conducted the differential expression analysis using DESeq2 v1.24.0 with default parameters, and we downloaded the CSV file from GEO.

Data Representation and Analysis

Morpheus (https://software.broadinstitute.org/morpheus/) was used to generate the heatmaps and to perform the clustering analyses. The Venn diagram was constructed using the Van de Peer Lab software (http://bioinformatics.psb.ugent.be/webtools/Venn/). Biorender was used to design cell images (https://app.biorender.com/).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession numbers can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

This study was a retrospective analysis based on the reanalysis of existing publicly available datasets available from the GEO (https://www.ncbi.nlm.nih.gov/geo/; GSE9984 and GSE17995), Human Cell Atlas Data Portal (https://data.humancellatlas.org/; Vento-Tormo et al. (2018)), and EMBL-EBI, Express (https://www.ebi.ac.uk/arrayexpress; E-MTAB-6701, E-MTAB-8221, E-MTAB-7407, and E-MTAB-8581). Therefore, ethics approval and patient consent are not required.

AUTHOR CONTRIBUTIONS

Conceptualization, FC, RC, LJ; methodology, FC, RC, SC, LJ; formal analysis, FC, CN, RC, SC, LJ; investigation, FC, CN, RC, SC, LJ; resources, RC, LJ; data curation, FC, CN, RC, SC, LJ; writing—original draft preparation, FC, RC, SC, LJ.; writing-review and editing, all authors; supervision, RC, LJ; project administration, RC, LJ; funding acquisition, RC, LJ.

FUNDING

This research was supported by the National Council for Scientific and Technological Development, CNPq (Process 311530/2019-2 to RC, and Process 310663/2018-0 to LJ, and scholarship #870415/1997-2 to SC), and by Fundação de Amparo à Pesquisa do Estado de São Paulo (Process 2012/13961-6 to RC, Process 2017/01063-7 to LJ, and scholarship 2017/08716-6 to FBC).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at bioRxiv (Constantino et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.614728/full#supplementary-material

REFERENCES

Ashary, N., Bhide, A., Chakraborty, P., Colaco, S., Mishra, A., Chhabria, K., et al. (2020). Single-Cell RNA-Seq Identifies Cell Subsets in Human Placenta that Highly Expresses Factors Driving Pathogenesis of SARS-CoV-2. *Front. Cel Dev. Biol.* 8, 783. doi:10.3389/fcell.2020.00783

Barbeito-Andrés, J., Pezzuto, P., Higa, L. M., Dias, A. A., Vasconcelos, J. M., Santos, T. M. P., et al. (2020). Congenital Zika Syndrome Is Associated with Maternal Protein Malnutrition. *Sci. Adv.* 6, eaaw6284. doi:10.1126/sciadv.aaw6284

Celik, O., Saglam, A., Baysal, B., Derwig, I. E., Celik, N., Ak, M., et al. (2020). Factors Preventing Materno-Fetal Transmission of SARS-CoV-2. *Placenta* 97, 1–5. doi:10.1016/j.placenta.2020.05.012

Chen, H., Guo, J., Wang, C., Luo, F., Yu, X., Zhang, W., et al. (2020). Clinical Characteristics and Intrauterine Vertical Transmission Potential of COVID-19 Infection in Nine Pregnant Women: A Retrospective Review of Medical Records. *The Lancet* 395, 809–815. doi:10.1016/S0140-6736(20)30360-3

Constantino, F. B., Cury, S. S., Nogueira, C. R., Carvalho, R. F., and Justulin, L. A. (2020). Prediction of Non-Canonical Routes for SARS-CoV-2 Infection in Human Placenta Cells. *bioRxiv*. doi:10.1101/2020.06.12.148411

Dong, E., Du, H., and Gardner, L. (2020a). An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. *Lancet Infect. Dis.* 20, 533–534. doi:10.1016/S1473-3099(20)30120-1

Dong, L., Tian, J., He, S., Zhu, C., Wang, J., Liu, C., et al. (2020b). Possible Vertical Transmission of SARS-CoV-2 from an Infected Mother to Her Newborn. *JAMA* 323 (18), 1846–1848. doi:10.1001/jama.2020.4621

Generous, A., Thorson, M., Barcus, J., Jacher, J., Busch, M., and Sleister, H. (2014). Identification of Putative Interactions between Swine and Human Influenza A Virus Nucleoprotein and Human Host Proteins. *Virol. J.* 11, 228. doi:10.1186/s12985-014-0228-6

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., et al. (2015). Understanding Multicellular Function and Disease with Human Tissue-Specific Networks. *Nat. Genet.* 47, 569–576. doi:10.1038/ng.3259

Hashimoto, Y., Sheng, X., Murray-Nerger, L. A., and Cristea, I. M. (2020). Temporal Dynamics of Protein Complex Formation and Dissociation during Human Cytomegalovirus Infection. *Nat. Commun.* 11, 806. doi:10.1038/s41467-020-14586-5

Hoffmann, A. K., Naj, X., and Linder, S. (2014). Daam1 Is a Regulator of Filopodia Formation and Phagocytic Uptake of Borrelia Burgdorferi by Primary Human Macrophages. *FASEB j.* 28, 3075–3089. doi:10.1096/fj.13-247049

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181, 271–280. e8. doi:10.1016/j.cell.2020.02.052

Hosier, H., Farhadian, S., Morotti, R. A., Deshmukh, U., Lu-Culligan, A., Campbell, K. H., et al. (2020). SARS-CoV-2 Infection of the Placenta. medRxiv. doi:10.1101/2020.04.30.20083907

Kimberlin, D. W., and Stagno, S. (2020). Can SARS-CoV-2 Infection Be Acquired In Utero?: More Definitive Evidence Is Needed. *JAMA* 323 (18), 1788–1789. doi:10.1001/jama.2020.4868

Koi, H., Zhang, J., and Parry, S. (2001). The Mechanisms of Placental Viral Infection. *Ann. N.Y Acad. Sci.* 943, 148–156. doi:10.1111/j.1749-6632.2001.tb03798.x

Lasso, G., Mayer, S. V., Winkelmann, E. R., Chu, T., Elliot, O., Patino-Galindo, J. A., et al. (2019). A Structure-Informed Atlas of Human-Virus Interactions. *Cell* 178, 1526–1541. e16. doi:10.1016/j.cell.2019.08.005

Lei, J., Kusov, Y., and Hilgenfeld, R. (2018). Nsp3 of Coronaviruses: Structures and Functions of a Large Multi-Domain Protein. *Antiviral Res.* 149, 58–74. doi:10.1016/j.antiviral.2017.11.001

L. Enjuanes (Editor) (2005). *Coronavirus Replication and Reverse Genetics* (Berlin, Heidelberg: Springer Berlin Heidelberg). doi:10.1007/b138038

Letko, M., Marzi, A., and Munster, V. (2020). Functional Assessment of Cell Entry and Receptor Usage for SARS-CoV-2 and Other Lineage B Betacoronaviruses. *Nat. Microbiol.* 5, 562–569. doi:10.1038/s41564-020-0688-y

Li, Y., Zhang, Z., Yang, L., Lian, X., Xie, Y., Li, S., et al. (2020). The MERS-CoV Receptor DPP4 as a Candidate Binding Target of the SARS-CoV-2 Spike. *iScience* 23, 101160. doi:10.1016/j.isci.2020.101160

Liu, T., Luo, S., Libby, P., and Shi, G.-P. (2020). Cathepsin L-Selective Inhibitors: A Potentially Promising Treatment for COVID-19 Patients. *Pharmacol. Ther.* 213, 107587. doi:10.1016/j.pharmthera.2020.107587

Lu-Culligan, A., Chavan, A. R., Vijayakumar, P., Irshaid, L., Courchaine, E. M., Milano, K. M., et al. (2021). Maternal Respiratory SARS-CoV-2 Infection in Pregnancy Is Associated with a Robust Inflammatory Response at the Maternal-Fetal Interface. *Med* 2, 591–610. e10. doi:10.1016/j.medj.2021.04.016

Mikheev, A. M., Nabekura, T., Kaddoumi, A., Bammler, T. K., Govindarajan, R., Hebert, M. F., et al. (2008). Profiling Gene Expression in Human Placentae of Different Gestational Ages: An OPRU* Network and UW SCOR Study. *Reprod. Sci.* 15, 866–877. doi:10.1177/1933719108322425

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., et al. (2020). Characterization of Spike Glycoprotein of SARS-CoV-2 on Virus Entry and its Immune Cross-Reactivity with SARS-CoV. *Nat. Commun.* 11, 1620. doi:10.1038/s41467-020-15562-9

Patanè, L., Morotti, D., Giunta, M. R., Sigismondi, C., Piccoli, M. G., Frigerio, L., et al. (2020). Vertical Transmission of Coronavirus Disease 2019: Severe Acute Respiratory Syndrome Coronavirus 2 RNA on the Fetal Side of the Placenta in Pregnancies with Coronavirus Disease 2019-positive Mothers and Neonates at Birth. *Am. J. Obstet. Gynecol. MFM* 2, 100145. doi:10.1016/j.ajogmf.2020.100145

Penfield, C. A., Brubaker, S. G., Limaye, M. A., Lighter, J., Ratner, A. J., Thomas, K. M., et al. (2020). Detection of Severe Acute Respiratory Syndrome Coronavirus 2 in Placental and Fetal Membrane Samples. *Am. J. Obstet. Gynecol. MFM* 2, 100133. doi:10.1016/j.ajogmf.2020.100133

Pereira, L. (2018). Congenital Viral Infection: Traversing the Uterine-Placental Interface. *Annu. Rev. Virol.* 5, 273–299. doi:10.1146/annurev-virology-092917-043236

Pierson, T. C., and Diamond, M. S. (2018). The Emergence of Zika Virus and its New Clinical Syndromes. *Nature* 560, 573–581. doi:10.1038/s41586-018-0446-y

Pique-Regi, R., Romero, R., Tarca, A. L., Luca, F., Xu, Y., Alazizi, A., et al. (2020). Does the Human Placenta Express the Canonical Cell Entry Mediators for SARS-CoV-2? *eLife* 9, e58716. doi:10.7554/eLife.58716

Qadri, F., and Mariona, F. (2020). Pregnancy Affected by SARS-CoV-2 Infection: A Flash Report from Michigan. *J. Maternal-Fetal Neonatal Med.* ahead of print. doi:10.1080/14767058.2020.1765334

Schoenmakers, S., Snijder, P., Verdijk, R. M., Kuiken, T., Kamphuis, S. S. M., Koopman, L. P., et al. (2020). SARS-CoV-2 Placental Infection and Inflammation Leading to Fetal Distress and Neonatal Multi-Organ Failure in an Asymptomatic Woman. *medRxiv*. doi:10.1101/2020.06.08.20110437

Schwartz, D. A., and Graham, A. L. (2020). Potential Maternal and Infant Outcombes from Coronavirus 2019-nCoV (SARS-CoV-2) Infecting Pregnant Women: Lessons from SARS, MERS, and Other Human Coronavirus Infections. *Viruses* 12, 194. doi:10.3390/v12020194

Schwartz, D. A., and Morotti, D. (2020). Placental Pathology of COVID-19 with and without Fetal and Neonatal Infection: Trophoblast Necrosis and Chronic Histiocytic Intervillositis as Risk Factors for Transplacental Transmission of SARS-CoV-2. *Viruses* 12, 1308. doi:10.3390/v12111308

Shanes, E. D., Mithal, L. B., Otero, S., Azad, H. A., Miller, E. S., and Goldstein, J. A. (2020). Placental Pathology in COVID-19. *Am. J. Clin. Pathol.* 154, 23–32. doi:10.1093/ajcp/aqaa089

Shende, P., Gaikwad, P., Gandhewar, M., Ukey, P., Bhide, A., Patel, V., et al. (2021). Persistence of SARS-CoV-2 in the First Trimester Placenta Leading to Transplacental Transmission and Fetal Demise from an Asymptomatic Mother. *Hum. Reprod.* 36, 899–906. doi:10.1093/humrep/deaa367

Simmons, G., Gosalia, D. N., Rennekamp, A. J., Reeves, J. D., Diamond, S. L., and Bates, P. (2005). Inhibitors of Cathepsin L Prevent Severe Acute Respiratory Syndrome Coronavirus Entry. *Proc. Natl. Acad. Sci.* 102, 11876–11881. doi:10.1073/pnas.0505577102

Singh, M., Bansal, V., and Feschotte, C. (2020). A Single-Cell RNA Expression Map of Human Coronavirus Entry Factors. *Cel Rep.* 32, 108175. doi:10.1016/j.celrep.2020.108175

Sungnak, W., Huang, N., Huang, N., Bécavin, C., Berg, M., Queen, R., et al. (2020). SARS-CoV-2 Entry Factors Are Highly Expressed in Nasal Epithelial Cells Together with Innate Immune Genes. *Nat. Med.* 26, 681–687. doi:10.1038/s41591-020-0868-6

Valdespino-Vázquez, M. Y., Helguera-Repetto, C. A., León-Juárez, M., Villavicencio-Carrisoza, O., Flores-Pliego, A., Moreno-Verduzco, E. R., et al. (2021). Fetal and Placental Infection with SARS-CoV-2 in Early Pregnancy. *J. Med. Virol.* 93, 4480–4487. doi:10.1002/jmv.26965

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., et al. (2018). Single-Cell Reconstruction of the Early Maternal-Fetal Interface in Humans. *Nature* 563, 347–353. doi:10.1038/s41586-018-0698-6

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., et al. (2020). Factors Associated with COVID-19-related Death Using OpenSAFELY. *Nature* 584, 430–436. doi:10.1038/s41586-020-2521-4

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., et al. (2020). A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* 579, 265–269. doi:10.1038/s41586-020-2008-3

Yang, Z., and Liu, Y. (2020). Vertical Transmission of Severe Acute Respiratory Syndrome Coronavirus 2: A Systematic Review. *Am. J. Perinatol* 37, 1055–1060. doi:10.1055/s-0040-1712161

Zeng, H., Xu, C., Fan, J., Tang, Y., Deng, Q., Zhang, W., et al. (2020). Antibodies in Infants Born to Mothers with COVID-19 Pneumonia. *JAMA* 323 (18), 1848–1849. doi:10.1001/jama.2020.4861

Zhang, Y., Zhang, J., Chen, Y., Luo, B., Yuan, Y., Huang, F., et al. (2020). The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through Potently Downregulating MHC-I. *bioRxiv*. doi:10.1101/2020.05.24.111823

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature* 579, 270–273. doi:10.1038/s41586-020-2012-7

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nat. Commun.* 10, 1523. doi:10.1038/s41467-019-09234-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership