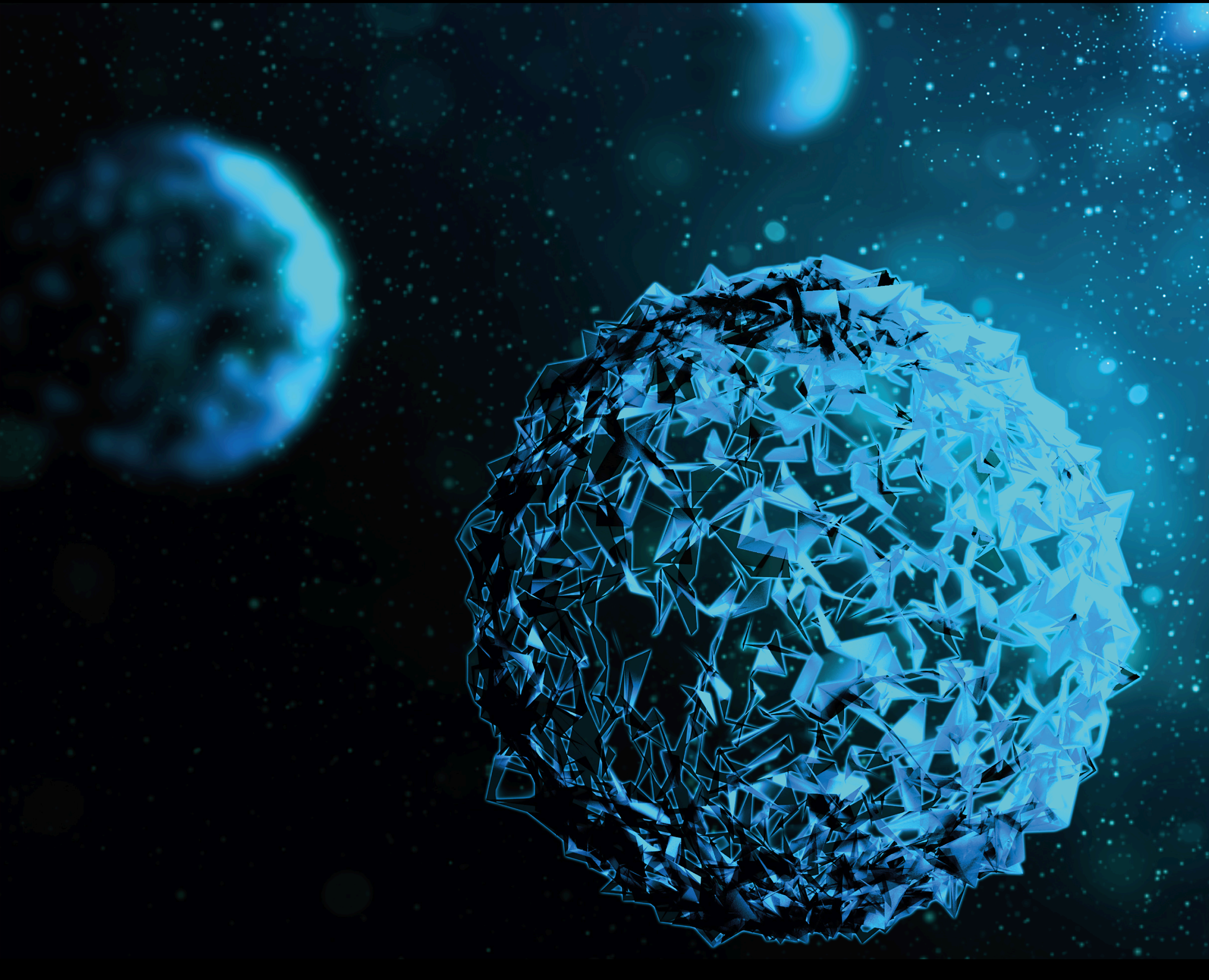
BioMed Research International

Current High-throughput Approaches in Bioinformatics for Big Data

Lead Guest Editor: Jian Zhang
Guest Editors: Jun Shen and Jiangning Song



Current High-throughput Approaches in Bioinformatics for Big Data

BioMed Research International

Current High-throughput Approaches in Bioinformatics for Big Data

Lead Guest Editor: Jian Zhang
Guest Editors: Jun Shen and Jiangning Song

Section Editors

Academic Editors

Bioinformatics

Contents

Hindawi
BioMed Research International
Volume 2023, Article ID 5469258, 32 pages
https://doi.org/10.1155/2023/5469258

Research Article

In Silico Identification and Analysis of Potentially Bioactive Antiviral Phytochemicals against SARS-CoV-2: A Molecular Docking and Dynamics Simulation Approach

Sajal Kumar Halder,[1] **Ive Sultana**,[2] **Md Nazmussakib Shuvo**,[3] **Aparna Shil**,[3] **Mahbubul Kabir Himel**,[3] **Md. Ashraful Hasan**,[1] **and Mohammad Mahfuz Ali Khan Shawan**[1]

[1]*Department of Biochemistry and Molecular Biology, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh*
[2]*Department of Microbiology, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh*
[3]*Department of Botany, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh*

Correspondence should be addressed to Mohammad Mahfuz Ali Khan Shawan; mahfuz_026shawan@juniv.edu

Received 1 January 2023; Revised 7 February 2023; Accepted 5 April 2023; Published 11 May 2023

Academic Editor: Jian Zhang

SARS-CoV-2, a deadly coronavirus sparked COVID-19 pandemic around the globe. With an increased mutation rate, this infectious agent is highly transmissible inducing an escalated rate of infections and death everywhere. Hence, the discovery of a viable antiviral therapy option is urgent. Computational approaches have offered a revolutionary framework to identify novel antimicrobial treatment regimens and allow a quicker, cost-effective, and productive conversion into the health center by evaluating preliminary and safety investigations. The primary purpose of this research was to find plausible plant-derived antiviral small molecules to halt the viral entrance into individuals by clogging the adherence of Spike protein with human ACE2 receptor and to suppress their genome replication by obstructing the activity of Nsp3 (Nonstructural protein 3) and 3CLpro (main protease). An in-house library of 1163 phytochemicals were selected from the NPASS and PubChem databases for downstream analysis. Preliminary analysis with SwissADME and pkCSM revealed 149 finest small molecules from the large dataset. Virtual screening using the molecular docking scoring and the MM-GBSA data analysis revealed that three candidate ligands CHEMBL503 (Lovastatin), CHEMBL490355 (Sulfuretin), and CHEMBL4216332 (Grayanoside A) successfully formed docked complex within the active site of human ACE2 receptor, Nsp3, and 3CLpro, respectively. Dual method molecular dynamics (MD) simulation and post-MD MM-GBSA further confirmed efficient binding and stable interaction between the ligands and target proteins. Furthermore, biological activity spectra and molecular target analysis revealed that all three preselected phytochemicals were biologically active and safe for human use. Throughout the adopted methodology, all three therapeutic candidates significantly outperformed the control drugs (Molnupiravir and Paxlovid). Finally, our research implies that these SARS-CoV-2 protein antagonists might be viable therapeutic options. At the same time, enough wet lab evaluations would be needed to ensure the therapeutic potency of the recommended drug candidates for SARS-CoV-2.

1. Introduction

The World Health Organization (WHO) reported that the prevalence of SARS-CoV-2 is spreading at an alarming rate, posing severe health problems. The most recent outburst of second wave of SARS-CoV-2 has turned into a worldwide catastrophe. Following the coronavirus (CoV) epidemic in China in December 2019, WHO classified SARS-CoV-2, as the newest candidate of the *Coronaviridae* family within Nidovirales order [1]. As of 3rd May 2023, WHO has received a report from around 765,222,932 diagnosed COVID-19 infections worldwide, with 6,921,614 fatalities [2, 3]. According to available information, the virus can be transmitted often by close, indirect, or direct exposure to infectious persons, as well as contaminated secretions such as nasal droplets and saliva, and respiratory secretions released when an infected individual sneezes, coughs, or speaks [3]. It has been linked to a wide range of signs and symptoms, consisting of minor to severe illness, which varies from patient to patient. Complications might appear anywhere from two to fourteen days after the virus has been infected. Fever, fatigue, chronic cough, sore throat, difficulty breathing, impairment of taste/odor, nausea, sputum production, headache, expectoration, diarrhea, anorexia, and some other symptoms might occur at separate phases of the disease [1, 4].

SARS-CoV-2 is a membrane-encased positive-sense single-stranded RNA ((+) ssRNA) virus having a diameter ranging from 60 to 140 nanometers [4, 5]. The envelope is surrounded by spike-shaped glycoprotein protrusions that resemble crowns under the electron microscope [6]. The spike (S), nucleocapsid (N), envelope (E), and membrane (M) proteins are among the four crucial targets encoded by the SARS-CoV-2 genome. Main protease (3CLpro), RNA-dependent RNA polymerase (RdRp), and papain-like protease (PLpro) are some of the nonstructural proteins synthesized by the viral DNA [7]. Nonstructural protein 3 (Nsp3) proteins containing macrodomains are pervasive and evolutionarily conserved and responsible for the transcription process [8]. Previous study has established that human angiotensin converting enzyme 2 (ACE2) receptor has a greater affinity for the RBD region of the spike protein [9]. The attachment of favorable ligands to the active pockets of human ACE2 receptor might alter the protein's structure. As a result, the viral ACE2 entrance region might be a feasible object for therapeutic advancement. Since the main protease of SARS-CoV-2 is vital for its growth and the consequent expression of the replicase polyproteins, it has turned into an obvious target for anti-COVID-19 therapeutic design [10]. As a result, focusing on these proteins might help with long-term COVID-19 infection management and eradication.

The viral disease is spreading at a surprising pace worldwide, and researchers are racing to develop effective drugs to use as therapeutic agents. The most promising choices appear to be natural compounds with substantial bioavailability and minimal cytotoxicity [1]. Clinically approved antiviral drugs are effective; however, some people become resistant to drugs. In contrast, it has been claimed that phytochemicals have more acceptable side effects and can be a satisfactory substitute for synthetic antiviral compounds for the suppression of viral life-cycle and penetration [10].

Humans have always relied on natural compounds, especially phytochemicals, to treat health problems since the dawn of time. Recently, Shawan et al. presented luteolin and abyssinone II as possible phytochemicals against SARS-CoV-2 [1]. Besides, Manojkumar et al. reported ervoside had anticoronavirus properties [11]. Similarly, Emran et al. identified phytochemicals medicagol, faradiol, and flavanthrin as the potential barrier of SARS-CoV-2 [12]. Computer-assisted drug development (CAD) entails the usage of computerized techniques to discover, design, and evaluate therapeutics and associated pharmacologically active substances [13]. CAD techniques have improved compound screening significantly over time, aimed at targeting structure prediction and model development, active site determination, comprehending the protein-ligand complex, testing a huge dataset of substances by estimating their pharmacokinetics characteristics, and analyzing the dynamics of proteins binding with ligands within biological settings [14]. Existing medicines like Molnupiravir and Paxlovid have been authorized by the FDA for utilization in emergencies; the treatment may be used either alone or combined with others [15]. For COVID-19 patients, the antiviral drug Molnupiravir has been recommended as a therapeutic for SARS-CoV-2 because it increases the likelihood of viral RNA alterations while also inhibiting viral replication [16]. Through inhibition of proteasome breakdown of viral proteins, Paxlovid inhibits protein production (RNA-dependent RNA polymerase, helicase, exoribonuclease, RNA-binding protein endoribonuclease). Consequently, the viral transcription and replication are halted [7].

The main focus of this in silico work was to utilize computational tools, i.e., molecular docking and MD simulation to examine the effective binding interactivity and affinities of repurposed antiviral phytochemicals with the human ACE2 receptor, Nsp3 macrodomain, and the main protease of the SARS-CoV-2 virus and identify the finest ligand hit [17]. Among all other crucial characteristics, absorption, distribution, metabolism, toxicity, and excretion (ADMET) were evaluated, and the best of them were selected. Finally, the most effective phytochemicals with higher binding energy to the target receptor and stronger stabilizing capacity were confirmed by employing molecular dynamics simulation.

2. Materials and Methods

Virtual screening of natural bioactive molecules has become the standard method in the present therapeutic development workflow [18]. In this study, a wide range of repurposed phytochemicals were used from the NPASS (http://bidd.group/NPASS/) and PubChem (https://pubchem.ncbi.nlm.nih.gov/) servers as prospective ligands for SARS-CoV-2. The recently approved COVID-19 antiviral drugs Molnupiravir and Paxlovid were used as control drugs [19]. The workflow of our work was provided in Figure 1.

2.1. Characterization of Drug-Likeness Properties. A drug-like molecule can be considered a drug candidate by assessing its drug-like properties. The canonical SMILE sequence of the 1163 small molecules was fetched from the PubChem drug web server. The free accessible SwissADME was employed to compute the major physicochemical descriptors, pharmacokinetic properties, drug-like parameters, and associated factors [20]. To analyze the results, this application employs five principles of Lipinski's rule [21], Ghose's rule [22], Veber's rule [23], Egan rule [24], Muegge's rule [24], the number of rotatable bonds, and TPSA.

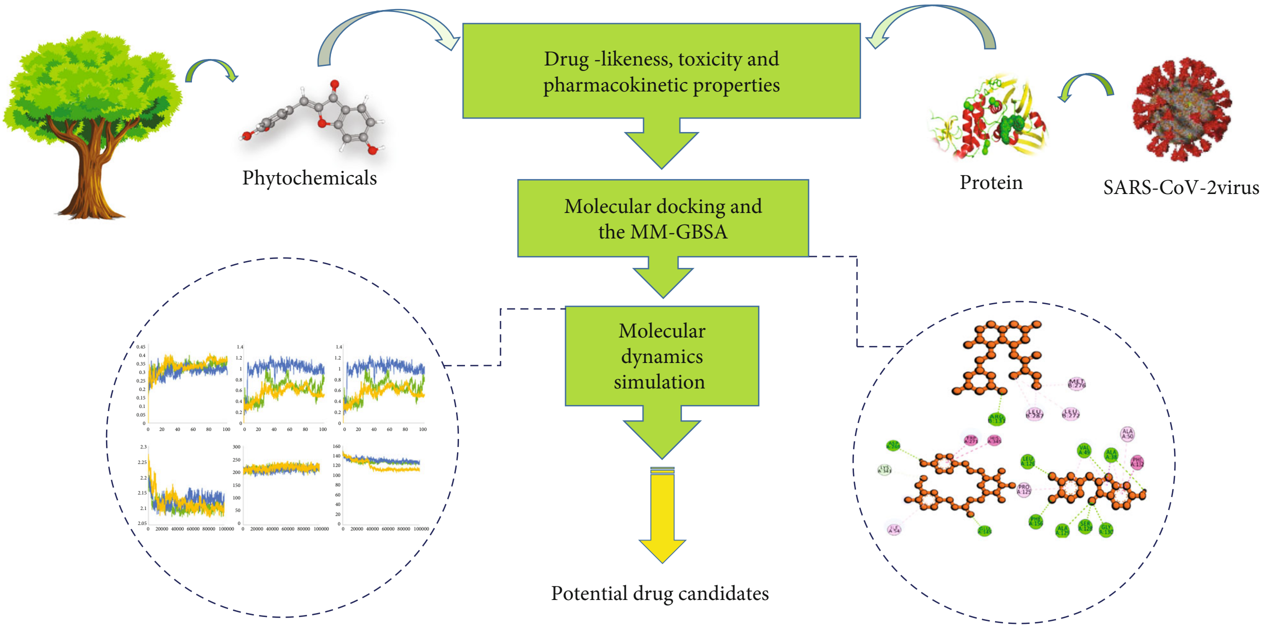

Drug -likeness, toxicity and pharmacokinetic properties
Phytochemicals
Protein
SARS-CoV-2virus
Molecular docking and the MM-GBSA
Molecular dynamics simulation
Potential drug candidates

Figure 1: Complete work flow of the structure-based virtual screening study.

2.2. Characterization of ADMET Properties. pkCSM is an online tool that employs graph-based structural signatures for determining and improving pharmacokinetic characteristics and toxicity in small molecules. To devise an ADMET prediction benchmark for in silico drug discovery, pkCSM applies a cut-off scanning strategy [25]. The chosen criteria for the prediction model were hepatotoxicity, Ames toxicity, oral rat acute toxicity, human intestinal absorption (HI), hERG I inhibitor, hERG II inhibitor, P-glycoprotein I inhibitor, P-glycoprotein II inhibitor, P-glycoprotein substrate, BBB permeability (log BB), Caco-2 permeability, CYP2D6 substrate, CYP3A4 substrate, CYP2C19 inhibitor, CYP1A2 inhibitor, CYP3A4 inhibitor, CYP2C9 inhibitor, and CYP2D6 inhibitor.

2.3. Molecular Docking by AutoDock vina

2.3.1. Ligand Preparation. At pH 7.4, polar hydrogen atoms were introduced to the downloaded 3D molecular ligands in SDF (spatial data file) format using the build module of the Avogadro 1.2.0. The same program was then used to conduct geometry optimization and energy reduction employing the MMFF94 force field and steepest descent option. These structures were retained in the PDB [26] extension for additional investigation. To add polar hydrogens and fix torsions of the ligands, AutoDockT tools-MGLTool 1.5.6 was used [27].

2.3.2. Protein Preparation. The preferred structures of SARS-CoV-2 main protease in complex with FSCU015 (PDB ID: 7NT3), Nsp3 macrodomain in complex with ADP-ribose (PDB ID: 7KQP), and inhibitor bound human ACE2-related carboxypeptidase (PDB ID: 1R4L) were taken from RCSB repository (https://www.rcsb.org/). Initially, the 3D structures were prepared in the PyMOL program [1]. Swiss-PdbViewer was subsequently used to minimize the energy of the selected proteins [28]. Next, the energy-minimized structures were loaded into AutoDock-MGLTools 1.5.6 to incorporate polar hydrogen and convert the PDB to PDBQT format.

2.3.3. Active Site Detection and Grid Box Preparation. Finding a ligand-binding region on a protein is the basic strategy for the molecular docking technique [29]. The possibility of protein-ligand attachment relies on numerous factors such as hydrogen bonds, hydrophobic or hydrophilic interactions, electrostatic and salt bridges. CASTp 3.0 website (http://sts.bioe.uic.edu/castp/) was employed to detect the active region of target proteins [30]. It applies an alpha shape detection approach to determine topographic properties and estimate protein area and volume for identifying ligand-binding cavities.

2.3.4. Binding Affinity Prediction by AutoDock vina. Virtual screening via docking studies is extensively used in computer-led pharmaceutical research to uncover promising drug-like substances. Initially, AutoDock vina was exploited to conduct rigid molecular docking among the proteins and selected compounds (ligands and control drugs), including a search area of 27,000 m^3 and exhaustiveness 10, and ligands being flexible while receptors remained rigid [18]. AutoDock vina calculates the binding energy and fixes the binding poses using the Lamarckian genetic algorithm. Here, in this study, 149 small molecules were docked with three target proteins (coordinates of geometry-optimized ligands of the best hits provided in Supplementary Table 5).

2.4. Glide Docking and MM-GBSA Analyses. Schrodinger was employed to perform glide docking and MM-GBSA analyses (Maestro 12.5, Schrodinger Suites 2020-3).

Previously screened ligands having higher affinity for target proteins than the reference drugs were explored in this step.

2.4.1. Preparation of Ligand Structures. The LigPrep module yields top hits of 3D configurations for small molecules, beginning from 1-dimensional/2-dimensional/3-dimensional structures in Maestro, Mol2, SMILES, or SD format [31]. By introducing hydrogens, ionizing at pH (7 ± 2.0), and subtracting salts, the LigPrep tool builds molecules and constructs 3D structures of them. Following that energy minimized and geometrically refined ligands were prepared by employing a built-in OPLS3e force field in Schrödinger Maestro 12.5 [32].

2.4.2. Preparation of Protein Structures. The protein structures (main protease, Nsp3, human ACE2 receptor) were loaded straight into the protein preparation wizard [32]. Protein structures were preprocessed by setting up bond orders, adding hydrogens and cap termini, and filling the missing atoms by prime module. At pH 7.0, the PROPKA application was used to calculate the protonation phases. Following that, the water portion around the protein was eliminated above 3.0 Å, and restrained minimization was executed utilizing the OPLS3e force field.

2.4.3. Preparation of Receptor Grid Box. The grid region directs small molecules to the binding center of the protein, making it an important part of molecular docking research. The grid model was created with the standard options of a Van der Waals radius scaling marker of 1.0 and a charge threshold score of 0.25 in the receptor grid generation package. The attached ligands UQZ, AR6, and XX5 within the protein structures main protease, Nsp3, and human ACE2 receptor, respectively, were used to define the region for the grid map.

2.4.4. Glide Docking and MM-GBSA Studies. Glide is a combined molecular docking technology that can facilitate both ligand and receptor flexibility [33]. Glide XP was developed to retrieve the finest docking poses having the greatest-scoring compounds. For drug molecules, a minimum scoring of 0.15 and a Van der Waals radius scaling marker of 0.80 was applied.

$$\begin{aligned}\text{Docking score} = a \times \text{VdW} + b \times \text{Coul} + \text{Hbond} + \text{Metal}\\ + \text{Lipo} + \text{BuryP} + \text{RotB} + \text{Site}.\end{aligned} \tag{1}$$

Here, a and b are coefficient constants for VdW and Coul, respectively, VdW is the Van der Waals energy, Coul is the Coulomb energy, H-bond is the hydrogen bonding with the receptor, Metal is the binding with metal, Lipo is the constant term for lipophilic, BuryP is the buried polar group penalty, RotB is the rotatable bond penalty, and Site is the active site polar interaction [1].

The binding free energy among the receptor and the collection of small molecules was measured using the prime MM-GBSA module. The binding energy of the ligand-protein constructs was estimated utilizing the OPLS3e force field, and the docked conformations were minimized utilizing Prime's native optimization tool.

$$\Delta\text{Gbind}\,(\text{Binding Free Energy}) = \Delta\text{EMM} + \Delta\text{Gsolv} + \Delta\text{GSA}. \tag{2}$$

Here, ΔEMM represents lowered energy deviations among the totality of the energies of the protein and ligand and protein-ligand complex. ΔGsolv displays the divergence in the GBSA solvation energy of the complex structure and the aggregate of the salvation energies for the ligand and protein. ΔGSA describes the deviation in the energies for the surface area of a complex and the total surface area of the ligand and protein complex [34].

2.5. Molecular Dynamics Simulation by GROMACS. The molecular dynamics program simulates the movements of a protein molecule utilizing the interaction potential to compute interatomic energies and equations of motion that regulate the machinery's dynamics in the drug design study. It illustrates the stability and flexibility data of ligand binding to a flexible target protein. GROMACS (https://simlab.uams.edu/) service was exploited to simulate the protein-ligand conformations, and the GROMOS96 43a1 force field was employed to produce the topological data of the complex constructs [35]. The PRODRG (http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg) Server was employed to render small molecule topology and coordinate information [36]. The aqueous phase of macromolecules was produced sequentially using the SPC water model (simple point-charge) and subsequently neutralized using 0.15 M NaCl solution [37]. A triclinic box was used to contain the bimolecular environment, and 5000 iterations of steepest descent strategies were used to minimize energy. The equilibrium of ion molecules around the macromolecule was accomplished at 310 K and 1.0 bar utilizing NPT (constant pressure) and NVT (constant volume) setups. After 100 nanoseconds of simulation, it provided trajectories of simulated structures, including the root-mean-square deviation (RMSD), the root-mean-square fluctuation (RMSF), the solvent-accessible surface area (SASA), hydrogen bonds (HBs), and the radius of gyration (Rg) [38].

2.6. Molecular Dynamics Simulation and Post MM-GBSA Evaluation by Desmond. The molecular dynamics simulation provides evidence regarding the mobility and stability of the bound protein-ligand complex. On Desmond software, the MD simulation and post-MMGBSA analysis of the main-protease_ligand, Nsp3_ligand, and human ACE2 receptor_ligand complexes were performed [39]. These compounds were solvated on a cubic TIP3P water model using the system builder package. A minimal spacing of 10 was maintained between the protein and the solvated region. Subsequently, Na+ salts were supplied until the final system strength reached 0.15 M, which is the physiological salt concentration present in the human body. The integrated OPLS3e force field was used to optimize the final system's energy. To complete the MDS, we used the isothermal isobaric ensemble (NPT) at 1.013 bar and 310° K. The total

Table 1: Drug-like properties of the best hit phytochemicals and control drugs.

Phytochemicals/ Drugs	MW (g/mol)	Rotatable bonds	H-bond acceptors	H-bond donors	Lipinski violation	Ghose violation	Veber violation	Egan violation	Muegge violation
CHEMBL503 (Lovastatin)	404.54	7	5	1	0	0	0	0	0
CHEMBL490355 (Sulfuretin)	270.24	1	5	3	0	0	0	0	0
CHEMBL4216332 (Grayanoside A)	476.47	10	10	5	0	0	0	0	0
Molnupiravir	329.31	6	8	4	0	1	1	1	0
Paxlovid	501.54	12	8	4	1	1	1	0	0

Table 2: ADMET properties of the best hit phytochemicals and control drugs.

Phytochemicals/ Drugs	Human intestinal absorption (% absorbed)	Caco-2 permeability (log Papp in 10-6 cm/s)	BBB permeability (log BB)	CYP2D6 substrate	CYP1A2 inhibitor	AMES toxicity	hERG I inhibitor	hERG II inhibitor	Hepatotoxicity
CHEMBL503 (Lovastatin)	94.656	0.924	-0.366	Yes	No	No	No	No	No
CHEMBL490355 (Sulfuretin)	98.77	1.795	-0.618	Yes	No	No	No	No	No
CHEMBL4216332 (Grayanoside A)	49.67	48.681	-1.266	No	No	No	No	No	No
Molnupiravir	53.464	0.531	-1.057	No	No	No	No	No	Yes
Paxlovid	61.975	0.081	-0.907	No	No	Yes	No	No	Yes

Table 3: Binding affinity and nonbonded interaction between the main protease (PDB ID: 7NT3) and the best hit phytochemical and control drugs.

Phytochemicals/ Drugs	Affinity (kcal/mol)	No. of H bonds	Interacting amino acids	No. of hydrophobic bonds	Interacting amino acids	No. of halogen bonds	Interacting amino acids	No. of electrostatic bonds	Interacting amino acids
Lovastatin	-7.2	1	ARG131 (2.30102 Å)	2	LEU28, PRO293	×	×	×	×
Molnupiravir	-6.4	3	ASP197, THR199 (2.04463 Å), LEU287	3	LEU27, TYR23, TYR239	×	×	1	ASP289
Paxlovid	-6.6	6	THR26, HIS41, ASN11, ASN14, GLY143 (1.98365 Å), CYS145	8	HIS49, MET49, ILE249, PRO29, HIS41	3	GLY10, GLN11, ASN203	1	GLU166

Table 4: Binding affinity and nonbonded interaction between the Nsp3 (PDB ID: 7KQP) and the best hit phytochemical and control drugs.

Phytochemicals/ Drugs	Affinity (kcal/mol)	No. of H bonds	Interacting amino acids	No. of hydrophobic bonds	Interacting amino acids	No. of halogen bonds	Interacting amino acids
Sulfuretin	-8.8	7	VAL49, LEU126, SER128, ALA129, GLY130, PHE156, ALA38	7	ALA38, PHE132, VAL49, ALA38, ALA50, VAL49, PRO125	×	×
Molnupiravir	-7.7	6	ASN40, GLY47, VAL49, ALA50, LYS44, ALA38 (1.90623 Å)	7	ALA38, PHE132, ALA52, ILE23, VAL49, PHE156	×	×
Paxlovid	-7.5	4	LYS158, LEU160, TYR161 (1.23877 Å)	8	ALA38, VAL49, ALA129, VAL155, LEU160, LEU126, LEU160	1	GLY48

Table 5: Binding affinity and nonbonded interaction between the human ACE2 receptor (PDB ID: 1R4L) and the best hit phytochemical and control drugs.

Phytochemicals/ Drugs	Affinity (k cal/Mol)	No of H bonds	Interacting amino acids	No of hydrophobic bonds	Interacting amino acids	No of halogen bonds	Interacting amino acids
Grayanoside A	-7.8	3	ARG273, ARG273, GLU406 (1.84129 Å)	7	Val209, LYS562, TRP566, LEU95, LYS562, ALA99	×	×
Molnupiravir	-7.6	5	ASP206, HIS378, ASN394, ARG514, LYS562 (2.198 Å)	4	TYR51, HIS401, PHE504, TYR510	×	×
Paxlovid	-7	6	ASP206, ALA348, TRP349 (1.978 Å), ASP350, HIS378, ARG514	×	×	1	SER43

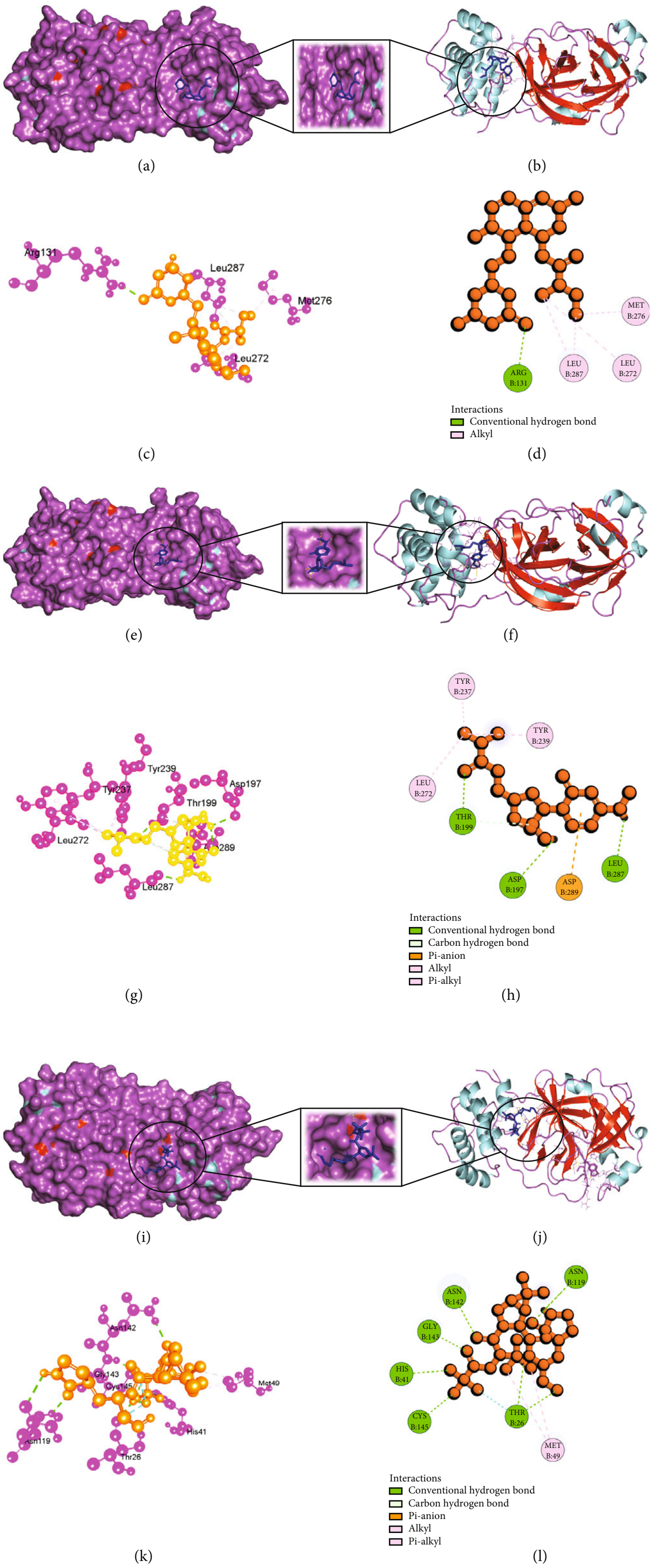


Figure 2: Schematic illustration of 7NT3_CHEMBL503 (Lovastatin), 7NT3_Molnupiravir, and 7NT3_Paxlovid complexes. (a, b) Share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors and ligand is in blue color. (c, d) Share 3D and 2D interactions of protein and ligand complex. Magenta color represents proteins, and yellow color presents ligands. (e, f) Share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors and ligand is in blue color. (g, h) Share 3D and 2D interactions of protein and ligand complex. Here, protein is in agenta color and ligand is in yellow color. (i, j) Share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors and ligand is in blue color. (k, l) Share 3D and 2D interactions of protein and ligand complex. Here, protein in magenta color and ligand in yellow color.

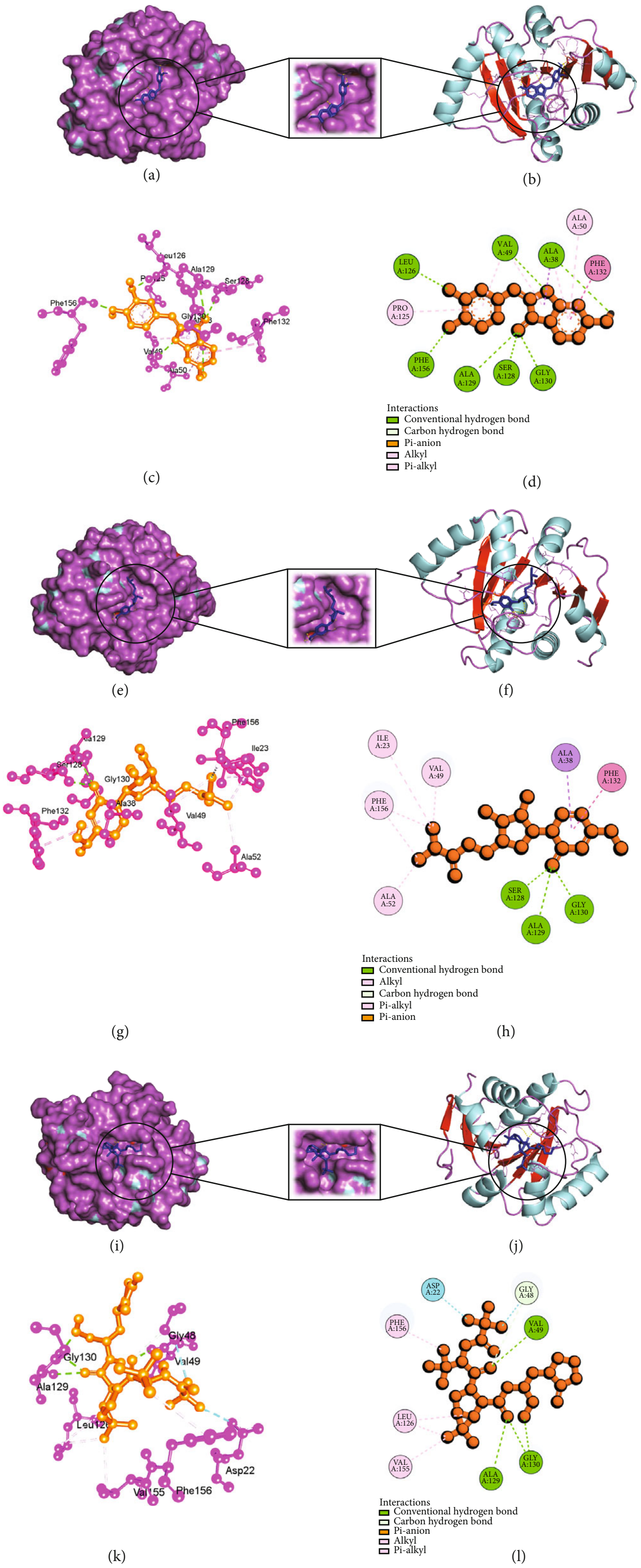


Figure 3: Schematic illustration of 7KQP_CHEMBL490355 (Sulfuretin), 7KQP_Molnupiravir, and 7KQP_Paxlovid complexes. (a, b) Share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors and ligand is in blue color. (c, d) Share 3D and 2D interactions of protein and ligand complex. Magenta color represents proteins and yellow color presents ligands. (e, f) Share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors and ligand is in blue color. (g, h) Share 3d and 2D interactions of protein and ligand complex. Here, protein in magenta color and ligand in yellow color. (i, j) share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors and ligand is in blue color. (k, l) Share 3D and 2D interactions of protein and ligand complex. Here, protein is in magenta color and ligand is in yellow color.

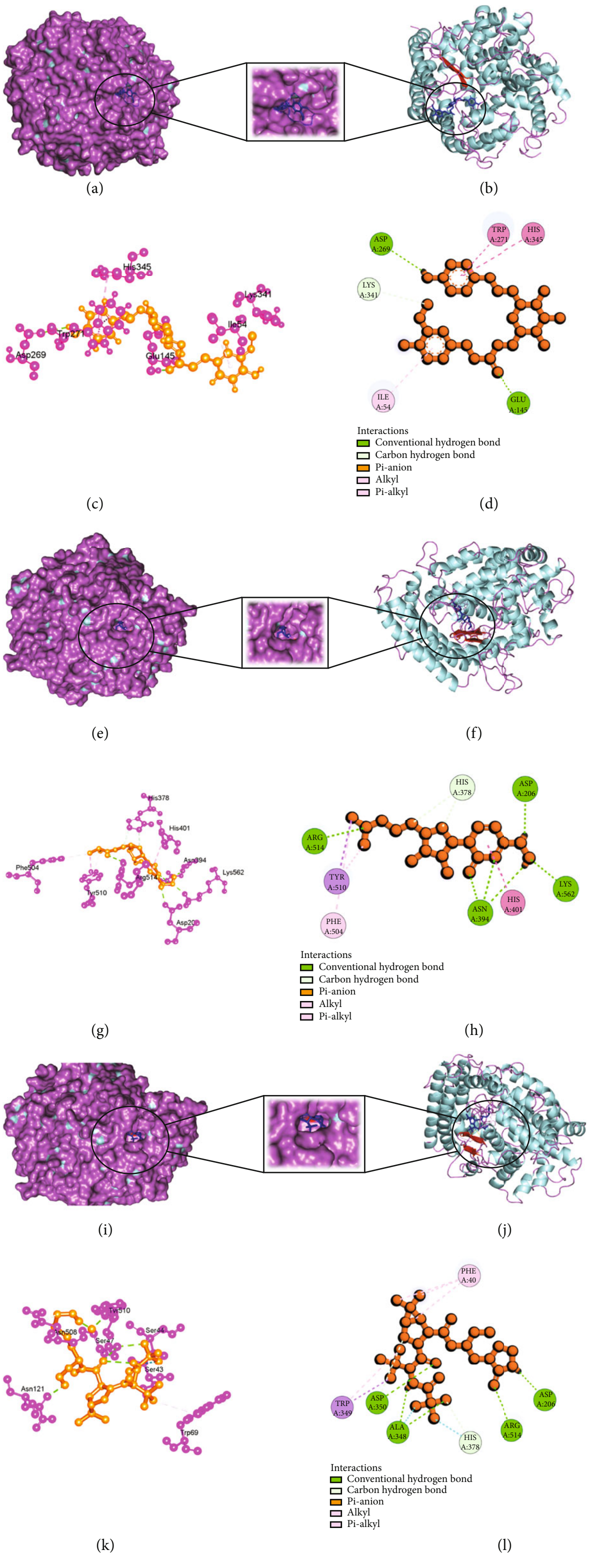


Figure 4: Schematic illustration of 1R4L_CHEMBL4216332 (Grayanoside A), 1R4L_Molnupiravir, and 1R4L_Paxlovid complexes. (a, b) Share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors and ligand is in blue color. (c,, d) Share 3D and 2D interactions of protein and ligand complex. Magenta color represents proteins and yellow color presents ligands. (e, f) Share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors, and ligand is in blue color. (g, h) Share 3D and 2D interactions of protein and ligand complex. Here, protein in magenta color and ligand in yellow color. (i, j) Share the pose and surface view of protein and ligand complex. Here, protein is in purple and cyan colors and ligand is in blue color. (k, l) Share 3D and 2D interaction of protein and ligand complex. Here, protein in magenta color and ligand in yellow color.

Table 6: XP Gscore and MM-GBSA values between the main protease (PDB ID: 7NT3) and the best hit phytochemical and control drugs.

Drug	XP Gscore (kcal Mol−1)	MM-GBSA scores (kcal Mol−1)	Hydrogen bonds	Hydrophobic bonds
Lovastatin	-6.01	-52.85	HIS163, GLU166, GLN189	LEU27, CYS44, MET49, TYR54, PHE140, LEU141, CYS145, GLY154, MET165
Molnupiravir	-5.035	-43.48	GLU166	CYS44, MET49, PRO52, TYR54, CYS145, MET165, LEU167, PRO168
Paxlovid	-5.185	-43.34	GLU166, ASN142	CYS44, MET49, PRO52, TYR54, PHE140, LEU141, GLY143, MET165, LEU167

Table 7: XP Gscore and MM-GBSA values between the Nsp3 (PDB ID: 7KQP) and the best hit phytochemical and control drugs.

Drug	XP Gscore (kcal Mol−1)	MM-GBSA scores (kcal Mol−1)	Hydrogen bonds	Hydrophobic bonds
Sulfuretin	-9.563	-52.85	ALA38, ASN40, GLY47, ALA50	ALA39, VAL49, PRO125, LEU126, LEU127, ALA129, ILE131, PHE132, PHE156
Molnupiravir	-7.604	-43.48	VAL49, ALA39, LEU126	ALA38, ALA39, PRO125, LEU126, LEU127, ALA129, ILE131, PHE132, PHE156
Paxlovid	-2.727	-43.34	GLY48, GLY130, LEU126	ALA38, VAL49, PRO125, LEU126, LEU127, ALA129, ILE131, VAL155, PHE156

Table 8: XP Gscore and MM-GBSA values between the human ACE2 receptor (PDB ID: 1R4L) and the best hit phytochemical and control drugs.

Drug	XP Gscore (kcal mol−1)	MM-GBSA scores (kcal mol−1)	Hydrogen bonds	Hydrophobic bonds
Grayanoside A	-7.87	-63.54	ARG273, HIS345, ALA348, GLN375	TYR127, LEU144, TRP271, PHE274, CYS344, PRO346, ALA348, PHE504, TYR510, TYR515
Molnupiravir	-6.02	-40.53	ALA348, GLN375, ARG514	PRO346, TRP349, PHE504, TYR510, TYR515
Paxlovid	-5.679	-32.02	ARG273, HIS345, ALA348, GLN375,	TYR127, LEU144, TRP271, PHE274, CYS344, PRO346, ALA348, PHE504, TYR510, TYR515

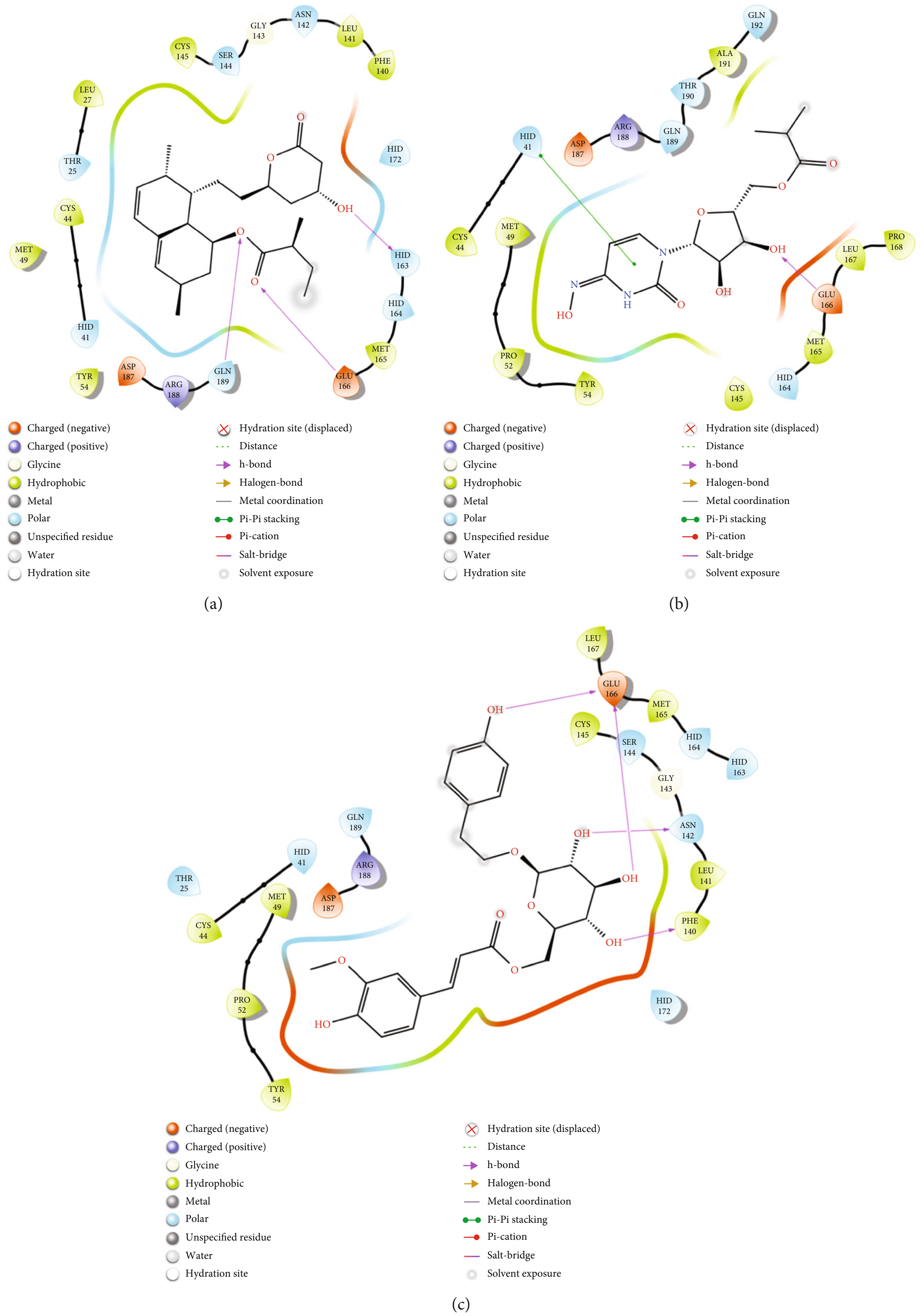
duration of the simulation run was 100 nanoseconds (ns). It was paired with a recording duration of 100 picoseconds (ps), during which 1000 frames were incorporated into the trajectory. Next, we studied the trajectories in the simulation interaction diagram (SID) program, and the reported results comprised RMSD, RMSF, protein-ligand contact outline, and biophysical properties of ligands. After running the simulations, MM-GBSA was evaluated employing the thermal MM-GBSA.py program. During the assessment, a 0-1000 periodic frame was incorporated for the analysis [40].

2.7. Prediction of Molecular Target with SwissTargetPrediction Server. The anticipation of a molecular target for a small-molecule is vital for drug research and development. These studies are essential for assessing the potential for adverse reactions or cross-reactivity in *Homo sapiens* caused by the action of that bioactive small molecule. We employed SwissTargetPredcition (http://www.swisstargetprediction.ch/) to determine the human body receptors for small compounds that had previously been identified by molecular docking and shown stability via MD simulation investigations [41]. The canonical smiles of the small compounds were used in the server and analyzed.

2.8. Prediction of Biological Activity by PASS-Way2Drug Tool. The PASS-Way2Drug webserver (http://www.pharmaexpert.ru/passonline/) was employed to the prediction of biological activity scales for Lovastatin, Sulfuretin, and Grayanoside A [42]. For PASS recommendations to be reliable, the Pa (likelihood to be effective) threshold should be set at 70% or above. This is because surpassing the Pa>70% threshold yields very reliable predictions [42]. Calculated ligand activity was based on Pi and Pa scores.

3. Results

3.1. Analysis of Drug-Like Properties. In this experiment, 1163 drug-like substances were checked for their drug-like activities. All of them have been filtered using five principles of Lipinski's filtration technique, which included molecular mass (recommended value: <500), the number of hydrogen bond donors (ideal value: ≤5), the number of hydrogen bond acceptors (standard range: ≤10), lipophilicity (represented as LogP, normal value: <5), and molar refractivity (preferable range: 40–130). Additionally, the ligands were screened based on the criteria of Ghose, Veber, Egan, and Muegge's rule. Subsequently, 497 out of 1163 compounds were



(a)

(b)

(c)

FIGURE 5: 2D interaction of (a) 7NT3_Lovastatin, (b) 7NT3_Molnupiravir, and (c) 7NT3_Paxlovid complexes.

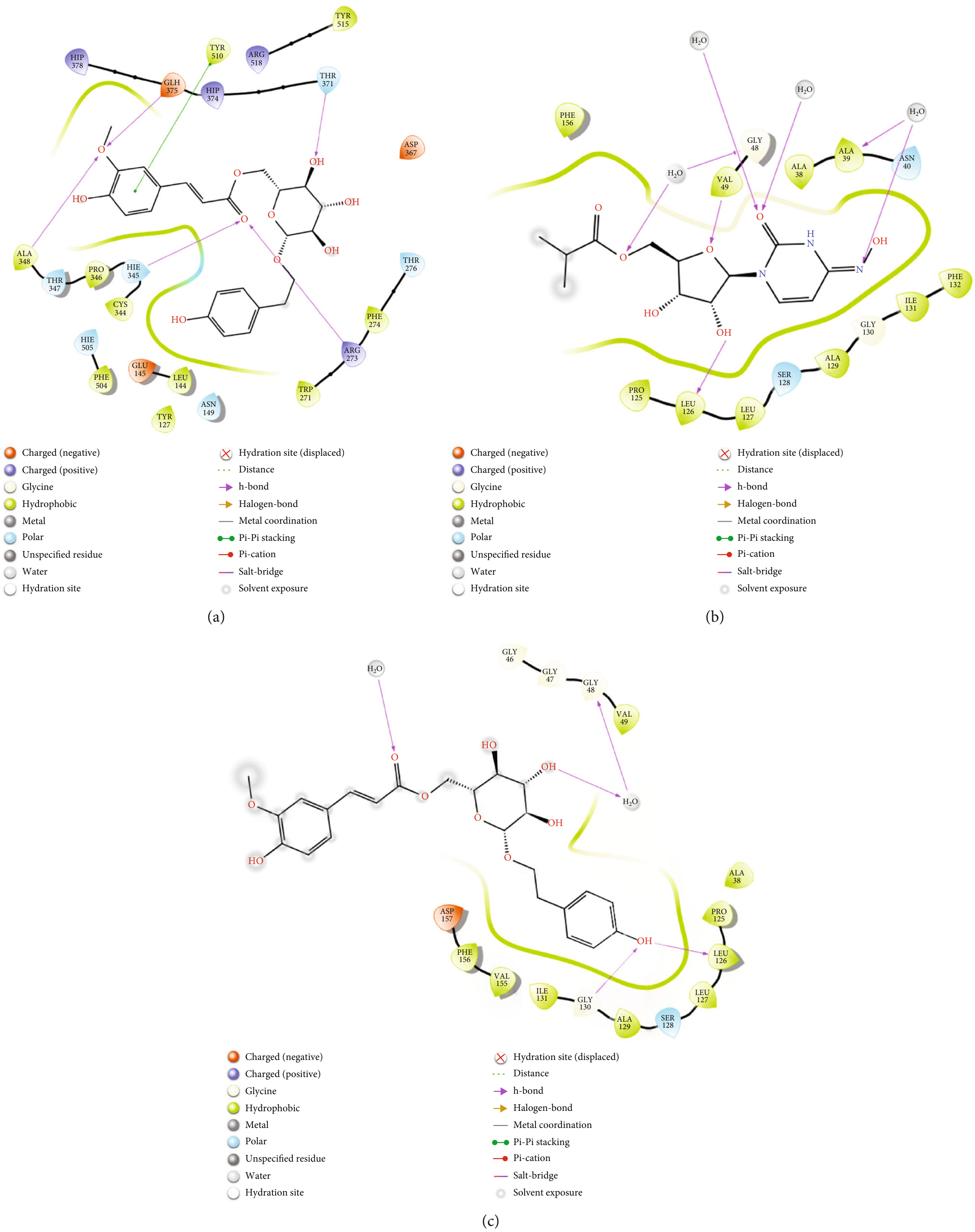
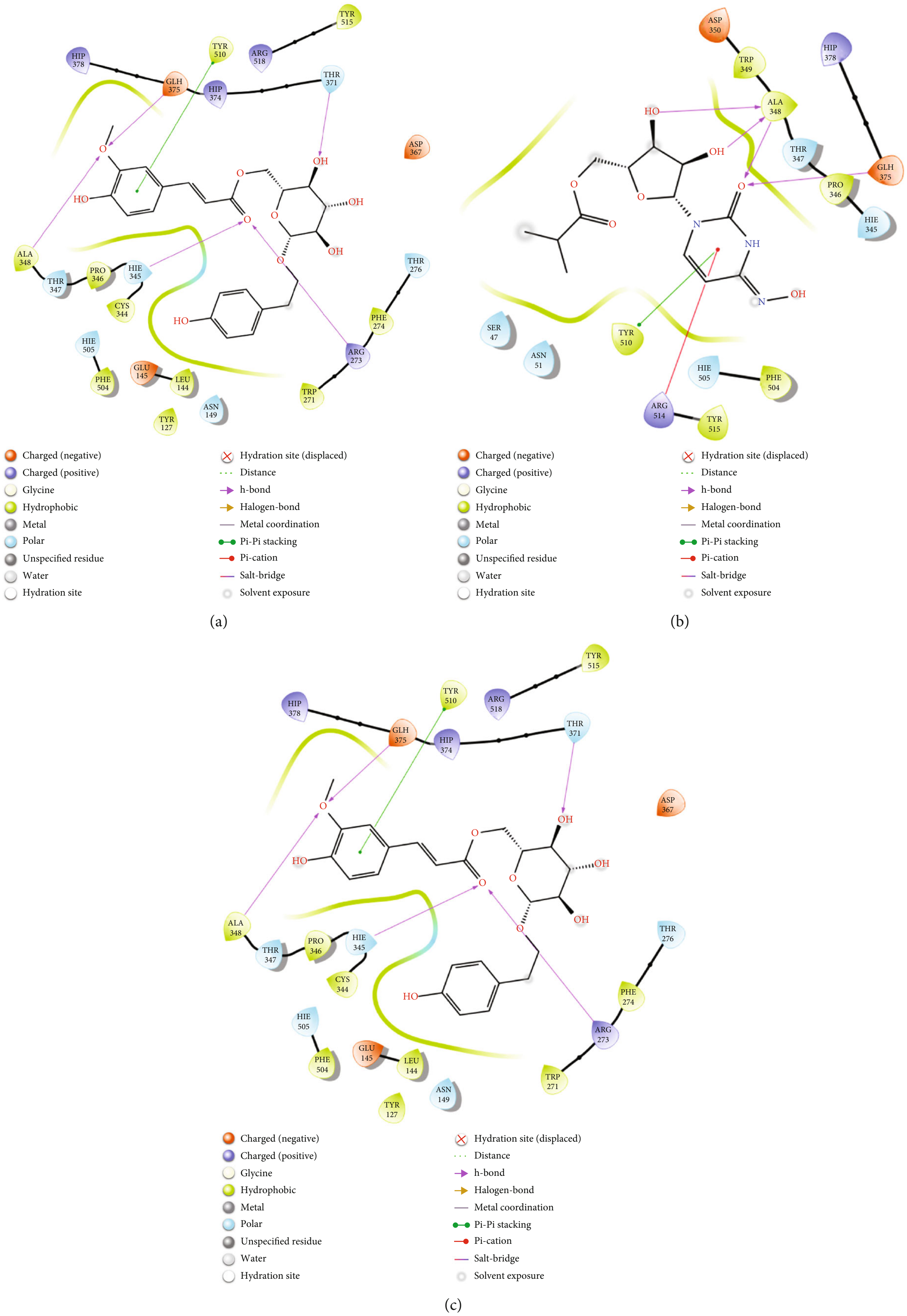


Figure 6: 2D interaction of (a) 7KQP_Sulfuretin, (b) 7KQP_Molnupiravir, and (c) 7KQP_Paxlovid complexes.



(a)

(b)

(c)

Figure 7: 2D interaction of (a) 1R4L_Grayanoside A, (b) 1R4L_Molnupiravir, and (c) 1R4L_Paxlovid complexes.

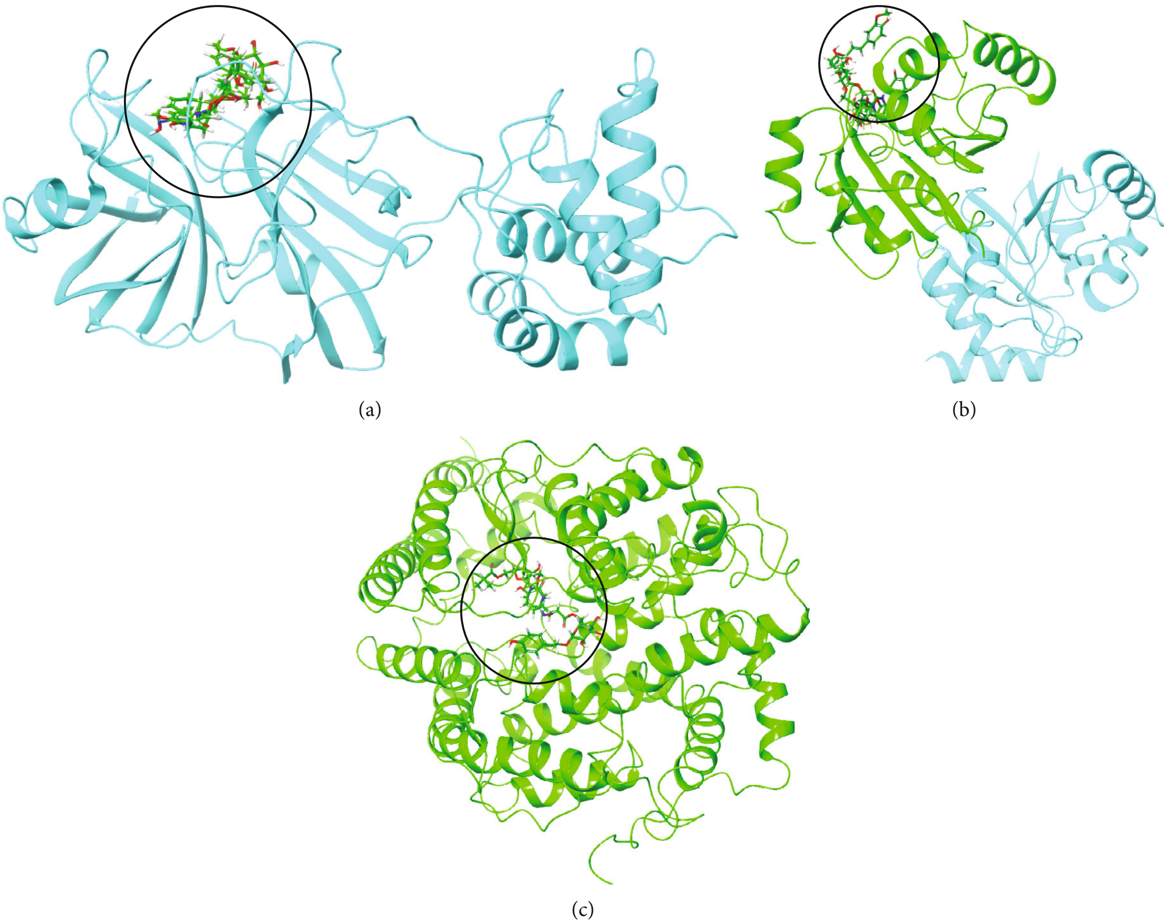


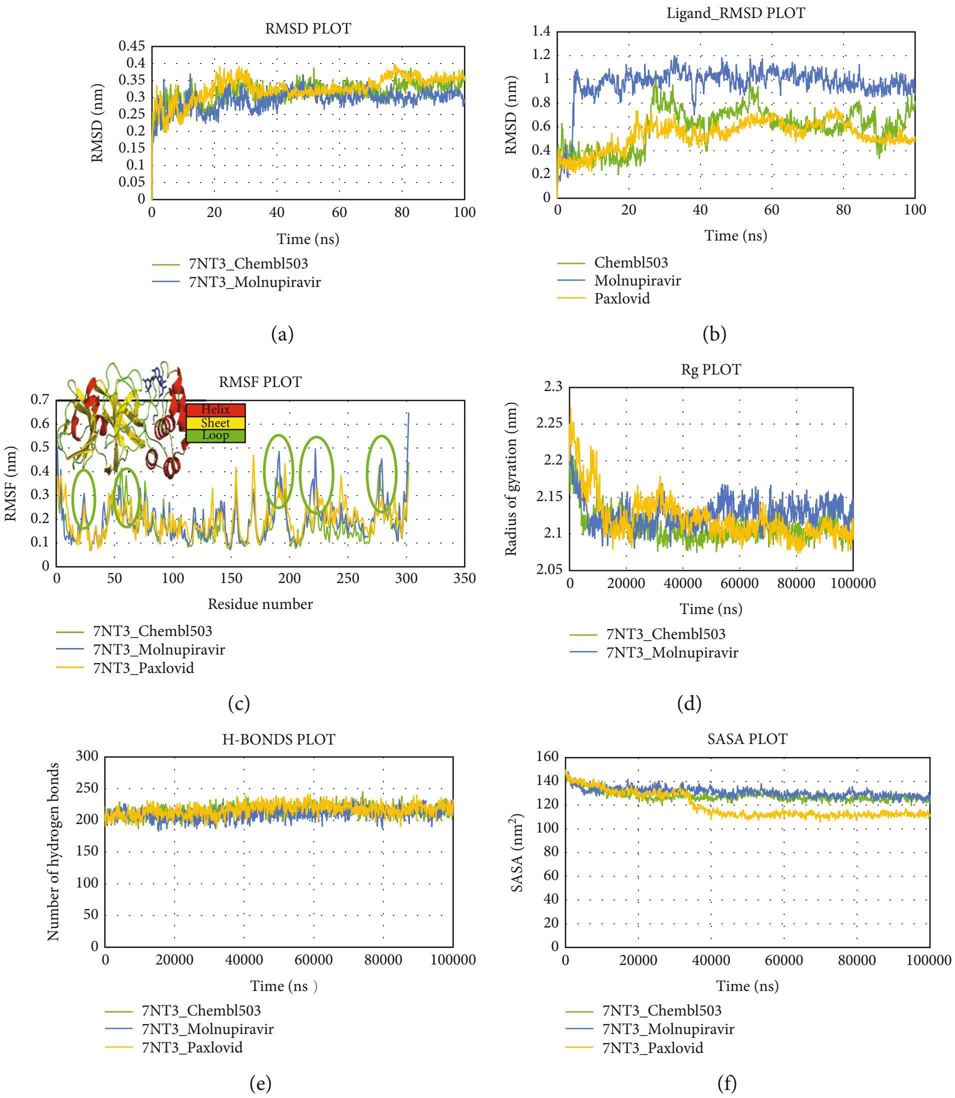
Figure 8: Illustration of 3D representation of (a) 7NT3_complexes, (b) 7KQP_complexes, and (c) 1R4L_complexes. Black circle portrays the binding pockets and incorporates ligands and cocrystallized compounds.

shortlisted for the following evaluation (Supplementary table1). Table 1 represented the drug-like properties of the best-hit phytochemicals and control drugs.

3.2. Analysis of ADMET Properties. A total of 149 drug-like substances were qualified after this analysis. Moreover, from estimating distribution levels, all the compounds are impermeable to the blood-brain barrier. Metabolic inability could cause lower bioavailability and excretion, high toxicity, and drug-drug interactions. These 149 small substances function as isoforms of the CYP 2D6 and 3A4 enzymes. Diverse computational algorithms are used to evaluate toxicity: hERG inhibitors, AMES toxicity, maximum tolerated dosage Hepatotoxicity. Ligands with a negative value in these models were chosen for the following step. ADMET properties of the best-hit phytochemicals and control drugs were presented in Table 2. Finally, we filtered out 149 drug-like substances from this analysis (Supplementary table2).

3.3. Analysis of Molecular Docking Results by AutoDock vina. In structure-based pharmaceutical research, molecular docking is a commonly used strategy to identify the finest ligand hits against a particular protein. The docking method predicts the ligand orientation, location, conformation in the protein's active site, binding interaction, and affinity. AutoDock vina determines the binding energy and poses of trial ligands by employing a grid-based technique. Previously selected small molecules were docked with three SARS-CoV-2 target proteins. Out of the 149 small compounds, 97 small molecules exhibited a higher binding affinity for the main protease (Supplementary Table 3a), 75 small molecules for the Nsp3 (Supplementary Table 3b), and 106 small molecules for human ACE2 receptor compared to the control therapeutics (Supplementary Table 3c).

Lovastatin's binding energy for the main protease was -7.2 kcal/mol, which was considerably higher than that of the control ligands, Molnupiravir (-6.4 kcal/mol), and Paxlovid (-6.6 kcal/mol) (Table 3). Lovastatin produced a robust hydrogen interaction with the amino acids ARG131 (2.30102 Å) of the main protease, whereas Molnupiravir and Paxlovid formed three and six hydrogen bonds with the target protein, respectively, with THR26, HIS41, ASN119, ASN142, GLY143 (1.98365 Å), and CYS145 residues (AutoDock vina). Sulfuretin had binding energy of -8.8 kcal/mol for Nsp3 compared to the control ligands molnupiravir and Paxlovid, which had binding energies of -7.7 and -7.5 kcal/mol, respectively (Table 4). Sulfuretin formed seven strong hydrogen bonds with the Nsp3 protein (VAL49, LEU126, SER128, ALA129, GLY130, PHE156, and ALA38), whereas Molnupiravir and Paxlovid created



(a)

(b)

(c)

(d)

(e)

(f)

Figure 9: Schematic illustration of 100 ns molecular dynamics simulation of 7NT3_CHEMBL503 (Lovastatin) (green), 7NT3_Molnupiravir (blue), and 7NT3_Paxlovid complexes (yellow). Representations (a, b, c, d, e, and f) share the RMSD, RMSF, Rg, hydrogen bonds, and SASA values of 7NT3_CHEMBL503 (Lovastatin), 7NT3_Molnipiravir, and 7NT3_Paxlovid complexes. Representation b shares ligand RMSD value of Chembl503, Molnupiravir and Paxlovid.

six (ASN40, GLY47, VAL49, ALA50, LYS44, and ALA38 (1.90623 Å)) and three (LYS158, LEU160, and TYR161 (1.23877 Å)) amino acid residues. Sulfuretin also created seven hydrophobic bonds (ALA38, PHE132, VAL49, ALA38, ALA50, VAL49, and PRO125) with the same protein. For human ACE2 receptor, Grayanoside A showed a binding affinity of 7.8 kcal/mol compared to the control molecules Molnupiravir (-7.6 kcal/mol) and Paxlovid (-7.0 kcal/mol) (Table 5). Molnupiravir and Paxlovid formed five (ASP206, HIS378, ASN394, ARG514, and LYS562 (2.198 Å)) and six (ASP206, ALA348, TRP349 (1.978 Å), ASP350, HIS378, and ARG514) hydrogen bonds with the target protein, human ACE2 receptor, respectively. Grayanoside A formed three strong hydrogen bonds (ARG273, ARG273, and GLU406) and six hydrophobic bonds (VAL209, LYS562, TRP566, LEU95, LYS562, and ALA99) with the human ACE2 receptor.

3.4. Analysis of Glide and MM-GBSA Scores. Glide incorporates high-throughput virtual screening (HVS), estimating protein-ligand interacting sites and grading ligands using experimental score systems. Out of the 149 small compounds, 120 small molecules exhibited greater binding energy for SARS-CoV-2 main protease (Supplementary Table 4a), 75 small molecules for Nsp3 (Supplementary Table 4b), and 99 small molecules for human ACE2 receptor compared to control therapeutics (Supplementary Table 4c) (Figures 2–4), and it showed the comparative representation of protein-ligand complexes of the best hit ligands and control drugs. Here, Tables 6–8 summarized the Glide score and MM-GBSA scores between three target proteins and the best hit phytochemicals and control drugs. Analysis of glide XP score and MMGBSA values, it was evident that Lovastatin is better candidate than other potential ligands. It formed three hydrogen bonds (HIS163,

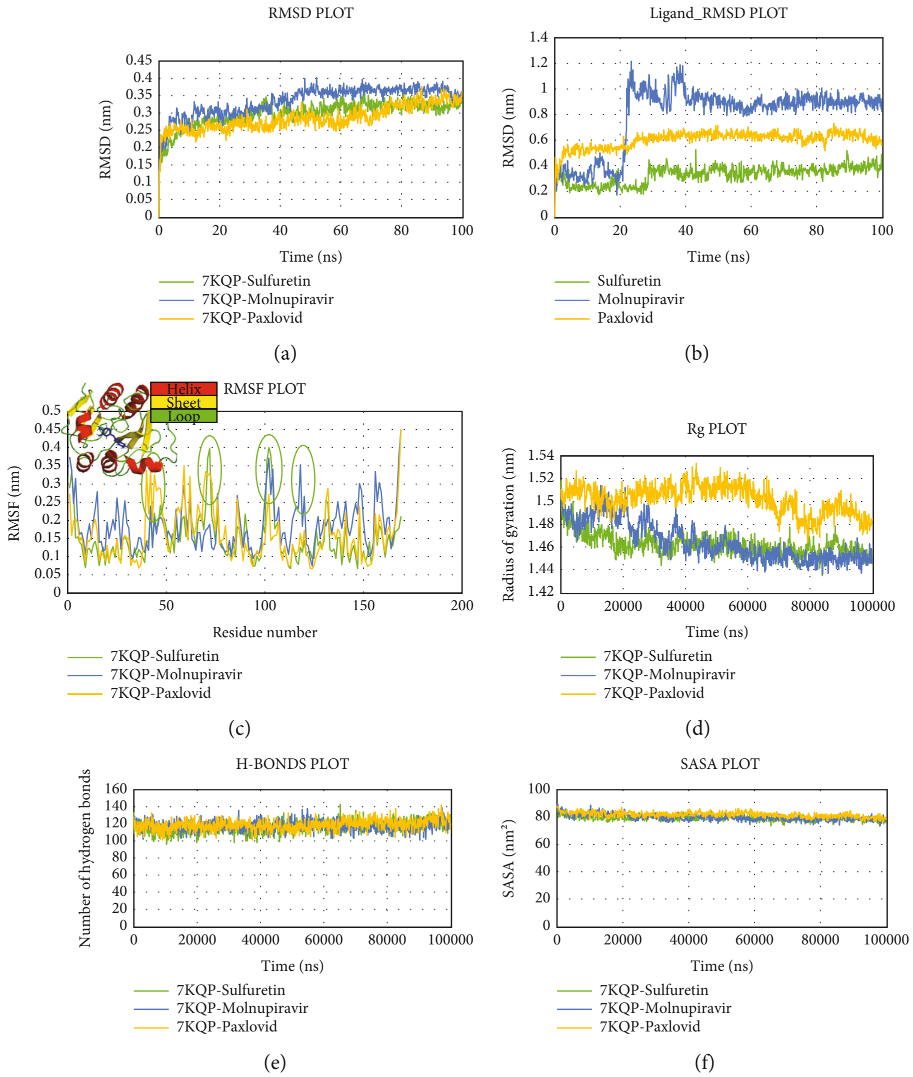


Figure 10: Schematic illustration of 100 ns molecular dynamics simulation of 7KQP_CHEMBL490355 (Sulfuretin) (green), 7KQP_Molnupiravir (blue), and 7KQP_Paxlovid complexes (yellow). Representations (a, b, c, d, e, and f) shares the RMSD, RMSF, Rg, hydrogen bonds, and SASA values of 7KQP_CHEMBL490355 (Sulfuretin), 7KQP_Molnipiravir, and 7KQP_Paxlovid complexes. Representation b share Ligand RMSD value of CHEMBL490355, Molnupiravir, and Paxlovid.

GLU166, and GLN189) and nine hydrophobic bonds (LEU27, CYS44, MET49, TYR54, PHE140, LEU141, CYS145, GLY154, and MET165) with the main protease (PDB ID: 7NT3). Sulfuretin showed glide and MMGBSA scores of -9.563 and -52.85 (kcal/mol). It created four hydrogen bonds (ALA38, ASN40, GLY47, and ALA50) and nine hydrophobic bonds (ALA39, VAL49, PRO125, LEU126, LEU127, ALA129, ILE131, PHE132, and PHE156). Grayanoside A managed a glide and MMGBSA scores of -7.87 and -63.54 (kcal/mol). It maintained four hydrogen bonds (ARG273, HIS345, ALA348, and GLN375) and ten hydrophobic bonds (TYR127, LEU144, TRP271, PHE274, CYS344, PRO346, ALA348, PHE504, TYR510, and TYR515) with the human ACE2 receptor (PDB ID: 1R4L) (Figures 5–7). Lovastatin, Sulfuretin, and Grayanoside A were found inside the binding cavity with the cocrystallized compound (Figure 8). As a result, they were proved to be the best candidate for main protease and Nsp3 of SARS-CoV-2 and human ACE2 proteins, respectively.

3.5. *Analysis of Molecular Dynamics Simulation.*

To circumvent the fundamental drawback of molecular docking, we ran a computational MD simulation, which incorporated the dynamic character of the protein following inhibitor binding. This experiment produced statistical figures for the RMSD, RMSF, hydrogen bonds, SASA, and Rg values of protein-ligand complexes. The average RMSD of main protease_Lovastatin, main protease_Molnupiravir, and main protease_Paxlovid complexes for the main protease was 0.312696293 nm, 0.291836715 nm, and 0.326214306 nm, respectively, indicating that the chosen drug candidate Lovastatin exhibited an identical result compared to Molnupiravir and Paxlovid (Figure 9). As per Figure 9, the main protease_Lovastatin and main protease_Molnupiravir complexes were stable with a fixed RMSD value less than 0.35 from 30 to 80 ns, but the main protease_Paxlovid complex had an increased RMSD value more than 0.35 after 75 ns. Similarly, the predicted average RMSD values of the ligands

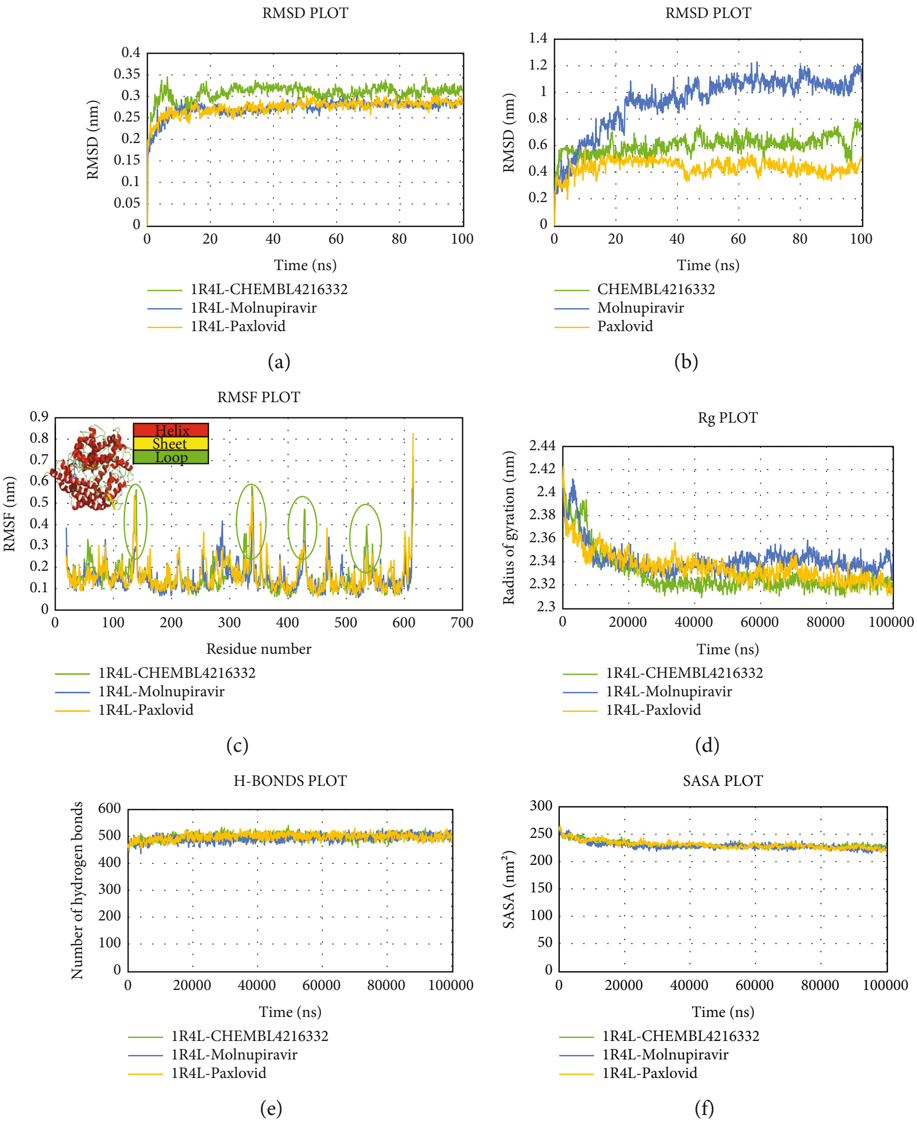
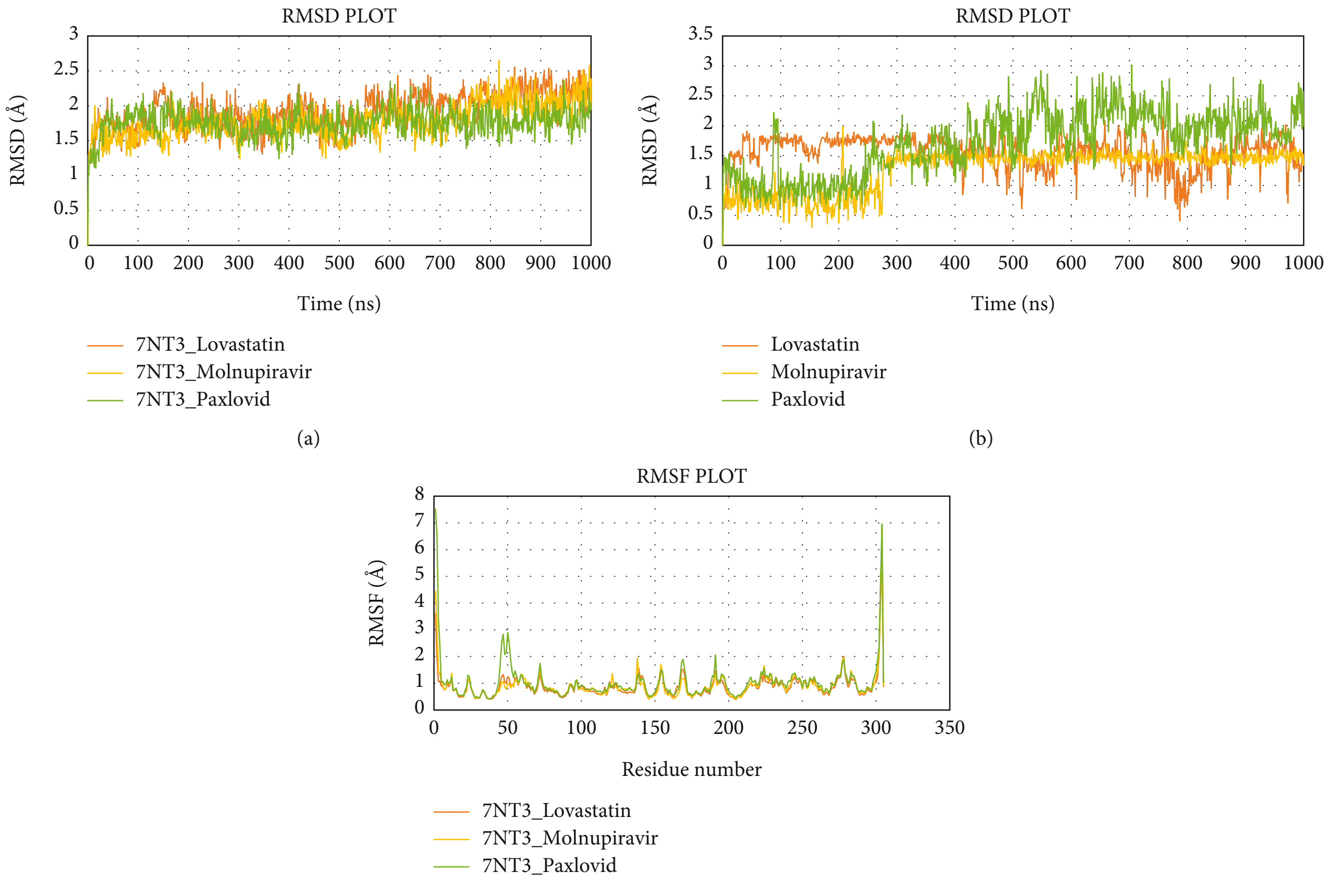


Figure 11: Schematic illustration of 100 ns molecular dynamics simulation of 1R4L_CHEMBL4216332 (Grayanoside A) (green), 1R4L_Molnupiravir (blue), and 1R4L_Paxlovid (yellow). Representations (a, b, c, d, e, and f) share the RMSD, RMSF, Rg, hydrogen bonds, and SASA values of 1R4L_CHEMBL4216332 (Grayanoside A), 1R4L_Molnipiravir, and 1R4L_Paxlovid complexes. Representation b share ligand RMSD value of CHEMBL4216332, Molnupiravir and Paxlovid.

Lovastatin, Molnupiravir, and Paxlovid were 0.594898993 nm, 0.96096714 nm, and 0.531292037 nm, respectively. Throughout 100 ns simulation, the RMSF value of amino acids for backbone components of the main protease_Lovastatin complex was less than 0.4 nm, but the main protease_Molnupiravir and main protease_Paxlovid complexes showed some inconsistent higher fluctuation. The RMSD fluctuation of the ligands inside the first three loop regions between 50 and 80 amino acids was less than 0.40 nm. However, the RMSF oscillation inside the other three considerably larger loop areas was higher than 0.40 for Molnupiravir and Paxlovid. The average Rg values of the complexes main protease_Lovastatin, main protease_Molnupiravir, and main protease_Paxlovid were 2.109437 nm, 2.128492 nm, and 2.122325654 nm, respectively, describing the increased compactness of the Lovastatin complex. Main protease_Lovastatin, main protease_Molnupiravir, and main protease_Paxlovid complexes had an average of 215.0, 209.0, and 216.0 hydrogen bonds, respectively, showing a significant dynamic interaction of the main protease_Lovastatin complex. Figure 9(f) represented the solvent-accessible surface area (SASA) of structures. While the main protease_Lovastatin and main protease_Molnupiravir complexes had an average SASA value of 127.8404086 nm^2 and 130.891962 nm^2, the main protease_Paxlovid complex had a lower value of 119.4976923 nm^2.

The average RMSD value of the Nsp3_Sulfuretin and Nsp3_Paxlovid complexes for SARS-CoV-2 Nsp3 protein was 0.297815 nm and 0.284552759 nm, respectively, though the Nsp3_Molnupiravir complex displayed an increased variation of RMSD value exceeding 0.35 nm after 45 ns (Figure 10). During the 100 ns simulation timeline with Nsp3 protein, control drugs molnupiravir and Paxlovid



(a)

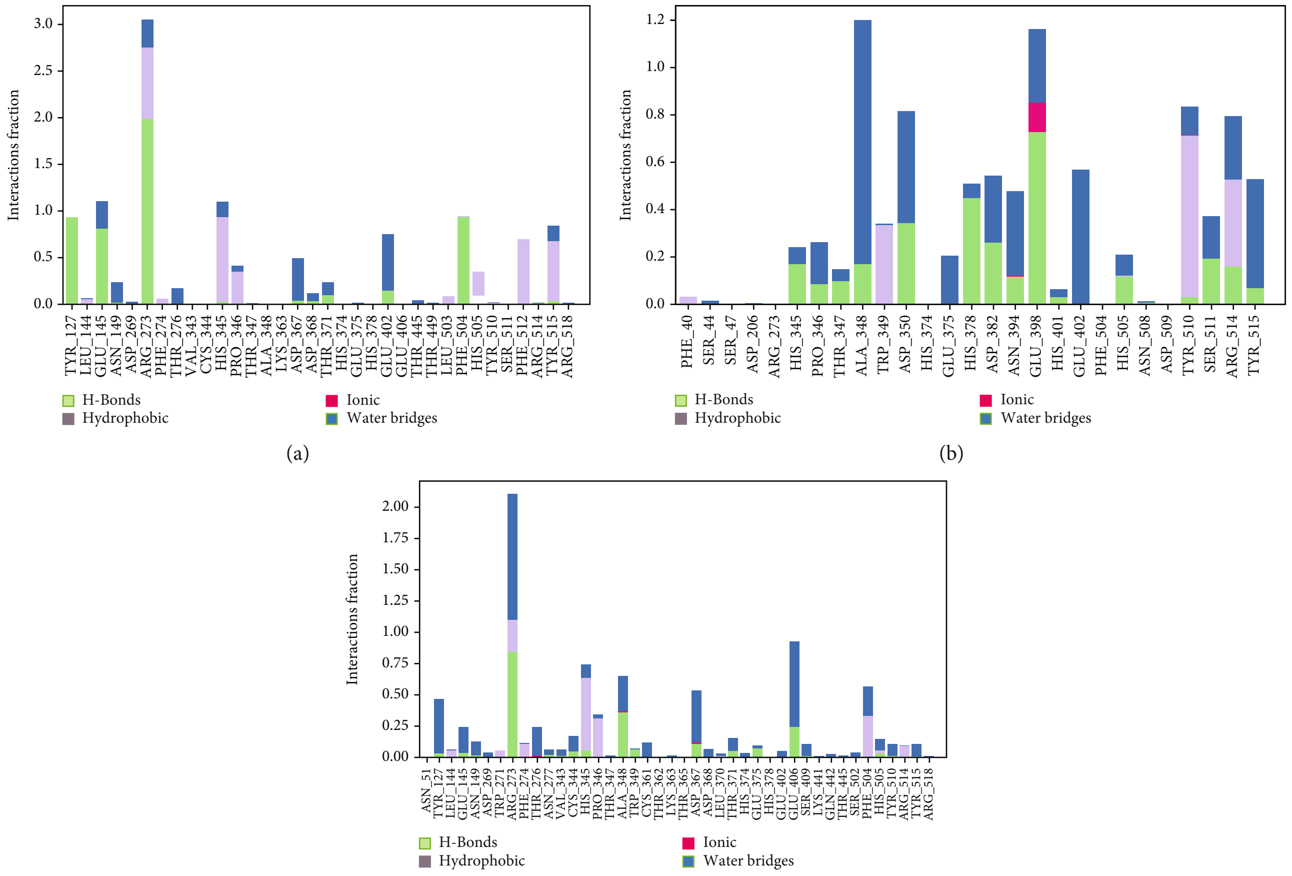
(b)

(c)

Figure 12: Simulation graph of root-mean-square deviation (RMSD) showing Lovastatin_7NT3 (orange), Molnupiravir_7NT3 (yellow), and Paxlovid_7NT3 (green). (b) Simulation graph of root-mean-square deviation (RMSD) showing Lovastatin (orange), Molnupiravir (yellow), and Paxlovid (green). (c) Simulation findings showing of root-mean-square fluctuation (RMSF) of Lovastatin_7NT3 (orange), Molnupiravir_7NT3 (yellow), and Paxlovid_7NT3 (green).

had an RMSD value above 0.6 nm and 0.8 nm. However, after an initial equilibration phase, Sulfuretin stayed below 0.4 nm, indicating the most stable of the three ligands. Except for the C-terminal and N-terminal areas, the RMSF value of the Nsp3_Sulfuretin, Nsp3_Molnupiravir, and Nsp3_Paxlovid complexes was less than 0.4 nm. Furthermore, there was higher fluctuation among the structures inside larger loop sections between 41 and 46, 83 and 91, 97 and 105, and 116 and 135 amino acids. The Rg values of the Nsp3_Sulfuretin and Nsp3_Molnupiravir complexes stabilized after initial equilibration steps, but the Rg value of Nsp3_Paxlovid complexes oscillated more during the whole run time. According to Figure 10(e), the average count of hydrogen bonds among the complexes Nsp3_Sulfuretin, Nsp3_Molnupiravir, and Nsp3_Paxlovid were 116.0, 117.0, and 119.0, indicating a similar course of interaction within the 100 ns timeframe. The SASA value of the Nsp3_Sulfuretin, Nsp3_Molnipiravir, and Nsp3_Paxlovid complexes were stable with an average value of 79.26847 nm, 79.74635 nm, and 81.74065634 nm respectively.

The RMSD value of the human ACE2 receptor_Grayanoside A, human ACE2 receptor_Molnupiravir, and human ACE2 receptor_Paxlovid complexes for human ACE2 protein stayed under 0.35 nm, and stable throughout the 100 ns run (Figure 11). Likewise, the ligands Grayanoside A, Molnupiravir, and Paxlovid had average RMSD values of 0.601344 nm, 0.933326 nm, and 0.43800 nm, respectively. The RMSF of backbone heteroatoms per residue of the human ACE2 receptor_Grayanoside A complex stayed within 0.4 nm, with higher RMSF oscillation inside loops from 137 to 139 and 338 to 340 residues. On the other hand, peaks inside loop regions beyond 0.4 nm were evident from 137 to 140 and 331 to 345 residues for human ACE2 receptor_Molnupiravir and human ACE2 receptor_Paxlovid complexes, respectively. The average Rg values of human ACE2 receptor_Grayanoside A, human ACE2 receptor_Molnipiravir, and human ACE2 receptor_Paxlovid complexes were 2.329435, 2.342172667, and 2.335405325 nm, respectively. The average hydrogen bond interactions for the complexes human ACE2 receptor_Grayanoside A and human ACE2 receptor_Molnupiravir were 499.0 and 492.0, respectively, whereas the complex human ACE2 receptor_Paxlovid had a higher value of 498.0. Figure 11 shows that the SASA values of the human ACE2 receptor_Grayanoside A, human ACE2 receptor_Molnupiravir, and human ACE2 receptor_Paxlovid complexes were stable throughout the simulation, suggesting that the protein's surface area remained unaltered.



(a)

(b)

(c)

FIGURE 13: Contact maps of Lovastatin_7NT3 (a), Molnupiravir_7NT3 (b), and Paxlovid_7NT3 (c) complexes.

3.6. Evaluation of MD Simulation and Post-MD Simulation MM-GBSA Results from Desmond. Analyzing the simulation trajectory, we plotted the RMSF, RMSD, biophysical properties of ligands, and protein-ligand network. We found an average RMSD plot of 1.92, 1.78, and 1.75 Å for Lovastatin_7NT3, Molnupiravir_7NT3, and Paxlovid_7NT3 complexes. The protein structure of the Sulfuretin_7NT3 complex remained under 3 Å throughout the simulation. The ligands Sulfuretin, Molnupiravir, and Paxlovid had average RMSD of 1.55, 1.27, and 1.71 Å respectively, indicating a stable conformation with protein. Similarly, the average RMSF of Lovastatin_7NT3, Molnupiravir_7NT3, and Paxlovid_7NT3 complexes was 0.87, 0.91, and 1.04 Å respectively. Except for N-terminal and C-terminal zones, all complexes stayed under 3 Å (Figure 12). Sulfuretin interacted with 7NT3 creating bonds with THR26 (hydrogen bonds and water bridges), GLY143 (hydrogen bonds and water bridges), SER144 (hydrogen bonds and water bridges), CYS145 (hydrogen bonds and water bridges), and GLU166 (hydrogen bonds, water bridges, and ionic bonds) amino acids for 30%, 20%, 30%, 40%, 20%, and 100% of 100 ns timeframe. Molnupiravir interacted with HIS41 (hydrophobic), GLU166 (water bridges), VAL186 (hydrogen bonds and water bridges), and GLN189 (hydrogen bonds and water bridges) for 80%, 100%, 70%, and 90% of 100 ns timescale. Paxlovid had bonds with HIS41 (hydrophobic, hydrogen bonds, and water bridges), GLY143 (hydrogen bonds and water bridges), SER144 (hydrogen bonds and water bridges), and GLU166 (hydrogen bonds, water bridges, and ionic bonds) for 50%, 90%, 40%, and 300% of the simulation period (Figure 13).

Protein structures of Sulfuretin_7KQP, Molnupiravir_7KQP, and Paxlovid_7KQP showed an average RMSD value of 1.97, 1.77, and 1.65 Å. All the complex structures remained under 3 Å which suggested that the ligands were tightly bound inside the binding pocket of receptor structures. The ligands Sulfuretin, Molnupiravir, and Paxlovid displayed an average RMSD of 0.19, 1.33, and 2.37 Å respectively. RMSF plot presented an average of 0.94, 1.97, and 0.92 Å for Sulfuretin_7KQP, Molnupiravir_7KQP, and Paxlovid_7KQP complexes implying structural stability (Figure 14). Sulfuretin made bonds with ASN40 (hydrogen bonds and water bridges), LYS44 (hydrogen bonds, water bridges, and ionic bonds), HIS45 (hydrogen bonds and water bridges), GLY48 (hydrogen bonds and water bridges), PHE156 (hydrogen bonds and water bridges) residues of 7KQP for 17%, 30%, 25%, 30%, and 20% of simulation timeframe. Molnupiravir_7KQP complex formed bonds with THR57 (hydrogen bonds and water bridges), ASN58 (hydrogen bonds and water bridges), HIS86 (hydrophobic, hydrogen bonds, and water bridges) residues for 20%, 11%, and 26% of simulation timeframe. Paxlovid_7KQP complex

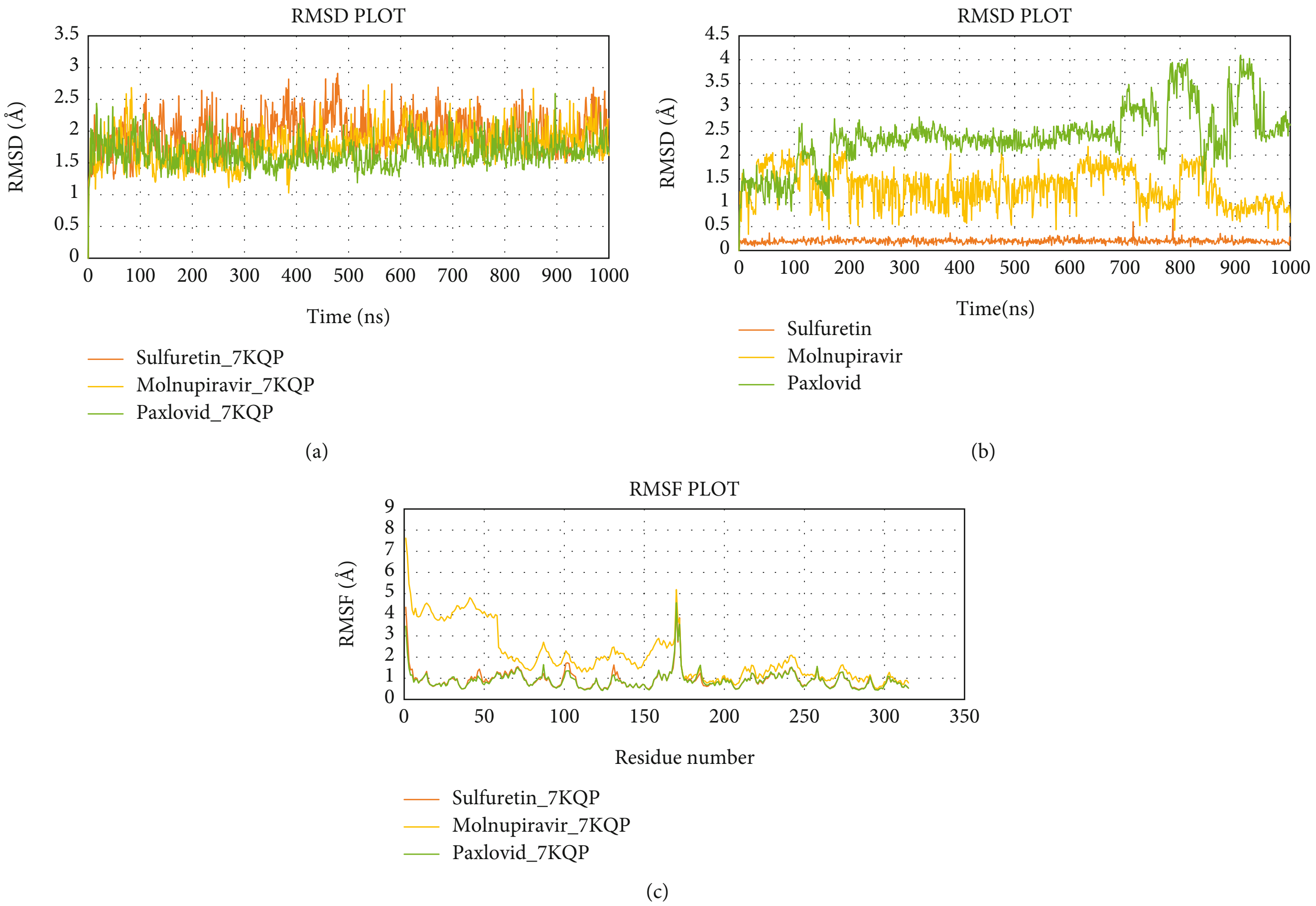
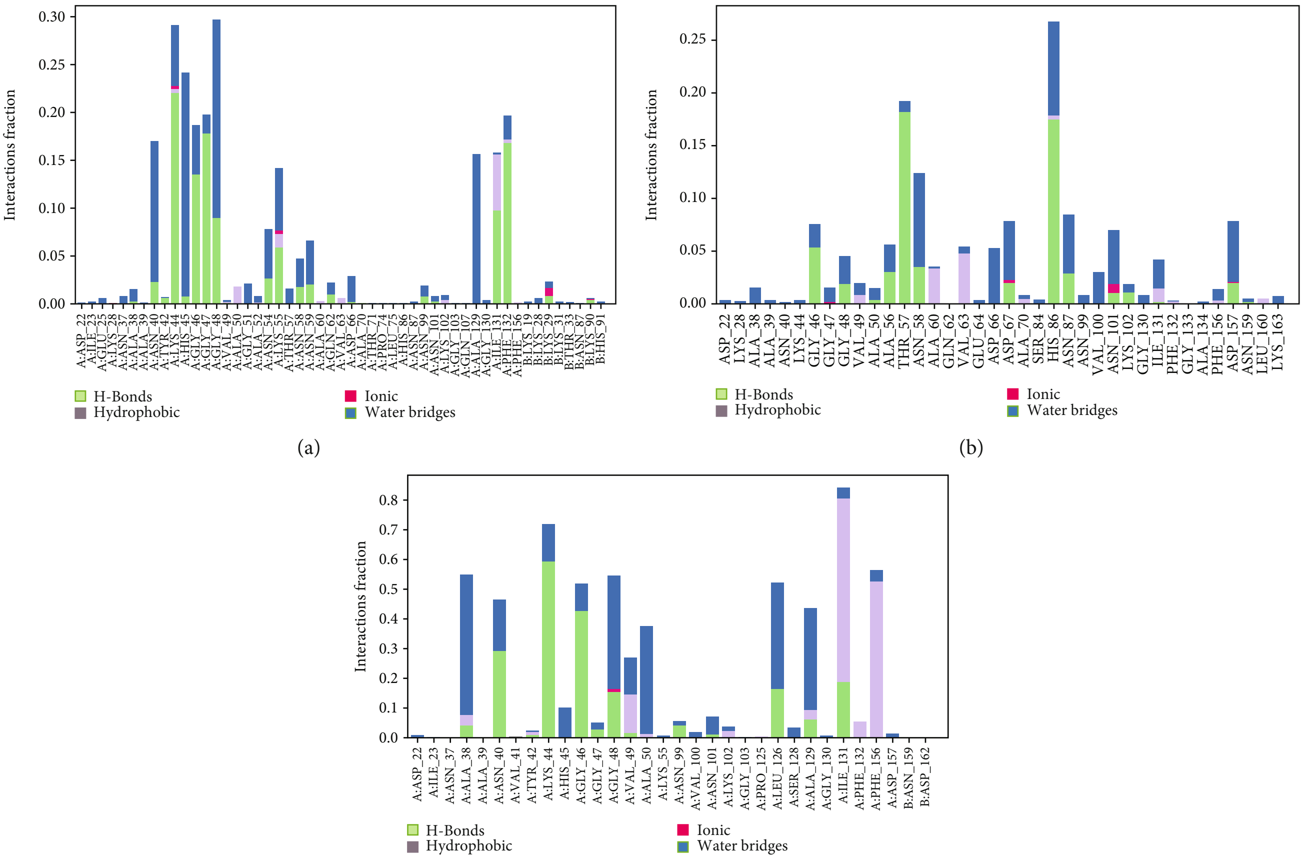


Figure 14: Simulation graph of root-mean-square deviation (RMSD) showing Sulfuretin_7KQP (orange), Molnupiravir_7KQP (yellow), and Paxlovid_7KQP (green). (b) Simulation graph of root-mean-square deviation (RMSD) showing Sulfuretin (orange), Molnupiravir (yellow), and Paxlovid (green). (c) Simulation findings showing of root-mean-quare fluctuation (RMSF) of Sulfuretin_7KQP (orange), Molnupiravir_7KQP (yellow), and Paxlovid_7KQP (green).

maintained binding network with ALA38 (hydrophobic, hydrogen bonds, and water bridges), ASN40 (hydrogen bonds and water bridges), LYS44 (hydrogen bonds and water bridges), GLY46 (hydrogen bonds and water bridges), GLY48 (hydrogen bonds, water bridges, and ionic bonds), ILE131 (hydrophobic, hydrogen bonds, and water Bridges), and GLU156 (hydrophobic and water bridges) residues for 55%, 45%, 70%, 52%, 55%, 80%, and 55% of the simulation run (Figure 15).

Next, the Grayanoside A_1R4L, Molnupiravir_1R4L, and Paxlovid_1R4L complex structures maintained an average RMSD of 1.97, 1.80, and 2.22 Å respectively. Grayanoside A_1R4L complex remained 2.7 Å throughout the timeframe demonstrating a stable protein-ligand association. The average RMSD of the ligands Grayanoside A, Molnupiravir, and Paxlovid was 2.15, 1.23, and 2.03 Å respectively. The proteins of Grayanoside A_1R4L, Molnupiravir_1R4L, and Paxlovid_1R4L complexes maintained an average RMSD of 0.83, 0.93, and 1.31 Å respectively. A small steep was observed between 115 to 125 residues for Grayanoside A_1R4L and Molnupiravir_1R4L complexes (Figure 16). Grayanoside A had TYR127 (hydrogen bonds), GLU145 (hydrogen bonds and water bridges), ARG273 (hydrophobic, hydrogen bonds, and water bridges), HIS345 (hydrophobic and water bridges), GLU402 (hydrogen bonds and water bridges), PHE504 (hydrogen bonds), and HIS505 (hydrophobic, hydrogen) binding residues with 1R4L protein for 90%, 110%, 330%, 110%, 80%, 100%, and 40% of simulation cycle. Molnupiravir_1R4L complex formed interaction with ALA348, ASP350, GLU398, TYR510, and ARG514 for 120%, 80%, 119%, 82%, and 80% of the simulation period. On the other hand, Paxlovid_1R4L complex had interactions with ARG273, HIS345, ALA348, and GLU406 residues for 200%, 70%, 65%, and 90% of the simulation timescale (Figure 17). We also superimposed the pre_MD and post_MD structures of protein-ligand complexes in Desmond and found less than 2 Å deviation (Figure 18).

The average postsimulation MM-GBSA of Lovastatin_7NT3, Molnupiravir_7NT3, and Paxlovid_7NT3 complexes were −52.56 ± 8.93, −50.52 ± 12.75, and −49.68 ± 16.27 kcal/mol, respectively. Sulfuretin_7KQP, Molnupiravir_7KQP, and Paxlovid_7KQP complexes had average postsimulation MM-GBSA scores of −66.17 ± 11.62, -36.51 ± 13.74, and −54.30 ± 15.45 kcal/mol, respectively. Grayanoside A_1R4L, Molnupiravir_1R4L, and Molnupiravir_1R4L complexes showed an average MM-GBSA value of −74.94 ± 8.50, −34.23 ± 12.82, and −57.50 ± 24.35 kcal/mol, respectively, (Tables 9–11).



(a)

(b)

(c)

FIGURE 15: Contact maps of Sulfuretin_7KQP (a), Molnupiravir_7KQP (b), and Paxlovid_7KQP (c) complexes.

3.7. Analysis of Target within Human. The target sites in humans where Lovastatin binds (in humans) are cytochrome p450, oxidoreductase, kinase, family A of G protein-coupled receptor, enzyme, and membrane receptor and the possibility of binding, respectively, are 16%, 12%, 8%, 8%, 8%, and 4% respectively. For Sulfuretin, they they may bind with kinase (52%), enzyme (24%), and membrane receptor (4%); and for Grayanoside A, they they may bind with protease (20%), kinase (20%), surface antigen (4%), enzyme (12%), and family A of G protein-coupled receptor (4%). Control drug Paxlovid provides the binding possibility in target sites are protease (60%), enzyme (16%), family A of G protein-coupled receptor (8%), membrane receptor (4%), and surface antigen (4%). The prediction tool did not show any human target for Molnupiravir (Figure 19).

3.8. Analysis of Activity Spectra of the Phytochemicals. Using the identified compounds, prediction of activity spectra for substances (PASS) was carried out and is shown in Supplementary Tables 6a, 6b, 6c. In our study, we used PASS to build a predictive model, and we kept the Pa (likelihood of activity) that was higher than 70%; since an absolutely durable forecast may be made using the Pa > 70% (0.7) criteria. Lovastatin had 18 biological activities, Sulfuretin showed 27 activities, and Grayanoside A possessed 30 biological features. Lovastatin, Sulfuretin, and Grayanoside A present Pa values greater than 0.70 across the board to be considered for use as an active biological agent in the treatment of SARS-CoV-2.

4. Discussion

In recent years, pandemics and epidemics caused by viruses have become one of the most prevalent reasons for infections and mortality worldwide. SARS-CoV-2, the updated variant of coronaviruses, has led to a catastrophe, with 665,518,891 and 6,714,212 confirmed cases and fatalities, respectively (11th January,2023; https://covid19.who.int/). Surprisingly, currently, a limited amount of effective anti-SARS-CoV-2 therapeutics are available, and most of them are under investigation [43].

Following the outbreak of SARS-CoV-2, Mpro, also regarded as 3CLpro (main protease), became a viable therapeutic focus due to its involvement in the development of replication-translation mechanisms. Furthermore, given the accessibility of high-resolution protein structures, these proteins have a highly conserved sequence and no homology with human proteins [44]. Nsp3 is a multidomain protein with a Glu-rich acidic domain at the N-terminus, an X domain, a SUD domain, a papain-like protease domain, and a transmembrane domain. Nsp3 is responsible for viral multiplication and pathogenesis in humans and facilitates immune evasion via its hydrolyzing capability [43]. The attachment of the SARS-CoV-2 Spike protein to human

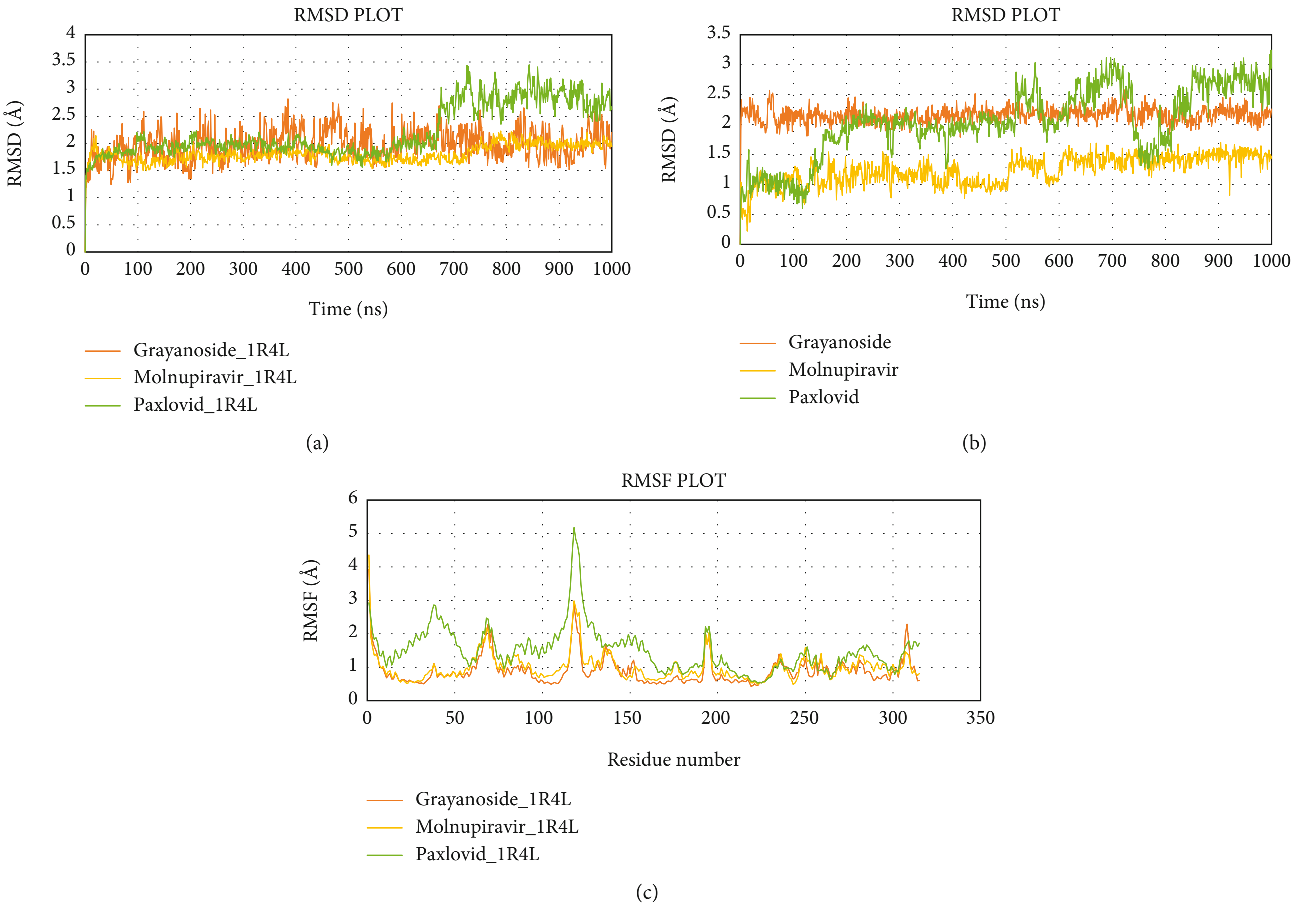
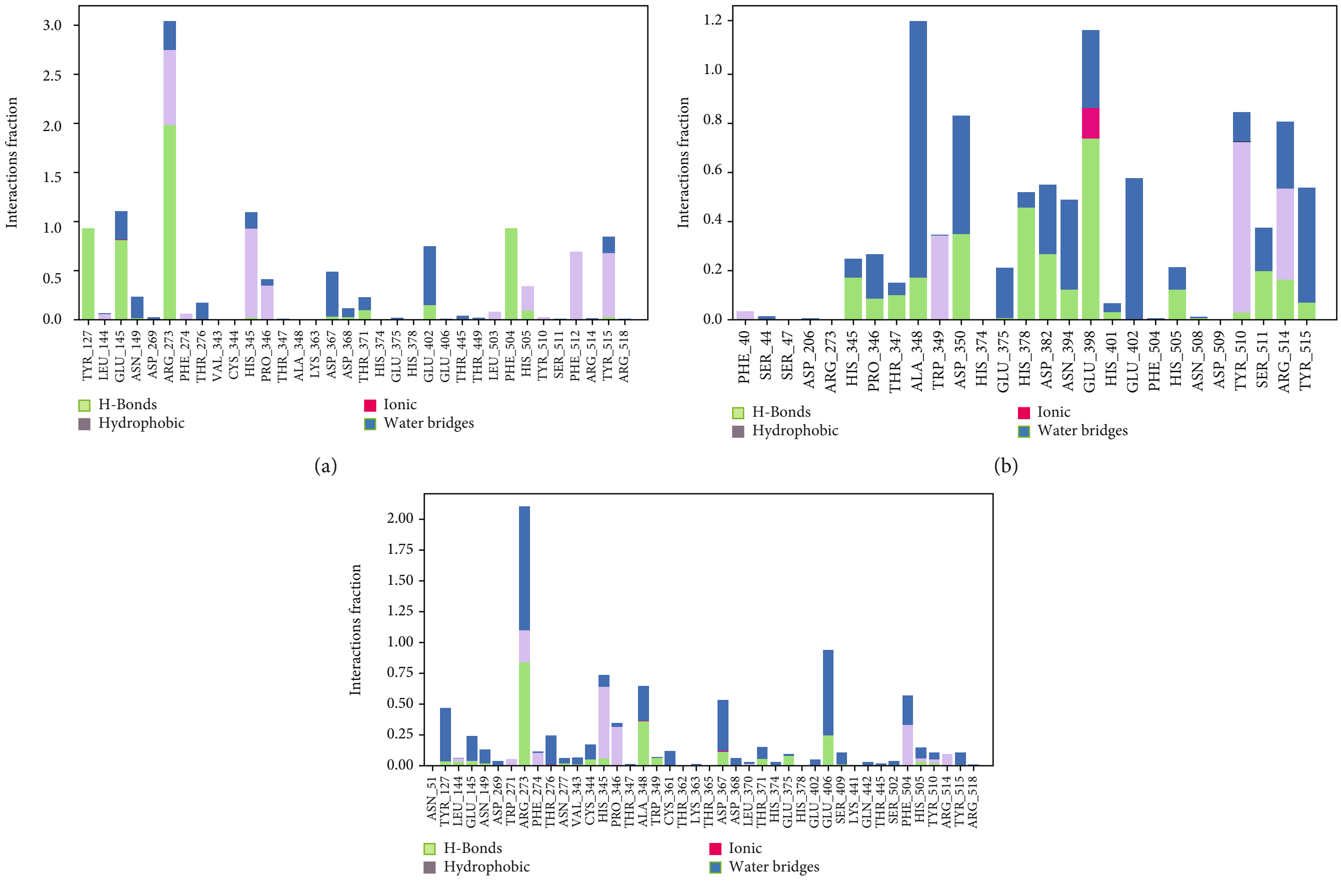


Figure 16: Simulation graph of root-mean-square deviation (RMSD) showing Grayanoside A_1R4L (orange), Molnupiravir_1R4L (yellow), and Paxlovid_1R4L (green). (b) Simulation graph of root-mean-square deviation (RMSD) showing Grayanoside A (orange), Molnupiravir (yellow), and Paxlovid (green). (c) Simulation findings showing of root-mean-square fluctuation (RMSF) of Grayanoside A_1R4L (orange), Molnupiravir_1R4L (yellow), and Paxlovid_1R4L (green).

ACE2 receptor on the cellular wall permits the virus to enter cells, which is required for infection to begin [45]. To inhibit these viral proteins, we utilized phytochemicals with drug-like properties.

This research was divided into three sections, namely, virtual screening (VS) of the physicochemical and pharmacokinetic features of drug-like compounds, virtual screening by molecular docking of proteins and ligands, and simulation of the best hit complexes. In the initial stage, we studied the drug-like characteristics of the ligands utilizing the five principles of Lipinski. We stuck to the established guidelines, i.e., hydrogen bond donors ≤ 5, hydrogen bond acceptors ≤ 10, molecular mass < 500, and logP < 5. The molecular weight of a small molecule can influence its absorption, bile excretion ratio, blood-brain barrier passage, and engagements involving receptors [46]. Likewise, hydrogen donor and hydrogen acceptor groups are mostly responsible for the permeability and polarity of the drug-like molecules. Lipophilicity is an indicator that influences the metabolism and solubility of those molecules. A lower or higher score might impede this characteristic [47]. TPSA refers to the area belonging to polar atoms like nitrogen, oxygen, and their associated hydrogens [48]. Out of 1163 small molecules, 497 of them passed the criteria. We tested the pharmacokinetic figure of the ligands before analyzing their binding affinity and orientation. The characteristics of a small molecule in terms of ADMET properties make it a viable candidate. Using the human intestinal absorption (HI) prediction score and the Caco-2 permeable theory, the probability that the small molecules would reach systemic circulation and exert their function was calculated [49]. P-glycoprotein serves as a drug carrier and eliminating compounds from different organs [50]. The main subfamily (2D6, 2C9, and 3A4) of cytochrome P450 monooxygenase (CYP) enzymes is crucial in the metabolism of the drug-like molecules [51]. In the initial phases of pharmaceutical research, AMES mutagenicity is commonly used to determine the probability of genotoxicity and teratogenicity [52]. Cardiovascular poisoning might be caused by inhibiting the cardiac human ether-a-go-go-related (hERG) gene [53]. We also tested the maximum tolerated dose of chemical substances for the human body. Eventually, only 149 drug-like molecules passed the ADMET evaluation.

The importance of virtual screening employing molecular docking has grown significantly in the field of drug development over time. According to the study, 24 small molecules had a greater binding affinity against the main protease (7NT3) than the reference compounds:



(a)

(b)

(c)

Figure 17: Contact maps of Grayanoside A_1R4L (a), Molnupiravir_1R4L (b), and Paxlovid_1R4L (c) complexes.

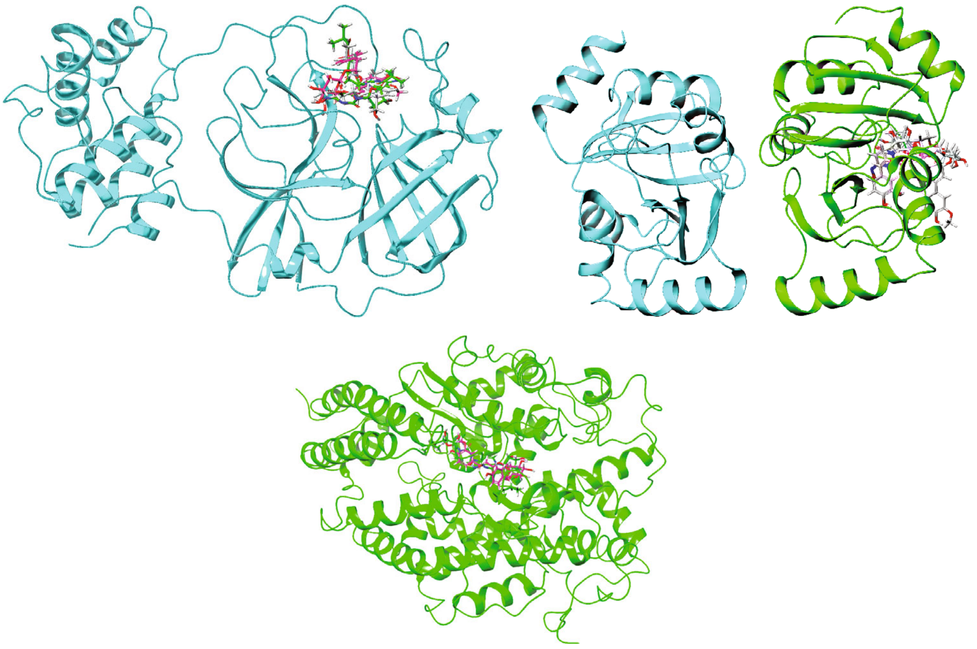


Figure 18: Superimposed representation of the pre-MD and post-MD structures of Ligand_7NT3, Ligand_7KQP, and Ligand_1R4L complexes.

Table 9: Post MD MM-GBSA based binding free energy evaluation for main protease (3CLpro) (PDB ID: 7NT3) and the best hit phytochemical along with control drugs.

Name of complex	MM-GBSA (kcal/mol)	
	ΔG_{Bind}	ΔG_{Bind} range
Lovastatin_7NT3	−52.56 ± 8.93	-61.49 to −43.63
Molnupiravir_7NT3	−50.52 ± 12.75	-63.27 to −37.77
Paxlovid_7NT3	−49.68 ± 16.27	-65.95 to −33.41

Table 10: Post MD MM-GBSA based binding free energy evaluation for Nsp3 (PDB ID: 7KQP) and the best hit phytochemical along with control drugs.

Name of complex	MM-GBSA (kcal/mol)	
	ΔG_{Bind}	ΔG_{Bind} range
Sulfuretin_7KQP	−66.17 ± 11.62	-77.79 to −54.55
Molnupiravir_7KQP	−36.51 ± 13.74	-50.25 to −22.77
Paxlovid_7KQP	−54.30 ± 15.45	-81.85 to −33.15

Table 11: Post MD MM-GBSA based binding free energy evaluation for human ACE2 receptor (PDB ID: 1R4L) and the best hit phytochemical along with control drugs.

Name of complex	MM-GBSA (kcal/mol)	
	ΔG_{Bind}	ΔG_{Bind} range
Grayanoside A_1R4L	−74.94 ± 8.50	-83.57 to −66.40
Molnupiravir_1R4L	−34.23 ± 12.82	-46.22 to −21.72
Paxlovid_1R4L	−57.50 ± 24.35	-81.85 to −33.15

Molnupiravir and Paxlovid (-5.035 and -5.185 kcal/mol, respectively). The MM-GBSA approach is recognized for its impressive precision in estimating the free binding energy of small molecules to target proteins. Both analyses indicated that Lovastatin (CHEMBL503) had a higher glide score and binding-free energy value of -6.01 kcal/mol and -52.85 kcal/mol, respectively. Recently, Mashraqi et al. found fenoterol had a glide score and MM-GBSA values of −7.14 and -38.733 kcal/mol [54]. We found Lovastatin attached to the active site residues (His41, Cys145) in the docking study and after MD simulation. Though CHEMBL182992, CHEMBL1909923, CHEMBL1972346, CHEMBL249454, CHEMBL477778, CHEMBL557501, and CHEMBL4216332 had better glide XP scores over 6 kcal/mol, the binding-free energy is higher for CHEMBL503 (Lovastatin). Similarly, two small compounds, CHEMBL490355 (Sulfuretin) and CHEMBL226683, showed binding energies greater than 9.0 Kcal/mol than control therapeutics against Nsp3 (7KQP). But the binding free energy (-46.31 kcal/mol) and the number of hydrogen bonds were higher for CHEMBL490355 (Sulfuretin). So, Sulfuretin was selected as the best candidate against Nsp3. Recently, Mishra et al. reported ZINC82673 as the potential inhibitor of Nsp3 with glide and MM-GBSA values of −9.348 and 50.175 kcal/mol [55]. It was also found inside the binding pocket (Asp22, Ile23, Gly48, Val49, Gly130, or Phe156) [43]. A total of 20 phytochemicals had higher glide XP scores over 6 kcal/mol and 2 of them showed binding-free energy above −50 kcal/mol against human ACE2 receptor. Based on the glide and MMGBSA scores, as well as the number of hydrogen bonds, we selected Grayanoside A as the lead candidate against human ACE2 receptor (1R4L). Most of the residues of the active site (Tyr515, Arg514, His505, Phe504, Glu402, His378, Glu375, His374, Asp368, Cys361, His345, Cys344, and Glu145) were found attached to Grayanoside A [56]. Pai et al. found that isochlorogenic acid showed inhibition activity against human ACE2 receptor with a glide score MM-GBSA values of −8.799 and −44.248.

MD simulations offer a plethora of energetic data on protein and ligand binding, as well as a wealth of structural figures on biological macromolecules. This type of knowledge is crucial for comprehending the structural and functional correlation of the receptor and the basis of protein-ligand association, and also for steering the therapeutic research [51]. During the simulation trajectory, the RMSD of the protein Cα and RMSF of the amino acids, also the ligand-protein H-bonding association, the solvent-accessible surface area (SASA), and the radius of gyration (Rg), were assessed to determine the steadiness of the complex structures [52]. The RMSD value is considered to indicate the flexibility and dynamic character of the protein. It showcased the movement of amino acids along with the MD simulation [38]. Thus, a relatively large RMSD value suggested more motion, whereas a relatively low RMSD value of protein showed less movement. The RMSD results suggested that the RMSD value of the main protease_Lovastatin backbone was identical to those of the reference complexes main protease_Molnupiravir and main protease_ Paxlovid. The ligands Lovastatin and Paxlovid remained steady, with two short peaks. As a result, the protein might remain stable during the simulation, after the attachment of the Lovastatin molecule. A detailed investigation of the RMSF demonstrated the specific fluctuation of amino acids in the catalytic and noncatalytic areas of the protein-ligand complexes. The RMSF value confirmed that the attachment of Lovastatin to the receptor might not increase flexibility. The Rg values display the compactness of the protein with folding and unfolding nature by the thermodynamic concept. The interaction of the ligand Lovastatin did not modify the structure of the protein. Hydrogen bonds are another vital determinant of protein-ligand stability. Protein-ligand interaction is stronger with more hydrogen bonds. When compared to the reference complexes, the main protease_Lovastatin complex had a similar amount of hydrogen bonds, indicating a stable protein-ligand construct. The unfolding of the protein during the denaturation process exposes nonpolar hydrophobic interactions to the aqueous system. As a result, the protein's structure is disrupted. The SASA value computing determines the fluctuation in the accessibility of protein to solvent [57]. The SASA analysis revealed a similar tendency, with both main protease_Lovastatin and main protease_Molnupiravir complexes exhibiting significant similarities throughout the simulation.

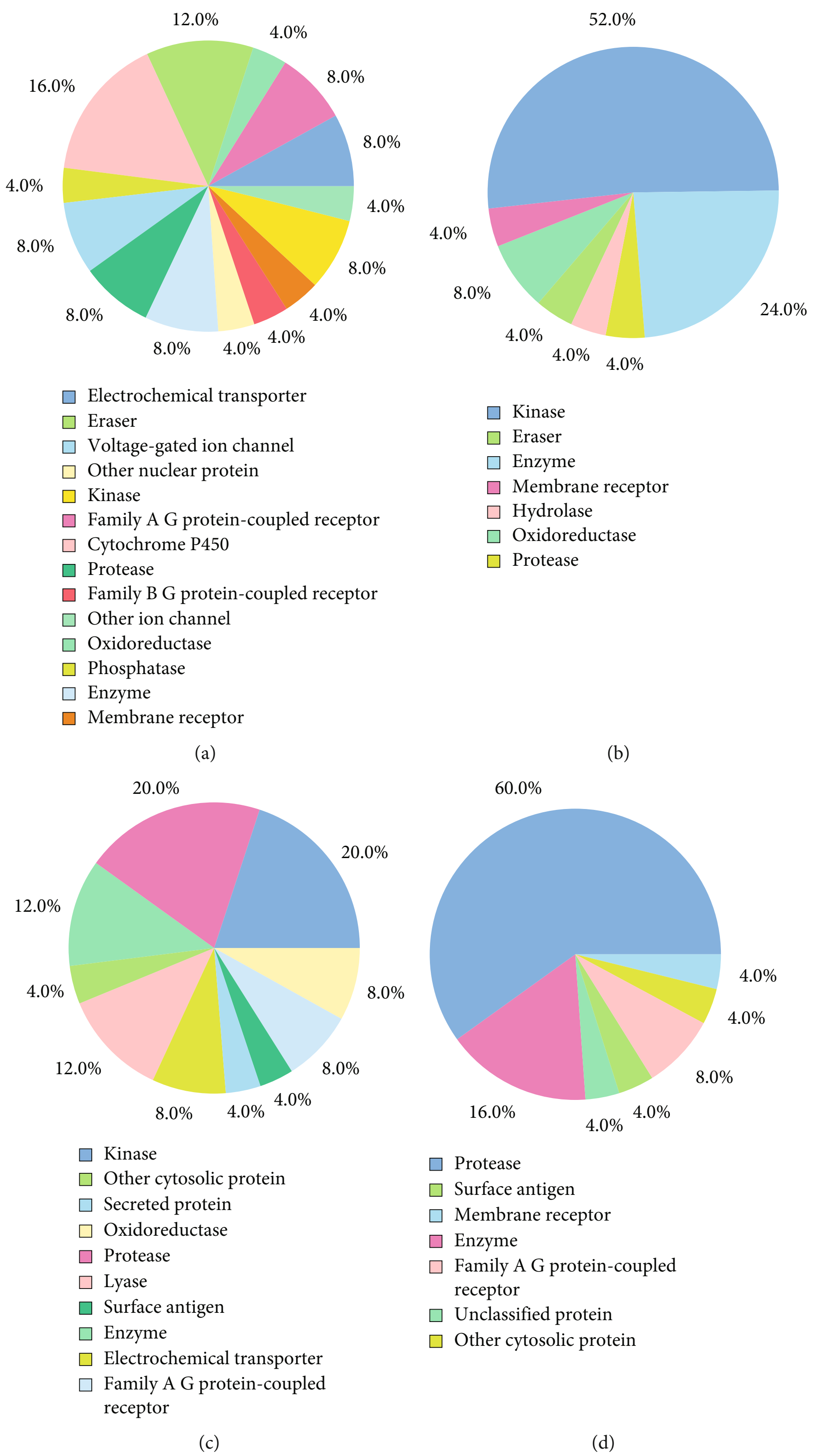


Figure 19: Predicted top 25 classes of *H. sapiens* molecular targets for (a) Lovastatin, (b) Sulfuretin, (c) Grayanoside A, and (d) Paxlovid.

In the context of Nsp3, the ligand Sulfuretin did not produce conformational changes to the protein. Firstly, the RMSD value revealed that Nsp3_Sulfuretin was consistently stable compared to the reference complexes. Throughout the simulation, the Sulfuretin molecule remained relatively stable. Upon binding of the Sulfuretin molecule, the Nsp3_Sulfuretin complex displayed lesser fluctuation in comparison to the reference complexes. According to the Rg value of the Nsp3_Sulfuretin complex, it remained steady during the simulation timeframe, suggesting the compactness of the protein following inhibitor binding. Nsp3_Sulfuretin complex displayed a similar amount of hydrogen interactions as the reference complexes demonstrating excellent protein-ligand stability. Similarly, the SASA value revealed that the Nsp3_Sulfuretin complex remained unchanged throughout the simulation, supporting earlier findings. The simulation results for human ACE2 protein showed that the binding of the Grayanoside A molecule caused a small consequence on the structure of human ACE2 protein. The RMSD graph of protein-ligand complexes and ligands implied that the ligand (CHEMBL1909923) might not destabilize the protein. The RMSF results revealed that there was a similar fluctuation, suggesting the identical nature of the binding of the three ligands (CHEMBL1909923, Molnupiravir, and Paxlovid). The plots of Rg, hydrogen bond, and SASA value also confirmed the previous viewpoint, indicating that the Grayanoside A molecule's attachment did not impair the stability of human ACE2 protein. In case of Desmond, we found similar results that further validate our findings. The RMSD values suggested that Lovastatin, Sulfuretin, and Grayanoside A were firmly bound to the proteins. The RMSF plots implied that the main protease (3CLpro), Nsp3, and human ACE2 receptor were structurally stable while bound with respective ligands. The protein-ligand attachment maps continuously showed that the proposed ligands-maintained contact with active site residues. Lastly, the postsimulation MM-GBSA results stated that Lovastatin, Sulfuretin, and Grayanoside A had a higher free-binding affinity for their respective proteins.

Previous structure-based computational research yielded similar findings, demanding wet-lab investigation. According to study lead by Gurung et al., bonducellpin D was found as a potential inhibitor for SARS-CoV-2 3CLpro protein [58]. In another study, Eissa et al. identified vidarabine as prospective antiviral agent for SARS-Cov-2 nonstructural protein-10 [59]. Ottavia Spiga et al. found simeprevir and lumacaftor the most potent inhibitors of Spike protein on the basis of their computational findings [60]. Kusumaningsih et al. found luteolin and naringenin as the probable drug candidates for main protease of SARS-CoV-2 [61]. Lovastatin, Sulfuretin, and Grayanoside A have been reported as antiviral agents [62–64]. Our structure-based strategy again showed antiviral activity of these small substances against SARS-CoV-2 critical protein targets. However, these compounds should be examined further in the pharmaceutical research facility to evaluate their potency, inhibitory power, and toxicity against their respective targets. While there is no denying the enormous success of drug repurposing, the in silico approach is not without its limitations. In a similar fashion, one disadvantage of molecular docking is the lack of proper scoring functions and algorithms, which may compromise molecular screening. Another challenge that researchers face is the difficulty in selecting the most effective target combinations due to the absence of quantifiable data for assessing network dynamics, as well as the inability to construct the molecular network of the disease [18, 65]. Apart from those certain constraints due to data reliability, biasness and irregularities in the available current data, our study shows a comparison between established compounds and screened compounds using several bioinformatics tools and introduces in silico models that can swiftly present a summary of prospective therapeutic options economically and expediently for a microorganism such as SARS-CoV-2, which is constantly mutating and without any established therapy.

5. Conclusion

Repurposing drug-like phytochemicals is a secure means of building new therapeutics, with the main benefit of lowering the cost and duration of preclinical trials for novel candidates. The COVID-19 infectious disease induced by the SARS-CoV-2 outbreak has caused a worldwide medical catastrophe and finding a suitable cure for the virus continues to be a primary concern. The findings of our work indicated that using a structure-based strategy such as molecular docking and MD simulations, novel therapeutic candidates may be developed that selectively address the nonstructural protein 3, the main protease from SARS-CoV-2, and the human ACE2 protein. A preliminary screening of 1163 small phytochemicals combining drug-likeness and ADMET characteristics resulted in the identification of 149 of them. The degree of binding interaction and energy between the filtered compounds and the main protease, nonstructural protein 3, and human ACE2 receptor was estimated utilizing the docking procedure on the AutoDock vina and Schrodinger Suites. Compounds Lovastatin, Sulfuretin, and Grayanoside A outperformed Molnupiravir and Paxlovid in terms of binding score and hydrogen bond numbers against the main protease, Nsp3, and human ACE2 receptor, respectively. Eventually, 100 ns MD simulation studies of 3CLpro_ligand, Nsp3_ligand, Grayanoside A_ ligand complexes were completed to evaluate and improve our proposed design. This investigation is aimed at determining the promising inhibitors and devise protocols for continual improvement of COVID-19 medications. To summarize, all the repurposed compounds suggested here may provide a holistic understanding of structure-based drug development for SARS-CoV-2 given that they continue to remain potent in further drug development processes.

Data Availability

Availability of additional data (supplementary files) will be provided on request.

Conflicts of Interest

There are no conflicts of interest to declare.

Acknowledgments

The research was carried out on a self-funded basis.

Supplementary Materials

Supplementary 1. Supplementary Table 1: drug-like properties of all the selected phytochemicals.

Supplementary 2. Supplementary Table 2: ADMET properties of the selected phytochemicals.

Supplementary 3. Supplementary Table 3a: binding affinity of the main protease (PDB ID: 7NT3) and the phytochemicals and control drugs. Supplementary Table 3b: binding affinity of the NSP3 (PDB ID: 7KQP) and the phytochemicals and control drugs. Supplementary Table 3c: binding affinity of human ACE2 receptor(PDB ID: 1R4L) and phytochemicals.

Supplementary 4. Supplementary Table 4a: XP Gscore and MM-GBSA values between the main protease (PDB ID: 7NT3) and phytochemicals and control drugs. Supplementary Table 4b: XP Gscore and MM-GBSA values between the NSP3 (PDB ID: 7KQP) and the phytochemicals and control drugs. Supplementary Table 4c: XP Gscore and MM-GBSA values between the human ACE2 receptor (PDB ID: 1R4L) and phytochemicals and control drugs.

Supplementary 5. Supplementary Table 5a: coordinates of geometry-optimized ligand (Lovastatin) after molecular docking study. Supplementary Table 5b: coordinates of geometry-optimized ligand (Sulfuretin) after molecular docking study. Supplementary Table 5c: coordinates of geometry-optimized ligand (Grayanoside A) after molecular docking study.

Supplementary 6. Supplementary Table 6a: predicted biological activities of Lovastatin. Supplementary table 6b: predicted biological activities of Sulfuretin. Supplementary Table 6c: predicted biological activities of Grayanoside A.

References

[1] M. M. A. K. Shawan, S. K. Halder, and M. A. Hasan, "Luteolin and abyssinone II as potential inhibitors of SARS-CoV-2: an in silico molecular modeling approach in battling the COVID-19 outbreak," *Bulletin of the National Research Centre*, vol. 45, no. 1, p. 27, 2021.

[2] "WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard with Vaccination Data," https://covid19.who.int/.

[3] J. Liu, X. Liao, S. Qian et al., "Community transmission of severe acute respiratory syndrome coronavirus 2, Shenzhen, China, 2020," *Emerging Infectious Diseases*, vol. 26, pp. 1320–1323, 2020.

[4] A. K. Srivastava, A. Kumar, and N. Misra, "On the inhibition of COVID-19 protease by Indian herbal plants: an in silico investigation," 2020, https://arxiv.org/abs/2004.03411.

[5] N. Zhu, D. Zhang, W. Wang et al., "A novel coronavirus from patients with pneumonia in China, 2019," *The New England Journal of Medicine*, vol. 382, no. 8, pp. 727–733, 2020.

[6] N. Petrosillo, G. Viceconte, O. Ergonul, G. Ippolito, and E. Petersen, "COVID-19, SARS and MERS: are they closely related?," *Clinical Microbiology and Infection*, vol. 26, no. 6, pp. 729–734, 2020.

[7] I. Khan, Z. Ahmed, A. Sarwar, A. Jamil, and F. Anwer, "The potential vaccine component for COVID-19: a comprehensive review of global vaccine development efforts," *Cureus*, vol. 12, article e8871, 2020.

[8] A. Malik, M. Kohli, N. A. Jacob et al., "*In silico* screening of phytochemical compounds and FDA drugs as potential inhibitors for NSP16/10 5' methyl transferase activity," *Journal of Biomolecular Structure & Dynamics*, vol. 41, no. 1, pp. 221–233, 2023.

[9] J. Lan, J. Ge, J. Yu et al., "Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor," *Nature*, vol. 581, no. 7807, pp. 215–220, 2020.

[10] P. V'kovski, A. Kratzel, S. Steiner, H. Stalder, and V. Thiel, "Coronavirus biology and replication: implications for SARS-CoV-2," *Nature Reviews. Microbiology*, vol. 19, pp. 155–170, 2020.

[11] D. R. Sherin, N. Sharanya, and T. K. Manojkumar, "Potential drug leads for SARS-CoV2 from phytochemicals of Aerva lanata: a machine learning approach," *Virus*, vol. 32, no. 4, pp. 757–765, 2021.

[12] I. Muhammad, N. Rahman, Gul-E-Nayab et al., "Screening of potent phytochemical inhibitors against SARS-CoV-2 protease and its two Asian mutants," *Computers in Biology and Medicine*, vol. 133, article 104362, 2021.

[13] W. Yu and A. D. MacKerell, "Computer-aided drug design methods," *Methods in Molecular Biology*, vol. 1520, pp. 85–106, 2017.

[14] A. Tiwari and S. Singh, "Computational approaches in drug designing," in *Bioinformatics*, pp. 207–217, Academic Press, 2022.

[15] Q. M. S. Jamal, V. Ahmad, A. H. Alharbi et al., "Therapeutic development by repurposing drugs targeting SARS-CoV-2 spike protein interactions by simulation studies," *Saudi Journal of Biological Sciences*, vol. 28, no. 8, pp. 4560–4568, 2021.

[16] F. Kabinger, C. Stiller, J. Schmitzová et al., "Mechanism of molnupiravir-induced SARS-CoV-2 mutagenesis," *Nature Structural & Molecular Biology*, vol. 28, no. 9, pp. 740–746, 2021.

[17] R. Islam, M. R. Parves, A. S. Paul et al., "A molecular modeling approach to identify effective antiviral phytochemicals against the main protease of SARS-CoV-2," *Journal of Biomolecular Structure & Dynamics*, vol. 39, no. 9, pp. 3213–3224, 2021.

[18] S. K. Halder and F. Elma, "In silico identification of novel chemical compounds with antituberculosis activity for the inhibition of InhA and EthR proteins from mycobacterium tuberculosis," *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, vol. 24, article 100246, 2021.

[19] W. A. Fischer, J. J. Eron, W. Holman et al., "A phase 2a clinical trial of molnupiravir in patients with COVID-19 shows accelerated SARS-CoV-2 RNA clearance and elimination of infectious virus," *Science Translational Medicine*, vol. 14, no. 628, 2022.

[20] A. Daina, O. Michielin, and V. Zoete, "SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules," *Scientific Reports*, vol. 7, no. 1, article 42717, 2017.

[21] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced Drug Delivery Reviews*, vol. 46, no. 1-3, pp. 3–26, 2001.

[22] A. K. Ghose, V. N. Viswanadhan, and J. J. Wendoloski, "A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases," *Journal of Combinatorial Chemistry*, vol. 1, no. 1, pp. 55–68, 1999.

[23] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple, "Molecular properties that influence the oral bioavailability of drug candidates," *Journal of Medicinal Chemistry*, vol. 45, no. 12, pp. 2615–2623, 2002.

[24] I. Muegge, S. L. Heald, and D. Brittelli, "Simple selection criteria for drug-like chemical matter," *Journal of Medicinal Chemistry*, vol. 44, no. 12, pp. 1841–1846, 2001.

[25] D. E. V. Pires, T. L. Blundell, and D. B. Ascher, "pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures," *Journal of Medicinal Chemistry*, vol. 58, no. 9, pp. 4066–4072, 2015.

[26] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform," *Journal of Cheminformatics*, vol. 4, no. 1, p. 17, 2012.

[27] G. M. Morris, R. Huey, and A. J. Olson, "Using AutoDock for ligand-receptor docking," *Current Protocols in Bioinformatics*, vol. 8, p. 14, 2008.

[28] N. Guex and M. C. Peitsch, "SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling," *Electrophoresis*, vol. 18, no. 15, pp. 2714–2723, 1997.

[29] M. Brylinski, "Local alignment of ligand binding sites in proteins for polypharmacology and drug repositioning," *Methods in Molecular Biology*, vol. 1611, pp. 109–122, 2017.

[30] W. Tian, C. Chen, X. Lei, J. Zhao, and J. Liang, "CASTp 3.0: computed atlas of surface topography of proteins," *Nucleic Acids Research*, vol. 46, no. W1, pp. W363–W367, 2018.

[31] A. Uniyal, M. K. Mahapatra, V. Tiwari, R. Sandhir, and R. Kumar, "Targeting SARS-CoV-2 main protease: structure based virtual screening, in silico ADMET studies and molecular dynamics simulation for identification of potential inhibitors," *Journal of Biomolecular Structure & Dynamics*, vol. 40, no. 8, pp. 3609–3625, 2022.

[32] M. I. Choudhary, M. Shaikh, A. tul-Wahab, and A. ur-Rahman, "In silico identification of potential inhibitors of key SARS-CoV-2 3CL hydrolase (Mpro) via molecular docking, MMGBSA predictive binding energy calculations, and molecular dynamics simulation," *PLoS One*, vol. 15, no. 7, article e0235030, 2020.

[33] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, "Molecular docking: a powerful approach for structure-based drug discovery," *Current Computer-Aided Drug Design*, vol. 7, no. 2, pp. 146–157, 2011.

[34] J. J. Sahayarayan, K. S. Rajan, R. Vidhyavathi et al., "*In-silico* protein-ligand docking studies against the estrogen protein of breast cancer using pharmacophore based virtual screening approaches," *Saudi Journal of Biological Sciences*, vol. 28, no. 1, pp. 400–407, 2021.

[35] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[36] A. W. Schüttelkopf and D. M. F. van Aalten, "PRODRG: a tool for high-throughput crystallography of protein-ligand complexes," *Acta Crystallographica. Section D, Biological Crystallography*, vol. 60, no. 8, pp. 1355–1363, 2004.

[37] J. Zielkiewicz, "Structural properties of water: comparison of the SPC, SPCE, TIP4P, and TIP5P models of water," *The Journal of Chemical Physics*, vol. 123, no. 10, article 104501, 2005.

[38] S. K. Halder, M. O. Rafi, E. B. Shahriar et al., "Identification of the most damaging nsSNPs in the human *CFL1* gene and their functional and structural impacts on cofilin-1 protein," *Gene*, vol. 819, article 146206, 2022.

[39] Schrodinger, D. E. S., *Schrödinger Release 2022-3: Desmond Molecular Dynamics System, D*, E. Shaw Research, New York, NY, 2021.

[40] S. K. Halder, M. M. Mim, M. M. H. Alif et al., "Oxa-376 and Oxa-530 variants of β-lactamase: computational study uncovers potential therapeutic targets of Acinetobacter baumannii," *RSC Advances*, vol. 12, no. 37, pp. 24319–24338, 2022.

[41] S. K. Enmozhi, K. Raja, I. Sebastine, and J. Joseph, "Andrographolide as a potential inhibitor of SARS-CoV-2 main protease: an in silico approach," *Journal of Biomolecular Structure & Dynamics*, vol. 39, no. 9, pp. 1–7, 2020.

[42] D. A. Filimonov, A. A. Lagunin, T. A. Gloriozova et al., "Prediction of the biological activity spectra of organic compounds using the pass online web resource," *Chemistry of Heterocyclic Compounds*, vol. 50, no. 3, pp. 444–457, 2014.

[43] I. L. Shytaj, M. Fares, L. Gallucci et al., "The FDA-approved drug Cobicistat synergizes with Remdesivir to inhibit SARS-CoV-2 replication in vitro and decreases viral titers and disease progression in Syrian hamsters," *MBio*, vol. 13, no. 2, article e0370521, 2022.

[44] C. Wu, Y. Liu, Y. Yang et al., "Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods," *Acta Pharmaceutica Sinica B*, vol. 10, no. 5, pp. 766–788, 2020.

[45] M. Schuller, G. J. Correy, S. Gahbauer et al., "Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking," *Science Advances*, vol. 7, no. 16, 2021.

[46] M. P. Gleeson, A. Hersey, D. Montanari, and J. Overington, "Probing the links between *in vitro* potency, ADMET and physicochemical parameters," *Nature Reviews. Drug Discovery*, vol. 10, no. 3, pp. 197–208, 2011.

[47] A. Alex, D. S. Millan, M. Perez, F. Wakenhut, and G. A. Whitlock, "Intramolecular hydrogen bonding to improve membrane permeability and absorption in beyond rule of five chemical space," *MedChemComm*, vol. 2, no. 7, pp. 669–674, 2011.

[48] S. Prasanna and R. Doerksen, "Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR," *Current Medicinal Chemistry*, vol. 16, no. 1, pp. 21–41, 2009.

[49] M. P. Gleeson, "Generation of a set of simple, interpretable ADMET rules of thumb," *Journal of Medicinal Chemistry*, vol. 51, no. 4, pp. 817–834, 2008.

[50] L. Chen, Y. Li, H. Yu, L. Zhang, and T. Hou, "Computational models for predicting substrates or inhibitors of P-glycoprotein," *Drug Discovery Today*, vol. 17, no. 7-8, pp. 343–351, 2012.

[51] U. M. Zanger and M. Schwab, "Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation," *Pharmacology & Therapeutics*, vol. 138, no. 1, pp. 103–141, 2013.

[52] X. Liu, D. Shi, S. Zhou, H. Liu, H. Liu, and X. Yao, "Molecular dynamics simulations and novel drug discovery," *Expert Opinion on Drug Discovery*, vol. 13, no. 1, pp. 23–37, 2018.

[53] O. M. Salo-Ahen, I. Alanko, R. Bhadane et al., "Molecular dynamics simulations in drug discovery and pharmaceutical development," *Processes*, vol. 9, p. 71, 2021.

[54] A. Alzamami, N. A. Alturki, Y. S. Alghamdi et al., "Hemi-Babim and fenoterol as potential inhibitors of MPro and papain-like protease against SARS-CoV-2: an in-silico study," *Medicina*, vol. 58, 2022.

[55] A. Mishra, V. Mulpuru, and N. Mishra, "Identification of SARS-CoV-2 inhibitors through phylogenetics and drug repurposing," *Structural Chemistry*, vol. 33, no. 5, pp. 1789–1797, 2022.

[56] P. Towler, B. Staker, S. G. Prasad et al., "ACE2 X-ray structures reveal a large hinge-bending motion important for inhibitor binding and catalysis," *The Journal of Biological Chemistry*, vol. 279, no. 17, pp. 17996–18007, 2004.

[57] D. Zhang and R. Lazim, "Application of conventional molecular dynamics simulation in evaluating the stability of apomyoglobin in urea solution," *Scientific Reports*, vol. 7, no. 1, article 44651, 2017.

[58] T. L. Gower and B. S. Graham, "Antiviral activity of lovastatin against respiratory syncytial virus in vivo and in vitro," *Antimicrobial Agents and Chemotherapy*, vol. 45, no. 4, pp. 1231–1237, 2001.

[59] M. Kowalczyk, A. Golonko, R. Świsłocka et al., "Drug design strategies for the treatment of viral disease. Plant phenolic compounds and their derivatives," *Frontiers in Pharmacology*, vol. 12, article 709104, 2021.

[60] A. B. Gurung, M. A. Ali, J. Lee, M. A. Farah, and K. M. Al-Anazi, "Unravelling lead antiviral phytochemicals for the inhibition of SARS-CoV-2 M^{pro} enzyme through in silico approach," *Life Sciences*, vol. 255, no. 255, article 117831, 2020.

[61] I. H. Eissa, M. M. Khalifa, E. B. Elkaeed, E. E. Hafez, A. A. Alsfouk, and A. M. Metwaly, "In silico exploration of potential natural inhibitors against SARS-CoV-2 nsp10," *Molecules*, vol. 26, no. 20, p. 6151, 2021.

[62] A. Trezza, D. Iovinelli, A. Santucci, F. Prischi, and O. Spiga, "An integrated drug repurposing strategy for the rapid identification of potential SARS-CoV-2 viral inhibitors," *Scientific Reports*, vol. 10, no. 1, article 13866, 2020.

[63] W. E. Prasetyo, H. Purnomo, M. Sadrini, F. R. Wibowo, M. Firdaus, and T. Kusumaningsih, "Identification of potential bioactive natural compounds from Indonesian medicinal plants against 3-chymotrypsin-like protease (3CLpro) of SARS-CoV-2: molecular docking, ADME/T, molecular dynamic simulations, and DFT analysis," *Journal of Biomolecular Structure and Dynamics*, vol. 19, pp. 1–8, 2022.

[64] C.-H. Li, J.-Y. Zhang, X.-Y. Zhang, S.-H. Li, and J.-M. Gao, "An overview of grayanane diterpenoids and their biological activities from the Ericaceae family in the last seven years," *European Journal of Medicinal Chemistry*, vol. 166, pp. 400–416, 2019.

[65] A. Sacan, S. Ekins, and S. Kortagere, "Applications and limitations of in silico models in drug discovery," *Bioinformatics and Drug Discovery*, vol. 910, pp. 87–124, 2012.

Hindawi
BioMed Research International
Volume 2022, Article ID 9039377, 13 pages
https://doi.org/10.1155/2022/9039377

Research Article

Novel Target Study to Cure Cardiovascular Disease regarding Proprotein Converse Subtilisin/Kexin Type 9

Yingjing Zhao,[1] Weihang Li,[2] Weiye Li,[3] Hong Tao,[1] Yuting Li,[1] Bo Wu,[4] Xinhui Wang,[5] Huasong Zhou,[6] and Bo Gao[2]

[1] *Department of Critical Care Medicine, Nanjing First Hospital, Nanjing Medical University, Nanjing, Jiangsu, China 210006*
[2] *Department of Orthopedic Surgery, Xijing Hospital, Air Force Medical University, Xi'an, China*
[3] *Clinical Medical School, China-Japan Union Hospital of Jilin University, 126 Xian Street Changchun 130033, Changchun, Jilin, China*
[4] *Department of Orthopaedics, The First Hospital of Jilin University, Changchun, Jilin, China*
[5] *Department of Oncology, First People's Hospital of Xinxiang & The Fifth Affiliated Hospital of Xinxiang Medical College, Street Yiheng, 63 Xinxiang, China*
[6] *Department of General Surgery, Xi'an Hospital of Traditional Chinese Medicine, Xi'an, Shaanxi, China*

Correspondence should be addressed to Xinhui Wang; wxh_xxyxy@126.com, Huasong Zhou; windzs115420@163.com, and Bo Gao; gaobofmmu@hotmail.com

Received 5 July 2022; Revised 29 August 2022; Accepted 12 September 2022; Published 11 October 2022

Academic Editor: Jian Zhang

Objective. This study is aimed at screening the potential ideal lead compounds from natural drug library (ZINC database), which had potential inhibition effects against proprotein converse subtilisin/kexin type 9 (PCSK9), and contributing to enrich the practical basis of PCSK9 inhibitor screening. *Methods*. A series of computer-aided virtual screening techniques were used to identify potential inhibitors of PCSK9. Structure-based virtual screening by LibDock was carried out to calculate the LibDock scores, followed by ADME (absorption, distribution, metabolism, and excretion) and toxicity predictions. Molecule docking was next employed to demonstrate the binding affinity and mechanism between the candidate ligands and PCSK9 macromolecule. Finally, molecular dynamics simulation was performed to evaluate the stability of ligand-PCSK9 complex under natural circumstance. *Results*. Two novel natural compounds ZINC000004099069 and ZINC000014952116 from the ZINC database were found to bind with PCSK9 with a higher binging affinity together with more favorable interaction energy. Also, they were predicted to be non-CYP2D6 inhibitors, together with low rodent carcinogenicity and AMES mutagenicity as well as hepatotoxicity. Molecular dynamics simulation analysis demonstrated that these two complex ZINC000004099069- and ZINC000014952116-PCSK9 had more favorable potential energy compared to the reference ligand, which could exist stably whether in vivo or in vitro. *Conclusion*. This study elucidated that ZINC000004099069 and ZINC000014952116 were finally screened as safe and potential drug candidates, which may have great significance in the development of PCSK9 inhibitor development.

1. Introduction

Atherosclerosis is a chronic inflammatory disease with large/medium size of arteries, characterized by the detention of modified lipoproteins in the arterial wall, which could lead to ischemic heart disease and strokes as well as peripheral vascular disorders, collectively named as cardiovascular disease (CVD) [1, 2]. Lipoprotein is involved in the formation of atherosclerosis and plays a pivotal role in plaque rupture, which is a common pathophysiological indicator of acute ischemic syndrome [3], among which low-density lipoprotein cholesterol (LDLc) is a type of the lipoprotein; lowering LDLc could decrease the risk of CVD [4, 5], such as stroke, which is the fifth leading cause of death in 2017 in the United States [6]; atherosclerosis; and myocardial infarction [7, 8]. Human PCSK9 gene, namely, proprotein converse subtilisin/kexin type 9, is mainly synthesized and secreted by the liver and is one of the key modulators of LDLc; besides, PCSK9 is also found to be closely connected with series of pathophysiological processes, like brain development, platelet activation, intestinal physiology, pancreas, and adipose tissue as well as neoplasms [2], suggesting that PCSK9 is the key regulatory target among different diseases. Existing genetic and interventional researches have fully reported that reducing the levels of PCSK9 corresponds to CVD benefits [2].

LDLc is eliminated through LDL-R recycling [9], while this process could be altered negatively by PCSK9 through degrading LDL-R. When PCSK9 capture the LDL-R/LDLc complex, it could further combine with the complex closely and then form a novel complex PCSK9/LDL-R/LDLc, which is then internalized through the cell membrane and sent to the lysosome for degradation, resulting in the degradation of LDL-R, thus preventing LDL-R from recycling to the cell membrane [10, 11]. Consequently, decreasing the degradation of LDL-R by PCSK9 inhibitor could help LDLc cleaning and eventually reduce the risk of atherosclerosis [12, 13]. These findings implied that PCSK9 inhibition could be a potential and effective therapeutic target to cure or prevent CVD in individuals with high levels of PCSK9.

Recent researches showed that the current primary pharmacological inhibitors of PCSK9 such as evolocumab and alirocumab were monoclonal antibodies, which were potent in the LDL-lowering process, together with good tolerance by patients [14]. However, monoclonal antibodies still have some disadvantages like the high cost, injection site adverse reaction, and no oral administration approach. The expensive as well as inconvenient situation makes it hard for patients to receive this therapeutic approach widely. Current methods in lowering LDLc level include inhibiting the function and influencing the synthesis as well as processing of PCSK9 [15], interfering the PCSK9/LDLR protein-protein interaction, and silencing PCSK9 gene expression by genetic alteration such as siRNA [16]. However, most of these approaches did not emerge promising effects because of limitations, such as potential off-target mutagenesis for disrupting PCSK9 by gene genome editing and instability in plasma parenteral administration for small peptide [13, 16].

Nature products and their derivatives play a crucial role in today's pharmacologic market, small molecules are pivotal aspect if not the first means to tackle an emergent or unpredictable diseases, and they have still made a great contribution to medication design and improvement [2, 17, 18]. Novel nature inhibitors targeting PCSK9 may benefit from these aspects: newly aromatic compounds from the fruiting body of Sparassis crispa, berberine, and inclisiran are reported to be potent PCSK9 inhibitors, which can influence PCSK9 mRNA expression [15, 19–21]. Imidazole-based minimalist peptidomimetic and truncated LDL-R EGF-A-domain peptides can disrupt the PCSK9/LDLR protein-protein interaction [22]. However, a suitable novel nature inhibitor targeting PCSK9 was hard to discover without comprehensive and professional evaluation, not to mention further in vivo studies. Currently, only small amount of PCSK9 inhibitor researches were found to be relatively mature, such as polydatin and tetrahydroxydiphenylethylene-2-O-glucoside [23, 24]. Therefore, there still needs more study to screen the potential PCSK9 inhibitors as well as analyze possible mechanism of the interactions.

Structural biology study is an effective way on the basis of high-throughput techniques, to screen nature compounds targeting specialized protein molecules from huge of ligands, which avoid the large amount of manpower, materials, and financial resources required for traditional drug screening (manual drug addition experiments). Current computational simulation study on PCSK9 inhibitors include Exploring Key Orientations (EKO) and computational GOLD algorithm analysis [25, 26]. This study performed different chemical molecule database and computational methods to discover potential candidate compounds, aiming to screen potential lead compounds with well binding affinity and effective functions as well as existence of stability under natural environment. A set of virtual screening, molecular docking, toxicity prediction, and ADME model was fully performed to screen the promising compounds targeting PCSK9; then, ligand binding analysis and molecular dynamics simulation were used to understand the mechanism further. A reported inhibitor of PCSK9 was chosen as reference to make comprehensive evaluation for novel ligands and existing inhibitors of PCSK9 [27].

2. Results

2.1. High-Throughput Screening of Natural Product Database against PCSK9. Chemical structure of PCSK9 is displayed in Figures 1(a) and 1(b), the existed ligand-binding pocket was an essential active regulatory site of PCSK9, and small ligands binding to this region could change the conformation of the protein and thus inactivate the activity of PCSK9, so the initial ligand from PCSK9 complex was extracted and the region was set as the binding sphere. Totally, 17776 purchasable-natural-named products were obtained from ZINC repository for research. With high-throughput screening, each of these ligands was put into the binding sphere to bind with the protein, and finally, 13430 compounds were found to bind eligibly with PCSK9 through the screening algorithm; among those, 2081 ligands had higher LibDock scores than the reference ligand (LibDock score: 124.227). The top 20 compounds with the highest LibDock scores are listed in Table 1, together with these chemical structures of these potential lead compounds (Figure 2).

2.2. Pharmacological Properties and Toxicity Prediction. Pharmacological properties of these ligands were fully evaluated through ADME (absorption, distribution, metabolism, and excretion) algorithm; these indicators include solubility level, brain/blood barrier (BBB) level, cytochrome P450 2D6 (CYP2D6) prediction, hepatotoxicity, absorption level, and toxicity properties. As shown in Table 2, all compounds could pass through the BBB indicated by BBB level (score: 4); solubility level showed that all compounds were soluble in water except ZINC000008220036; three compounds were predicted to be inhibitors of CYP2D6, which had an important role in drug metabolism; and seven compounds were

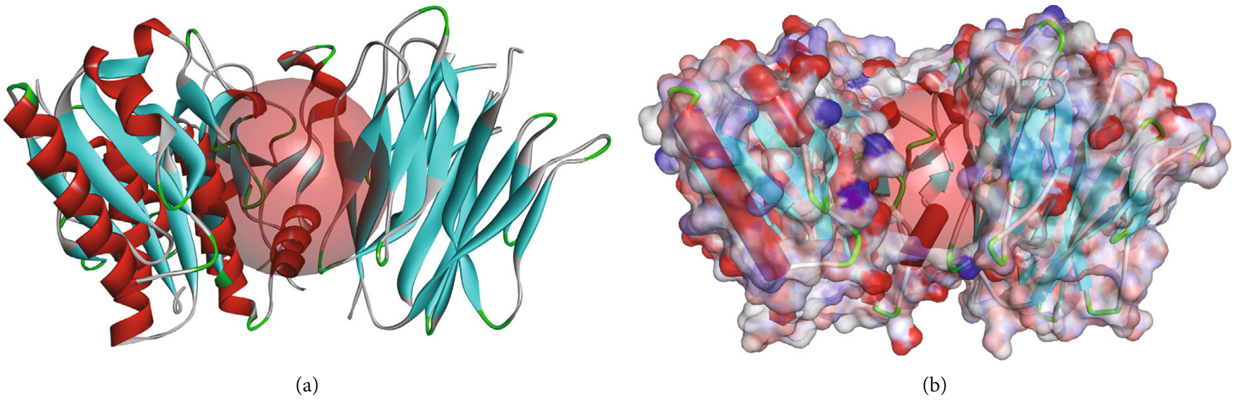


Figure 1: Molecular structure of proprotein converse subtilisin/kexin type 9 (PCSK9): (a) initial molecular structure; (b) surface of binding area was added, blue indicated positive charge, and red indicated negative charge.

Table 1: Top 20 ranked compounds with higher LibDock scores than PCSK9.

Number	ZINC ID	Compounds	LibDock score
1	ZINC000062238222	5-Methyltetrahydropteroyltri-L-glutamate	239.613
2	ZINC000085544839	Thf-polyglutamate	238.17
3	ZINC000095620524	Tetra-*trans*-P-coumaroylspermine	226.594
4	ZINC000004099069	S-(pga1)-glutathione	206.719
5	ZINC000004654845	3-Hexaprenyl-4-hydroxybenzoate	205.355
6	ZINC000085541163	(+-)-Grossamide	204.523
7	ZINC000008552069	Thf-L-glutamate	202.145
8	ZINC000013513540	1,14-Bis(dihydrocaffeoyl)spermine	201.87
9	ZINC000004228293	5-Formiminotetrahydrofolate	201.743
10	ZINC000009212428	Leucal	200.227
11	ZINC000014952116	Enkephalin	197.698
12	ZINC000014712793	Kukoamine B	196.511
13	ZINC000011616636	Hibon	196.025
14	ZINC000008220036	2-Hexaprenyl-3-methyl-6-methoxy-1,4 benzoquinone	195.582
15	ZINC000004096653	Dhhpba	195.52
16	ZINC000012494317	Isodesmosin	192.893
17	ZINC000014951658	Endomorphin 1	192.844
18	ZINC000085826835	(+-)-Grossamide	192.709
19	ZINC000042805482	Grossamide	192.559
20	ZINC000004097774	Lithospermic acid	192.238

predicted to be hepatotoxic; seventeen compounds were found to have a higher absorption level compared to the rest of three compounds; as for the reference ligand, it was predicted to be toxic to the liver and non-CYP2D6 inhibitors.

Toxicity of candidate drugs also needs to be considered when screening potential compounds, through TOPKAT module (Table 3); indicators like AMES mutagenicity (AMES), developmental toxicity potential (DTP), and rodent carcinogenicity (based on the US. National Toxicology Program dataset) were included to ensure the safety of these potential drugs. Results revealed almost all drugs had developmental toxicity potential except ZINC000008220036; the reference ligand had high probability of DTP and AMES mutagenicity.

2.3. Ligand Binding Analysis. To further understand the mechanism of the interaction between these candidate compounds with PCSK9, CDOCKER modules were conducted to make a precise docking algorithm, which could generate more accurate chemical bonds between ligand and protein and caused more running time. After the reference ligand redocking into the binding region of PCSK9, RMSD between initial ligand and docked posture was calculated as 0.92 Å, proving that the docking program applied in this study was highly reliable. Then, CDOCKER interaction energy was calculated to verify the binding affinity of ligands and PCSK9. CDOCKER module provided a 3D structure of the interaction between compounds and PCSK9, and CDOCKER interaction energy showed the affinity of

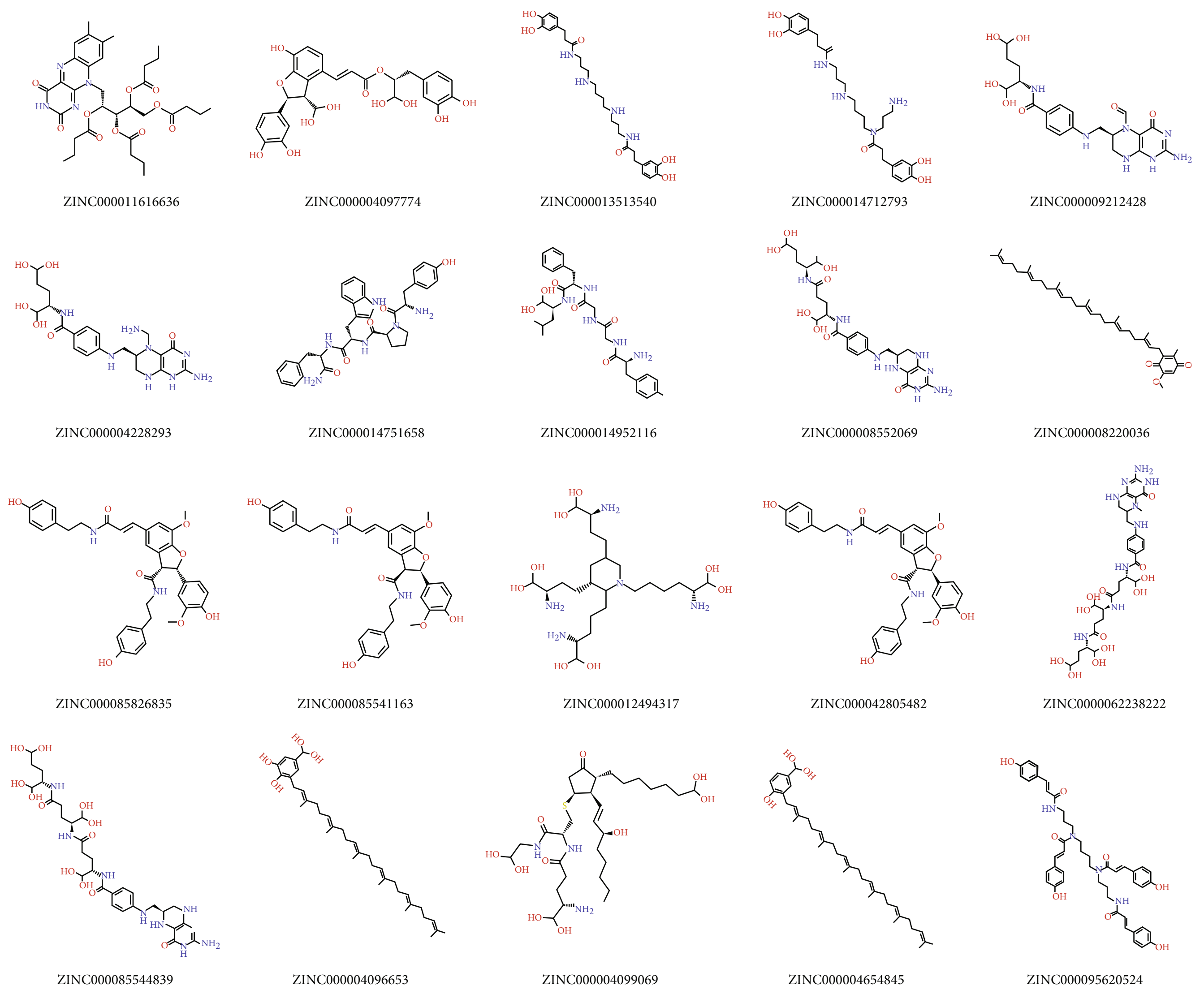


Figure 2: The chemical structures of the top 20 compounds.

potential compounds with PCSK9. The CDOCKER interaction energy of ZINC000004099069 with PCSK9 is -87.8609 Kcal/mol, lower than the CDOCKER interaction energy of ZINC000014952116 with PCSK9, -65.9632 Kcal/mol, which meant that the former complex could bind with PCSK9 better (Table 4). The hydrogen bonds and hydrophobic interactions formed by PCSK9, and these two compounds were visualized (Figures 3(a)–3(c)) and analyzed (Figures 4(a)–4(c)). Table 5 displays that ZINC000004099069 formed 22 pairs of hydrogen bonds with PCSK9, and ZINC000014952116 formed 18 pairs of hydrogen bonds with PCSK9. The interaction between ZINC000004099069 and PCSK contains 1 hydrophobic interaction, and the interaction between PCSK and the other promising chemical molecular includes 4 hydrophobic interactions, as shown in Table 6.

2.4. Molecular Dynamics Simulation. Molecular dynamics simulation had been performed to further evaluate the stabilities of PCSK9-ligand complexes under natural situation. The initial conformation of these complexes was obtained from CDOCKER module, an accurate molecular docking program. RMSD curves as well as potential energy of ZINC000004099069- and ZINC000014952116-PCSK9 complexes are shown in Figure 5, the trajectory of each complex got stable gradually at about 20 ps, and then, they both went to equilibrium. It was convincing that through these MD simulation procedures, hydrogen bonds and π-related chemical bonds between ligands and PCSK9 could enhance the interactions within the complex and thereby contributed the stabilities of complexes. MD simulation results suggested both of the two compounds could interact stably with PCSK9, and ligand-PCSK9 complexes could exist steadily under natural situation. Considering all results above, ZINC000004099069 and ZINC000014952116 were finally selected as potential lead compounds with less rodent carcinogenicity, hepatotoxicity, AMES mutagenicity, and good solubility and intestinal absorption level; they were not toxic to the liver and did not behave as CYP2D6 inhibitors. Additionally, in terms of molecular dynamics simulation, the

Table 2: ADME (adsorption, distribution, metabolism, and excretion) properties of compounds.

Number	ZINC ID	Solubility level[a]	BBB level[b]	CYP2D6 prediction[c]	Hepatotoxicity[d]	Absorption level[e]	PPB prediction[f]
1	ZINC000062238222	3	4	0	1	3	0
2	ZINC000085544839	3	4	0	1	3	0
3	ZINC000095620524	4	4	0	1	3	0
4	ZINC000004099069	3	4	0	0	3	0
5	ZINC000004654845	1	4	1	0	3	1
6	ZINC000085541163	2	4	0	0	2	0
7	ZINC000008552069	4	4	0	1	3	0
8	ZINC000013513540	4	4	0	1	3	0
9	ZINC000004228293	4	4	0	1	3	0
10	ZINC000009212428	4	4	0	1	3	0
11	ZINC000014952116	4	4	0	0	3	0
12	ZINC000014712793	4	4	0	0	3	0
13	ZINC000011616636	2	4	0	0	3	0
14	ZINC000008220036	0	4	1	0	3	1
15	ZINC000004096653	1	4	0	0	3	1
16	ZINC000012494317	1	4	1	0	3	0
17	ZINC000014951658	3	4	0	0	3	0
18	ZINC000085826835	2	4	0	0	2	0
19	ZINC000042805482	2	4	0	0	2	0
20	ZINC000004097774	2	4	0	0	3	0
21	6U3X	/	/	0	1	/	0

[a] Aqueous solubility level: 0 (extremely low); 1 (very low, but possible); 2 (low); and 3 (good); [b] blood-brain barrier level: 0 (very high penetrant); 1 (high); 2 (medium); 3 (low); and 4 (undefined); [c] cytochrome P450 2D6 level: 0 (noninhibitor) and 1 (inhibitor); [d] hepatotoxicity: 0 (nontoxic) and 1 (toxic); [e] human intestinal absorption level: 0 (good); 1 (moderate); 2 (poor); and 3 (very poor); [f] plasma protein binding: 0 (binding is <90%); 1 (binding is >90%); and 2 (binding is >95%).

formed complex could also behave stable performance under natural situation.

2.5. Validation of the Effects of the Candidate Compounds. This study recruited the top 20 compounds with the highest LibDock scores conducted by high-throughput screening and then put them into a more reliable algorithm "analyze ligand poses,", to test the binding affinity of ZINC000004099069 and ZINC000014952116 compounds. After calculating the residues and PCSK9 receptor, there displayed favorable and unfavorable count and hydrophobic count as well as hydrogen count, respectively. As shown in Figure 6, results validated that ZINC000004099069 and ZINC000014952116 compounds had the most active residues and the least unfavorable residues relatively, proving the reliability of these two candidate compounds.

2.6. Possible Pharmacophore Modification of the Lead Compounds. After screening the potential lead compounds of PCSK9, this study further analyzed the editable skeleton of these two compounds through pharmacophore properties, to observe the possible modification site. As shown in Figures 7(a) and 7(b), results visualized that on the skeleton of these compounds, there were 106 editable features in ZINC000004099069 and 49 editable features in ZINC000014952116, among which, ZINC000004099069 had 40 hydrogen bond acceptors, 61 hydrogen bond donors, and 5 positive ionizable; ZINC000014952116 possessed 20 hydrogen bond acceptors, 19 hydrogen bond donors, 5 hydrophobic centers, and 1 positive ionizable as well as 4 ring aromatics, which could be further improved targeting these editable sites.

3. Discussion

PCSK9, mainly synthesized by the hepatocytes, is a pivotal regulator of LDL-R, which could reverse the clearance of LDLc [12, 13]. Many hypercholesterolemia patients possess a high level of PCSK9. Therefore, targeting the function of PCSK9 is a promising direction for lowering LDLc in order to further cure CVD. Up to now, there have been some targeted drugs of PCSK9, such as monoclonal antibodies (evolocumab and alirocumab) and nature inhibitors (polydatin and tetrahydroxydiphenylethylene-2-O-glucoside) [14, 23, 24]. The high cost of using monoclonal antibodies and the slow progression in putting natural inhibitors into clinical trial require more candidate natural drug research in this field. However, it is worth noting that most natural products generally lead to low aqueous solubility and poor stability as well as bioavailability like less cellular absorption and low intestinal absorption and also high molecular weight due to their physiochemical properties, which may block the development of natural drug screening [28, 29]; these aspects all prompt researchers to fully evaluate each compound when

Table 3: Toxicities of compounds.

Number	ZINC ID	Rat NTP[a]		Mouse NTP[a]		AMES[b]	DTP[c]
		Male	Female	Male	Female		
1	ZINC000062238222	0.969	0.000	0.000	0.000	0.989	1.000
2	ZINC000085544839	0.964	0.000	0.080	0.000	0.999	1.000
3	ZINC000095620524	0.999	1.000	0.000	1.000	0.000	1.000
4	ZINC000004099069	0.000	0.000	0.000	0.000	0.002	0.864
5	ZINC000004654845	0.000	1.000	1.000	0.000	1.000	1.000
6	ZINC000085541163	0.998	1.000	1.000	0.186	0.000	1.000
7	ZINC000008552069	0.997	0.000	0.000	0.015	1.000	1.000
8	ZINC000013513540	0.840	0.020	0.000	0.139	0.000	1.000
9	ZINC000004228293	1.000	0.000	0.000	0.165	0.000	1.000
10	ZINC000009212428	1.000	0.000	0.000	1.000	0.356	1.000
11	ZINC000014952116	0.000	0.000	0.000	0.001	0.000	0.928
12	ZINC000014712793	0.632	1.000	0.000	0.640	0.000	1.000
13	ZINC000011616636	1.000	1.000	1.000	0.000	1.000	1.000
14	ZINC000008220036	0.000	1.000	1.000	0.000	0.064	0.000
15	ZINC000004096653	0.000	1.000	1.000	0.000	1.000	1.000
16	ZINC000012494317	0.000	0.000	0.000	0.222	0.000	1.000
17	ZINC000014951658	0.000	1.000	0.000	1.000	0.000	1.000
18	ZINC000085826835	0.998	1.000	1.000	0.186	0.000	1.000
19	ZINC000042805482	0.998	1.000	1.000	0.186	0.000	1.000
20	ZINC000004097774	0.053	0.987	0.916	0.000	0.002	1.000
21	6U3X	1.000	0.139	0.012	0.002	1.000	1.000

[a] 0 (noncarcinogen) and 1 (carcinogen); [b] 0 (nonmutagen) and 1 (mutagen); [c] 0 (nontoxic) and 1 (toxic).

Table 4: CDOCKER interaction energy of compounds with PCSK9.

ZINC ID	CDOCKER interaction energy (Kcal/mol)
ZINC000004099069	-87.8609
ZINC000014952116	-65.9632

drug screening to discover the best effective candidate drugs. The high-throughput method used in this study could reduce the cost of medicine research and development, such as manpower and materials. Some studies have suggested a high priority of using structural biology method [2, 18, 30]. To the best knowledge, the researches using this analytical method to screen PCSK9 have not been reported so far; thus, this study could provide a novel insight for exploring targeted drug therapy of PCSK9 and contribute in this field.

In this study, a total of 13430 compounds of purchasable-natural-named products taken from the ZINC database were found to bind with PCSK9 eligibly. The top 20 of these compounds based on LibDock score were screened firstly and selected for further ADME and toxicity prediction. Through results, LibDock scores represented their binding affinity with protein PCSK9; compounds with higher LibDock score indicated a better energy optimization and more stable structure in its complex. This study chose the top 20 compounds with the best LibDock score for first screening, which were pooled for the following study.

ADME and toxicity predictions were employed for compounds to access their pharmacological properties. The results demonstrated that ZINC000004099069 and ZINC000014952116 were ideal compounds for high solubility, strong plasma protein binding affinity, and high absorption levels. At the same time, they were predicted to be non-CYP2D6 inhibitors, with low rodent carcinogenicity and AMES mutagenicity as well as hepatotoxicity. The high solubility and intestinal absorption level could promote the drug dissolution and absorption process, which could benefit from oral medication. Because of their noninhibition to CYP2D6, they were not easy to be accumulated in the liver. Additionally, these two compounds were assessed to have less AMES mutagenicity and rodent carcinogenicity, which presented their preferable safety characteristic. However, the two compounds still had a relatively high DTP, showing the possible risk of further usage; more refinements needed to be conducted in these skeletons to avoid further toxicity. Through trimming the molecular groups to overcome the deficiency of these compounds, they still had potential in PCSK9 inhibitor research and development.

Ligand binding analysis elucidated the mechanism of the interaction between ZINC000004099069 and ZINC000014952116 with PCSK9. The main interaction of ZINC000004099069 with PCSK9 was hydrogen bonds, while the interaction of ZINC000014952116 with PCSK9 was composed of hydrogen bonds and π-π bonds. From illustrations, we could observe that these two compounds formed many chemical bonds with the protein PCSK9, and more bonds

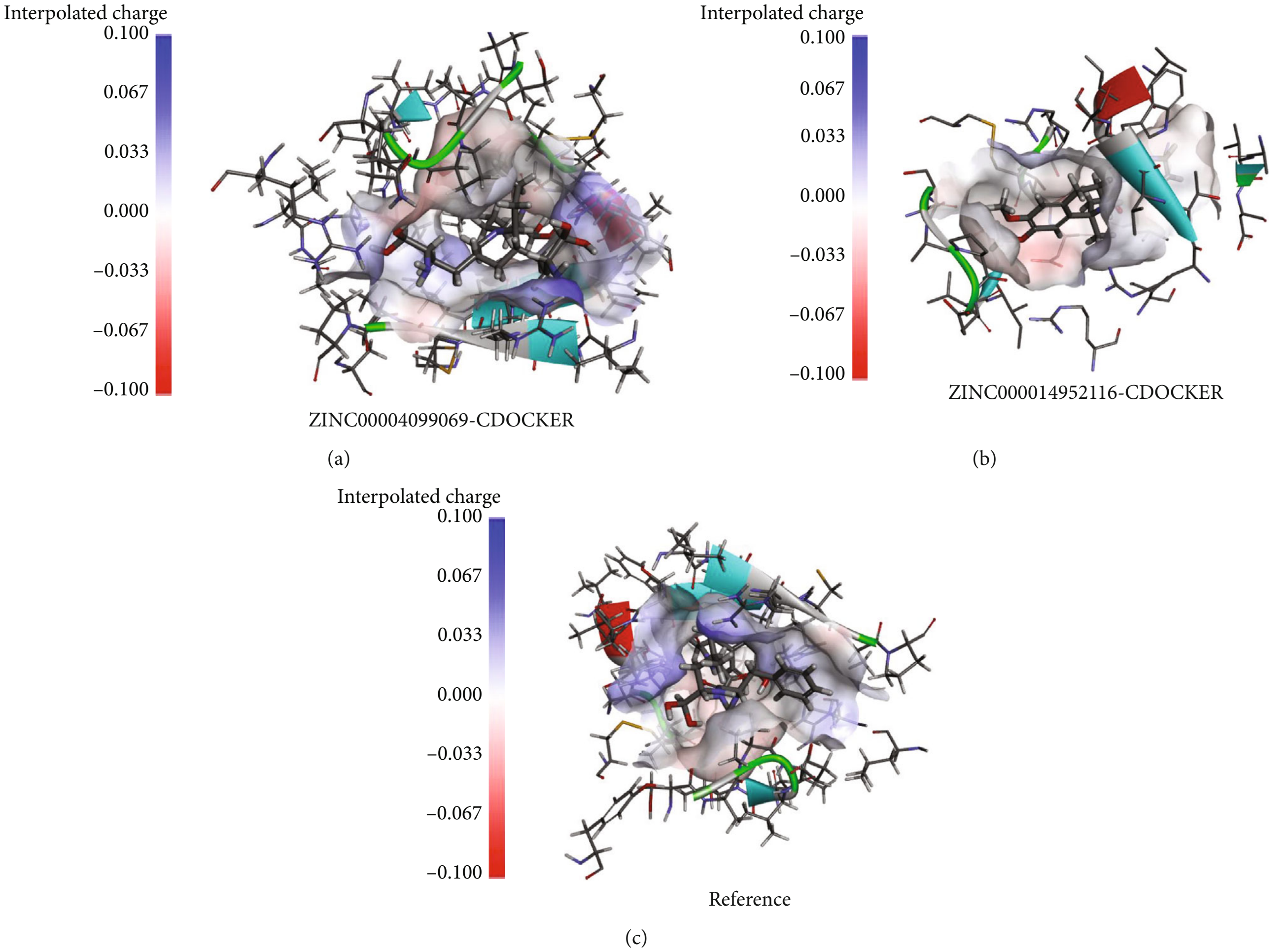


Figure 3: Schematic diagram of PCSK9 and the interaction of ZINC000004099069 (a), ZINC000014952116 (b), and reference (c) with PCSK9.

suggested harder structure, which could not be easily broken by other factors. Additionally, the CDOCKER interaction energy illustrated that both ZINC000004099069 and ZINC000014952116 had a high affinity with PCSK9, of which the former is higher, which confirmed the two complexes' stability. For this reason, by binding with different ligands which had high affinity with PCSK9, the conformation of PCSK9 could be changed, resulting in inactivation of protein function and leading into the clearance of LDLc, and ultimately played a major role in the treatment of CVD disease.

Lastly, molecular dynamics simulation was used to assess the stability of these complexes under natural circumstances. The stable existence of a complex in the natural environment indicated that it could be metabolized in the body as a whole unit and the ligand could not be separated from the protein by some metabolism processes, inactivating the function of the ligand. RMSD curves as well as potential energy of these ligand-PCSK9 complexes suggested the stability alteration of this conformation; as the progress of molecular dynamics went on, such as heating and equilibrium procedure, some groups and chemical bonds from the complex might change slightly, like bond rotation or fold conformation change; these alterations might cause the change of potential energy and RMSD value; high fluctuations indicated the instability of the conformation. From the illustrations, we observed that RMSD curves as well as potential energy got stable gradually, elucidating that these two ligand-PCSK9 complexes might exist stably. Consequently, these two complexes could keep stable and have promotion under nature environment. Besides, the ligand pose analysis as well as pharmacophore modification analysis all validated our results that ZINC000004099069 and ZINC000014952116 were potential lead compounds with activities.

Currently, the medication screening and design by a computational-aided method are mainly focused on tumor field, while targeted drug on basic diseases had hardly been studied. This study screened two ideal lead compounds targeting PCSK9 from natural products, which had effective activity and may inactivate the function of PCSK9, and finally decreased the accumulation of LDLc in the body and took place in the treatment of CVD disease. In conclusion, from a series of computer-aided studies, ZINC0000 04099069 and ZINC000014952116, two compounds, were finally selected as safe and potential candidate drugs.

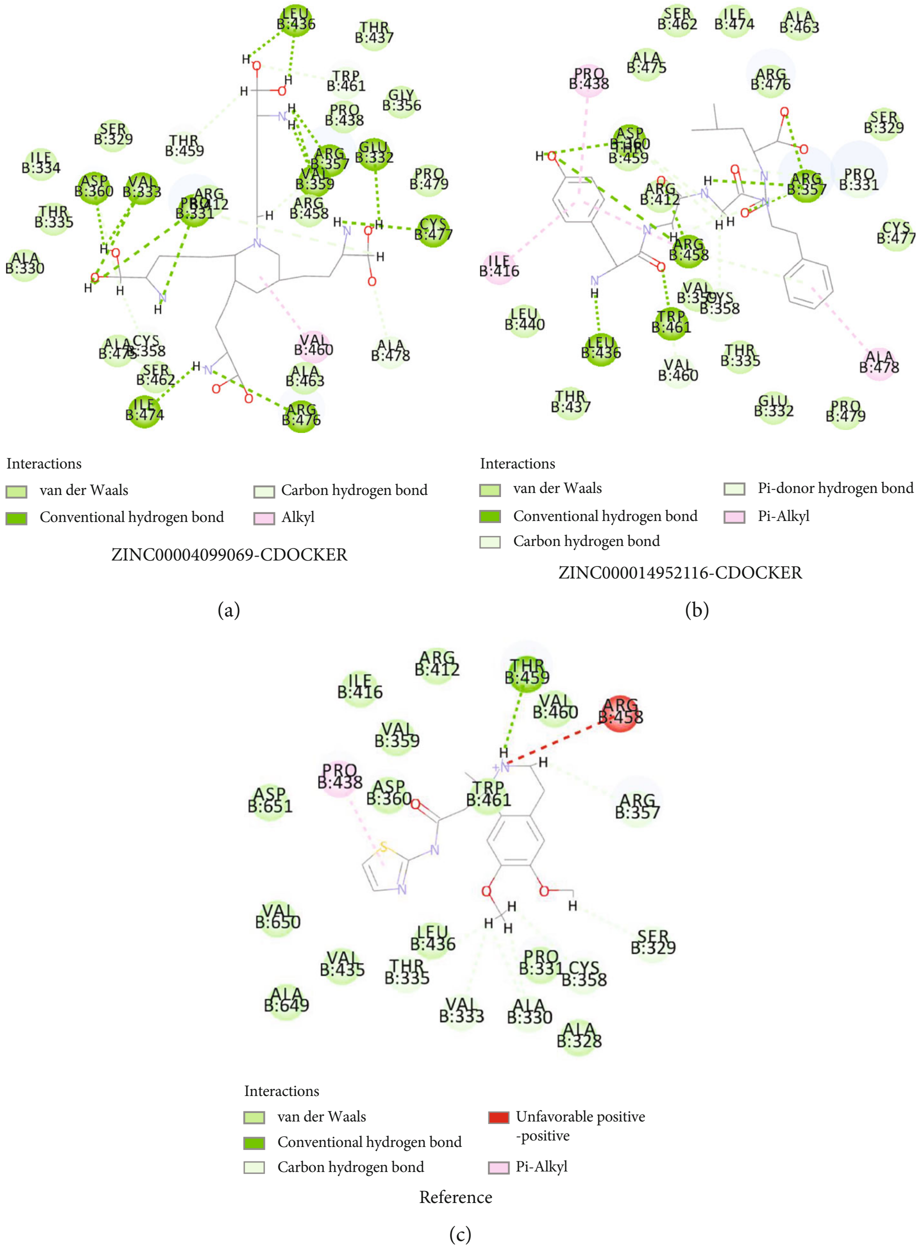


Figure 4: The intermolecular interaction of the predicted binding modes of ZINC00004099069 (a), ZINC000014952116 (b), and reference (c) with PCSK9.

Meanwhile, the information of other candidates provided in this study is listed in Tables 1–3, which can enrich the researches of PCSK9 and contribute a strong basis for PCSK9 inhibitor or other medication design and improvement. It is noteworthy that despite the disadvantages of some compounds analyzed from ADMET model in this study, it still provided a novel drug skeleton for medication design and refinement; different atoms or groups could be added or deleted from the skeleton to avoid the toxicity or other side effects. Thus, based on these two skeletons, more alterations and modifications on pharmacophore would be conducted to further improve the pesticide effects as well as reducing toxicity.

Although precise measurements and detailed designs had been conducted in this study, there were still some deficiencies due to the method of computational study. Animal model experiments were still needed to verify our research in the further. More indicators, like half-maximal inhibitory concentration and half-maximal effective concentration, need to be conducted to advance these two compounds to animals and eventually clinical application.

4. Methods and Materials

4.1. Docking Software and Ligand Library. Discovery Studio 4.5 software (BIOVIA, San Diego, California, US) applied

Table 5: Hydrogen bond interaction parameters for potential compounds and PCSK9.

Receptor	ZINC ID	Donor atom	Receptor atom	Distances (Å)
PCSK9	ZINC000004099069	B:ARG476:HH11	ZINC000012494317:N22	2.52581
		ZINC000012494317:H38	B:ARG357:O	2.05695
		ZINC000012494317:H38	B:VAL359:O	2.3965
		ZINC000012494317:H39	B:VAL359:O	2.0501
		ZINC000012494317:H57	B:CYS477:O	2.32261
		ZINC000012494317:H60	B:GLU332:OE2	2.18568
		ZINC000012494317:H71	B:ILE474:O	2.08114
		ZINC000012494317:H84	B:PRO331:O	2.25439
		ZINC000012494317:H86	B:PRO331:O	2.22788
		ZINC000012494317:H86	B:VAL333:O	2.64061
		ZINC000012494317:H87	B:VAL333:O	2.1696
		ZINC000012494317:H87	B:ASP360:OD2	2.22538
		ZINC000012494317:H89	B:LEU436:O	1.85936
		ZINC000012494317:H90	B:LEU436:O	2.0148
		B:PRO331:HA	ZINC000012494317:N31	2.49123
		B:TRP461:HD1	ZINC000012494317:O36	2.22741
		B:ALA478:HA	ZINC000012494317:O16	2.67059
		ZINC000012494317:H48	B:ARG357:O	2.56198
		ZINC000012494317:H59	B:PRO331:O	2.82566
		ZINC000012494317:H85	B:CYS358:O	2.49137
		ZINC000012494317:H85	B:ASP360:OD2	2.83867
		ZINC000012494317:H88	B:THR459:O	2.44502
	ZINC000014952116	B:ARG357:HH12	ZINC000014952116:O8	2.89478
		B:ARG357:HH22	ZINC000014952116:O40	1.86693
		B:ARG458:HH11	ZINC000014952116:O35	2.07486
		B:ARG458:HH12	ZINC000014952116:O23	2.05105
		B:TRP461:HN	ZINC000014952116:O27	2.7019
		ZINC000014952116:H60	ZINC000014952116:O23	1.92783
		ZINC000014952116:H63	B:ARG357:O	1.9849
		ZINC000014952116:H68	B:LEU436:O	1.85726
		ZINC000014952116:H74	B:ASP360:OD1	2.1021
		B:PRO331:HA	ZINC000014952116:O19	2.69646
		B:VAL460:HA	ZINC000014952116:O27	2.85792
		B:TRP461:HD1	ZINC000014952116:O27	2.13275
		ZINC000014952116:H61	B:ASP360:OD1	2.55504
		ZINC000014952116:H62	B:CYS358:O	2.45617
		ZINC000014952116:H62	B:ASP360:OD2	3.01446
		ZINC000014952116:H65	B:ARG357:O	2.40176
		ZINC000014952116:H77	ZINC000014952116:O8	2.41561
		B:ARG458:HH22	ZINC000014952116	2.79122

Table 6: Hydrophobic interaction parameters for compounds and PCSK9 residues.

Receptor	ZINC ID	Donor atom	Receptor atom	Distances (Å)
PCSK9	ZINC000004099069	B:VAL460	ZINC000012494317	4.51394
	ZINC000014952116	ZINC000014952116	B:ALA478	4.83821
		ZINC000014952116	B:ILE416	4.45717
		ZINC000014952116	B:PRO438	5.38467
		ZINC000014952116	B:ARG458	4.74229

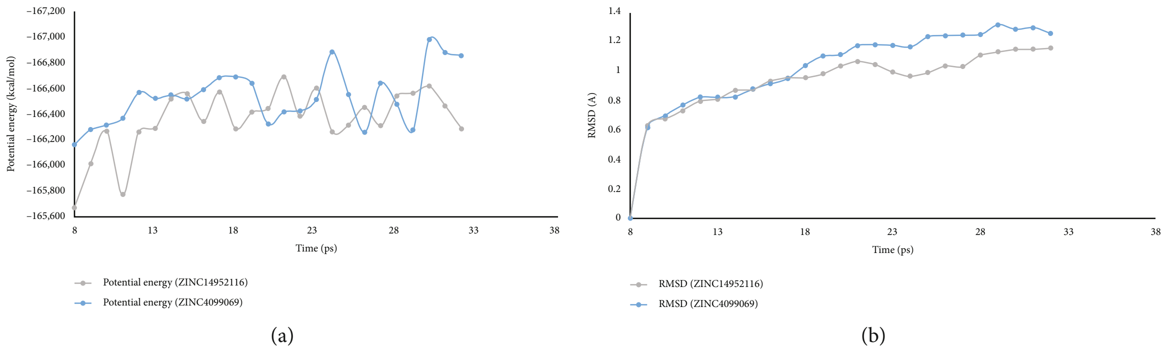


Figure 5: Results of MD simulation of ZINC000004099069 and ZINC000014952116.

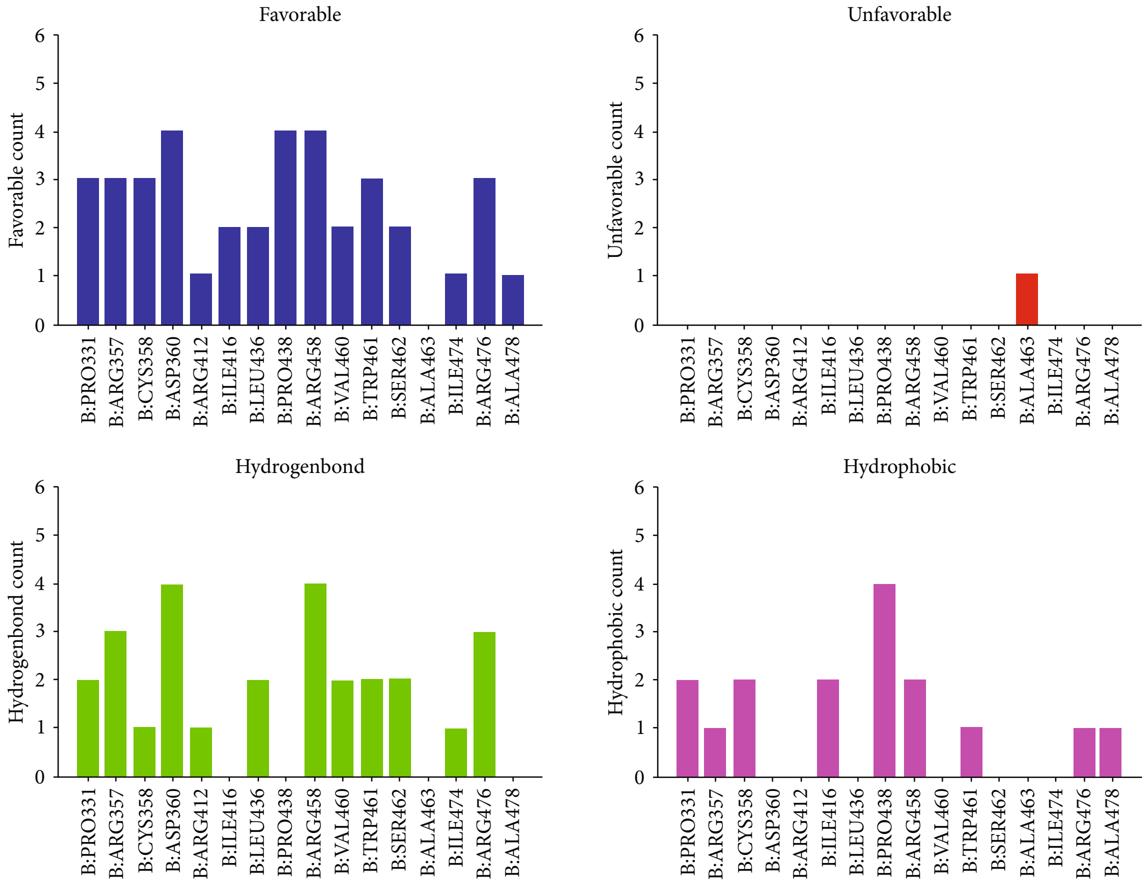
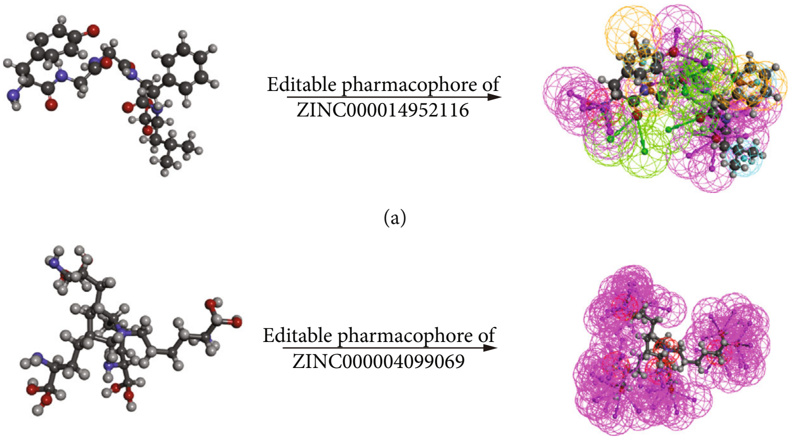


Figure 6: Schematic intermolecular interactions of the candidate compounds with PCSK9.

computation-aided structural biologic analysis to protein and other compounds for docking, modeling, prediction, etc. Natural inhibitors of PCSK9 used in this study were selected by analyzing information from the ZINC database, a free repository containing numerous ligands for commercial utility. And CDOCKER is used to explore the interaction of compounds and proteins.

4.2. Structure-Based Virtual Screening Using LibDock. Ligand-binding pocket region of PCSK9 was selected as the binding site and was used to screen compounds that could potentially bind with and then inhibit PCSK9. LibDock was a program which was applied to screen small molecules virtually. Using polar probes, nonpolar probes, and a grid placed into the binding site, hotspots were calculated by LibDock for the protein. Ligands were arranged to form favorable interactions using these hotspots. For ligand minimization, the Smart Minimiser algorithm and CHARMm force field were used. All the ligands were ranked according to the ligand score after minimization. The 2.0 Å crystal structure of human PCSK9 was downloaded from the protein data bank (PDB ID: 6U3X) and imported to the working circumstance of LibDock. To prepare the protein, crystal water and other heteroatoms around the protein



(a)

(b)

Figure 7: Pharmacophore predictions of ZINC000014952116 (a) and ZINC000004099069 (b) using 3D-QSAR. Green represents hydrogen acceptor, blue represents hydrophobic center, purple represents hydrogen donor, and orange represents aromatic ring.

were removed, and then, protonation, ionization, hydrogen, and energy minimization were added. Using the binding site of prepared protein and ligands, the active site for docking was generated. Using LibDock, the prepared ligands at the defined active site was docked virtually. Based on the LibDock score, all docked poses were generated and the ligands were ranked.

4.3. ADME (Absorption, Distribution, Metabolism, and Excretion) Properties and Toxicity Prediction. The ADME module and TOPKAT module of Discovery Studio 4.5 were employed to calculate ADME and toxicity properties including absorption, distribution, metabolism, and excretion. The four ADME aspects included BBB (blood-brain barrier) level, CYP2D6 prediction, hepatotoxicity, absorption (intestinal absorption) level, solubility (defined at 25°C in water) level, and PPB (plasma protein binding) prediction. These characteristics were fully assessed when selecting appropriate drugs for PCSK9.

4.4. Molecule Docking Analysis. CDOCKER module in Discovery Studio, an implementation of a CHARMm-based docking tool, was employed for precise docking study between ligands and protein. During the docking process, the receptor held rigid, while the ligands were allowed to be flexible. The CHARMm energy and the interaction energy, which demonstrated the ligand binding affinity, were calculated for each complex pose. Because the fixed crystal water molecules might affect the combination between the receptor and ligand, they were generally removed in a rigid and semiflexible docking process. Additionally, the water molecules were removed, and then, hydrogen atoms were added to the protein for protein optimization.

4.5. Molecular Dynamics Simulation. After the most appropriate ligand-PCSK9 complex was obtained and selected from the above calculation, molecular dynamics simulation would be done for valuing their stabilities. They were put into an orthorhombic box and solvated using an explicit periodic boundary solvated water model. Then, solid chloride was placed in this box with an ionic strength of 0.145 to simulate the natural environment. The box was subjected to the CHARMm forced field and relaxed by energy minimization (1000 steps of the steepest descent and 1000 steps of the conjugated gradient), with the final RMS gradient of 0.08326. The system's temperature was slowly driven from 50 K to 300 K for 2 ns, and equilibration simulation was run for 1 ns. With a time step of 2 fs, the whole MD simulations were run for 40 ns. The results were saved every 2 fs. Using Discovery Studio 4.5 software (BIOVIA, San Diego, California, US), the structural properties, potential energy, and RMSD of MD trajectory were determined. The CHARMm force field was used for both receptors and ligands. The binding site sphere of PCSK9 was defined as the region that came with radius 5 Å from the geometric centroid of the ligands. During the docking process, the ligands could bind with the residues within the binding spheres. After the parts of identified site were determined, the parts would be prepared into the binding site of PCSK9. After the docking process, each ligand generated 10 docking poses. And the posture with the highest docking score and best affinity would be selected. Besides, the CDOCKER interaction energy of different poses was also taken into calculation.

4.6. Validation of the Effects of the Candidate Compounds. To further validate if the selected compounds were the effective drugs in this study, we next performed ligand pose analysis based on the top 20 compounds in high-throughput module and analyzed the residue interactions between each compound and PCSK9. The candidate compounds with the best binding affinity could be evaluated by counts with favorable and unfavorable residues with PCSK9.

4.7. Pharmacophore Predictions of the Lead Compounds. After comprehensive assessment of these two compounds, this study then analyzed their pharmacophore characteristics and editable site through 3D-QSAR pharmacophore algorithm, which could provide up to 255 fits per molecule to represent a small molecule; only fits with energy values within the threshold of 10 Kcal/mol were retained.

5. Conclusions

This study employed a series of high-throughput methods based on structural biology, like virtual screening, precisely molecular docking, ADME, and toxicity prediction, as well as molecular dynamics simulation to find novel natural inhibitors regarding protein PCSK9, in order to treat cardiovascular disease by inhibiting the function of PCSK9. Totally, two compounds, ZINC000004099069 and ZINC000014952116, were finally screened as safe drug candidates, which had great significance in contributing to the development of PCSK9 inhibitor.

Data Availability

The data used and analyzed in this study are available upon reasonable request and can be found in the article/Supplementary Material.

Conflicts of Interest

All authors declare no conflicts of interest related to this manuscript.

Authors' Contributions

This study was completed with teamwork. Each author had made corresponding contribution to the study. Bo Gao, Huasong Zhou, and Xinhui Wang conceived the idea. Yingjing Zhao, Weihang Li, and Weiye Li wrote the main manuscript. Weihang Li, Xinhui Wang, Weiye Li, Yingjing Zhao, and Bo Wu used the software. Huasong Zhou, Hong Tao, Yuting Li, and Bo Wu downloaded and collected data. Yingjing Zhao, Weiye Li, Hong Tao, and Yuting Li analyzed the data. Yingjing Zhao, Weihang Li, and Weiye Li prepared figures. All authors redressed the manuscript. Bo Gao, Weihang Li, Huasong Zhou, and Xinhui Wang reviewed the manuscript. Yingjing Zhao and Weihang Li contributed equally as the co-first authors. All authors have approved the publication of this work

Acknowledgments

This study was supported by Flow Station of Principal Investigator Program (2021HYPI05) and Innovative Talents Program (2020SZRC1002) of Air Force Medical University.

References

[1] K. Kobiyama and K. Ley, "Atherosclerosis," *Circulation Research*, vol. 123, no. 10, pp. 1118–1120, 2018.

[2] C. Macchi, N. Ferri, C. R. Sirtori, A. Corsini, M. Banach, and M. Ruscica, "Proprotein convertase subtilisin/kexin type 9: a view beyond the canonical cholesterol-lowering impact," *The American Journal of Pathology*, vol. 191, no. 8, pp. 1385–1397, 2021.

[3] R. K. Myler, C. Ryan, R. Dunlap et al., "Dyslipoproteinemias in atherosclerosis, thrombosis and restenosis after coronary angioplasty," *The Journal of Invasive Cardiology*, vol. 7, no. 2, pp. 33–46, 1995.

[4] P. Mourikis, S. Zako, L. Dannenberg et al., "Lipid lowering therapy in cardiovascular disease: from myth to molecular reality," *Pharmacology & Therapeutics*, vol. 213, p. 107592, 2020.

[5] B. A. Ference, J. G. Robinson, R. D. Brook et al., "Variation inPCSK9andHMGCRand risk of cardiovascular disease and diabetes," *The New England Journal of Medicine*, vol. 375, no. 22, pp. 2144–2153, 2016.

[6] M. Heron, "Deaths: leading causes for 2017," *National Vital Statistics Reports*, vol. 68, no. 6, pp. 1–77, 2019.

[7] B. Genser and W. März, "Low density lipoprotein cholesterol, statins and cardiovascular events: a meta-analysis," *Clinical Research in Cardiology*, vol. 95, no. 8, pp. 393–404, 2006.

[8] N. G. Seidah and A. Prat, "The multifaceted biology of PCSK9," *Endocrine Reviews*, vol. 43, no. 3, pp. 558–582, 2022.

[9] M. S. Brown and J. L. Goldstein, "A receptor-mediated pathway for cholesterol homeostasis," *Science*, vol. 232, no. 4746, pp. 34–47, 1986.

[10] D. Cunningham, D. E. Danley, K. F. Geoghegan et al., "Structural and biophysical studies of PCSK9 and its mutants linked to familial hypercholesterolemia," *Nature Structural & Molecular Biology*, vol. 14, no. 5, pp. 413–419, 2007.

[11] K. N. Maxwell and J. L. Breslow, "Adenoviral-mediated expression of Pcsk9 in mice results in a low-density lipoprotein receptor knockout phenotype," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 18, pp. 7100–7105, 2004.

[12] S. Poirier, G. Mayer, S. Benjannet et al., "The proprotein convertase PCSK9 induces the degradation of low density lipoprotein receptor (LDLR) and its closest family members VLDLR and ApoER2," *The Journal of Biological Chemistry*, vol. 283, no. 4, pp. 2363–2372, 2008.

[13] A. J. P. Klein-Szanto and D. E. Bassi, "Keep recycling going: new approaches to reduce LDL-C," *Biochemical Pharmacology*, vol. 164, pp. 336–341, 2019.

[14] R. D. Santos, A. Ruzza, G. K. Hovingh et al., "Evolocumab in pediatric heterozygous familial hypercholesterolemia," *The New England Journal of Medicine*, vol. 383, no. 14, pp. 1317–1327, 2020.

[15] G. K. Hovingh, N. E. Lepor, D. Kallend, R. M. Stoekenbroek, P. L. J. Wijngaard, and F. J. Raal, "Inclisiran durably lowers low-density lipoprotein cholesterol and proprotein convertase subtilisin/kexin type 9 expression in homozygous familial hypercholesterolemia," *Circulation*, vol. 141, no. 22, pp. 1829–1831, 2020.

[16] Q. Ding, A. Strong, K. M. Patel et al., "Permanent alteration of PCSK9 with in vivo CRISPR-Cas9 genome editing," *Circulation Research*, vol. 115, no. 5, pp. 488–492, 2014.

[17] D. J. Newman, "Developing natural product drugs: supply problems and how they have been overcome," *Pharmacology & Therapeutics*, vol. 162, pp. 1–9, 2016.

[18] L. Yang, W. Li, Y. Zhao et al., "Computational study of novel natural inhibitors targeting O^6-methylguanine-DNA methyltransferase," *World Neurosurgery*, vol. 130, pp. e294–e306, 2019.

[19] S. Bang, H. S. Chae, C. Lee et al., "New aromatic compounds from the fruiting body of Sparassis crispa (Wulf.) and their

inhibitory activities on proprotein convertase subtilisin/kexin type 9 mRNA expression," *Journal of Agricultural and Food Chemistry*, vol. 65, no. 30, pp. 6152–6157, 2017.

[20] K. K. Ray, U. Landmesser, L. A. Leiter et al., "Inclisiran in patients at high cardiovascular risk with elevated LDL cholesterol," *The New England Journal of Medicine*, vol. 376, no. 15, pp. 1430–1440, 2017.

[21] K. Fitzgerald, S. White, A. Borodovsky et al., "A highly durable RNAi therapeutic inhibitor of PCSK9," *The New England Journal of Medicine*, vol. 376, no. 1, pp. 41–51, 2017.

[22] M. Stucchi, G. Grazioso, C. Lammi et al., "Disrupting the PCSK9/LDLR protein-protein interaction by an imidazole-based minimalist peptidomimetic," *Organic & Biomolecular Chemistry*, vol. 14, no. 41, pp. 9736–9740, 2016.

[23] L. Li, C. Shen, Y. X. Huang et al., "A new strategy for rapidly screening natural inhibitors targeting the PCSK9/LDLR interaction in vitro," *Molecules*, vol. 23, no. 9, p. 2397, 2018.

[24] P. Ahmad, S. S. Alvi, D. Iqbal, and M. S. Khan, "Insights into pharmacological mechanisms of polydatin in targeting risk factors-mediated atherosclerosis," *Life Sciences*, vol. 254, p. 117756, 2020.

[25] D. K. Min, H. S. Lee, N. Lee et al., "In silico screening of chemical libraries to develop inhibitors that hamper the interaction of PCSK9 with the LDL receptor," *Yonsei Medical Journal*, vol. 56, no. 5, pp. 1251–1257, 2015.

[26] J. Taechalertpaisarn, B. Zhao, X. Liang, and K. Burgess, "Small molecule inhibitors of the PCSK9.LDLR interaction," *Journal of the American Chemical Society*, vol. 140, no. 9, pp. 3242–3249, 2018.

[27] W. L. Petrilli, G. C. Adam, R. S. Erdmann et al., "From screening to targeted degradation: strategies for the discovery and optimization of small molecule ligands for PCSK9," *Cell Chemical Biology*, vol. 27, no. 1, pp. 32–40.e3, 2020.

[28] J. Sharifi-Rad, A. Sureda, G. C. Tenore et al., "Biological activities of essential oils: from plant chemoecology to traditional healing systems," *Molecules*, vol. 22, no. 1, p. 70, 2017.

[29] K. Kesarwani, R. Gupta, and A. Mukerjee, "Bioavailability enhancers of herbal origin: an overview," *Asian Pacific Journal of Tropical Biomedicine*, vol. 3, no. 4, pp. 253–266, 2013.

[30] B. Yang, J. Mao, B. Gao, and X. Lu, "Computer-assisted drug virtual screening based on the natural product databases," *Current Pharmaceutical Biotechnology*, vol. 20, no. 4, pp. 293–301, 2019.

Hindawi
BioMed Research International
Volume 2022, Article ID 7210909, 14 pages
https://doi.org/10.1155/2022/7210909

Research Article

Alterations of Microorganisms in Tongue Coating of Gastric Precancerous Lesion Patients with a Damp Phlegm Pattern

Xiangqun Xiao,[1] Renling Zhang,[2] Junhong Lu,[3] Yifeng Xu,[3] Zhujing Zhu,[2] Yiqin Wang,[3] Yaxiang Shi,[1] and Yiming Hao[3]

[1] *Zhenjiang Traditional Chinese Medicine Spleen and Stomach Disease Clinical Medicine Research Center, Zhenjiang Hospital Affiliated to Nanjing University of Chinese Medicine/Zhenjiang Hospital of Traditional Chinese Medicine, Zhenjiang, Jiangsu 212008, China*
[2] *Longhua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai 200032, China*
[3] *Shanghai Key Laboratory of Health Identification and Assessment/Laboratory of TCM Four Diagnostic Information, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China*

Correspondence should be addressed to Yiming Hao; hymjj888@163.com

Received 10 August 2022; Revised 1 September 2022; Accepted 5 September 2022; Published 22 September 2022

Academic Editor: Jian Zhang

Objective. In the research, the microbial changes in the tongue coating of patients with a damp phlegm pattern of gastric precancerous lesion (GPL) were investigated. *Methods*. This was a case-control study, in which 80 tongue coating samples were collected including 40 patients with a damp phlegm pattern of GPL, 20 patients with a nondamp phlegm pattern of GPL, and 20 healthy control people. The 16S rRNA microbiome technology was used to analyze the alterations of microorganisms in tongue coating of GPL patients with a damp phlegm pattern. *Results*. Microorganisms in the genus level were analyzed. Compared with the healthy control group, the relative abundance of 4 microorganisms (Solobacterium, Rothia, Oribacterium, and Alloprevotella) in the GPL group was significantly higher ($P < 0.05$). The relative abundance of 10 microorganisms (Terrisporobacter, Solobacterium, Porphyromonas, Parvimonas, Lactobacillus, Johnsonella, Gemella, Fusibacter, Azoarcus, and Acidothermus) in the GPL damp phlegm pattern group was significantly lower than that in the GPL nondamp phlegm pattern group ($P < 0.05$). In the comparison of phenotype "forms biofilms," the relative abundance of microorganisms in the GPL group was significantly higher than that in the healthy control group ($P < 0.05$). In the comparison of phenotype "contains mobile elements," the relative abundance of microorganisms in the GPL damp phlegm pattern group was significantly lower than that in the GPL nondamp phlegm pattern group ($P < 0.05$). In the comparison of microbial metabolic functions, the abundance ratio of "infectious diseases: bacterial" in the GPL group was significantly lower than that in the healthy control group ($P < 0.05$). The abundance ratio of the "excretory system" and "folding, sorting, and degradation" in the GPL group was significantly higher than that in the healthy control group ($P < 0.05$). *Conclusions*. Solobacterium may be a marker microorganism of the GPL damp phlegm pattern. The differential phenotype of microorganisms in tongue coating of the GPL damp tongue pattern is mainly reflected in "forms biofilms" and "contains mobile elements."

1. Introduction

Gastric cancer is the 3rd most common malignant tumor in the whole world in terms of incidence and mortality and is a serious threat to human health [1]. According to the global cancer statistics in 2018, new gastric cancer accounts for 5.7% of new cancer cases every year, of which more than 40% are in China. From normal tissue to gastric cancer, it usually goes through several stages: chronic gastritis, atrophy, gastrointestinal epithelial metaplasia, and dysplasia, and finally develops into gastric cancer [2]. Gastric precancerous lesions (GPL) are a kind of histopathological changes of gastric mucosa, mainly including intestinal metaplasia and dysplasia whose developments are considered reversible

[3, 4]. Therefore, early identification and active prevention and treatment of GPL can reduce the incidence rate of gastric cancer with a certain probability. However, modern medicine still lacks ideal treatment for GPL. Numerous clinical reports have confirmed that traditional Chinese medicine (TCM) treatment can reduce and remove some intestinal epithelial metaplasia and dysplasia and has certain curative effect on both symptom improvement and pathological reversal [5].

According to the theory of TCM, distinguishing the TCM patterns of GPL correctly is fundamental to treat the disease effectively by using Chinese medicine. One of the common TCM patterns of GPL is damp phlegm pattern. And the change of tongue coating appearance is one of the most significant diagnostic criteria of the GPL damp phlegm pattern. There are many pathogenic commensal bacteria in the human body. In a healthy state, these microorganisms exist in the form of symbiosis. It plays an important role in human's nutrition absorption, energy metabolism, immune function, and other physiological activities. The diversity and abundance of these microorganisms will also change relatively under unhealthy conditions, leading to the formation and development of multiple diseases, such as gastric cancer and other tumor diseases [6]. Oral microorganism is an important part to change the balance between oral and systemic health and disease. In the oral cavity, the morphological structure of the tongue surface allows the formation of a unique bacterial biofilm. Therefore, tongue coating has been considered the most complex ecological biofilm niche in the mouth [7]. In the previous study, it was found that there were different microorganisms in the tongue coating between GPL patients and healthy people, and the metabolites in the tongue coating of GPL patients with a damp phlegm pattern were significantly different from those of the nondamp phlegm pattern [8, 9]. Therefore, whether there are some special microorganisms in the tongue coating of GPL patients with a damp phlegm pattern may affect the formation of metabolites, which is a problem worthy of study.

16S ribosomal RNA (16S rRNA) gene sequencing has become the preferred method to study the composition and distribution of microbial communities [10]. In recent years, this technology has been widely used in the study of microbial diversity and relative abundance in the human body. Therefore, the 16S rRNA microbiome technology is adopted in this study; the changes of microorganisms in tongue coating of GPL patients with a damp phlegm pattern are explored. The significant fluctuations in the species and abundance of these microorganisms may help us better understand the formation and development mechanism of the GPL damp phlegm pattern from various aspects.

2. Materials and Methods

2.1. Samples. From December 2018 to October 2019, 60 patients with GPL voluntarily enrolled in Longhua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine were selected as the GPL group, including 40 cases of the damp phlegm pattern group and 20 cases of the nondamp phlegm pattern group. There were 20 people in the healthy control group who currently had no stomach discomfort and no history of stomach disease. Their routine physical examination indicators were normal. These indicators included blood cell analysis, liver and renal function, blood lipid, blood glucose, blood pressure, carcinoembryonic antigen, alpha-fetoprotein, color Doppler ultrasound of neck and abdomen, chest computed tomography, and X-ray barium meal. After the tongue coating samples of GPL patients were collected, the gastric mucosae were immediately examined by gastroscopy and pathology.

Table 1 summarizes the demographic and clinical information of all participants. It can be seen from this table that the numbers of GPL patients with a damp phlegm pattern with only mild, moderate, and severe intestinal metaplasia were 31, 7, and 1, respectively. The numbers of GPL patients with a nondamp phlegm pattern with only mild, moderate, and severe intestinal metaplasia were 13, 3, and 4, respectively. In addition, one GPL patient with a damp phlegm pattern had mild intestinal metaplasia and dysplasia at the same time. Among GPL patients with *Helicobacter pylori* (Hp) infection, there were 9 cases of a damp phlegm pattern and 1 case of a nondamp phlegm pattern.

In terms of treatment, there were 8, 10, 17, and 5 GPL patients with a damp phlegm pattern who were not treated, only treated with Western medicine, only treated with traditional Chinese medicine, and treated with integrated traditional and Western medicine, respectively, while there were 2, 5, 5, and 8 GPL patients with a nondamp phlegm pattern corresponding to the above four treatments.

2.2. Ethics Approval. In the study, all subjects gave written informed consent before collecting samples and the study was conducted in accordance with the Declaration of Helsinki. In addition, this study was approved by the Ethics Committee of Shanghai University of TCM in December 2018.

2.3. Criteria. The diagnostic criteria of GPL were as follows:

When the doctor used the endoscope to examine the patient's stomach, a small amount of gastric mucosa was removed from the suspected lesion sites (such as gastric antrum, gastric horn, gastric body, or cardia) for histopathological evaluation, which was conducted by two experienced pathologists according to the clinical guidelines of the "updated Sydney system" [11]. When the pathological evaluation of gastric mucosa showed atrophy with intestinal metaplasia or/and dysplasia, the patient was diagnosed with GPL [12].

The diagnostic criteria of the GPL damp phlegm pattern according to the "Diagnostics of Traditional Chinese Medicine" [13] were as follows:

GPL patients felt full stomach, even nausea, and/or vomiting, and their appetite decreases, accompanied by an unformed stool and greasy tongue coating.

The inclusion criteria were as follows:

(1) Patients that meet the above diagnostic criteria of the GPL damp phlegm pattern were included

Table 1: Summary of demographics and clinical information of the participants.

Demographics and clinical information	Damp phlegm pattern group	Nondamp phlegm pattern group	Healthy control group
Sample number	40	20	20
Ratio of male to female	1 : 0.82	1 : 1.22	1 : 1.86
Average age (year)	43.28 ± 14.73	42.9 ± 16.1	30.95 ± 11.68
Number (percentage) of samples diagnosed for less than 10 years	30 (75.00%)	15 (75.00%)	N/A
Number (percentage) of samples diagnosed for 10–20 years	6 (15.00%)	2 (10.00%)	N/A
Number (percentage) of samples diagnosed for 20–30 years	2 (5.00%)	1 (5.00%)	N/A
Number (percentage) of samples diagnosed for 30–40 years	2 (5.00%)	2 (10.00%)	N/A
Number (percentage) of samples only with intestinal metaplasia (mild)	31 (77.50%)	13 (65.00%)	N/A
Number (percentage) of samples only with intestinal metaplasia (moderate)	7 (17.50%)	3 (15.00%)	N/A
Number (percentage) of samples only with intestinal metaplasia (severe)	1 (2.50%)	4 (20.00%)	N/A
Number (percentage) of samples with intestinal metaplasia (mild) and dysplasia (mild)	1 (2.50%)	0 (0.00%)	N/A
Number (percentage) of samples with *Helicobacter pylori* infection	9 (22.50%)	1 (5.00%)	N/A
Number (percentage) of samples untreated	8 (20.00%)	2 (10.00%)	N/A
Number (percentage) of samples only taking Western medicine	10 (25.00%)	5 (25.00%)	N/A
Number (percentage) of samples only taking traditional Chinese medicine	17 (42.50%)	5 (25.00%)	N/A
Number (percentage) of samples taking Western medicine and traditional Chinese medicine	5 (12.50%)	8 (40.00%)	N/A

(2) The healthy controls were not found to have systemic organic diseases in routine physical examination

(3) The age range is 20–70 years old

(4) No antibiotics or probiotics were taken before collection

(5) All signed informed consent

The exclusion criteria were as follows:

(1) The patients had other digestive system diseases except gastritis

(2) Patients that are suffering from major organ diseases such as nervous system, circulatory system, and respiratory system were excluded

(3) Patients that are suffering from mental illness were excluded

(4) The female subjects were pregnant or lactating

(5) There were lesions in oral mucosa

(6) The subjects took antibiotics within half a year or probiotics as well as foods containing probiotics within one month before the sample was collected

(7) The subjects who smoke or drink alcohol were excluded

(8) The body mass index (BMI) exceeds 28 [14]

2.4. Sample Collection and Experimental Methods. We followed the sample collection and experimental methods of our previous research [8].

Tongue coating samples were collected in the morning, and all the subjects need not eat breakfast before being sampled. When collecting, first let the person to be collected gargle with sterile normal saline for 3 times to ensure that the residue in the mouth is removed as much as possible. Then, we used sterile sample collection swabs (CY-98000, iClean, Huachenyang Technology Co. Ltd., CN) to scrape tongue coating samples five times in the area with thick tongue coating. Finally, the swab head with a tongue coating sample was put into a sterile centrifuge tube and stored in an ultralow temperature refrigerator at −80°C. All tongue coating samples were scraped by the same person to ensure that the force used when scraping tongue coating was as consistent as possible. The patient underwent gastroscopy after the tongue coating sample was collected.

The Power Soil DNA Isolation Kit (MO BIO Laboratories) was used to extract microbial DNA from tongue coating samples. The quality and quantity of DNA were evaluated according to the ratio of 260 nm/280 nm and 260 nm/230 nm. First, the V3-V4 region of the microbial 16S rRNA gene was amplified by combining the adapter sequence and bar code sequence with common primer pairs (forward primer, 5′-ACTCCTACGGGAGGCAGCA-3′; reverse primer, 5′-GGACTACHVGGGTWTCTAAT-3′). Then, PCR amplification was performed. The initial denaturation lasted for 5 minutes at 95°C, followed by lasting for 1 minute cycles at 95°C (15 cycles), 1 minute at 50°C, 1 minute at 72°C, and finally 7 minutes at 72°C. The above is the first round of PCR. In this process, PCR products are

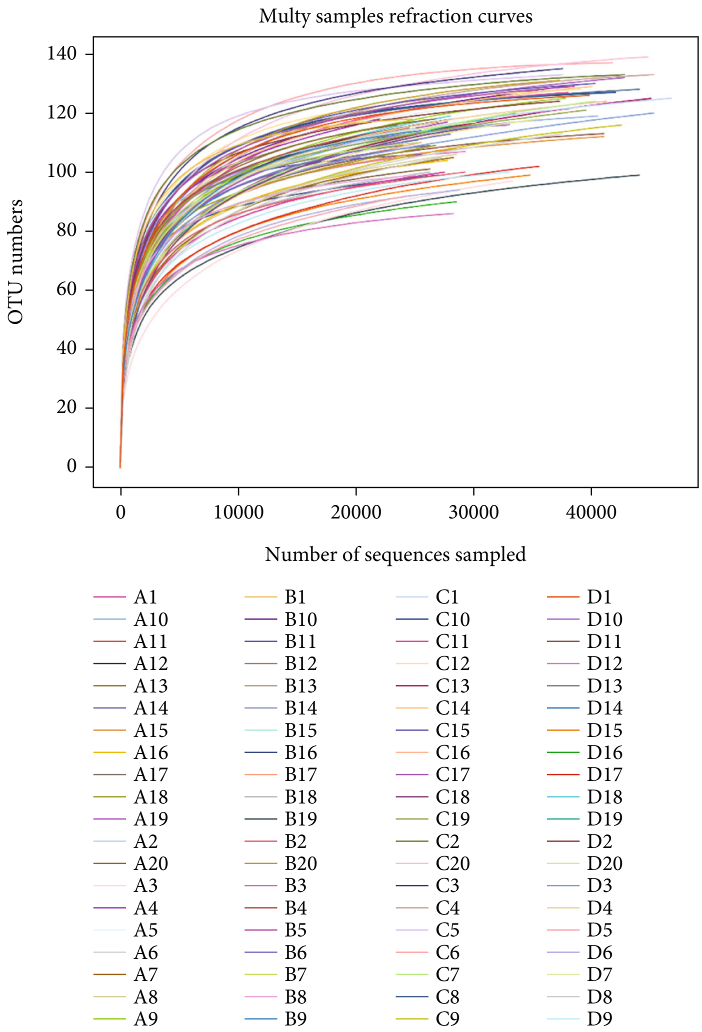


Figure 1: Rarefaction curve of each sample. The abscissa was the number of randomly selected sequencing strips, and the ordinate was the number of OTUs obtained by clustering based on the number of sequencing strips. Each curve represented a sample and was marked with different colors. The figure reflected the rate of emergence of new OTUs (new species) under continuous sampling: within a certain range, with the increase of the sequencing number, when the curve showed a sharp rise, it meant that a large number of species were found in the community. When the curve tended to be flat, it indicated that the species in this environment would not increase significantly with the increase of sequencing quantity.

purified by VAHTSTM DNA Clean Beads. Then, the second round of PCR was performed in the 40 μL reaction. The initial denaturation lasted for half a minute at 98°C, followed by lasting for 10 second at 98°C (10 cycles), half a minute at 65°C, half a minute at 72°C, and finally 5 minutes at 72°C. Finally, Quant-iT™ dsDNA HS Reagent was used to quantify and mix all PCR products. The Illumina Hiseq 2500 platform (2×250 paired ends) was used to perform high-throughput sequencing analysis of microbial rRNA genes on the purified mixed samples.

3. Statistical Analysis

The operational taxonomic unit (OTU) was analyzed by Trimmomatic (version 0.33), UCHIME (version 8.1), and USEARCH (version 10.0). Alpha diversity was calculated by mothur (version v.1.30). The Shannon index was used to measure the diversity of microorganisms. Beta diversity was calculated by QIIME. The unweighted algorithms named unweighted UniFrac was used to calculate the distance between samples to obtain the beta value. Microbial relative abundance between samples was compared using the Mann–Whitney U test. BugBase algorithm was used to predict the biological level coverage and biointerpretable phenotype of functional pathways between the two groups. Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to analyze the differences in metabolic pathways of functional genes between the two groups of microbial communities. The P value was corrected by the false discovery rate (FDR) of the rank sum test ($P < 0.05$).

4. Results

4.1. OTU Analysis. Through 16S rRNA gene sequencing of 80 tongue coating samples and subsequent splicing, filtering, and evaluation of tags, we finally obtained 151 OTUs.

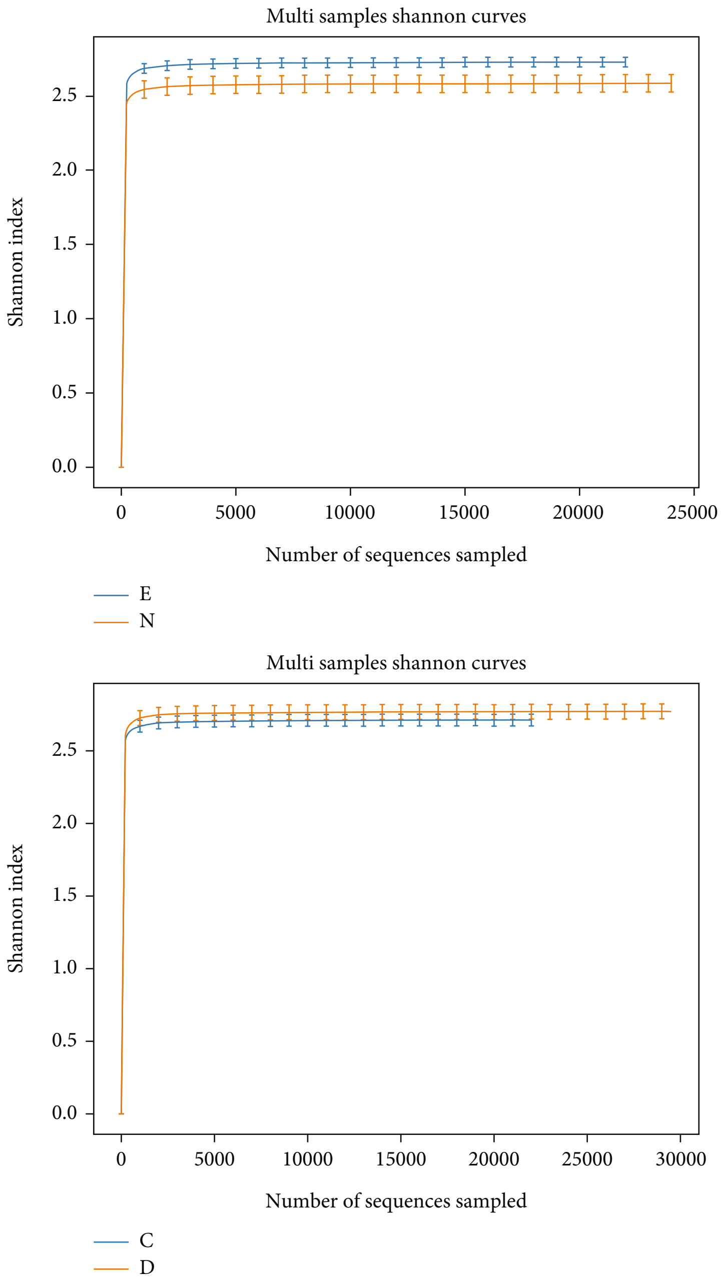
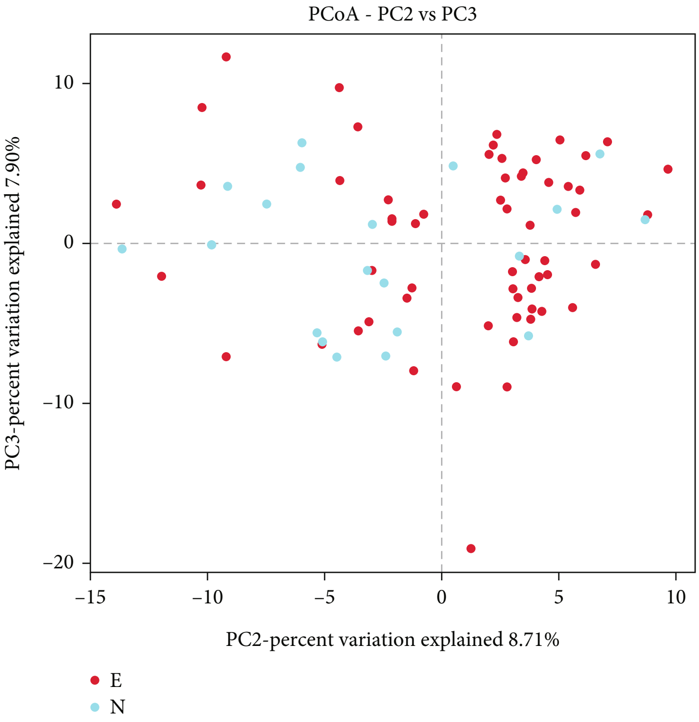


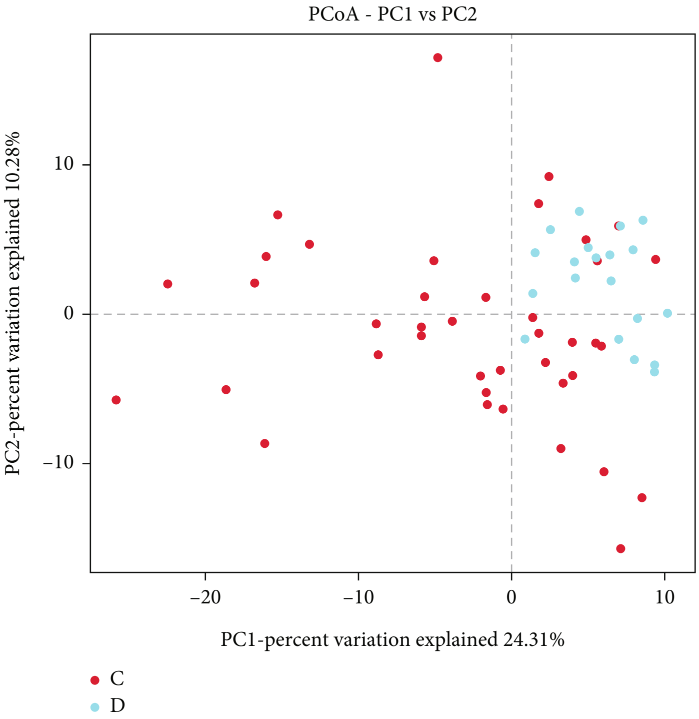
Figure 2: Shannon index curve of each group (E: GPL group, N: healthy control group, C: GPL damp phlegm pattern group, and D: GPL non-damp phlegm pattern group). The abscissa was the number of sequencing strips randomly extracted from samples, and the ordinate was the Shannon index. With the increase of sequencing quantity, more species were found. Until the species were saturated, increasing the number of sampling strips could not find new OTUs (new species). The microbial diversity of tongue coating in the GPL group was significantly higher than that in the healthy control group. There was no significant difference in microbial diversity of tongue coating between the GPL damp phlegm pattern group and GPL nondamp phlegm pattern group.

4.2. Alpha Diversity Analysis. The rarefaction curve was plotted before analyzing the microbial diversity of tongue coating. The rarefaction curve was formed by randomly sampling a certain number of sequences from the samples, counting the number of species represented by these sequences, and constructing the sequence number and the number of species. The curve was used to verify whether the amount of sequencing data was sufficient to reflect the species diversity in the samples and indirectly reflect the richness of species in the samples [15]. As shown in Figure 1, the curve representing each sample gradually tended to be gentle, indicating that the sequencing amount of each sample was sufficient and the data diversity analysis can be conducted.

In the alpha diversity analysis, the Shannon index was used to analyze the diversity of the microbiota between the two groups. The Shannon index was affected by species abundance and community evenness in samples. Under

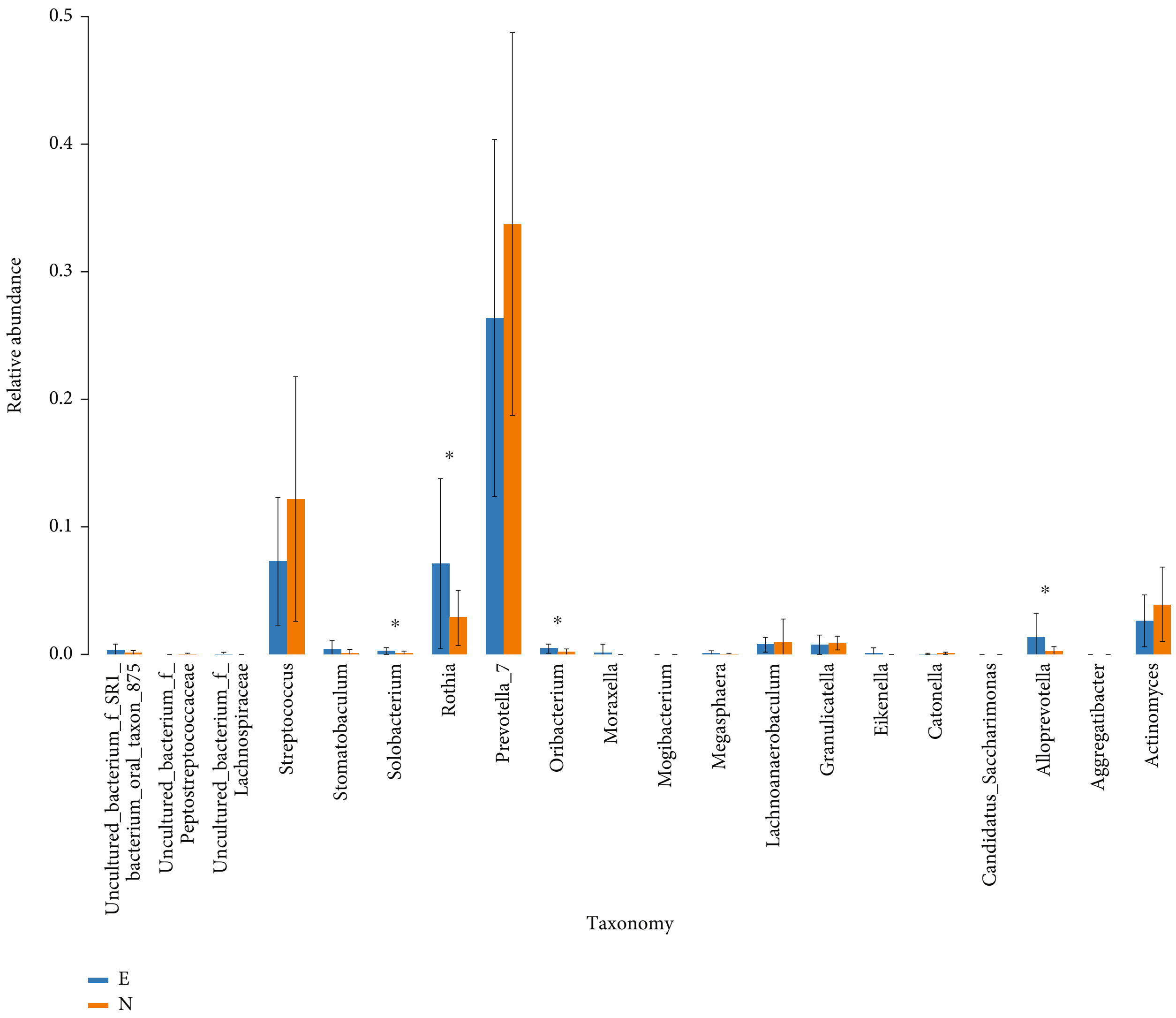


(a)



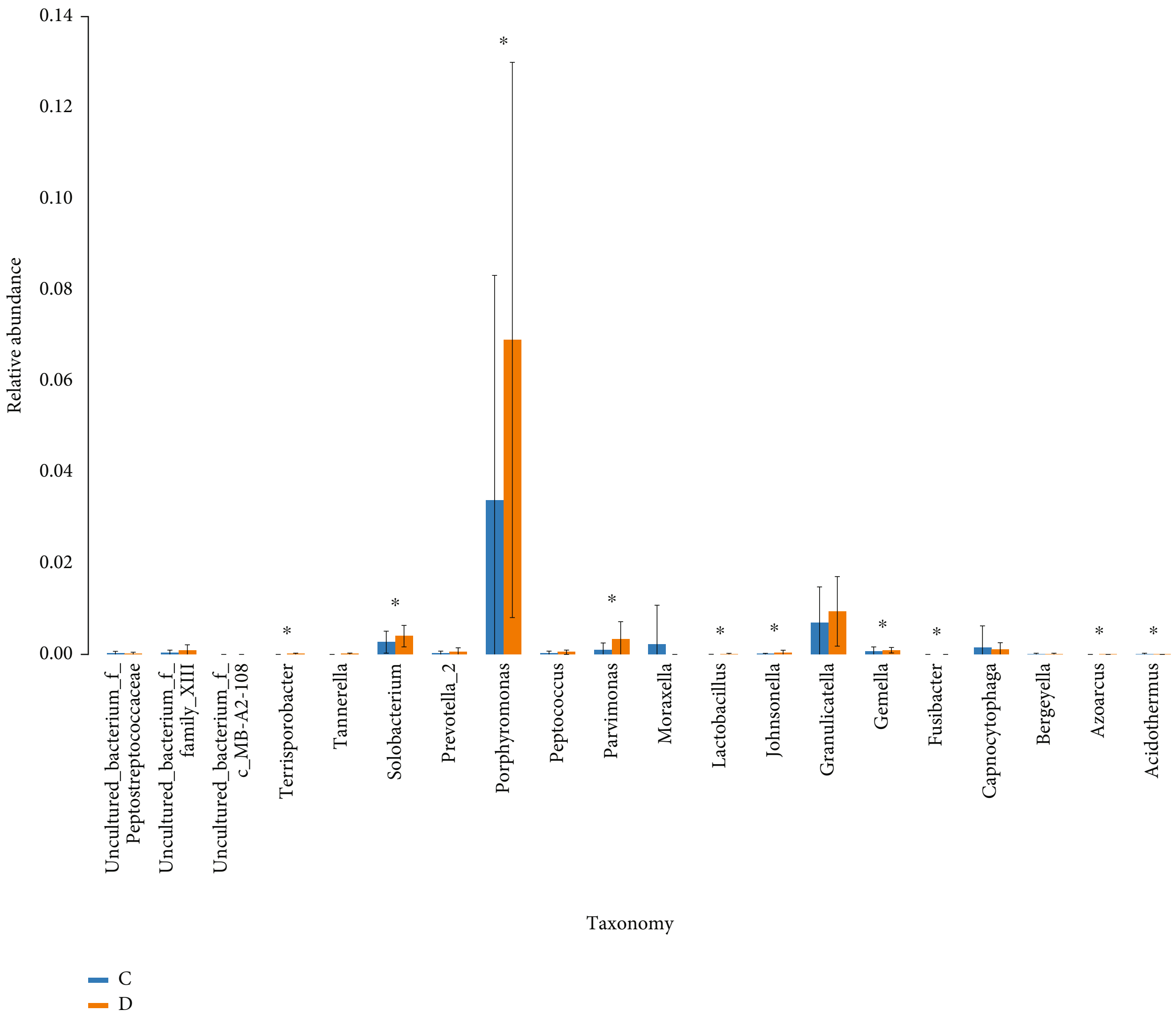
(b)

Figure 3: (a) PCoA diagram of the GPL group compared with the healthy control group (E: GPL group, N: healthy control group). (b) PCoA diagram of the GPL damp phlegm pattern group compared with the GPL non-damp phlegm pattern group. (C: GPL damp phlegm pattern group, D: GPL nondamp phlegm pattern group). The dots represented each sample. The abscissa and ordinate were the two characteristic values that lead to the largest difference between samples, and the main influence degree was expressed in the form of percentage. (a) The GPL group and healthy control group had a certain degree of differentiation. (b) The GPL damp phlegm pattern group and GPL non-damp phlegm pattern group had obvious differentiation.



(a)

Figure 4: Continued.



(b)

Figure 4: (a) Column figure of differential microorganism of the GPL group compared with the healthy control group (E: GPL group, N: healthy control group). (b) Column figure of differential microorganism of the GPL damp phlegm pattern group compared with the GPL nondamp phlegm pattern group (C: GPL damp phlegm pattern group, D: GPL non-damp phlegm pattern group). The abscissa represented the species of microorganisms (showing the top 20 species with the lowest P value), the ordinate represented the relative abundance of species, the columns with different colors represented each sample, and the "∗" on the column represented significant difference ($P < 0.05$). (a) The microorganisms named Solobacterium, Rothia, Oribacterium, and Alloprevotella with significant differences in the relative abundance between the GPL and healthy control groups at the genus level. Compared to healthy controls, the relative abundances of four microorganisms were significantly higher in the GPL group ($P < 0.05$). (b) The microorganisms named Terrisporobacter, Solobacterium, Porphyromonas, Parvimonas, Lactobacillus, Johnsonella, Gemella, Fusibacter, Azoarcus, and Acidothermus with significant differences in the relative abundance between the GPL damp phlegm pattern group and the GPL nondamp phlegm pattern group at the genus level. The relative abundances of ten microorganisms were significantly lower in the GPL damp phlegm pattern group compared to those in the GPL nondamp phlegm pattern group ($P < 0.05$).

the same species abundance, the greater the evenness of each species in the community, the greater the diversity of the community. The larger the Shannon index value, the higher the species diversity of the samples [16]. According to the Shannon index (2.73 ± 0.25 vs. 2.59 ± 0.26, $P = 0.03$), there was significant difference in microbiota diversity between the GPL group and healthy control group, with higher diversity being present in the GPL group. However, the Shannon index (2.71 ± 0.26 vs. 2.77 ± 0.23, $P = 0.41$) between the GPL damp phlegm pattern group and GPL nondamp phlegm pattern group indicated that there was no significant difference between the two groups (Figure 2).

4.3. Beta Diversity Analysis. In the beta diversity analysis, principal coordinates analysis (PCoA) was used to analyze the diversity of the microbiota between the two groups. PCoA was a dimension reduction sorting method. By assuming that there was data to measure the difference or

Table 2: Microorganisms with significant differences in the relative abundance between the GPL and healthy control groups at the genus level.

Microorganism	GPL group	Healthy control group	P corrected
Solobacterium	$3.19E-03 \pm 2.43E-03$	$1.36E-03 \pm 1.53E-03$	$5.13E-03$
Rothia	$7.16E-02 \pm 6.66E-02$	$2.93E-02 \pm 2.16E-02$	$4.85E-02$
Oribacterium	$4.61E-03 \pm 3.58E-03$	$2.22E-03 \pm 2.73E-03$	$6.47E-03$
Alloprevotella	$1.36E-02 \pm 1.89E-02$	$2.63E-03 \pm 3.56E-03$	$1.08E-03$

Table 3: Microorganisms with significant differences in the relative abundance between the GPL damp phlegm pattern and nondamp phlegm pattern groups at the genus level.

Microorganism	Damp phlegm pattern group	Nondamp phlegm pattern group	P corrected
Terrisporobacter	$1.85E-05 \pm 5.30E-05$	$1.16E-04 \pm 1.20E-04$	$8.71E-04$
Solobacterium	$2.74E-03 \pm 2.36E-03$	$4.09E-03 \pm 2.38E-03$	$2.47E-02$
Porphyromonas	$3.38E-02 \pm 4.95E-02$	$6.91E-02 \pm 6.09E-02$	$2.24E-02$
Parvimonas	$9.25E-04 \pm 1.70E-03$	$3.25E-03 \pm 4.07E-03$	$1.11E-02$
Lactobacillus	$3.51E-05 \pm 1.38E-04$	$1.08E-04 \pm 1.73E-04$	$5.33E-03$
Johnsonella	$8.68E-05 \pm 2.47E-04$	$3.61E-04 \pm 6.93E-04$	$1.18E-02$
Gemella	$6.88E-04 \pm 9.99E-04$	$9.85E-04 \pm 5.77E-04$	$2.40E-02$
Fusibacter	$1.74E-05 \pm 5.49E-05$	$4.29E-05 \pm 4.92E-05$	$1.03E-02$
Azoarcus	$2.65E-05 \pm 8.36E-05$	$6.27E-05 \pm 6.58E-05$	$1.33E-02$
Acidothermus	$6.24E-05 \pm 1.91E-04$	$1.02E-04 \pm 7.06E-05$	$2.49E-03$

distance between the samples, a rectangular coordinate system could be found to represent the samples as points and make the square of the Euclidean distance between the points equal to the original difference data, so as to realize the quantitative conversion of the qualitative data and extract the most important elements and structures from the multidimensional data. Through the principal coordinate analysis, the classification of multiple samples can be realized to further display the species diversity differences among the samples.

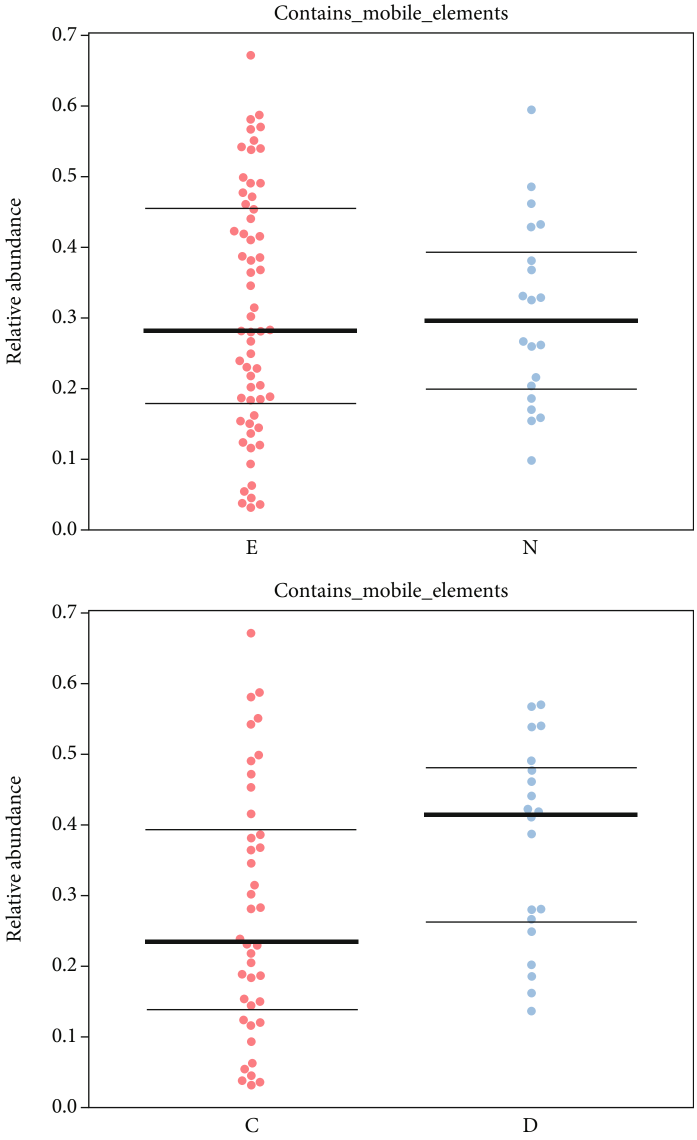
Figure 3(a) was the PCoA diagram of the GPL group compared with healthy control group. This figure showed that the samples (red dots) of the GPL group were relatively concentrated on the right side, while the samples (blue dots) of the healthy control group were relatively scattered on the left side. The two groups had a certain degree of differentiation. This result told that there were some differences in the microbial diversity of tongue coating between the GPL group and healthy control group.

Figure 3(b) was the NMDS diagram of the GPL damp phlegm pattern group compared with the GPL nondamp phlegm pattern group. From this figure, we can see that the samples (red dots) of the damp phlegm pattern group and the samples (blue dots) of the nondamp phlegm pattern group were concentrated in the lower left and upper right of the figure, respectively. The two groups had obvious differentiation. The results showed that there were significant differences in microbial diversity of tongue coating between the GPL damp phlegm pattern group and GPL nondamp phlegm pattern group.

4.4. Microbial Relative Abundance Analysis. In the comparison of the differences of microbial community abundance in the tongue coating samples between GPL patients and healthy controls, there were significant differences in the abundance of four kinds of microorganisms at the genus level. In Figure 4(a) and Table 2, the relative abundances of Solobacterium, Rothia, Oribacterium, and Alloprevotella were significantly higher in the GPL group compared to those in the healthy control group ($P < 0.05$).

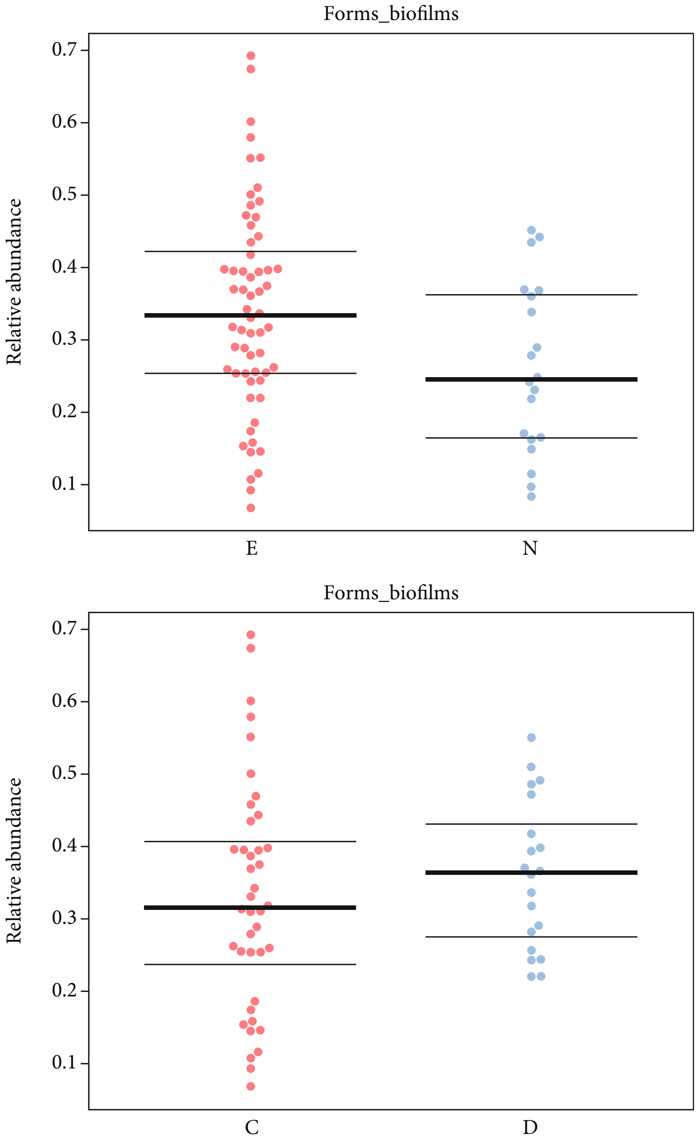
In the comparison of the differences of microbial community abundance in the tongue coating samples between GPL damp phlegm pattern group and GPL nondamp phlegm pattern group, there were significant differences in the abundance of ten kinds of microorganisms at the genus level. In Figure 4(b) and Table 3, the relative abundances of Terrisporobacter, Solobacterium, Porphyromonas, Parvimonas, Lactobacillus, Johnsonella, Gemella, Fusibacter, Azoarcus, and Acidothermus were significantly lower in the damp phlegm pattern group compared to those in the nondamp phlegm pattern group ($P < 0.05$).

It can be seen from the above results that, compared with that of the healthy control group, the overall level of the relative abundance of microorganisms named Solobacterium was increased in the tongue coating of GPL patients, but it was at a relatively low level in GPL patients with a damp



(a)

Figure 5: Continued.



(b)

Figure 5: (a) BugBase phenotype "contains mobile elements" prediction diagram (E: GPL group, N: healthy control group). (b) BugBase phenotype "forms biofilms" prediction diagram (C: GPL damp phlegm pattern group, D: GPL non-damp phlegm pattern group). The abscissa was the group name, the ordinate was the relative abundance percentage, and the three lines from bottom to top were the lower quartile, the average, and the upper quartile. (a) In the comparison of phenotype "contains mobile elements," the relative abundance of microorganisms in the GPL damp phlegm pattern group was significantly lower than that in the GPL nondamp phlegm pattern group ($P < 0.05$). (b) In the comparison of phenotype "forms biofilms," the relative abundance of microorganisms in the GPL group was significantly higher than that in the healthy control group ($P < 0.05$).

Table 4: Comparison of relative abundance of microorganisms between the GPL and healthy control groups with different phenotypes.

Phenotype	GPL group	Healthy control group	*P* corrected
Contains mobile elements	0.31 ± 0.17	0.31 ± 0.13	0.99
Forms biofilms	0.34 ± 0.14	0.26 ± 0.12	0.02

phlegm pattern. The results suggested that Solobacterium may be used as a microbial marker to identify the GPL damp phlegm pattern.

4.5. Microbial Phenotype Prediction. In the prediction and analysis of microbial phenotype of tongue coating, as shown in Figure 5 and Tables 4 and 5, in the comparison of phenotype "contains mobile elements," there was no significant difference between the GPL group and the healthy control group; however, the relative abundance of microorganisms in the GPL damp phlegm pattern group was significantly lower than that in the GPL nondamp phlegm pattern group

Table 5: Comparison of relative abundance of microorganisms between the GPL damp phlegm pattern and nondamp phlegm pattern groups with different phenotypes.

Phenotype	Damp phlegm pattern group	Nondamp phlegm pattern group	P corrected
Contains mobile elements	0.28 ± 0.18	0.37 ± 0.14	0.04
Forms biofilms	0.33 ± 0.16	0.36 ± 0.10	0.39

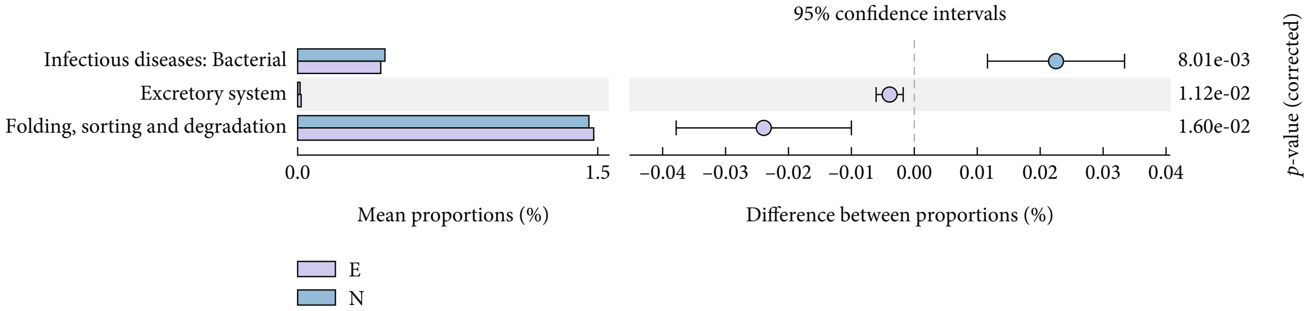


Figure 6: KEGG metabolic pathway difference analysis diagram (E: GPL group, N: healthy control group). The left part showed the abundance ratio of different metabolic functions in the two groups of tongue coating microorganisms, the middle figure showed the difference ratio of functional abundance within the 95% confidence interval, and the rightmost value was the P value. The abundance ratio of metabolic function “infectious diseases: bacterial” in the GPL group was significantly lower than that in the healthy control group ($P < 0.05$). The abundance ratio of the metabolic function “excretory system” and “folding, sorting, and degradation” in the GPL group was significantly higher than that in the healthy control group ($P < 0.05$).

Table 6: Comparison of abundance ratio (%) between the GPL and healthy control groups with different metabolic functions.

Metabolic function	GPL group	Healthy control group	P corrected
Infectious diseases: bacterial	0.41 ± 0.02	0.43 ± 0.02	$8.01E-03$
Excretory system	0.02 ± 0.01	$0.01 \pm 0.05E-02$	$1.12E-02$
Folding, sorting, and degradation	1.48 ± 0.04	1.45 ± 0.02	$1.60E-02$

($P < 0.05$). In the comparison of phenotype “forms biofilms,” the relative abundance of microorganisms in the GPL group was significantly higher than that in healthy control group ($P < 0.05$), but there was no significant difference between GPL damp phlegm pattern group and nondamp phlegm pattern group.

4.6. Microbial Metabolic Function Prediction. In the prediction and analysis of microbial metabolic function of tongue coating, as shown in Figure 6 and Table 6, the abundance ratio of metabolic function “infectious diseases: bacterial” in the GPL group was significantly lower than that in the healthy control group ($P < 0.05$). The abundance ratio of metabolic function “excretory system” and “folding, sorting, and degradation” in the GPL group was significantly higher than that in the healthy control group ($P < 0.05$). However, the significant difference in the metabolic pathway of the functional genes of the tongue coating microorganisms between the GPL damp phlegm pattern group and the GPL nondamp phlegm pattern group was not found.

5. Discussion

GPL is an important stage in the development of chronic gastritis into gastric cancer. The damp phlegm pattern is one of the most common TCM patterns of GPL. The appearance change of tongue coating is one of the important criteria for TCM doctors to diagnose the GPL damp phlegm pattern, but the related research on tongue coating microorganisms of GPL and its patients with damp phlegm pattern is rarely reported.

In this study, we used 16S rRNA technology to detect microbial changes in patients’ tongue coating. From the results of alpha and beta diversity analysis, it can be seen that there are differences in the microbial diversity of tongue coating between the GPL group and healthy control group as well as the GPL damp phlegm pattern group and GPL nondamp phlegm pattern group. In the further comparison of the relative abundance of microorganisms in each group, we found that there were significant differences in the relative abundance of 4 microorganisms between the GPL group and the healthy control group, which were Solobacterium, Rothia, Oribacterium, and Alloprevotella. The relative abundance of these four microorganisms in the GPL group was significantly higher than that in the healthy control group. There are significant differences in the relative abundance of 10 microorganisms between the GPL damp phlegm pattern group and GPL nondamp phlegm pattern group, including Terrisporobacter, Solobacterium, Porphyromonas, Parvimonas, Lactobacillus, Johnsonella, Gemella, Fusibacter, Azoarcus, and Acidothermus. The relative abundance of these 10 microorganisms in the GPL damp phlegm pattern group was significantly lower than that in the GPL nondamp phlegm pattern group.

Among these differential microorganisms, Solobacterium deserves our attention. Its relative abundance in the tongue coating of GPL patients increased, and its relative abundance in GPL damp phlegm pattern patients was significantly higher than that in GPL nondamp phlegm pattern patients. Studies have shown that Solobacterium is a Gram-positive, non-spore-forming obligate anaerobic bacterium from human feces [17]. This bacterium can cause halitosis and affect the development of digestive tract cancer, and the malodor causing cancer is hydrogen sulfide and acetaldehyde produced by Solobacterium [18]. Among the other three microorganisms whose relative abundance in the GPL group is significantly higher than that in the healthy control group, Rothia is a member of Gram-positive cocci of the Micrococcus family, which is considered to be an opportunistic pathogen, mainly affecting people with low immune function [19]. However, there is no report showing that this bacterium is found in patients with chronic gastritis or gastric cancer. Oribacterium has been found in saliva samples from patients with reflux esophagitis that its abundance is higher than that of healthy people [20]. Alloprevotella is also rarely found in patients with chronic gastritis or gastric cancer, but it was found to be increased in stool samples of ulcerative colitis or canceration [21].

Among the other 9 microorganisms with different relative abundance expression found in the comparison between the GPL damp phlegm pattern group and nondamp phlegm pattern group, Terrisporobacter has not been found in patients with stomach disease, but its relative abundance is different from that of healthy people in the feces of irritable bowel syndrome patients [22]. Porphyromonas, as an anaerobic bacterium, not only has local effects on its natural oral cavity but also has systemic tumorigenic effects, which may be related to GPL. Porphyromonas gingivalis can promote distant metastasis of cancer cells and resistance to anticancer drugs. This mechanism is mainly through affecting the gene expression of defensins, peptidyl arginine deaminase, and noncanonical activation of β-catenin. In addition, the microorganism can also convert ethanol into acetaldehyde, which is a carcinogenic intermediate [23]. Parvimonas, as an aerobic bacterium, is also related to the occurrence of tumors [24]. A study claimed that the presence of this microorganism in gastric mucosa could be used as one of the biomarkers to distinguish superficial gastritis from gastric cancer [25]. Lactobacillus has been confirmed in many studies that its relative abundance will increase in the development of gastric cancer [26]. Johnsonella was found in the oral cavity of patients with gastric internal metaplasia, and its enrichment degree was significantly higher than that of healthy people, which was related to the regulation of inflammation-related pathways [27]. Gemella has a high degree of centrality in the progression of precancerous lesions of gastric cancer [28]. Fusibacter, Azoarcus, and Acidothermus have not been reported in the study of digestive system diseases.

In the prediction and analysis of the microbial phenotype and metabolic function, we found that there was significant difference between the GPL group and the healthy control group in terms of phenotype “forms biofilms” and there was significant difference between the GPL damp phlegm pattern group and the GPL nondamp phlegm pattern group in terms of phenotype “contains mobile elements.” In addition, there were significant differences between the GPL damp phlegm pattern group and the GPL nondamp phlegm pattern group in the abundance ratio of metabolic function “infectious diseases: bacterial,” “excretory system,” and “folding, sorting, and degradation.” Although there is no relevant research on gastric diseases of the microbial phenotype and metabolic function which we found, our research results also provide some evidence for the microbial characteristics of GPL and its tongue coating of dampness syndrome and will guide us to further explore.

However, there are still some deficiencies in our research results. A total of 10 cases of Hp infection were distributed in the GPL damp phlegm pattern group and nondamp phlegm pattern group. In the current study, whether Hp affects oral microbiota is still controversial [29]. And we previously analyzed the differential microorganisms in the tongue coating of 60 patients with GPL in this study compared with 15 healthy people. The inclusion of these patients with Hp infection does not affect our final screening of differential microorganisms [8]. Even so, in future research, we will still carefully consider the factors of Hp infection. In addition, we will also expand the number of samples and use cohort research methods, metagenomics methods, and multiomics methods to further explore the formation mechanism of GPL and its damp phlegm pattern.

Data Availability

The 16S rRNA data used to support the findings of this study have been deposited in the OMIX repository (file ID: OMIX001727-01).

Disclosure

Xiangqun Xiao and Renling Zhang share the first authorship.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors’ Contributions

Yiming Hao designed the study. Xiangqun Xiao and Yiming Hao wrote the manuscript. Renling Zhang, Yiming Hao, Junhong Lu, and Yifeng Xu helped with the clinical sample collection. Zhujing Zhu helped with the experimentation. Yiqin Wang and Yaxiang Shi helped with the ideas of the study. All authors of this study agreed to be accountable for all aspects of the work.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant nos. 82174279 and 81703982), Shanghai Science and Technology Planning Program (Grant no. 21DZ2271000), Shanghai Education Commission Program (Grant no. 2021 Science and Technology 02-37), Jiangsu Administration of Traditional Chinese Medicine Key Program (Grant no. ZD201911), and Zhenjiang Science and Technology Bureau Program (Grant no. SS2021005). We were also grateful to Shanghai Biotree Biotech Co. Ltd. for its support in microbiological experiments.

References

[1] W. Xu, B. Li, M. Xu, T. Yang, and X. Hao, "Traditional Chinese medicine for precancerous lesions of gastric cancer: a review," *Biomedicine & Pharmacotherapy*, vol. 146, article 112542, 2022.

[2] Y. Zhang, X. Wu, C. Zhang et al., "Dissecting expression profiles of gastric precancerous lesions and early gastric cancer to explore crucial molecules in intestinal-type gastric cancer tumorigenesis," *The Journal of Pathology*, vol. 251, no. 2, pp. 135–146, 2020.

[3] I. Gullo, F. Grillo, L. Mastracci et al., "Precancerous lesions of the stomach, gastric cancer and hereditary gastric cancer syndromes," *Pathologica*, vol. 112, no. 3, pp. 166–185, 2020.

[4] D. Y. Graham, M. Rugge, and R. M. Genta, "Diagnosis: gastric intestinal metaplasia - what to do next?," *Current Opinion in Gastroenterology*, vol. 35, no. 6, pp. 535–543, 2019.

[5] X. Chen, Y. K. Dai, Y. Z. Zhang et al., "Efficacy of traditional Chinese medicine for gastric precancerous lesion: a meta-analysis of randomized controlled trials," *Complementary Therapies in Clinical Practice*, vol. 38, p. 101075, 2020.

[6] Z. Dong, B. Chen, H. Pan et al., "Detection of microbial 16S rRNA gene in the serum of patients with gastric cancer," *Frontiers in Oncology*, vol. 9, p. 608, 2019.

[7] S. Bernardi, M. A. Continenza, A. Al-Ahmad et al., "Streptococcus spp. and Fusobacterium nucleatum in tongue dorsum biofilm from halitosis patients: a fluorescence in situ hybridization (FISH) and confocal laser scanning microscopy (CLSM) study," *The New Microbiologica*, vol. 42, no. 2, pp. 108–113, 2019.

[8] Y. Hao, R. Zhang, R. Morris et al., "Metabolome and microbiome alterations in tongue coating of gastric precancerous lesion patients," *Expert Review of Gastroenterology & Hepatology*, vol. 15, no. 8, pp. 949–963, 2021.

[9] Y. Xu, R. Zhang, R. Morris et al., "Metabolite characteristics in tongue coating from damp phlegm pattern in patients with gastric precancerous lesion," *Evidence-based Complementary and Alternative Medicine*, vol. 2021, Article ID 5515325, 16 pages, 2021.

[10] J. E. Claridge 3rd., "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases," *Clinical Microbiology Reviews*, vol. 17, no. 4, pp. 840–862, 2004.

[11] M. F. Dixon, R. M. Genta, J. H. Yardley, and P. Correa, "Classification and grading of gastritis," *The American Journal of Surgical Pathology*, vol. 20, no. 10, pp. 1161–1181, 1996.

[12] P. Correa and M. B. Piazuelo, "The gastric precancerous cascade," *Journal of Digestive Diseases*, vol. 13, no. 1, pp. 2–9, 2012.

[13] J. C. He, *Diagnostics of Traditional Chinese Medicine*, Shanghai Pujiang Education Press, Shanghai, China, 2018.

[14] National Health and Family Planning Commission of PRC, *Criteria of Weight for Adults*, vol. 2, Standards Press of China, Beijing, China, 2013.

[15] Y. Wang, H. F. Sheng, Y. He et al., "Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags," *Applied and Environmental Microbiology*, vol. 78, no. 23, pp. 8264–8271, 2012.

[16] E. A. Grice, H. H. Kong, S. Conlan et al., "Topographical and temporal diversity of the human skin microbiome," *Science*, vol. 324, no. 5931, pp. 1190–1192, 2009.

[17] I. Barrak, A. Stájer, M. Gajdács, and E. Urbán, "Small, but smelly: the importance of _Solobacterium moorei_ in halitosis and other human infections," *Heliyon*, vol. 6, no. 10, p. e05371, 2020.

[18] K. Hampelska, M. M. Jaworska, Z. Ł. Babalska, and T. M. Karpiński, "The role of oral microbiota in intra-oral halitosis," *Journal of Clinical Medicine*, vol. 9, no. 8, p. 2484, 2020.

[19] M. Fatahi-Bafghi, "Characterization of the _Rothia_ spp. and their role in human clinical infections," *Infection, Genetics and Evolution*, vol. 93, p. 104877, 2021.

[20] B. Wang, Y. Zhang, Q. Zhao et al., "Patients with reflux esophagitis possess a possible different oral microbiota compared with healthy controls," *Frontiers in Pharmacology*, vol. 11, p. 1000, 2020.

[21] C. S. Wang, W. B. Li, H. Y. Wang et al., "VSL#3 can prevent ulcerative colitis-associated carcinogenesis in mice," *World Journal of Gastroenterology*, vol. 24, no. 37, pp. 4254–4262, 2018.

[22] T. Wang, I. Rijnaarts, G. D. A. Hermes et al., "Fecal microbiota signatures are not consistently related to symptom severity in irritable bowel syndrome," *Digestive Diseases and Sciences*, 2022.

[23] I. Olsen and Ö. Yilmaz, "Possible role of Porphyromonas gingivalis in orodigestive cancers," *Journal of Oral Microbiology*, vol. 11, no. 1, p. 1563410, 2019.

[24] J. Sun, Q. Tang, S. Yu et al., "Role of the oral microbiota in cancer evolution and progression," *Cancer Medicine*, vol. 9, no. 17, pp. 6306–6321, 2020.

[25] O. O. Coker, Z. Dai, Y. Nie et al., "Mucosal microbiome dysbiosis in gastric carcinogenesis," *Gut*, vol. 67, no. 6, pp. 1024–1032, 2018.

[26] Z. P. Li, J. X. Liu, L. L. Lu et al., "Overgrowth of Lactobacillus in gastric cancer," *World of Journal Gastrointest Oncology*, vol. 13, no. 9, pp. 1099–1108, 2021.

[27] F. Wu, L. Yang, Y. Hao et al., "Oral and gastric microbiome in relation to gastric intestinal metaplasia," *International Journal of Cancer*, vol. 150, no. 6, pp. 928–940, 2022.

[28] D. Liu, S. Chen, Y. Gou et al., "Gastrointestinal microbiota changes in patients with gastric precancerous lesions," *Frontiers in Cellular and Infection Microbiology*, vol. 11, article 749207, 2021.

[29] S. Z. Bakhti and S. Latifi-Navid, "Oral microbiota and helicobacter pylori in gastric carcinogenesis: what do we know and where next?," *BMC Microbiology*, vol. 21, no. 1, p. 71, 2021.

Hindawi
BioMed Research International
Volume 2022, Article ID 9015123, 14 pages
https://doi.org/10.1155/2022/9015123

Research Article

BERT-PPII: The Polyproline Type II Helix Structure Prediction Model Based on BERT and Multichannel CNN

Chuang Feng,[1] Zhen Wang ,[1,2] Guokun Li,[1] Xiaohan Yang,[1] Nannan Wu,[1] and Lei Wang [1]

[1] *School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China*
[2] *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China*

Correspondence should be addressed to Zhen Wang; wzh@sdut.edu.cn

Received 1 July 2022; Revised 1 August 2022; Accepted 3 August 2022; Published 24 August 2022

Academic Editor: Jian Zhang

Predicting the polyproline type II (PPII) helix structure is crucial important in many research areas, such as the protein folding mechanisms, the drug targets, and the protein functions. However, many existing PPII helix prediction algorithms encode the protein sequence information in a single way, which causes the insufficient learning of protein sequence feature information. To improve the protein sequence encoding performance, this paper proposes a BERT-based PPII helix structure prediction algorithm (BERT-PPII), which learns the protein sequence information based on the BERT model. The BERT model's *CLS* vector can fairly fuse sample's each amino acid residue information. Thus, we utilize the *CLS* vector as the global feature to represent the sample's global contextual information. As the interactions among the protein chains' local amino acid residues have an important influence on the formation of PPII helix, we utilize the CNN to extract local amino acid residues' features which can further enhance the information expression of protein sequence samples. In this paper, we fuse the *CLS* vectors with CNN local features to improve the performance of predicting PPII structure. Compared to the state-of-the-art PPIIPRED method, the experimental results on the unbalanced dataset show that the proposed method improves the accuracy value by 1% on the strict dataset and 2% on the less strict dataset. Correspondingly, the results on the balanced dataset show that the AUCs of the proposed method are 0.826 on the strict dataset and 0.785 on less strict datasets, respectively. For the independent test set, the proposed method has the AUC value of 0.827 on the strict dataset and 0.783 on the less strict dataset. The above experimental results have proved that the proposed BERT-PPII method can achieve a superior performance of predicting the PPII helix.

1. Introduction

Cowan et al. firstly discovered a special protein secondary structure the polyproline II (PPII) helix [1] which differs from the conventional protein secondary structure such as α-helix, β-pleated sheet, and random coil. The PPII helix consists of almost 3~8 amino acid residues, and it occupies only about 2% in the protein. The PPII helix has special biological characteristics and plays a crucial role in biochemical fields such as signal transduction, cell movement, and immune response [2, 3]. There are many interactions between the PPII helix and proteins or nucleic acids, such as SH3, WW, EVH1, GYF, UEV, and inhibitor proteins, which interact with the PPII helix [4–6]. Meanwhile, the PPII helix relates to many difficult diseases, such as the Alzheimer's disease and Parkinson's disease [7, 8]. Thus, it is very important to correctly predict the PPII helix. At present, the prediction of conventional secondary structures has made great achievements. But, a few researchers focused on the prediction of PPII helix. Furthermore, the PPII helix is very rare, which makes it become difficult to predict the PPII helix.

Anfinasen et al. [9] proposed the famous conclusion that protein sequence determines its spatial structure on the basis of experiments in 1961. Similarly, PPII structure is the same. The protein structure determination methods can be divided into two categories: traditional research methods of protein structure analysis and computational biology prediction methods. The traditional research methods use the X-ray crystal diffraction technology and the nuclear magnetic resonance imaging technology to predict the protein structure. It is hard for human to recognize, and the determination time is long. To solve the above problem, researchers proposed to predict PPII helices using protein sequence data in the bioinformatics field. However, the sequence based prediction models manually extract the features, and it usually leads to an inferior prediction result. Fortunately, the deep learning networks have powerful built-in feature extractors and have been widely used to extract protein feature information [10–12].

Recently, the researchers proposed to further improve the proteins features by using the natural language processing (NLP) technology. Proteins and languages are similar in concept [13], and Ofer et al. have descripted the relationship among the natural language processing, machine learning, and protein sequences. Ofer considers the protein sequence as an unknown language. Correspondingly, the amino acid is a word in biological vocabulary, and the biological sequence (such as DNA sequence and protein sequence) is text information. More and more natural language processing (NLP) techniques have been applied to solve the sequence prediction problems in bioinformatics [14–17].

The Bidirectional Encoder Representation from Transformers (BERT) [18] is a simple but powerful language model. We can pretrain BERT with the natural language corpus and use the trained BERT to transfer learning the biological sequences. Ho et al. [19] proposed the FAD-BERT model to predict the flavin adenine dinucleotide (FAD) binding sites, which can overcome the problem of insufficient feature learning caused by the shortage of training data. Charoenkwan et al. [20] used BERT4Bitter model to predict bitter peptides without system designing and feature coding selection. BERT4Bitter model automatically generate feature descriptors based on the original protein sequence. Li et al. [21] used the pretrained BERT model to learn both the protein sequence features and the amino acid hydrophilic features. As a result, it can improve the performance of predicting the missense mutations in protein sequences. To improve the encoding performance, Ali Shah et al. [22] utilized the pretrained BERT language model to extract the protein sequences features, which can effectively distinguish the three kinds of glucose transporter families. Le et al. [23] regarded DNA sequence as a natural language sentence and used BERT model to represent the DNA sequence information. It can capture the information which is equivalent to human language. BERT-m7G model [24] used the BERT model to convert RNA sequence information into feature matrix and select the optimal feature based on an elastic network. Finally, BERT-m7G model can effectively improve the prediction performance of RNA N7-methylguanosine.

As a special protein structure, many methods have been proposed to predict the PPII helix. Siermala et al. [25] firstly used the feed-forward neural network and back propagation algorithms to predict PPII helix structure. The prediction accuracy in reaches 75% on the datasets which has been eliminated more than 65% redundant sequences. Wang et al. [26] proposed to predict the PPII helix based on the support vector machine, and the prediction accuracy reached 70% on the dataset that further reduced homologous protein sequences. Lu et al. improved the artificial neural network [27] by jointly using the adjacent amino acid residue information and the one-hot encoding. Thus, Lu simultaneously use the improved artificial neural network, the support vector machine (SVM) [28], and the genetic neural network [29] to predict the PPII helix. O'Brien et al. [30] predict the PPII helix structure based on bidirectional recurrent neural network (BRNN). Its takes into account that the formation of PPII helix is affected by the remote residues, and other sequences are compared with the sequence to obtain a position-specific scoring matrix (PSSM) containing evolutionary information as a feature representation.

The existing PPII helix structure prediction methods usually adopt one kind of protein sequence code and only use the local or global protein sequence features. This will lead to an inferior performance. To solve the above problems, this paper uses the pretrained BERT model to improve the performance of protein sequences code. Each protein sequence is regarded as a sentence, and each amino acid is regarded as a word. This paper predicts the PPII helix structure by jointly using the local and global features. The flowchart of this algorithm is shown in Figure 1.

The proposed algorithm mainly includes three steps: learning global features, learning local features, and feature fusion.

(1) In the learning global features, we segment the protein amino acid sequences into many datasets with different sizes of sliding windows [34]. To further get the input of the BERT model, we separate each protein sequence sample into the amino acid residue by a space. After encoded by the BERT embedding layer, each amino acid residue is represented as a 768 dimensional context embedding vector. Then, each protein sequence sample is represented as n (n is window size) 768 dimensional vectors and 1 *CLS* vector. (2) In the learning local features, we use the multichannel CNN to extract n embedding vectors with 768 dimensions. The sizes of the multichannel CNN kernels are 3, 4, and 5, respectively. (3) In the feature fusion, we fuse the global *CLS* vector with the local features output by the multichannel CNN. Then, we use the softmax function to classify the fusion features.

In this paper, the BERT-PPII algorithm has the following innovations:

(i) The proposed method automatically extracts the feature extraction using protein primary sequences. This process has abandoned the system designing process and the feature selection procedure. Thus, it can avoid to manually extract the feature from raw amino acid sequences

(ii) We use the pretrained BERT model to improve the protein sequence encoding, and features to enhance the ability of feature representation

(iii) We design the comparative experiments on both the Strict_data dataset and the NonStrict_data dataset.

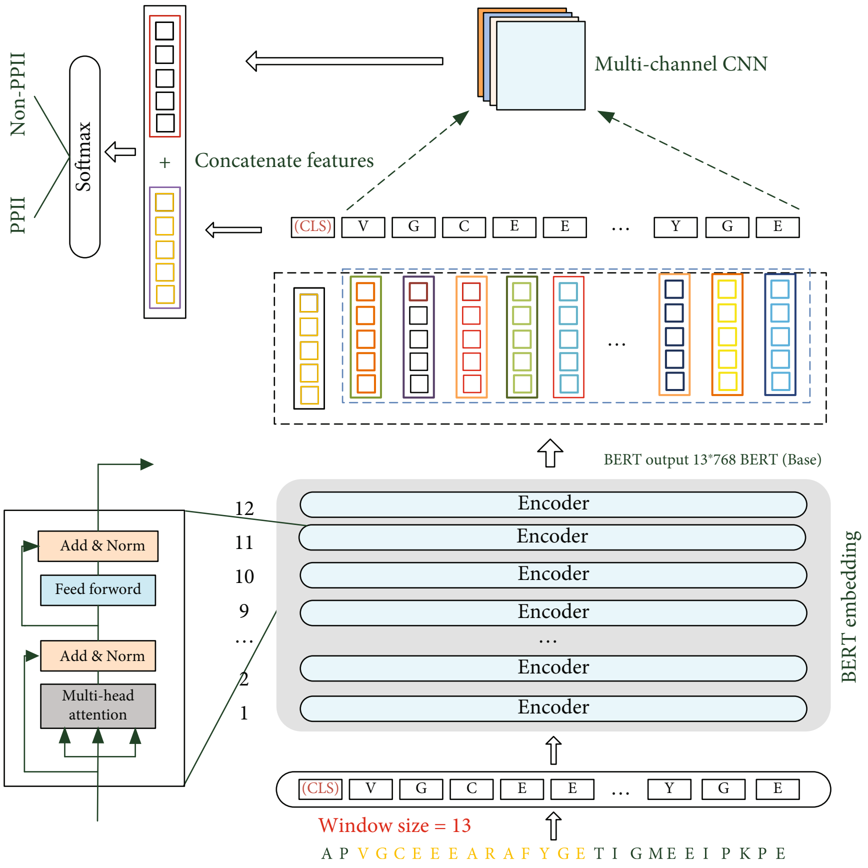


Figure 1: The flowchart of BERT-PPII model including input protein sequence samples, BERT embedding encoding and global feature extraction, local feature extraction by multichannel convolution, multifeature fusion, and prediction. It is assumed that the sliding window size is 13 and the amino acid residues in the sample are separated by space in the figure.

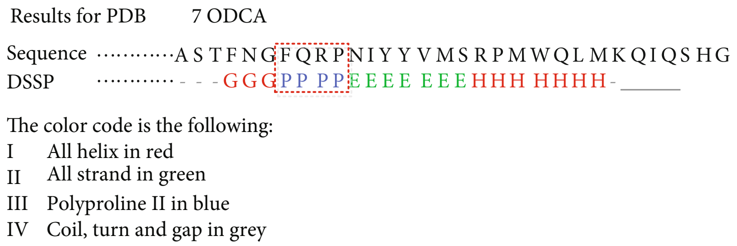
The final experimental results show that the proposed BERT-based model is better than the existing algorithms

2. Materials and Methods

2.1. Problem Description. The PPII helix is a local spatial conformation between amino acid residues in the protein polypeptide chain. It usually consists of 3~8 amino acids. Its prediction task maps the protein sequence composed of 20 amino acids to the corresponding the PPII helix structure sequence. As shown in Figure 2, FQRP, the partial amino acid residues of protein sequence, is mapped to PPII helix structure. The existing PPII secondary structure prediction algorithms adopt only one kind of the protein encoding method, which causes the problem of insufficient learning features. The PPII helix is determined by both the local and the long-range among the amino acid residues in the protein chain. If the prediction process only uses local or global features, it will ignore the important PPII helix formation information and decrease the prediction accuracy.

To solve the problem of encoding protein sequence, this paper employs the BERT to improve the code of amino acids. Moreover, the *CLS* feature of the protein sequence obtained by BERT and the local feature of the protein sequence obtained by multichannel CNN are further integrated to effectively improve the expression ability of sample features. Our model mainly includes BERT embedding encoding and global feature extraction, local feature extraction by multichannel convolution, and multifeature fusion, which are described in Sections 2.2, 2.3, and 2.4, respectively.

2.2. Bert Embedding Encoding and Global Feature Extraction. More and more natural language processing (NLP) techniques have been employed to learn the feature descriptors of protein sequences, DNA sequences, and RNA sequences [14–17]. The BERT embedding layer can obtain semantic and syntactic information from the context of a sentence or paragraph, which enables to learn better features. Recently, most PPII helix structure prediction algorithms usually adopt only one kind of protein sequence feature encoding method. In order to learn the better features, the pretrained BERT model is used to improve the of the PPII helix structure prediction performance. We break this limitation by pretraining the model based on bidirectional encoder representation from transformers (BERT). The BERT model uses the multiattention mechanism to obtain the *CLS* feature vector. The *CLS* feature vector can fairly integrate the information of each amino acid residue in the sample. Finally, the *CLS* feature is considered as the



Results for PDB 7 ODCA

Sequence …………A S T F N G F Q R P N I Y Y V M S R P M W Q L M K Q I Q S H G
DSSP ………… - - - G G G P P P P E E E E E E E H H H H H H H - ______

The color code is the following:
I All helix in red
II All strand in green
III Polyproline II in blue
IV Coil, turn and gap in grey

Figure 2: Some primary sequences of protein sequence (PDB id: 7ODCA) are assigned secondary structure conformations by DSSP algorithm. This graph is derived from the online PPII and secondary structure assignment database developed by Chebrek et al. [35]. In the graph, a letter represents a specific conformation, and its color relates to different secondary structure categories.

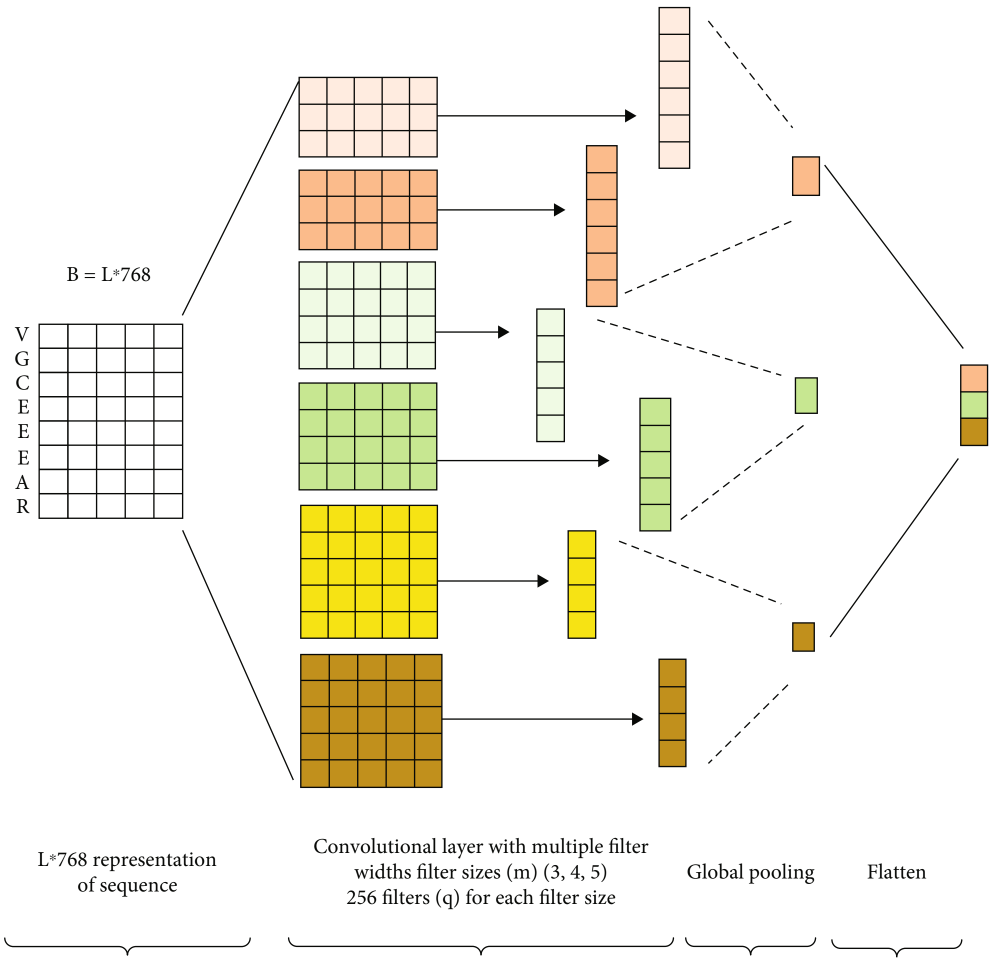

B = L*768
V
G
C
E
E
E
A
R
L*768 representation of sequence
Convolutional layer with multiple filter widths filter sizes (m) (3, 4, 5) 256 filters (q) for each filter size
Global pooling
Flatten

Figure 3: The Multichannel CNN model.

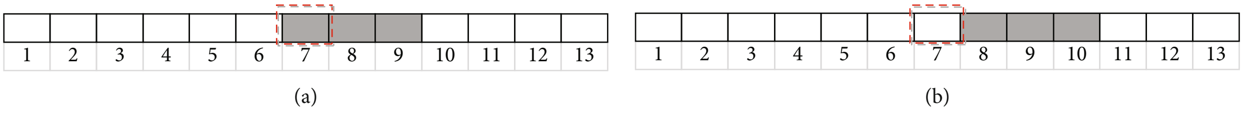


Figure 4: (a) Positive sample; (b) negative sample.

global feature. The BERT model handles the migration task's input samples by the position encoding, self-attention mechanism, and residual connection.

Position encoding: Generally, the same characters with different locations are assigned the same feature description. Thus, they cannot capture the location information of the

Table 1: The dataset under strict definition (Strict_data).

Dataset	Number of sequence	Number of PPII	Number of non-PPII	Total
Training set	6561	36622	1494487	1531109
Test set	1640	9068	382819	391887
Independent test set	920	4855	201537	206392

Table 2: The dataset under less strict definition (NonStrict_data).

Dataset	Number of sequence	Number of PPII	Number of non-PPII	Total
Training set	7121	64490	1554142	1618432
Test set	1781	15880	379276	395156
Independent test set	1001	8639	208785	217424

input text. To solve the above problem, the input samples are encoded according to the position of the character, as shown in Equation (1). *PE* denotes the position code of each input character. *pos* denotes the position of the character in the sequence. *dmodel* denotes the dimension of $WQ(x)$. When the same characters appear in the input amino acid residues, they will have different feature codes obtained by the self-attentive mechanism due to the different position codes.

$$PE(pos, j) = \begin{cases} \sin\left(\dfrac{pos}{10000^{j}/d_{\text{model}}}\right), j = 2i \\ \cos\left(\dfrac{pos}{10000^{j-1}/d_{\text{model}}}\right), j = 2i + 1 \end{cases}. \quad (1)$$

After that, the protein sequence sample $X = (x_1, x_2, x_3, \cdots, x_n)$ will be processed by word embedding query (WQ) and position coding (PE), as shown in Equation (2). X_{input} represents the input vector of BERT:

$$X_{\text{input}} = WQ(X) + PE. \quad (2)$$

Self-attention mechanism: This paper utilizes the self-attention mechanism to capture the relationship among the amino acid residues of the input sample sequence, as shown in Equation (3). As a result, each character contains the information of the other characters, where $Q = X_{\text{input}} W^Q$, $K = X_{\text{input}} W^K$, $V = X_{\text{input}} W^V$. Q, V, and K are the query vector, value vector, and key vector, respectively. W^Q, W^K, and W^V are the weight matrices of Q, K, and V, respectively.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V. \quad (3)$$

Residual connection: To avoid the problems of gradient disappearance and explosion during the training process, we establish the residual connection for the output of the self-attentive mechanism [36], as shown in Equation (4).

$$X_{\text{output}} = X_{\text{input}} + \text{Attention}(Q, K, V). \quad (4)$$

During training the model, we normalize the data [37, 38] as shown in Equation (5). Thus, the algorithm can quickly and smoothly converge to the optimal solution. μ is the mean value of X_{output} and σ is the standard deviation of X_{output}. When σ becomes 0, ε can avoid the denominator being 0. The training parameters α and β can compensate the information lost during the normalization process:

$$\text{LayerStandary} = \alpha \frac{X_{\text{output}} - \mu}{\sigma + \varepsilon} + \beta. \quad (5)$$

To obtain the amino acid residues, we put the standardized features into the fully connected neural network followed by a residual connection and a standardization procedure.

To ensure the transformer's self-attention mechanism [39] has excellent representation ability, BERT model employs two pretraining tasks [18]: the "masked language model" (MLM) and the "next sentence prediction" (NSP). As a result, it can provide a better generalization result for the downstream tasks.

2.3. Local Feature Extraction by Multichannel Convolution.

The interaction among the local amino acid residues in the protein chain has an important influence on the formation of the PPII helix. The protein sequences' features can be represented as matrices, and the local spatial correlations exist among the amino acids' features in the sequences. Moreover, the convolutional neural networks (CNNs) can handle the spatial correlation among the dense data in the network. In this paper, to obtain the relationships among the local amino acid residues, we further use the CNN to learn the feature of Bert's output vectors. The convolution neural networks capture the important local information of the protein sequence sample's features. Correspondingly, the pooling procedure learns the important local features. Thereafter, we obtain

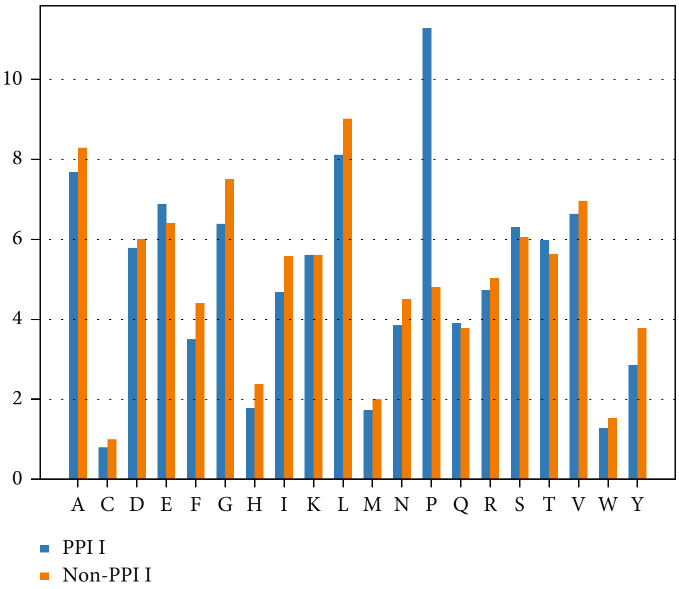


Figure 5: The amino acid composition of PPII and Non-PPII.

the final vector η by splicing the output vectors of the CNN layers.

In this paper, we design the CNN models with convolutional kernels of 3, 4, and 5, respectively (Section 3.5). As shown in Figure 3, the sample's local feature learning process mainly consists of the convolution operation and the pooling operation.

Convolution operation: We use the convolution operation to process the BERT layer's output matrix $B = \{H_1, H_2, \cdots H_n\}$. Assuming the convolution kernel's size is m, each time the convolution is computed based on m word vectors. Generally, we slide the convolution kernel 1 step from top to bottom and divide B into $\{H_{1:m}, H_{2:m+1}, \cdots, H_{n-m+1:n}\}$. Where $H_{i:j}$ represents the concatenated vectors of $\{H_i \cdots H_j\}$. The vector $C = \{c_1, c_2, \cdots, c_{n-m+1}\}$ and the value ci is obtained by convolving $H_{i:i+m-1}$, as shown in Equation (6):

$$c_1 = W^{\mathrm{T}} H_{i:i+m-1} + b. \tag{6}$$

We initialize the convolution kernel's parameter (W) as a random uniform distribution. b is the bias variable.

Pooling operation: After the convolution operation, we perform a pooling operation on the text feature mapping vector $C = \{c_1, c_2, \cdots, c_{n-m+1}\}$. For the results obtained with q convolution kernels, we use a global maximum pooling, as shown in Equation (7).

$$\widehat{C}_m = \max\left(C_{m1}, C_{m2}, \cdots, C_{mq}\right). \tag{7}$$

We concentrate the features extracted with the kernel sizes $m = (3, 4, 5)$ as the local feature vector η, as shown in Equation (8):

$$\eta = \left\{\widehat{C}_3, \widehat{C}_4, \widehat{C}_5\right\}. \tag{8}$$

2.4. Multifeature Fusion. A survey about the PPII helix structures prediction shows that most algorithms use the traditional features and manually select features to combine. Most research works only adopt the local features [26–29] or the global features [30–33], which decreases the accuracy of PPII helix structure prediction. Both the local and long-range interactions among amino acid residues determine the PPII helix. Therefore, the local features and global features are equally important in prediction the PPII helix. In this paper, we propose to fuse the protein sequences' local features η and the global features CLS, and the joint feature in Equation (9) is used to predict the PPII helix structure:

$$M = \mathrm{concat}(CLS, \eta). \tag{9}$$

The global feature CLS is obtained by the BERT model, and the local feature η is obtained by the multichannel CNN. We utilize the concat() algorithm to generate the final feature vector $M = \{CLS, \eta\}$. In this paper, we use the fusion feature M to predict the PPII helix structure.

3. Results and Discussion

3.1. Sample and Dataset. In this paper, we design the comparative experiments on the PPIIPRED dataset [30]. The filtering rules which define the PPII helix dataset [41] include two kinds of definitions: the "strict" and "less strict." The filter criteria are percentage identity ≤30%, resolution ≤2.5, and R-value ≤0.25. The strict criteria include the trans filtering, the dihedral filtering, and the regularization filtering.

The trans filtering:

$$-145 < \alpha C - 70. \tag{10}$$

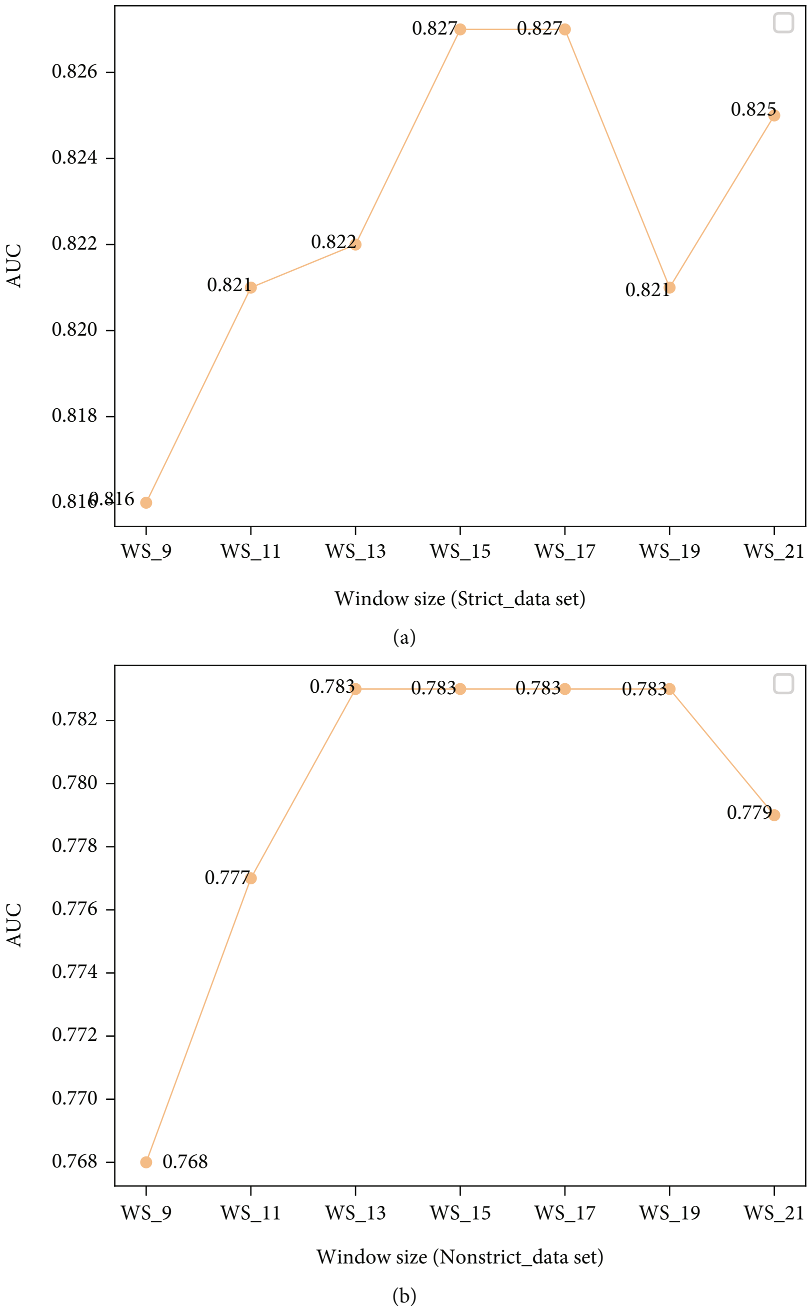


Figure 6: (a) The ROC of BERT-PPII model with different sliding window sizes on the balanced Strict_data test set, WS_9 means that the number of amino acid residues is 9. (b) The ROC plots of BERT-PPII model with different sliding window sizes on the balanced NonStrict_data test set.

The dihedral filtering:

$$-180 < \Psi < -160, \tag{11}$$

$$90 < \Psi < 180, \tag{12}$$

$$-105 < \Phi < -45. \tag{13}$$

The regularization filtering:

$$\frac{\sum_{k=1}^{n-1} d_{k,k+1}}{n}, \tag{14}$$

$$d_{k-1,k} = \sqrt{(\Psi_{i\text{-}1} - \Psi_i)^2 + (\Phi_i - \Phi_{i+1})^2}. \tag{15}$$

Compared with the strict definition, the less strict definition removes the requirement: $-105 < \Phi < -45$. Based on the

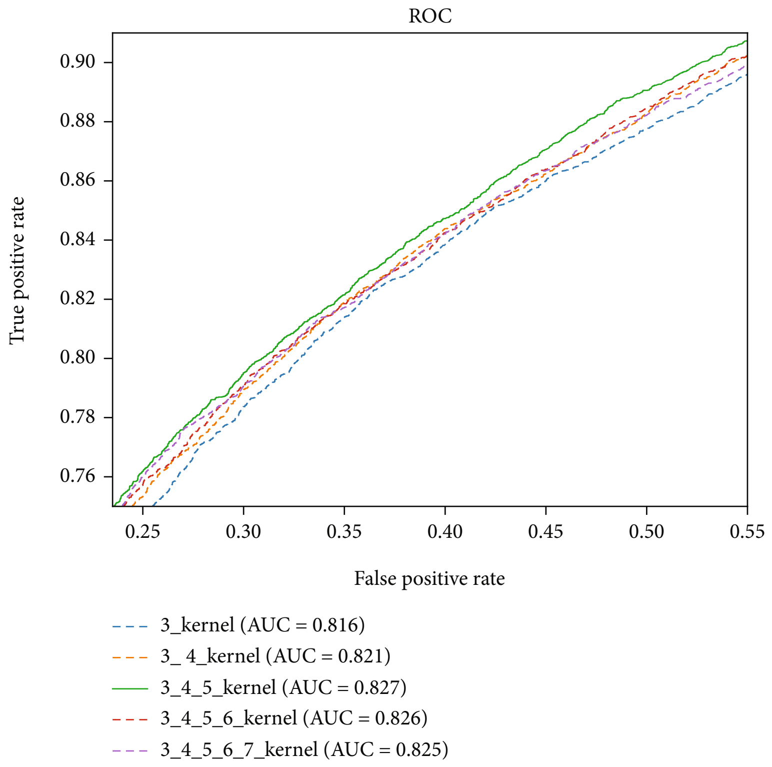


Figure 7: The ROC plots of the Multichannel CNN model with different integration of n-gram channels on the balanced Strict_data test set.

Table 3: The Comparative experiments of the BERT-PPII with different n-gram channel combinations on a balanced Strict_data test set.

Dataset	Window size	Sens	Spec	MCC	ACC
Stirct_data	3_kernel	0.636	0.846	0.491	0.741
	3_4_kernel	0.644	0.847	0.501	0.745
	3_4_5_kernel	0.661	0.841	0.510	0.751
	3_4_5_6_kernel	0.610	0.871	0.498	0.741
	3_4_5_6_7_kernel	0.640	0.854	0.510	0.747

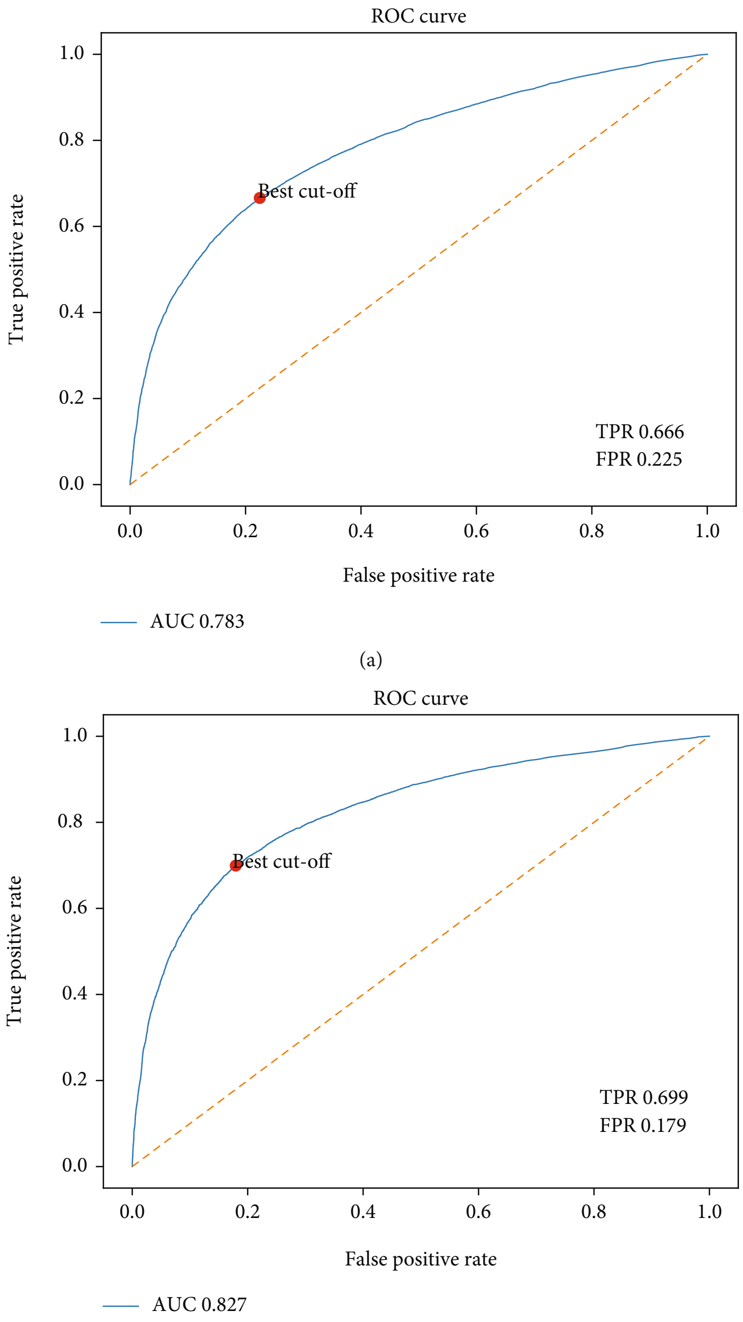
above the definitions of the strict and less strict, we obtained the strict and less strict PPII helix structure datasets.

We used the sliding window technique [34] to select sequences as input samples. Assuming a protein sequence of length L, we can obtain $2\,m + 1$ protein sequence fragment to represent a single amino acid sample. So, the number of samples is L. Given the sliding window size is 13, the positive samples (PPII helix structure) and negative samples (non-PPII helix structure) are shown as in Figures 4(a) and 4(b).

For the problem of protein secondary structure identification, we predict the PPII helix based on sample center residues, since the prediction results relate to the information of the neighbor amino acid residues. The datasets processed by the sliding window are divided into training sets, validation sets, and test sets. Table 1 is the dataset under strict definition (Strict_data), and Table 2 is the dataset under less strict definition (NonStrict_data).

To solve the serious imbalance problem between positive and negative samples, we employ the under-sampling method to randomly select the same number of negative samples as the positive samples in the original training data. We utilize both the negative samples and the positive samples as the training data. Furthermore, the training data is divided into training set and validation set, and their ratio is 4 : 1. The training set, the validation set, and the test set form a balanced dataset.

3.2. Analysis of Amino Acid Composition. We investigate the PPII helix structure and the non-PPII helix structure according to the relative frequency of the amino acid residues located in the center position of the PPII helix. In this study, the relative frequency of the various amino acids in the dataset is shown in Figure 5. It shows that A, E, L, and P are the amino acids in the PPII helical structure. A, G, L, and V are the main amino acids in the non-PPII helix structure. Compared with the non-PPII helix structure, amino acid P appears more frequently. Except the Proline (P), the other amino acids have no obvious characteristic in these two kinds of structure. The relative frequencies of the P in the middle of the PPII helix structure is about five



(a)

(b)

Figure 8: (a) The ROC plot of the BRET-PPII model on the Strict independent test set; (b) The ROC plot of the BERT-PPII model on the NonStrict independent test set. TPR represents the rate that is correctly judged to be positive, and FPR represents the rate that is wrongly judged to be positive.

times more than that in the middle of the non-PPII helix structure. Therefore, P can distinguish the PPII helix structure and the non-PPII helix structure effectively. Although P accounts for a large proportion, not all PPII helical structures contain P.

3.3. Evaluation Criteria. In this study, we adopt four commonly used metrics including sensitivity (Sens), specificity (Spec), accuracy (ACC) and Matthews correlation coefficient (MCC) to evaluate the performance. Their definitions are shown as follows:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{16}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{17}$$

Table 4: The comparative experiments on balanced Strict_data dataset.

Methods	Sens	Spec	MCC	ACC	AUC
ANN [25]	0.749	0.736	0.485	0.742	0.742
SVM [26]	0.673	0.841	0.493	0.744	0.822
RF	0.738	0.841	0.554	0.776	0.776
KNN	0.558	0.739	0.302	0.648	0.648
FAD-BERT [19]	0.660	0.821	0.492	0.741	0.752
EECL [10]	0.765	0.776	0.540	0.770	0.770
Adapt_Kcr [40]	0.792	0.767	0.559	0.779	0.855
BERT4Bitter [20]	0.661	0.825	0.493	0.744	0.762
OUR	**0.661**	**0.838**	**0.198**	**0.834**	**0.826**

Table 5: The comparative experiments on balanced NonStrict_data dataset.

Methods	Sens	Spec	MCC	ACC	AUC
ANN [25]	0.701	0.734	0.435	0.717	0.742
SVM [26]	0.629	0.789	0.423	0.709	0.822
RF	0.681	0.810	0.490	0.746	0.746
KNN	0.636	0.639	0.275	0.637	0.648
FAD-BERT [19]	0.581	0.797	0.411	0.732	0.733
EECL [10]	0.748	0.724	0.472	0.736	0.736
Adapt_Kcr [40]	0.751	0.736	0.487	0.744	0.823
BERT4Bitter [20]	0.590	0.798	0.397	0.695	0.743
OUR	**0.559**	**0.833**	**0.219**	**0.824**	**0.826**

Table 6: The comparative experiments with on unbalanced Strict_data dataset and NonStrict_data dataset.

Dataset	Methods	Sens	Spec	MCC	ACC
Strict_data	PPIIPRED	0.38	0.98	0.37	0.971
	OUR	0.30	0.99	0.44	0.980
NonStrict_data	PPIIPRED	0.43	0.97	0.38	0.949
	OUR	**0.30**	**0.99**	**0.43**	**0.966**

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{18}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{19}$$

Sensitivity represents the proportion of the positive samples which are correctly predicted. Specificity represents the proportion of the negative samples which are correctly predicted. ACC indicates the proportion of correctly classified samples; MCC represents the correlation coefficient between the observed category and the predicted binary classification. Its range is [−1,1]. We will get a better prediction result, when the MCC value is close to 1. TP represents the true positive. It is the number of positive samples correctly predicted. TF represents the true negative. It is the number of negative samples correctly predicted. FP represents the false positive. It is the number of negative samples incorrectly predicted. FN represents the false negative. It is the number of positive samples incorrectly predicted. AUC is the area under the ROC curve. We evaluate the generalization performance of the algorithm model based on AUC, and the value of a robust model is close to 1.

3.4. Optimal Sliding Window. To obtain the optimal window, we set up comparison experiments to measure the prediction performance with different windows. In this experiment, the step length is 2, and its value range is [11, 21]. The ROC of BERT-PPII model on the balanced Strict_data dataset and the NonStrict_data dataset is shown in Figures 6(a) and 6(b), respectively. Figure 6(a) shows that the model has the best performance with the window size of [15, 17] and the AUC is 0.827. Figure 6(b) shows that the model has the best performance with the window size of [13, 19] and the AUC is 0.783. Usually, the training time increases when the window size becomes. As a result, we set the window size as 15.

3.5. The Optimal Convolutional Kernel Combinations. To obtain the optimal channel number, we design the comparative methods combined with different *n*-gram channels as follows:

(1) 3_kernel: contains 3-gram CNN channels

(2) 3_4_kernel: a combination of 3-gram and 4-gram CNN channels

(3) 3_4_5_kernel: a combination of 3-gram, 4-gram, and 5-gram CNN channels

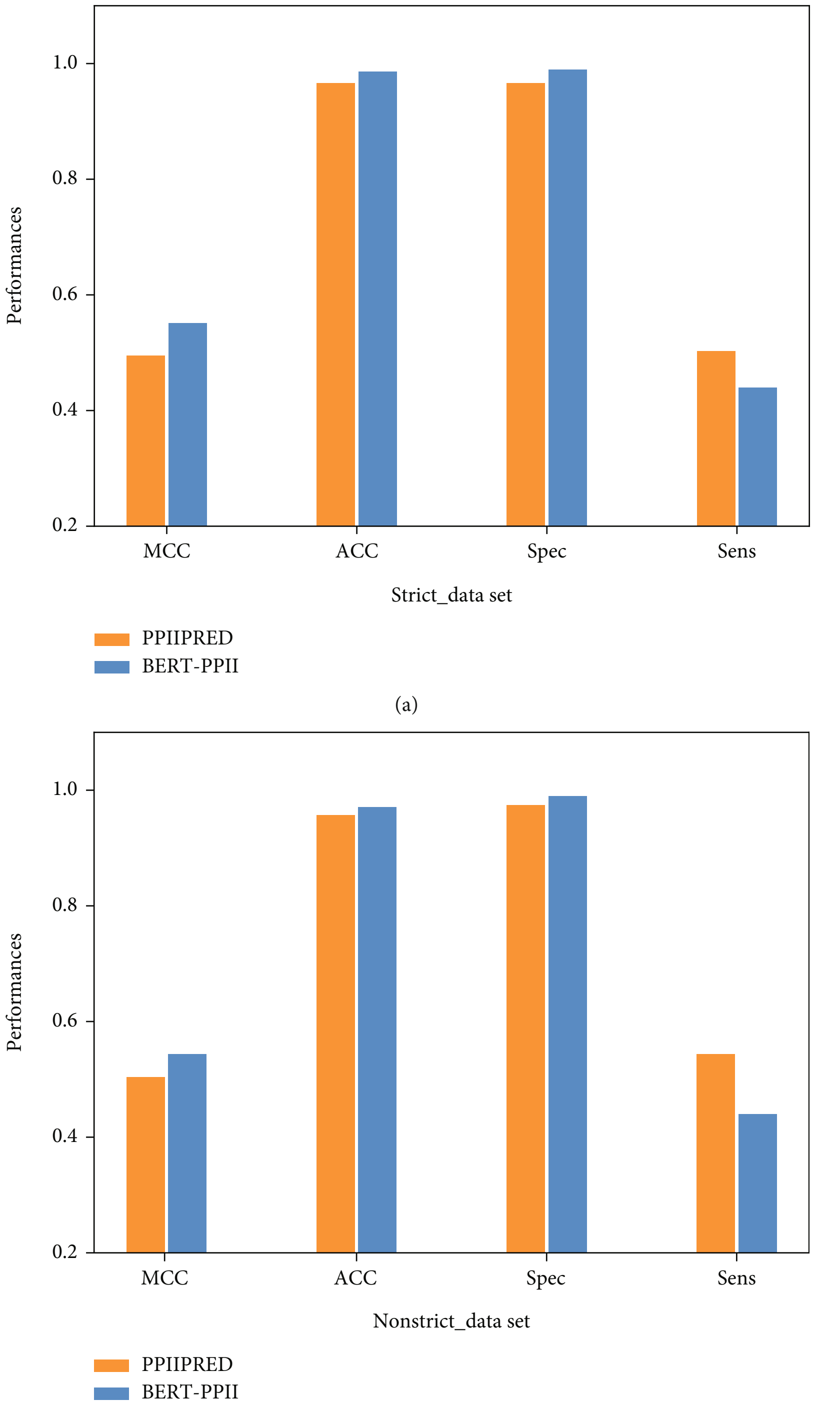
(4) 3_4_5_6_kernel: a combination of 3-gram, 4-gram, 5-gram, and 6-gram CNN channels

(5) 3_4_5_6_7_kernel:a combination of 3-gram, 4-gram, 5-gram, 6-gram, and 7-gram CNN channels

We test these five methods on the balanced Strict_data dataset. The ROC curves are shown in Figure 7, and other performances are shown in Table 3. The experimental results show that the 3_4_5_kernel method has the best performance, in the range of [0.2,0.7] of FPR and the range of [0.7,1.0] of TPR, which is the most meaningful part for performance comparison. We use the 3_4_5_kernel method in the following experiments.

3.6. Predictive Performance Experiments on an Independent Test Set. To further validate the generalization performance, we conduct the experiments on the independent Strict_data dataset and Nonstrict_data dataset. The ROC curves are shown in Figures 8(a) and 8(b). The AUC value of the BERT-PPII model is 0.827 on the independent Strict_data dataset, and the value is 0.783 on the independent NonStrict_data dataset.

3.7. The Comparative Experiments. In this paper, we compare BERT-PPII method with the following methods. To predict PPII helices on a balanced dataset, Siermala et al.



(a)

(b)

Figure 9: (a) Performance comparison between our algorithm and PPIIPRED on (a) Strict_data dataset and (b) NonStrict_data dataset, respectively.

[25] employs an artificial neural network (ANN), and Wang et al. [26] adopt a support vector machine (SVM). In contrast, O'Brien KT [30] proposed the PPIIPRED model, and it predicts PPII helix using a bidirectional recurrent neural network (BRNN) on an unbalanced dataset. We conduct the comparative experiments on both the balanced and unbalanced datasets, respectively. The experimental results are shown in Sections 3.7.1 and 3.7.2, respectively.

3.7.1. The Comparative Experiments on a Balanced Dataset. This section conducts the comparative experiments on the balanced dataset and the comparative methods including

ANN [25], SVM [26], random forest (RF), K-Nearest Neighbor (KNN), FAD-BERT [19], EECL [10], Adapt_Kcr [40], and BERT4Bitter [20]. All comparative methods use one-hot to encode the amino acid residues. The evaluation metrics are shown in Tables 4 and 5. On the dataset Strict dataset, compared to the best performing support vector machine algorithm (SVM), the BERT-PPII model improved the ACC value by 9.0% and the AUC by 0.4%, as shown in Table 4. On the NonStrict dataset, compared to the best performing support vector machine (SVM), the BERT-PPII model improved the ACC value by 11.5% and the AUC by 0.4%, as shown in Table 5. The BERT-PPII model has the best performance in predicting the PPII helix.

3.7.2. The Comparative Experiments on an Unbalanced Dataset. PPIIPRED model [30] uses a bidirectional recurrent neural network (BRNN) to predict the PPII helix, and we employ PPIIPRED model as the comparative method on the unbalanced dataset. We divide the unbalanced dataset (Strict_data, NonStrict_data) into training set, validation set and test set, and their ratio is 3 : 1 : 1. The experimental result is shown in Table 6 and Figure 9, and its shows that our model outperforms PPIIPRED in predicting the PPII helix. On the Strict_data dataset, the Spec, MCC, and ACC values of the proposed method are 0.99, 0.44, and 0.980, respectively. Compared to the PPIIPRED method, the values of Spec, MCC, and ACC have been improved about 1%, 7%, and 1%, respectively. On the NonStrict_data dataset, the Spec, MCC, and ACC values of the proposed method are 0.99, 0.43, and 0.966, respectively. Compared to the PPII PRED method, the values of Spec, MCC, and ACC have been improved about 2%, 5% and 1.7%, respectively. The above experiments show that our method can achieve the best performance in predicting the PPII helix structure.

4. Conclusions

The PPII helix plays a very important role in many biochemical processes, and it is necessary to quickly and accurately predict the PPII helix. However, it is a time-consuming and expensive work to identify PPII helix using traditional physical and chemical experimental methods. In this study, to some extent, protein sequences also have their own arrangement motifs, which constitute the structure of proteins in space and function in organisms. Due to the protein sequences are similar to the natural language, we can apply the natural language technology to the area of protein sequences. We propose a new model BERT-PPII to identify the PPII helix. The BERT-based BERT-PPII model automatically generates the feature descriptors according to the original amino acid sequence, and it does not need any system design and feature coding selection. We use BERT encoding mechanism to generate the *CLS* vector as the protein sequence feature and fuse it and the CNN local feature vector to enhance feature expression. A large number of experiments have shown that BERT-PPII achieves a better performance than the existing methods. In particular, our method is better than the PPIIPRED on the strict dataset. The ACC value of our method is 1% higher than that of PPIIPRED on the unbalanced datasets. Accuracy (ACC) is 2% higher than PPIIPRED on less stringent datasets. The high prediction performance of our model BERT-PPII enables it to provide robust performance and distinguish between PPII helix and non-PPII helix.

Data Availability

The data that support the findings of this study are openly available at https://github.com/Cambridge-F/BERT-PPII.git.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was funded by National Natural Science Foundation of China, grant number 61841602; the Natural Science Foundation of Shandong Province of China, grant numbers ZR2018PF005 and ZR2021MF017; the Youth Innovation Science and Technology Team Foundation of Shandong Higher School, grant number 2021KJ031; and the Fundamental Research Funds for the Central Universities, JLU, grant number 93K172021K12. The authors express their gratitude to the institutions that supported this research: Shandong University of Technology (SDUT) and Jilin University (JLU).

References

[1] P. M. Cowan, S. McGavin, and A. C. North, “The polypeptide chain configuration of collagen,” *Nature*, vol. 176, no. 4492, pp. 1062–1064, 1955.

[2] T. Ohgita, Y. Takechi-Haraya, K. Okada et al., “Enhancement of direct membrane penetration of arginine-rich peptides by polyproline II helix structure,” *Biochimica et Biophysica Acta - Biomembranes*, vol. 2020, no. 10, p. 183403, 2020, Epub 2020 Jun 23.

[3] A. Maaßen, J. M. Gebauer, E. Theres Abraham et al., “Triple-helix-stabilizing effects in collagen model peptides containing PPII-helix-preorganized diproline modules,” *Angewandte Chemie (International Ed. in English)*, vol. 59, no. 14, pp. 5747–5755, 2020.

[4] P. Zhou, S. Hou, Z. Bai et al., “Disrupting the intramolecular interaction between proto-oncogene c-Src SH3 domain and its self-binding peptide PPII with rationally designed peptide ligands,” *Artif Cells Nanomed Biotechnol.*, vol. 46, no. 6, pp. 1122–1131, 2018.

[5] J. R. Arndt, M. Chaibva, M. Beasley et al., “Nucleation inhibition of huntingtin protein (htt) by polyproline PPII helices: a potential interaction with the N-terminal α-helical region of Htt,” *Biochemistry*, vol. 59, no. 4, pp. 436–449, 2020.

[6] M. Mompeán, J. Oroz, and D. V. Laurents, “Do polyproline II helix associations modulate biomolecular condensates?,” *FEBS Open Bio*, vol. 11, no. 9, pp. 2390–2399, 2021.

[7] W. Niu, L. Xu, J. Li et al., “Polyphyllin II inhibits human bladder cancer migration and invasion by regulating EMT-associated factors and MMPs,” *Oncology Letters*, vol. 20, no. 3, pp. 2928–2936, 2020.

[8] A. A. Adzhubei, A. A. Anashkina, and A. A. Makarov, "Left-handed polyproline-II helix revisited: proteins causing proteopathies," *Journal of Biomolecular Structure & Dynamics*, vol. 35, no. 12, pp. 2701–2713, 2017.

[9] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr., "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 47, no. 9, pp. 1309–1314, 1961.

[10] Y. H. Qu, H. Yu, X. J. Gong, J. H. Xu, and H. S. Lee, "On the prediction of DNA-binding proteins only from primary sequences: a deep learning approach," *PLoS One*, vol. 12, no. 12, article e0188129, 2017.

[11] Y. Liu, P. Palmedo, Q. Ye, B. Berger, and J. Peng, "Enhancing evolutionary couplings with deep convolutional neural networks," *Cell Systems*, vol. 6, no. 1, pp. 65–74.e3, 2018.

[12] X. Pan and H. B. Shen, "Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks," *Bioinformatics*, vol. 34, no. 20, pp. 3427–3436, 2018, PMID: 29722865.

[13] D. Ofer, N. Brandes, and M. Linial, "The language of proteins: NLP, machine learning & protein sequences," *Computational and Structural Biotechnology Journal*, vol. 19, no. 19, pp. 1750–1758, 2021.

[14] A. Wahab, H. Tayara, Z. Xuan, and K. T. Chong, "DNA sequences performs as natural language processing by exploiting deep learning algorithm for the identification of N4-methylcytosine," *Scientific Reports*, vol. 11, no. 1, p. 212, 2021.

[15] S. M. Yusuf, F. Zhang, M. Zeng, and M. Li, "DeepPPF: a deep learning framework for predicting protein family," *Neurocomputing*, vol. 428, pp. 19–29, 2021.

[16] X. Pan and H.-B. Shen, "Learning distributed representations of RNA sequences and its application for predicting RNA-protein binding sites with a convolutional neural network," *Neurocomputing*, vol. 305, pp. 51–58, 2018.

[17] S. Seo, M. Oh, Y. Park, and S. Kim, "DeepFam: deep learning based alignment-free method for protein family modeling and prediction," *Bioinformatics*, vol. 34, no. 13, pp. i254–i262, 2018.

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," 2018, https://arxiv.org/abs/1810.04805.

[19] Q. T. Ho, T. T. Nguyen, N. Q. Khanh Le, and Y. Y. Ou, "FAD-BERT: improved prediction of FAD binding sites using pre-training of deep bidirectional transformers," *Computers in Biology and Medicine*, vol. 131, article 104258, 2021.

[20] P. Charoenkwan, C. Nantasenamat, M. M. Hasan, B. Manavalan, and W. Shoombuatong, "BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides," *Bioinformatics*, vol. 26, article btab133, 2021.

[21] K. Li, Y. Zhong, X. Lin, and Z. Quan, "Predicting the disease risk of protein mutation sequences with pre-training model," *Frontiers in Genetics*, vol. 11, no. 11, article 605620, 2020.

[22] S. M. Ali Shah, S. W. Taju, Q. T. Ho, T. T. Nguyen, and Y. Y. Ou, "GT-finder: classify the family of glucose transporters with pre-trained BERT language models," *Computers in Biology and Medicine*, vol. 131, article 104259, 2021.

[23] N. Q. K. Le, Q. T. Ho, T. T. Nguyen, and Y. Y. Ou, "A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information," *Briefings in Bioinformatics*, vol. 22, no. 5, p. -bbab005, 2021, PMID: 33539511.

[24] L. Zhang, X. Qin, M. Liu, G. Liu, and Y. Ren, "BERT-m7G: a transformer architecture based on BERT and stacking ensemble to identify RNA N7-Methylguanosine sites from sequence information," *Computational and Mathematical Methods in Medicine*, vol. 2021, 7764710 pages, 2021.

[25] M. Siermala, M. Juhola, and M. Vihinen, "On preprocessing of protein sequences for neural network prediction of polyproline type II secondary structures," *Computers in Biology and Medicine*, vol. 31, no. 5, pp. 385–398, 2001.

[26] M. L. Wang, H. Yao, and W. B. Xu, "Prediction by support vector machines and analysis by Z-score of poly-L-proline type II conformation based on local sequence," *Computational Biology and Chemistry*, vol. 29, no. 2, pp. 95–100, 2005.

[27] K. Z. Lu and W. B. Xu, "Support vector machine for prediction of polyproline type II secondary structures [J]," *China Journal of Bioinformatics*, vol. 1, pp. 26–29, 2005, (in Chinese).

[28] K. Z. Lu, Y. G. Hu, and W. B. Xu, "Prediction of Polyproline type II secondary structures by neural network based on genetic algorithm[J]," *Journal of Southern Yangtze University (Natural Science Edition)*, vol. 4, no. 3, pp. 244–247, 2005, (in Chinese).

[29] K. Z. Lu and W. B. Xu, "Prediction of polyproline binary structure based on improved coding [J]," *Journal of Chizhou Teachers College*, vol. 20, no. 5, pp. 11–13, 2006.

[30] K. T. O'Brien, C. Mooney, C. Lopez, G. Pollastri, and D. C. Shields, "Prediction of polyproline II secondary structure propensity in proteins," *Royal Society Open Science*, vol. 7, no. 1, article 191239, 2020.

[31] Y. Liu, W. Gong, Z. Yang, and C. Li, "SNB-PSSM: a spatial neighbor-based PSSM used for protein–RNA binding site prediction," *Journal of Molecular Recognition*, vol. 34, no. 6, article e2887, 2021Epub 2021 Jan 14.

[32] Y. Guo, J. Wu, H. Ma, S. Wang, and J. Huang, "EPTool: a new enhancing PSSM tool for protein secondary structure prediction," *Journal of Computational Biology*, vol. 28, no. 4, pp. 362–364, 2021.

[33] S. Wang, M. Li, L. Guo, Z. Cao, and Y. Fei, "Efficient utilization on PSSM combining with recurrent neural network for membrane protein types prediction," *Computational Biology and Chemistry*, vol. 81, pp. 9–15, 2019.

[34] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, no. 4, pp. 865–884, 1988.

[35] R. Chebrek, S. Leonard, A. G. de Brevern, and J. C. Gelly, "PolyprOnline: polyproline helix II and secondary structure assignment database," *Database: The Journal of Biological Databases and Curation*, vol. 2014, article bau102, 2014.

[36] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition[C]*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[37] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning.*, vol. 2015, pp. 448–456, 2015.

[38] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?," *Advances in Neural Information Processing Systems*, vol. 31, pp. 2483–2493, 2018.

[39] A. Vaswani, N. Shazeer, N. Parmar et al., “Attenttion is all you need[C],” *Proceedings of the 31st International Conference on Neural Information Processing*, vol. 30, pp. 6000–6010, 2017.

[40] Z. Li, J. Fang, S. Wang, L. Zhang, Y. Chen, and C. Pian, “Adapt-Kcr: a novel deep learning framework for accurate prediction of lysine crotonylation sites based on learning embedding features and attention architecture,” *Briefings in Bioinformatics*, vol. 23, no. 2, article bbac037, 2022.

[41] A. A. Adzhubei and M. J. Sternberg, “Left-handed polyproline II helices commonly occur in globular proteins,” *Journal of Molecular Biology*, vol. 229, no. 2, pp. 472–493, 1993.

Hindawi
BioMed Research International
Volume 2022, Article ID 8360481, 10 pages
https://doi.org/10.1155/2022/8360481

Research Article

Prognostic Values of BolA Family Member Expression in Hepatocellular Carcinoma

Dong Wang,[1] ZhiMing Wang,[2] and YiMing Tao[2]

[1] *Department of Liver Disease Center, The Affiliated Hospital of Qingdao University, Qingdao, Shandong, China*
[2] *Department of General Surgery, Xiangya Hospital, Central South University, Changsha, Hunan, China*

Correspondence should be addressed to YiMing Tao; yimingtao@csu.edu.cn

Received 21 June 2022; Revised 21 July 2022; Accepted 25 July 2022; Published 16 August 2022

Academic Editor: Jian Zhang

The BolA gene family member (BOLA1–3) plays an important role in regulating normal and pathological biological processes including liver tumorigenesis. However, their expression patterns as prognostic factors in hepatocellular carcinoma (HCC) patients have not to be elucidated. We examined the transcriptional expressions and survival data of BolA family member in patients with HCC from online databases including ONCOMINE, TCGA, UALCAN, Gene Expression Profiling Interactive Analysis (GEPIA), Kaplan-Meier plotter, SurvExpress, cBioPortal, and Exobase. Network molecular interaction views of BolA family members and their neighborhoods were constructed by the IntAct web server. In our research, we had found that the expression levels of BolA /2/3 mRNA were higher in HCC tissue than in normal liver tissues from TGCA databases. Moreover, the BolA family gene expression level is significantly associated with distinct tumor pathological grade, TMN stage, and overall survival (OS). The BolA family can be considered as prognostic risk biomarkers of HCC. A small number of BolA gene-mutated samples were detected in the HCC tissue. IntAct analysis revealed that BolA1/2/3 was closely associated with the GLRX3 expression in HCC, which is implicated in the regulation of the cellular iron homeostasis and tumor growth. Furthermore, prognostic values of altered BolAs and their neighbor GLRX3 gene in HCC patients were validated by SurvExpress analysis. In conclusion, the membrane BolA family identified in this study provides very useful information for the mechanism of hepatic tumorigenesis.

1. Introduction

Hepatocellular carcinoma (HCC) has very aggressive neoplasms and describes as a major health problem worldwide [1]. Genetic and epigenetic alterations, which lead to uncontrolled cellular proliferation and metastasis, are the characters of HCC development.

Recent research has revealed a critical role for cellular iron homeostasis in the clinical context of liver tumorigenesis [2, 3]. Although significant progress has been made in understanding the iron homeostasis disruption associated with HCC, the precise molecular signals that trigger initiation and progression of HCC remain to be identified.

The human BolA gene family consists of BOLA1, BOLA2, and BOLA3 [4]. It has been suggested that BolA family members serve as assembly factors for mitochondrial iron-sulfur (Fe/S) cluster proteins that has involvement in cancer cell biology [5, 6]; although, the functions of BOLA1 and BOLA3 are still undefined in cancer. Prior research has highlighted the importance role of BOLA3 in human endothelial metabolism and cardiovascular disease pathogenesis [7]. More specifically, evidence points out that BOLA2 has been shown to be highly correlated with hepatic iron homeostasis [8]. And yet, even the overexpression of BOLA2 is required to drive HCC tumor growth and tumor hemorrhage [9, 10], and high BOLA2 can promote tumor growth and predict the HCC prognosis [11]. BOLA1 plays a leading role in mitochondrial morphology by potential regulation and can induce diseases [12]. In the ovarian cancer, the BOLA2 and BOLA3 were higher in cancer tissues and may act as prognostic biomarkers [13], and in the lung adenocarcinoma, the BOLA3 was correlated with the immune cell infiltrates [14]. However, it has been poorly characterized whether the expression of BolA family members in HCC is correlated with clinical outcomes.

In our research, we analyzed the BolA family member mRNA level in HCC tissues and nontumor liver tissues by the public database. In addition, we investigated correlation between their expressions and clinical characteristics and performed SurvExpress analysis of prognostic risks for overall survival. The results showed that BOLA1\2\3 may be a promising biomarker for the prognosis in HCC.

2. Material and Methods

2.1. ONCOMINE Database Analysis. The difference mRNA expression level of the BolA family gene in human cancer was identified in the ONCOMINE online microarray database (http://www.oncomine.org). For each BolA family gene, the thresholds were set as the following values: P value of 0.01, fold change of 2, and gene ranking of all. Analysis type was set as follows: cancer vs. normal analysis.

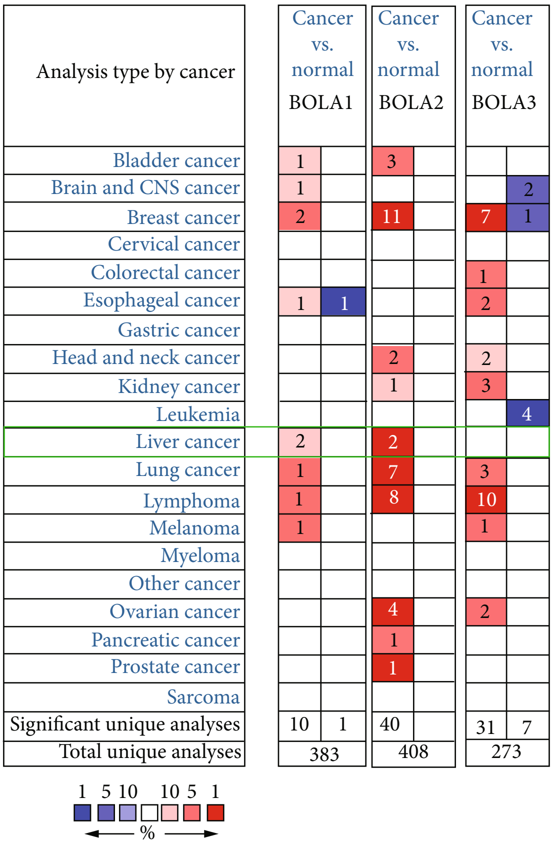
2.2. UALCAN Database Analysis. The UALCAN online database (http://ualcan.path.uab.edu) was used to calculate the BolA gene expression level and clinicopathologic parameters in the TCGA database on patient with LIHC (liver hepatocellular carcinoma) [15] .

2.3. cBioPortal and Exobase Database Analysis. The cBio Cancer Genomics Portal (http://www.cbioportal.org/) performed to estimate the cancer genomics data sets of BolA family gene using TCGA-LIHC data [16]. The exoRBase database (http://www.exoRBase.org) can analysis the human blood exosomes, including circRNA, lncRNA, and mRNA [17].

2.4. GEPIA Database Analysis. Gene Expression Profiling Interactive Analysis (GEPIA) web server (http://gepia.cancer-pku.cn/) was used to study the correlation mRNA expression of BolA family members and overall survival (OS) in LIHC [18]. A total of 331 LIHC patients were enrolled, and “median” was regarded as group cutoff value.

2.5. Kaplan-Meier Plotter Analysis. The Kaplan-Meier (KM) plotter database (http://kmplot.com/analysis) was used to calculate the survival time in LIHC patients [19]. Briefly, each BolA family member was individually analyzed to obtain KM plots. Group cutoff was set as “median.” Hazard ratios (HR) with 95% confidence intervals (CI) were extracted from the KM plotter webpage. Overall survival (OS) data from 364 patients with HCC were enrolled.

2.6. SurvExpress Database Analysis. SurvExpress (http://bioinformatica.mty.itesm.mx/SurvExpress) was used for obtaining survival data for the expression of BolA family members in patient with LIHC, for which information was not available on the GEPIA and KM plotter database [20]. Briefly, in the TCGA-LIHC datasets containing 381 samples, BOLA1, BOLA2, and BOLA3 were entered into the number-at-risk cases, median mRNA expression levels, HRs, 95% confidence interval (CI), and P values that were displayed.



Analysis type by cancer	Cancer vs. normal BOLA1		Cancer vs. normal BOLA2		Cancer vs. normal BOLA3	
Bladder cancer	1		3			
Brain and CNS cancer	1					2
Breast cancer	2		11		7	1
Cervical cancer						
Colorectal cancer					1	
Esophageal cancer	1	1			2	
Gastric cancer						
Head and neck cancer			2		2	
Kidney cancer			1		3	
Leukemia						4
Liver cancer	2		2			
Lung cancer	1		7		3	
Lymphoma	1		8		10	
Melanoma	1				1	
Myeloma						
Other cancer						
Ovarian cancer			4		2	
Pancreatic cancer			1			
Prostate cancer			1			
Sarcoma						
Significant unique analyses	10	1	40		31	7
Total unique analyses	383		408		273	

Figure 1: Transcriptional expression of BolA family members in 20 different types of cancer types (ONCOMINE database). Notes: the BOLA1\2\3 mRNA expression (cancer tissue vs. normal tissue) was compared by Students’ t-test. Cut-off of change values was as follows: P value: 0.01, fold change: 1.5, gene rank: 10%, and data type: mRNA.

2.7. IntAct Database Analysis. IntAct (http://www.ebi.ac.uk/intact) was applied to identify densely connected network components and BolA family members, for which protein-protein interaction enrichment analysis data populated by either curated from the literature or from direct data depositions [21].

2.8. Western Blot Analysis. Western blot analysis was performed as previously described [11]. The antibody dilutions were 1 : 1,000 for BOLA1 polyclonal antibody (Cat. # 18017-1-AP, Proteintech), 1 : 1,000 for BOLA2 polyclonal antibody (Cat. # ab169481, Abcam), 1 : 1,000 for BOLA3 polyclonal antibody (Cat. # ab185339, Abcam), and 1 : 5,000 for the β-actin mouse monoclonal antibody (Sigma-Aldrich, Cat. # A1978).

3. Results

3.1. BolA Family Members Are Frequently Upregulated in HCC. In order to analysis the expression differences of the BolA family, we first performed an analysis using the ONCOMINE database to investigate differences in the mRNA levels of each BolA family in cancers. As shown in Figure 1, the number of the upregulation BOLA1\2\3

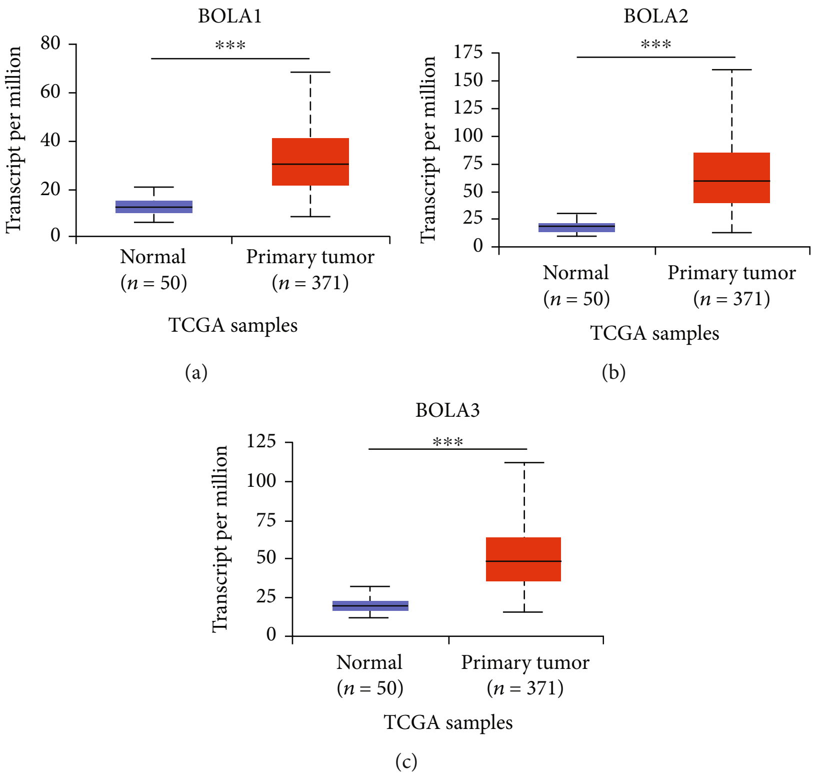


Figure 2: The BOLA1\2\3 mRNA expressions in HCC and adjacent nontumorous tissues (UALCAN database). Notes: BolA family gene mRNA was higher in HCC tissues compared to nontumorous tissues. Statistically significant changes were indicated with asterisks. $^{***}P < 0.001$.

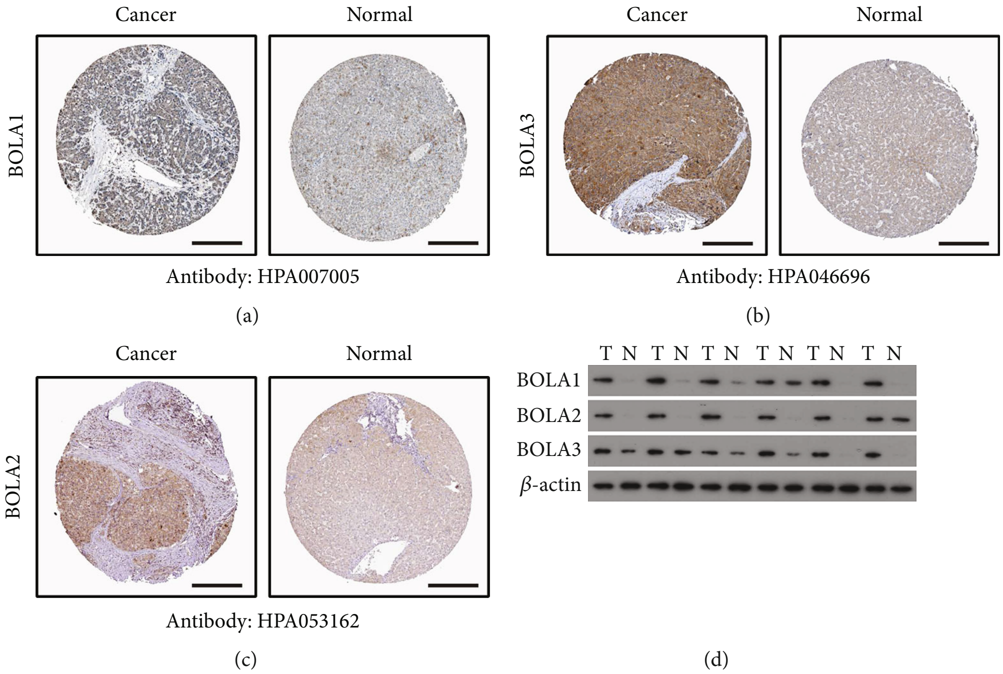
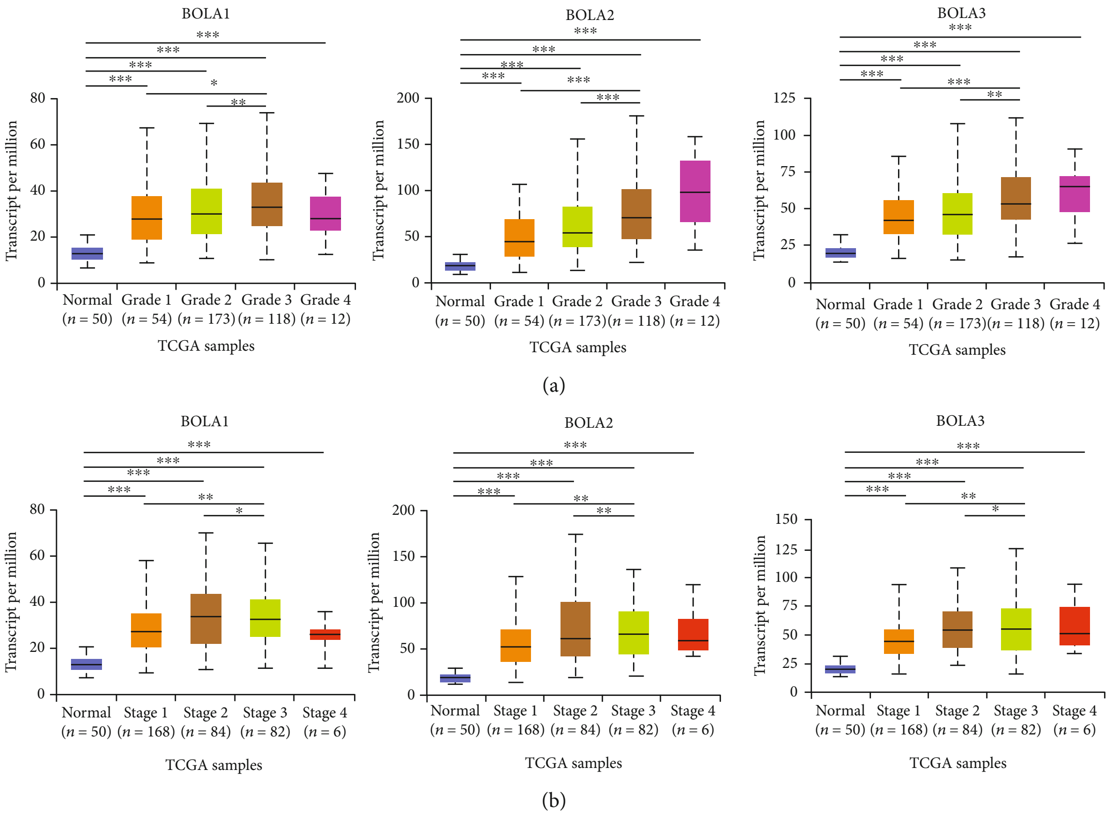


Figure 3: Protein expressions of BolA family members in human normal liver tissue and HCC (Human Protein Atlas database). Notes: BOLA1/2/3 proteins were lower in normal liver tissues than in HCC tissues. BOLA1 antibody HPA007005, BOLA2 antibody HPA046696, and BOLA3 antibody HPA053162. Scale bar is 100 μm (a)–(c). (d) The BOLA1/2/3 expression in the HCC and nontumor samples was as follows, and we had found that the BOLA1/2/3 BOLA1\2\3 expression level was higher in HCC than in the nontumor. T: tumor; NT: nontumor.



(a)

(b)

Figure 4: BolA family member mRNA expressions with clinicopathologic parameters in HCC (UALCAN database). Notes: mRNA expressions of BolA family members were significantly related to tumor grades, and as tumor grade increased, the mRAN expressions of BolAs tended to be higher (a). mRNA expressions of BolA family members were remarkably correlated with clinical stages, and patients who were in more advanced stages tended to express higher mRNA expression of BolAs (b). Data are mean ± SEM. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

expression was found in tumors compared with normal tissues in various types of cancers. Significantly higher mRNA expressions of BOLA1\2 were found in multiple HCC tissues datasets. The BOLA1\2 overexpression was found in HCC tissues compared with normal tissues in Roessler Liver 2 dataset (1.51-fold increase, $P = 3.26E - 33$; 2.67-fold increase, $P = 3.72E - 83$, respectively) [22], while were observed in Wurmbach liver dataset (1.65-fold increase, $P = 0.003$; 2.21-fold increase, $P = 4.67E - 4$, respectively) [23]. Significant upregulation of BOLA1\2 was also found in Chen Liver dataset (1.57-fold increase, $P = 1.00E - 8$; 1.56-fold increase, $P = 2.78E - 13$, respectively) [24]. The above-mentioned observations suggest that the overexpression of BOLA family members is associated with cancer progression and might be of clinical importance.

3.2. BolA Family Member Expression Was Higher in HCC. To further validate the observations made in the ONCOMINE database, TCGA-LIHC cohort performed a retrospective study. As shown in Figure 2, the BolA family expression level in HCC was higher than in the normal liver tissues ($P < 0.05$). In order to confirm this, we investigated protein levels of BolA family members by the Human Protein Atlas database (http://www.proteinatlas.org/pathology) [25]. As shown in Figure 3, BOLA1\2\3 proteins had lower level in the normal liver, while medium and high level were observed in HCC. And we also found that the BOLA1\2\3 expression level was higher in HCC than in the nontumor using our HCC samples. Human BolA proteins (BOLA1\2\3) are novel nonclassical secreted proteins [4]. In addition, a very low mutation rate of BOLA1\2\3 was observed in HCC patients (Figure S1A), the BOLA1 mutation rate was 4%, and there was no mutation in BOLA3. Intriguingly, using the Exosomes web-accessible database (http://www.exoRBase.org) analysis, the increased expression of BOLA2 may be used as circulating biomarkers for HCC patients (Figure S1B). Taken together, BOLA2 may had the potential ability for HCC diagnose.

3.3. Association between BolA Family Member and Tumor Grades and Stages. Both the mRNA and protein expression of BolA family members were found to be overexpressed in HCC; we next analyzed the relationship between mRNA expressions of each BolA family members with clinicopathological parameters of HCC patients by UALCAN. As was shown in Figure 4(a), we found that the elevated level of BOLA1\2\3 mRNA had a higher proportion of high-grade tumors (G3/G4). The BOLA1\2\3 mRNA level had significantly correlated with tumor stage in HCCs, which means that the advanced stage HCCs can express higher BolA mRNA (Figure 4(b)). The reason why mRNA expressions of BOLA1\2\3 in stage 3 seemed to be higher than that in stage 4 may be due to the small sample size (only 6 HCC patients were at stage 4). These findings indicated that the BOLA1\2\3 may accelerate HCC growth and progression.

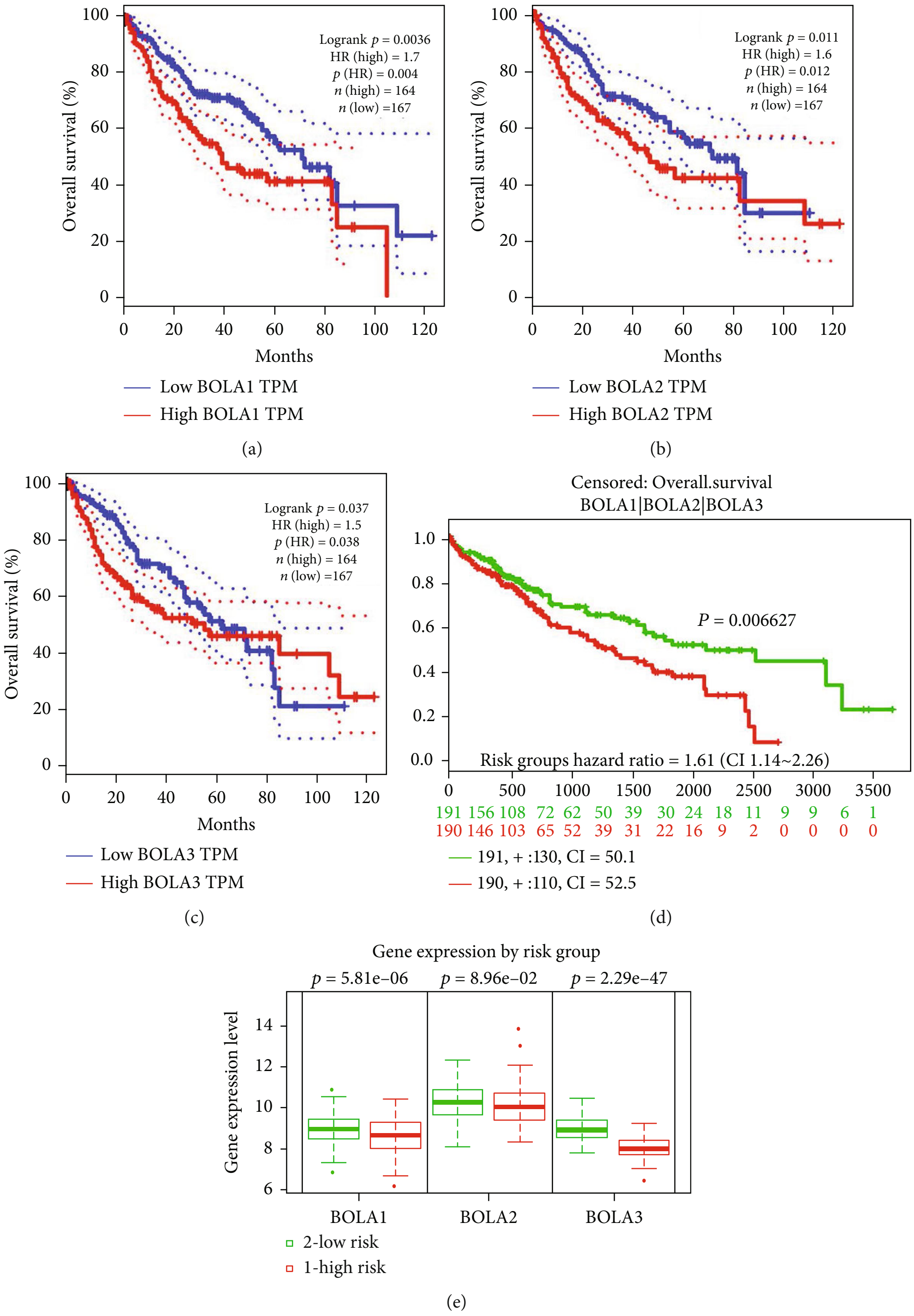


Figure 5: The BolA family members with clinical outcomes in HCC patients by Kaplan-Meier curves (GEPIA database and SurvExpress database). Notes: overall survival data of BolA family members are generated from the GEPIA web server (a)–(c). Prognostic risk of the mRNA expression of BolA family members in HCC patients (d). The concordance index and P value of log-rank testing equality of survival curves are indicated. The box plots indicate the difference in the expression of gene between risks groups, and P values are derived from t-test between both groups (e).

3.4. BolA Family Member Predicts the Prognosis in HCC Patients. We used GEPIA web server to analyze the prognostic values of BolAs in TCGA-LIHC patients. As were shown in Figure 5, upregulation of BOLA1, BOLA2, and BOLA3 were significantly associated with shorter OS (HR = 1.7, P = 0.0036; HR = 1.6, P = 0.012; HR = 1.5, P = 0.038, respectively,

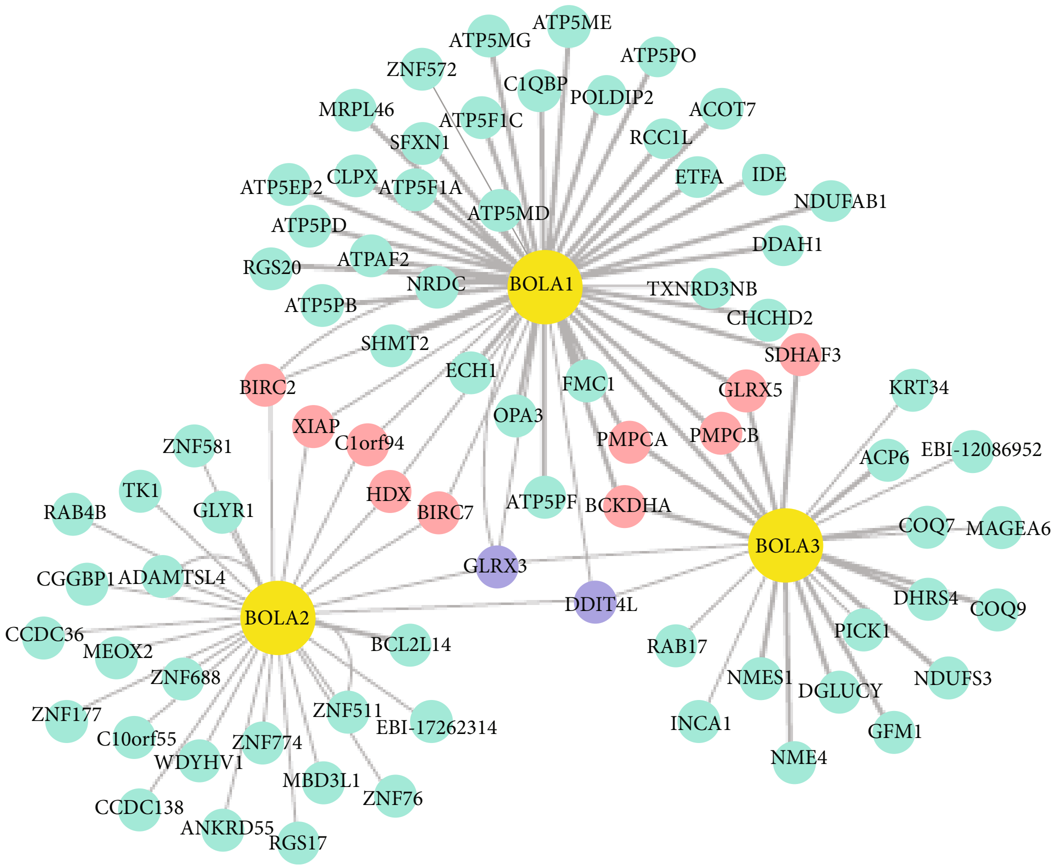


Figure 6: Network protein-protein interaction views of BolA family members and their neighborhood in TCGA-LIHC patients (IntAct database). Notes: network of molecular interaction was constructed by IntAct. Red marker indicates activation relationship of two intermediate related genes, including BIRC2, XIAP, C1orf94, HDX, BIRC7, BCKDHA, PMPCA, PMPCB, GLRX5, and SDHAF3. Purple marker indicates activation relationship of three intermediate related genes, including GLRX3 and DDIT4L.

Figures 5(a)–5(c)). The relationship between combinatory mRNA expressions of all 3 BolA family members and prognosis of liver cancer patients were further analyzed by SurvExpress. In our study, we also found that higher combinatory mRNA expressions of all 3 BolA family members were associated with poorer OS in LIHC patients ($HR = 1.61$, 95% CI: 1.14-2.26, and $P = 0.006627$, Figure 5(d)). And then, we anlayed the prognostic role of BolA family members in HCC patients.. As was shown in Figure 5(e), the higher mRNA expression of BOLA1 ($P = 5.81E - 06$), BOLA2 ($P = 8.96E - 02$), and BOLA3 ($P = 2.29E - 47$) was significantly associated with shorter OS of LIHC patients. These results indicated that mRNA expressions of BOLA1\2\3 may be exploited as useful biomarkers for prediction of HCC patient's survival.

3.5. Identification of Hub BolA Family Member and Their Clinical Value in HCC. After analyzing the genetic alterations in BolAs and their prognostic value in HCC patients, we further analyzed the protein-protein interaction network among BolAs using IntAct databases. The top hub genes were GLRX3, DDIT4L, BIRC7, HDX, C1orf94, XIAP, BIRC2, BCKDHA, PMPCA, PMPCB, GLRX5, and SDHAF3 (Figure 6). As was shown in Figure 7, Kaplan-Meier (KM) plotter survival analysis, based on clinical information from the TCGA liver cancer datasets, revealed that the low expression of BIRC2 ($HR = 0.67$, 95% CI: 0.46-0.96, and $P = 0.028$, Figure 7(c)), BCKDHA ($HR = 0.5$, 95% CI: 0.34-0.74, and $P = 0.00031$, Figure 7(d)), PMPCB ($HR = 0.69$, 95% CI: 0.49-0.99, and $P = 0.042$, Figure 7(e)), and GLRX5 ($HR = 0.7$, 95% CI: 0.5-1, and $P = 0.046$, Figure 7(f)) significantly correlated with shorter OS of LIHC patients. GLRX3 ($HR = 2.05$, 95% CI: 1.44-2.92, and $P = 4.7E - 5$, Figure 7(a)) and BIRC7 ($HR = 1.54$, 95% CI: 1.09-2.18, and $P = 0.015$, Figure 7(b)) were quite the contrary. Notably, higher combinatory mRNA expressions of BOLA2 with GLRX3 were associated with poorer OS in HCC patients ($HR = 1.56$, 95% CI: 1.1-2.22, $P = 8.1E - 4$ and $P = 2.7E - 8$, respectively, Figure 8). Many studies have investigated the expression of GLRX3 imply in regulating HCC cell proliferation, growth, and microvascular invasion via disruption of iron homeostasis [26]. Thus, we could guess that BOLA2 has the ability to promote the development of HCC and maintains cancer cell growth in the condition of metabolic stress.

4. Discussion

HCC is one of the leading causes of lethal, and there is great interest in understanding the underlying differentially expressed genes involved in the development and

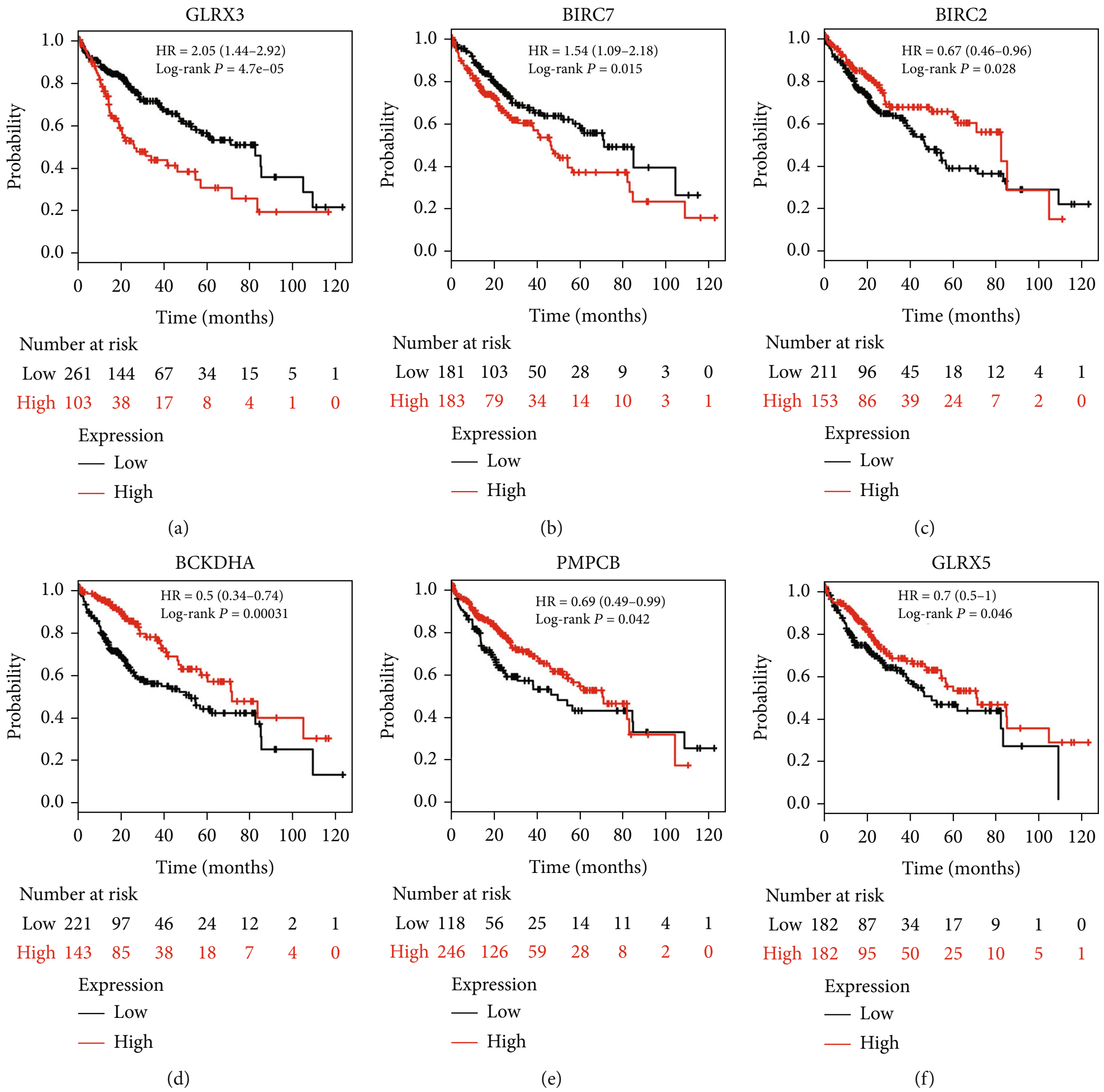


Figure 7: Prognostic values of BolA family altered neighbor genes in HCC patients (Kaplan-Meier plotter). Notes: representative altered neighbor genes of BolA family members, including GLRX3 (a), BIRC7 (b), BIRC2(c), BCKDHA (d), PMPC8 (e), and GLRX5 (f).

progression of individual tumors. In this study, we investigated the relationship between BolA family members and HCC patients using comprehensive data mining. We found that BolA family members are uniquely overexpressed in HCCs. Moreover, the mRNA expression levels of BolA family genes are associated with distinct tumor grade, TMN stage, and OS. Thus, BOLA1, BOLA2, and BOLA3 can predict the prognosis of HCC patients and may serve as oncogenes that promote HCC growth.

It has been proved that HCC development is a multistep process, including cell proliferation, adhesion, and metabolism. Iron metabolism plays an important role in both normal and cancer cells. In the process of HCC development, more iron is required to maintain the cancer cell proliferation, growth, and self-renewal in stem cells [27]. BOLA1, a mitochondrial protein, makes balances the effect of L-buthionine-(S, R)-sulfoximine (BSO)-induced glutathione (GSH) depletion on the mitochondrial thiol redox potential [12]. BOLA3 plays an important role in form [2Fe-2S] cluster-bridged dimeric heterocomplexes with the human monothiol glutaredoxin GRX5 [28]. A recent study indicated that BOLA1 and BOLA3 are associated with clinical outcomes in many diseases [5]. However, a thoughtful description of the relationship between expression level and cancer prognosis has not been analyzed. Although the increased expression of BOLA1/3 was obverse in present study, a correlation was observed between BOLA1/3 expression and defined genes in LIHC, such as oncogenic activity of BIRC2 [29] and tumor suppressor PMPCB [30]. Therefore, we can speculate that the BOLA1/3 expression in HCCs contributes to uncontrolled cell cellular proliferation. Further studies will be needed to clarify its role in HCC.

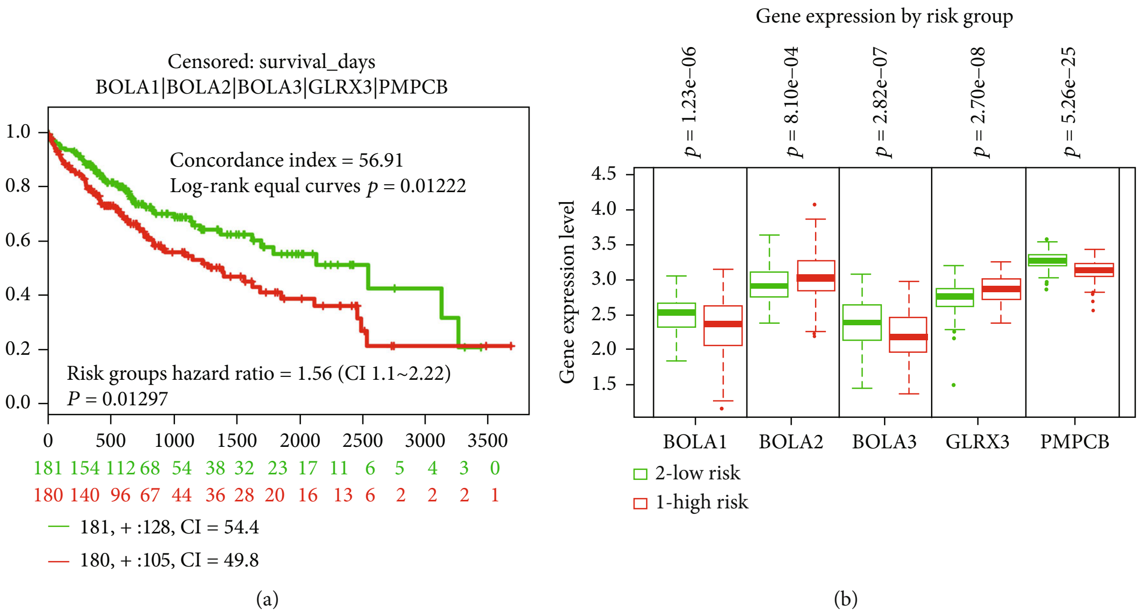


Figure 8: Combinatory BolA family members and GLRX3 and PMPCB predict survival of TCGA-LIHC patients (SurvExpress dataset). Notes: (a) Kaplan–Meier survival curve of 361 TCGA HCC samples using the SurvExpress database, based on the low or high risk for a poor outcome. (b) The high expression of five hub genes is correlated with high risk, poor prognosis, and shorter overall survival time. High- and low-risk groups are labeled with red and green curves, respectively. The box plots indicate the difference in expression of gene between risks groups, and P values are derived from t-test between both groups.

BOLA2, a gene associated with iron homeostasis, has been described in its biological function by the animal model [10]. The mechanisms of BOLA2 regulation are as follows: (i) specific in-frame fusion transcript regulation [31], (ii) monothiol CGFS glutaredoxin binding partners, (iii) GRX3-dependent anamorsin maturation pathway [32], and (iv) as c-Myc-regulated gene in HCC [10]. In our study, BOLA2 and GLRX3 are frequently overexpressed in HCC tumors tissues. Interestingly, our study revealed that the upregulation of BOLA2 and GLRX3 was associated with worse OS in patients with HCC. Up to now, more and more novel biomarkers, such as circular RNAs (circRNAs) [33], circulating microRNAs [34], and serum extracellular vesicles [35], had appeared for diagnosing HCC and predicting clinical outcomes. Our study analyzed the relationship between BOLA2 and serum extracellular vesicles. Hence, we postulate that the BOLA2 may have the potential for predicting the prognosis in HCC patients. Due to the limitations in our study, the relationship between the BOLA2 protein expression was not be clearly assessed, and further researches were needed to elaborate.

5. Conclusion

In our study, we found that BolA gene family members (BOLA1-3) may serve as prognostic biomarkers of HCC. In addition, BolA family members and their neighborhood GLRX3 play a leading role in HCC stage and tumor grade. These interesting results have important implications that can identify novel therapeutic targets in HCC.

Abbreviations

NP:	Normal person
CHD:	Coronary heart disease
CRC:	Colorectal cancer
HCC:	Hepatocellular carcinoma
PAAD:	Pancreatic adenocarcinoma
WhB:	Whole blood.

Data Availability

All the data used to support the findings of this study are available online.

Conflicts of Interest

The authors declare that they have no potential competing interests in this work.

Authors' Contributions

YiMing Tao designed the study and wrote the manuscript. Dong Wang analyzed the BOLA family members in all the database. Dong Wang and ZhiMing Wang performed all the figures in this study.

Acknowledgments

This work was supported by the Natural Science Foundation of Shandong Province (No. ZR202103020004), Clinical Medicine Plus X Project of Affiliated Hospital of Qingdao University (202102006), and National Nature Science

Foundation of China under Grant (Nos. 81372630, 81372631, 81702956). We thank the people for their efforts for providing data and databases.

Supplementary Materials

Figure S1: genetic alterations of 3 BolA family members were shown in HCC patients (cBioPortal). Notes: (a) OncoPrint of 3 BolA family member alterations in LIHC. (b) Using Exosomes web-accessible database (http://www.exoRBase.org) analysis, the increased expression of BOLA2 may be used as circulating biomarkers for HCC patients. *(Supplementary Materials)*

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics," *CA: a Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[2] Q. Ba, M. Hao, H. Huang et al., "Iron deprivation suppresses hepatocellular carcinoma growth in experimental studies," *Clinical Cancer Research*, vol. 17, no. 24, pp. 7625–7633, 2011.

[3] S. V. Torti and F. M. Torti, "Iron and cancer: more ore to be mined," *Nature Reviews. Cancer*, vol. 13, no. 5, pp. 342–355, 2013.

[4] Y. B. Zhou, J. B. Cao, B. B. Wan et al., "hBolA, novel non-classical secreted proteins, belonging to different BolA family with functional divergence," *Molecular and Cellular Biochemistry*, vol. 317, no. 1-2, pp. 61–68, 2008.

[5] M. A. Uzarska, V. Nasta, B. D. Weiler et al., "Mitochondrial Bol1 and Bol3 function as assembly factors for specific iron-sulfur proteins," *eLife*, vol. 5, 2016.

[6] H. Li, D. T. Mapolelo, S. Randeniya, M. K. Johnson, and C. E. Outten, "Human glutaredoxin 3 forms [2Fe-2S]-bridged complexes with human BolA2," *Biochemistry*, vol. 51, no. 8, pp. 1687–1696, 2012.

[7] Q. Yu, Y. Y. Tai, Y. Tang et al., "BOLA (BolA family member 3) deficiency controls endothelial metabolism and glycine homeostasis in pulmonary hypertension," *Circulation*, vol. 139, no. 19, pp. 2238–2255, 2019.

[8] X. Nuttle, G. Giannuzzi, M. H. Duyzend et al., "Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility," *Nature*, vol. 536, no. 7615, pp. 205–209, 2016.

[9] D. Wang, Z. M. Wang, S. Zhang, H. J. Wu, and Y. M. Tao, "Canopy homolog 2 expression predicts poor prognosis in hepatocellular carcinoma with tumor hemorrhage," *Cellular Physiology and Biochemistry*, vol. 50, no. 6, pp. 2017–2028, 2018.

[10] D. Hunecke, R. Spanel, F. Länger, S. W. Nam, and J. Borlak, "MYC-regulated genes involved in liver cell dysplasia identified in a transgenic model of liver cancer," *The Journal of Pathology*, vol. 228, no. 4, pp. 520–533, 2012.

[11] J. Luo, D. Wang, S. Zhang et al., "BolA family member 2 enhances cell proliferation and predicts a poor prognosis in hepatocellular carcinoma with tumor hemorrhage," *Journal of Cancer*, vol. 10, no. 18, pp. 4293–4304, 2019.

[12] P. Willems, B. F. J. Wanschers, J. Esseling et al., "BOLA1 is an aerobic protein that prevents mitochondrial morphology changes induced by glutathione depletion," *Antioxidants & Redox Signaling*, vol. 18, no. 2, pp. 129–138, 2013.

[13] M. Zhu and S. Xiao, "Expression profiles and prognostic values of BolA family members in ovarian cancer," *J Ovarian Res*, vol. 14, no. 1, p. 75, 2021.

[14] X. F. Wang, W. Lei, C. M. Liu, J. Yang, and Y. H. Zhu, "BOLA3 is a prognostic-related biomarker and correlated with immune infiltrates in lung adenocarcinoma," *International Immunopharmacology*, vol. 107, article 108652, 2022.

[15] D. S. Chandrashekar, B. Bashel, S. A. H. Balasubramanya et al., "UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses," *Neoplasia*, vol. 19, no. 8, pp. 649–658, 2017.

[16] E. Cerami, J. Gao, U. Dogrusoz et al., "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer Discovery*, vol. 2, no. 5, pp. 401–404, 2012.

[17] S. Li, Y. Li, B. Chen et al., "exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes," *Nucleic Acids Research*, vol. 46, no. D1, pp. D106–d112, 2018.

[18] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, "GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses," *Nucleic Acids Research*, vol. 45, no. W1, pp. W98–w102, 2017.

[19] A. Nagy, A. Lánczky, O. Menyhárt, and B. Győrffy, "Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets," *Scientific Reports*, vol. 8, no. 1, p. 9227, 2018.

[20] R. Aguirre-Gamboa, H. Gomez-Rueda, E. Martínez-Ledesma et al., "SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis," *PLoS One*, vol. 8, no. 9, article e74250, 2013.

[21] S. Orchard, M. Ammari, B. Aranda et al., "The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Research*, vol. 42, no. D1, pp. D358–D363, 2014.

[22] S. Roessler, H. L. Jia, A. Budhu et al., "A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients," *Cancer Research*, vol. 70, no. 24, pp. 10202–10212, 2010.

[23] E. Wurmbach, Y. B. Chen, G. Khitrov et al., "Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma," *Hepatology*, vol. 45, no. 4, pp. 938–947, 2007.

[24] X. Chen, S. T. Cheung, S. So et al., "Gene expression patterns in human liver cancers," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1929–1939, 2002.

[25] M. Uhlen, C. Zhang, S. Lee et al., "A pathology atlas of the human cancer transcriptome," *Science*, vol. 357, article eaan2507, no. 6352, 2017.

[26] A. Mollbrink, R. Jawad, A. Vlamis-Gardikas et al., "Expression of thioredoxins and glutaredoxins in human hepatocellular carcinoma: correlation to cell proliferation, tumor size and metabolic syndrome," *International Journal of Immunopathology and Pharmacology*, vol. 27, no. 2, pp. 169–183, 2014.

[27] A. Hamaï, T. Cañeque, S. Müller et al., "An iron hand over cancer stem cells," *Autophagy*, vol. 13, no. 8, pp. 1465-1466, 2017.

[28] V. Nasta, A. Giachetti, S. Ciofi-Baffoni, and L. Banci, "Structural insights into the molecular function of human [2Fe-2S] BOLA1-GRX5 and [2Fe-2S] BOLA3-GRX5 complexes," *Biochimica et Biophysica Acta - General Subjects*, vol. 1861, no. 8, pp. 2119–2131, 2017.

[29] A. Yamato, M. Soda, T. Ueno et al., "Oncogenic activity ofBIRC2 andBIRC3 mutants independent of nuclear factor-κB-activating potential," *Cancer Science*, vol. 106, no. 9, pp. 1137–1142, 2015.

[30] A. Takai, H. Dang, N. Oishi et al., "Genome-wide RNAi screen identifies PMPCB as a therapeutic vulnerability in EpCAM(+) hepatocellular carcinoma," *Cancer Research*, vol. 79, no. 9, pp. 2379–2391, 2019.

[31] H. Li and C. E. Outten, "Monothiol CGFS glutaredoxins and BolA-like proteins: [2Fe-2S] binding partners in iron homeostasis," *Biochemistry*, vol. 51, no. 22, pp. 4377–4389, 2012.

[32] L. Banci, F. Camponeschi, S. Ciofi-Baffoni, and R. Muzzioli, "Elucidating the molecular function of human BOLA2 in GRX3-dependent anamorsin maturation pathway," *Journal of the American Chemical Society*, vol. 137, no. 51, pp. 16133–16143, 2015.

[33] J. Hu, P. Li, Y. Song et al., "Progress and prospects of circular RNAs in hepatocellular carcinoma: novel insights into their function," *Journal of Cellular Physiology*, vol. 233, no. 6, pp. 4408–4422, 2018.

[34] H. R. Mirzaei, A. Sahebkar, M. Mohammadi et al., "Circulating microRNAs in hepatocellular carcinoma: potential diagnostic and prognostic biomarkers," *Current Pharmaceutical Design*, vol. 22, no. 34, pp. 5257–5269, 2016.

[35] A. Arbelaiz, M. Azkargorta, M. Krawczyk et al., "Serum extracellular vesicles contain protein biomarkers for primary sclerosing cholangitis and cholangiocarcinoma," *Hepatology*, vol. 66, no. 4, pp. 1125–1143, 2017.