Interdisciplinary Sciences: Computational Life Sciences (2020) 12:368–376
https://doi.org/10.1007/s12539-020-00376-6

SHORT COMMUNICATION

Deep Learning Based Drug Screening for Novel Coronavirus 2019-nCov

Haiping Zhang[1] · Konda Mani Saravanan[1] · Yang Yang[2] · Md. Tofazzal Hossain[1,6] · Junxin Li[3] · Xiaohu Ren[4] · Yi Pan[5] · Yanjie Wei[1]

Received: 1 February 2020 / Revised: 20 April 2020 / Accepted: 25 May 2020 / Published online: 1 June 2020

Abstract

A novel coronavirus, called 2019-nCoV, was recently found in Wuhan, Hubei Province of China, and now is spreading across China and other parts of the world. Although there are some drugs to treat 2019-nCoV, there is no proper scientific evidence about its activity on the virus. It is of high significance to develop a drug that can combat the virus effectively to save valuable human lives. It usually takes a much longer time to develop a drug using traditional methods. For 2019-nCoV, it is now better to rely on some alternative methods such as deep learning to develop drugs that can combat such a disease effectively since 2019-nCoV is highly homologous to SARS-CoV. In the present work, we first collected virus RNA sequences of 18 patients reported to have 2019-nCoV from the public domain database, translated the RNA into protein sequences, and performed multiple sequence alignment. After a careful literature survey and sequence analysis, 3C-like protease is considered to be a major therapeutic target and we built a protein 3D model of 3C-like protease using homology modeling. Relying on the structural model, we used a pipeline to perform large scale virtual screening by using a deep learning based method to accurately rank/identify protein–ligand interacting pairs developed recently in our group. Our model identified potential drugs for 2019-nCoV 3C-like protease by performing drug screening against four chemical compound databases (Chimdiv, Targetmol-Approved_Drug_Library, Targetmol-Natural_Compound_Library, and Targetmol-Bioactive_Compound_Library) and a database of tripeptides. Through this paper, we provided the list of possible chemical ligands (Meglumine, Vidarabine, Adenosine, D-Sorbitol, D-Mannitol, Sodium_gluconate, Ganciclovir and Chlorobutanol) and peptide drugs (combination of isoleucine, lysine and proline) from the databases to guide the experimental scientists and validate the molecules which can combat the virus in a shorter time.

Keywords Coronavirus · Deep learning · Drug screening · Homology modeling · 3C-like protease

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12539-020-00376-6) contains supplementary material, which is available to authorized users.

✉ Yanjie Wei
yj.wei@siat.ac.cn

[1] Center for High Performance Computing, Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, People's Republic of China

[2] Shenzhen Key Laboratory of Pathogen and Immunity, Guangdong Key Laboratory for Diagnosis and Treatment of Emerging Infectious Diseases, State Key Discipline of Infectious Disease, Second Hospital Affiliated to Southern University of Science and Technology, Shenzhen Third People's Hospital, Shenzhen 518112, People's Republic of China

[3] Shenzhen Laboratory of Human Antibody Engineering, Institute of Biomedicine and Biotechnology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Boulevard, University City of Shenzhen, XiliNanshan, Shenzhen 518055, People's Republic of China

[4] Institute of Toxicology, Shenzhen Center for Disease Control and Prevention, No 8 Longyuan Road, Nanshan District, Shenzhen 518055, China

[5] Department of Computer Science, Georgia State University, Atlanta 30302-5060, USA

[6] University of Chinese Academy of Sciences, No. 19(A) Yuquan Road, Shijingshan District, Beijing 100049, People's Republic of China

1 Introduction

In December 2019, a severe respiratory illness similar to severe acute respiratory syndrome coronavirus emerged in Wuhan, Hubei, China and is spreading all over the world with high mortality. In the past, beta coronaviruses, severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), respectively, have caused high mortality rates and became a threat to human life [1]. The most recent outbreak of the viral pneumonia was first disclosed by the Wuhan Municipal Health Commission [2, 3], and the World Health Organization (WHO) was alarmed about the outbreak of pneumonia announced by the Chinese Officials [4]. The novel coronavirus (2019-nCoV) was isolated from 27 patients who were initially reported and the number of patients was subsequently revised to 31,498 as of March 23, 2020, with 3267 deaths [5]. The current 2019-nCoV outbreak has some common features like the SARS outbreak: both have happened in winter, are linked to live animal markets, and caused by unknown coronaviruses [2, 5].

Fever, cough, and shortness of breath are the symptoms in common cases, whereas pneumonia, severe acute respiratory syndrome, and kidney failure are being reported as the symptoms in severe cases [4]. Most of the 2019-nCoV patients are linked to the Huanan Seafood Wholesale Market where several wildlife animals including bats, snakes as well as poultry are sold. So far, no specific wildlife animal is identified as the host of the novel coronavirus. Bat is considered as the native host of the novel coronavirus (2019-nCoV) although there are other hosts in transmission from bats to humans [5]. The Spring Festival travel rush has accelerated the spread, so it is of top priority to prevent the spread, develop a new drug to combat it, and cure the patients in time. Knowledge of current 2019-nCoV can be learned from previous SARS-CoV. For SARS-CoV, a variety of modern machine learning methods, in particular, deep neural networks were used for drug discovery and development. These methods take advantage of bigger datasets compiled from high-throughput screening data and perform prediction of bioactivities of a target with high accuracy [6].

The genetic sequences of 2019-nCoV have shown similarities to SARS-CoV (79.5%) [7, 8]. The *S*-protein and 3C-like protease are potential drug targets. The *S*-protein is the main target of neutralizing antibodies, and antibodies binding with this protein have the potential to stop the virus entry into host cells [9]. The 3C-like protease catalyzes a chemical reaction which is important in SARS coronavirus replicase polyprotein processing [10, 11]. The neutralizing antibodies against *S*-protein of SARS have been obtained from human patients and the anti-SARS-CoV S antibody triggered fusogenic conformational changes [9]. This provides an important clue to prevent virus entry into host cells by antibodies or peptides. The 3C-like protease inhibitors also have potential to prevent coronavirus maturation, and series of unsaturated esters inhibitors against 3C-like protease of SARS-CoV was deposited in PDB database (crystal structures of SARS-Cov 3C-like protease complexed with a series of unsaturated esters, Protein Databank Identifier: 3TIT).

One can also use these previous SARS inhibitors to design the inhibitor against 2019-nCoV. Based on the increasing protein–ligand complex structures, the deep learning algorithms for identifying/predicting potential binding compounds for a given target became possible [12, 13]. In addition to small molecular chemical compounds, scientists also rely on peptide/antibody to combat the virus due to stronger binding affinity. In the post-genomics era, a Dense Fully Convolutional Neural Network (DFCNN) model is more effective, faster, and cheaper for drug discovery, because the deep layers of the model can learn more features from the data and perform an accurate prediction. By using these techniques, an antimalarial drug "pyrimethamine" was discovered against Dihydrofolate reductase (DHFR) enzyme and another drug BPM31510 is in a phase II trial involving humans with advanced pancreatic cancer [14–16]. Hence we believe that the integrated applications of such machine learning models as a pipeline for drug discovery have implications in therapeutic drug targeting.

Considering all the above facts, in the present work, we considered 2019-nCov_3C-like protease as a potential target and built a structural model after systematically analyzing its sequence features. We built a pipeline with a deep learning based method developed in our group by representing molecules as vectors to identify potential drugs (peptides or small ligands) against the protein target of the 2019-nCoV virus [13]. Our method is extremely fast in virtual drug screening and it takes less than a day to finish the virtual screening over millions of protein–ligand or protein-peptide predictions, whereas traditional docking methods take several weeks with the help of a supercomputer. Although, 2019-nCoV outbreak is a major challenge for clinicians [17], we believe the proposed potential drug list can help them to validate the drug that relieves symptoms or even cures the disease rapidly.

2 Materials and Methods

2.1 Dataset and Sequence Alignment

We retrieved the virus RNA sequences from Global Initiative on Sharing All Influenza Data (GISAID) database [18] and the sequences are aligned with a focus on the interested

S-protein and ligand binding region of 2019-nCov_3C-like protease. The amino acid sequence is translated from the RNA sequence by Translate web tool (https://web.expasy.org/translate/). We used 18 patient's sequences in this work (EPI_ISL_402119 to EPI_ISL_404228). Details of the sequences and acknowledgement to the authors who submitted the data to the server is presented in the Supplementary Table S1. Multiple sequence alignment is performed by using Clustal Omega program [19].

2.2 Homology Modeling of 2019-nCov_3C-like Protease

The structural model of 2019-nCov_3C-like protease was built by using Modeller 9.9 [20]. The SARS coronavirus 3C-like protease was used as a template (PDB ID: 3TNT) which has about 96.07% amino acid sequence identity. The software outputs multiple predicted structures and they are ranked according to the discrete optimized protein energy (DOPE) score [21]. The quality of the model was validated by looking at the stereo chemical quality on Ramachandran map. The model was further optimized by PROCHECK [22], ERPAT [23] and Qmean [24] and the final optimized structural model is considered for further analysis.

2.3 A Deep Learning Model is Used to Virtual Screen Large Databases

In our previous work, we built a Dense Fully Convolutional Neural Network (DFCNN) deep learning model to reverse search drug targets. Here we apply this model to perform large-scale virtual screening. Since the method is shown to have relatively higher accuracy and efficiency, it is very suitable for applying to such an emerging disease outbreak. The DFCNN is a densely fully connected neural network, and the densely network (similar to DenseNet, but replace the convolution layer to fully connected layer) allows deep layer without the gradient vanishing problem. The deeper layers make it to learn more abstract features from the data. The training data of DFCNN is from PDB bind database [25], for which we define the crystal protein–ligand PDB complexes as positive and cross-docking complexes as negative. The detailed process to build the deep learning model is described in our recently published work to virtual screen targets by inputting a small molecule by using a vector type

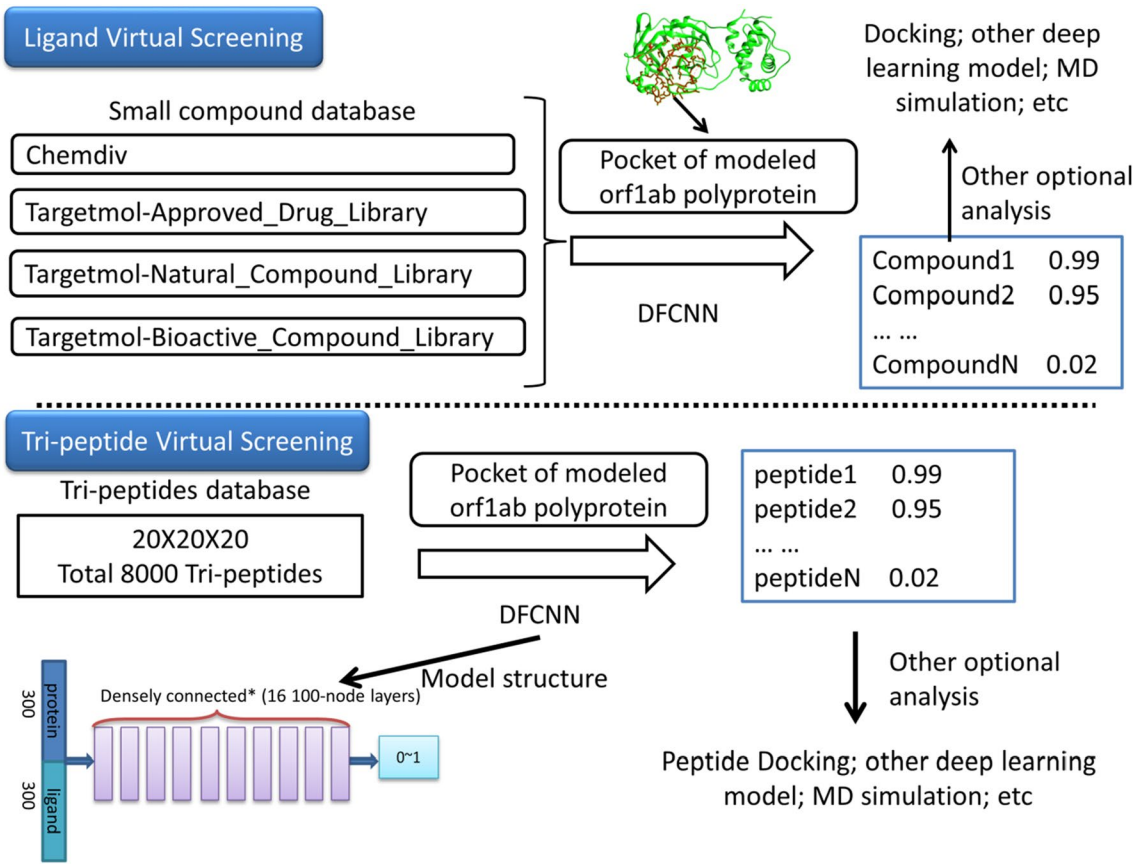

Fig. 1 The workflow of virtual screening of small chemical compounds and tripeptides against the 2019-nCov_3C-like protease

of representation [13]. The overall workflow of the proposed method is shown in Fig. 1. DFCNN model has two advantages over many other methods such as independent of docking simulation and the training dataset includes nonbinding decoys. The independence of the docking simulation makes it extremely fast, while the inclusion of nonbinding decoys during training makes the model robust in the real application scenarios.

2.4 Virtual Screening Against Chimdiv Database

The structural model of the ligand binding region of 2019-nCov_3C-like protease is used as the target protein structure. We define the residues with a cutoff distance of 1 nm from the known ligand as a pocket (binding site is defined based on the ligand from the template PDB 3TNT is used). The ligand database is taken from the chimdiv company (https://www.chemdiv.com/) which contains around 1000,000 compounds. We first used the DFCNN model to perform large-scale virtual screening. The mean and deviation of the training dataset were used during data normalization for a more stable performance. In the second stage, the top prediction by DFCNN model was chosen for an autodock vina-based docking simulation. The docking result was visualized and examined by the discovery studio visualizer [26]. Finally, we provide a proposed compound list that has the potential to bind protein pocket.

2.5 Virtual Screening Against Targetmol-Approved_Drug_Library, Targetmol-Natural_Compound_Library, and Targetmol-Bioactive_Compound_Library

The Targetmol-Approved_Drug_Library, Targetmol-Natural_Compound_Library, and Targetmol-Bioactive_Compound_Library contain about 2040, 1680, and 5370 compounds, respectively. We have applied DFCNN model to perform virtual screening against these three libraries for 2019-nCov_3C-like protease. The compounds with high DFCNN scores are recommended as the potential inhibitors for further experimental validation.

2.6 Virtual Screening Against Tripeptide Database

Tri-amino acid peptide database is first built with a total size of 8000. Each amino acid in the tripeptide database was converted into a molecule vector by Mol2vec [27]. For each peptide, the sum of its amino acid vector was used to represent this peptide's vector. Protein pocket is defined as residues with a cutoff distance of 1 nm from the known ligand. The pocket is then converted into Vector. The pocket and peptide vector are then concatenated into one line as input with a maximum dimension of 600. We will use the same model as DFCNN, a densely fully connected model that is trained by a protein–ligand dataset from the PDB bind database. Since the ligand and peptides are composed of chemical groups, the model trained on the protein–ligand complexes should also be suitable for protein–small peptide interaction.

3 Results

3.1 Sequence Alignment and Homology Modeling

Eighteen patient's RNA sequences obtained from GISAID public domain database are translated into protein sequences by using translate tool. The ligand-binding sites of the template protein (3TNT) is considered as reference to define pocket region of our homology model. We have checked the mutations in the pocket region of 2019-nCov_3C-like protease, and the sequences have 100% similarity with the virus from 18 different patients. This indicates the virus is highly conserved in this region, and it is suitable for designing drugs by targeting this site. The alignment of *S*-protein epitope regions also shows high conservation among the patients (Supplementary Figure S1). From the figure, it is observed that the RNA sequence EPI_ISL_402132 has a point mutation at 32nd position where the codon of phenylalanine is replaced by isoleucine. 2019-nCoV_3C-like protease is also aligned to SARS-CoV protease by Clustal Omega [19]. The aligned sequence is shown in Fig. 2. There are 276 amino acid residues in both of the proteins. The figure indicates high similarity between 2019-nCov and SARS-CoV, which is consistent with the findings by Xu et al. [5]. Using the X-ray crystallographic structure of SARS coronavirus 3C-like protease solved at 1.59Å resolution as the template, a theoretical protein model is built for 2019-nCoV_3C-like protease using modeler software. Figure 3a shows the crystallographic structure of SARS_coronavirus_3C-like protease and 3B shows the homology model of 2019-nCoV_3C-like protease. There are only four mutations (T35V, A46S, S94A and K180N) between SARS_coronavirus_3Clike protease and 2019-nCoV_3C-like protease shown in Fig. 3a and b. In the Figure, the mutated residues are marked with blue color. Figure 3c shows the model structure with known SARS_coronavirus_3C-like protease inhibitor. The binding pocket and two-dimensional ligand interaction pattern of the target protein is shown with reference to the template. There are 23 protein–ligand interactions observed including 15 hydrogen bonds, one disulphide bond and few pi stacking interactions which is shown in Fig. 3d. The pocket extracted from the model is used for further analysis of large-scale virtual screening.

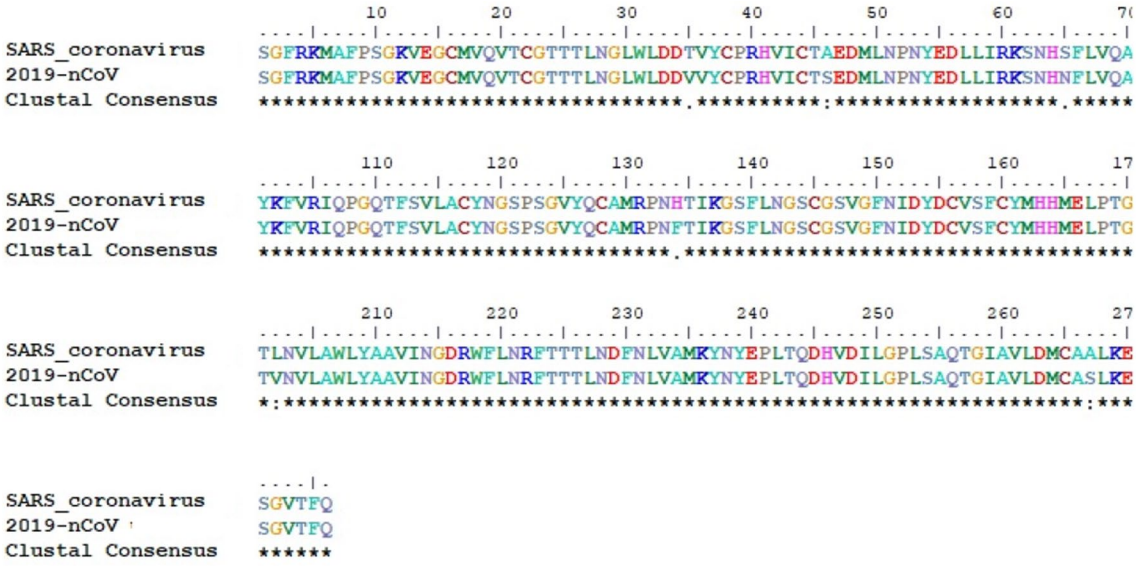

Fig. 2 The sequence alignment of SARS_coronaivrus_3C-like protease and 2019-nCov_3C-like protease

3.2 Virtual Screening Against Four Small Molecular Compound Databases

Chemdiv dataset, widely used for large-scale virtual screening, contains a large amount (~ 1,000,000) of drug-like compounds or drug leads. The potential drug candidates with the highest score (Autodock vina score and our deep learning model score) from the Chemdiv dataset are presented in Table 1. Interestingly, the compound with identifier "C998-0189" has a top vina score compared to other six compounds listed. The name of the compound is *N*~2~-(3,5-dimethylphenyl)-*N*~2~-(5,5-dioxido-3a,4,6,6a-tetrahydrothieno[3,4-d][1,3] thiazol-2-yl)-*N*~1~-[3-(trifluoromethyl)phenyl]glycinamide with molecular formula C22H22F3N3O3S2. The molecular weight of the compound is 497.6 g/mol and the compound satisfies most of the drug-likeness parameters including Lipinski's filters. The other five recommended compounds also have reasonable vina scores around 7.5 with important stabilizing interactions.

The top 100 predictions by our deep learning model against the database are shown in Supplementary Table S2. The top five compounds with Chimdiv identifier 8017-4328, 8017-4325, 8002-7777, 8004-0123 and 8010-0095, respectively, are listed with the high DFCNN score. Three other well-known compound libraries were screened in the present work, including Targetmol-Approved_Drug_Library, Targetmol-Natural_Compound_Library and Targetmol-Bioactive_Compound_Library. It is worth to test whether there is any natural compound that can combat the virus by inhibiting 2019-nCov_3C-like protease. Table 2 shows the screening result for Targetmol-Natural compound library. The compounds with a DFCNN score higher than 0.997 are listed in Table 2, and it is found that Adenosine, Vidarabine, Mannitol, Dulcitol, D-Sorbitol, D-Mannitol, Allitol, Sodium_gluconate are the top predictions (Table 2). Natural products are often active ingredients of known herb medicine, and relatively safe because of long history usage. If it is proved by an experiment that is effective to the target, patients can easily access it by taking corresponding herb medicine. There are about 8 compounds with the score of 0.999 and about 20 compounds with the score of 0.998 which are presented in Table S2. As indicated above, most of the drugs listed by our model are antiviral drugs and hence it can be tested against nCoV-2019 and can be validated in the clinical lab within a short time.

The screening result for Targetmol-Approved Drug library is shown in Table 3. The compounds with a DFCNN score higher than 0.997 are listed in Table 3. We randomly considered drugs from potential drugs list and performed a systematic literature search. It is found that Meglumine, Vidarabine, Adenosine, D-Sorbitol, D-Mannitol, Sodium_gluconate, Ganciclovir and Chlorobutanol, respectively, are top predictions according to the DFCNN score (Table 3). Interestingly, we found most of the drugs in the list such as meglumine, Ganciclovir and Vidarabine, respectively, show antiviral activity. The list of all the compounds above score 0.990 is provided in Table S4. The screening result for Targetmol-Bioactive_Compound_Library is shown in Table 4. The compounds with a DFCNN score higher than 0.997 is listed in Table 4. Bioactive compounds are a type of chemicals that can be found in plants and some foods and have been studied in the prevention of various diseases. It is worth to check whether any of them can act on the target protein. We

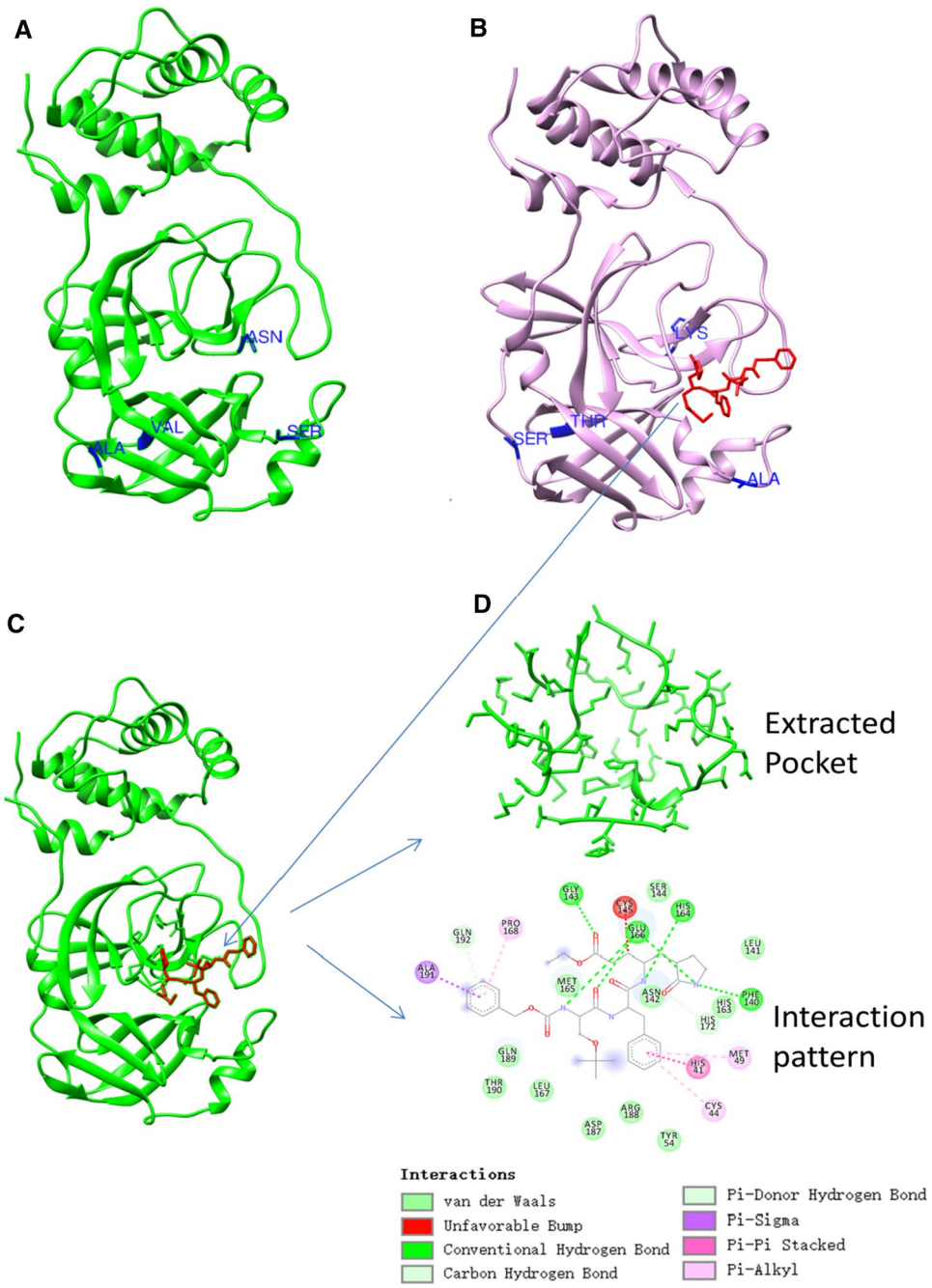

Fig. 3 The structural model of 2019-nCov_3C-like protease and its template. In **a** and **b**, the modeled 2019-nCov_3C-like protease and SARS_3C-like protease are shown with the mutated four residues marked with blue color. The ligand from the PDB 3TNT is transferred to the modeled structure (**c**) and based on residue distance from the transferred ligand, we define the pocket (**d**). The interaction between the ligand and the modeled 2019-nCov_3C-like protease is also shown (**d**)

Table 1 The selected compounds that may inhibit 2019-nCov_3C-like protease based on the DFCNN score and autodock vina score

Chemdiv ID	Vina score (kcal/mol)	DeepBindVec	Recommendation
C998-0189	−8.5	> 0.995	Recommended
C998-0197	−7.9	> 0.995	Can try
C998-0090	−7.8	> 0.995	Can try
C998-0948	−7.7	> 0.995	Recommended
C998-1046	−7.6	> 0.995	Recommended
D076-0195	−7.3	> 0.995	Recommended

found compounds such as Vidarabine, Adenosine, Dulcitol, D-Sorbitol, D-Mannitol, Ganciclovir and 5′-deoxyadenosine are the top predictions in the Targetmol-Bioactive compounds (Table 4). The list of compounds all the compounds above score 0.99 is provided in Table S5. The list in Table 4 has narrowed down the hit compounds for later drug development stages, such as molecular dynamics simulation, or even directly experimental validation for finding bioactive compounds against 2019-nCov_3C-like protease.

Table 2 The potential drug candidates selected from the Targetmol-Natural Compound Library

Natural compound	DFCNN Score
Adenosine; Vidarabine; Mannitol; Dulcitol; D-Sorbitol; D-Mannitol; Allitol; Sodium_gluconate	Score ≥ 0.999
L(-)-sorbose; D-(-)-Fructose; Guanosine; Inosine; Trichostatin_A; D-(-)-Ribose; DL-Xylose; Cordycepin; β-Glycerophosphate_disodium_salt_hydrate; Xanthosine; Zeatin; *N*6-methyladenosine; Atractylodin; Tubercidin; Glucosamine_sulfate; Panthenol; Dexpanthenol; Ubenimex; Phospho(enol)pyruvic_acid_monopotassium	0.999 > Score ≥ 0.998
Aztreonam; Cytidine; Cytarabine; D-Saccharic_acid_potassium_salt; D-Glucose_6-phosphate_sodium_salt; Quinic_acid; 2′-Deoxyadenosine_monohydrate; *N*-Sulfo-glucosamine_sodium_salt; 2′-Deoxyguanosine_monohydrate	0.998 > Score ≥ 0.997

Table 3 The potential drug candidates selected from the Targetmol-Approved Drug library

Approved Drug name	DFCNN Score
Meglumine; Vidarabine; Adenosine; D-Sorbitol; D-Mannitol; Sodium_gluconate; Ganciclovir; Chlorobutanol	Score ≥ 0.999
AICAR_(Acadesine); Mylosar; Inosine; D-Pantothenic_acid_sodium_salt; DL-Xylose; Ethambutol_dihydrochloride; Glucosamine; Myclobutanil; Sodium_etidronate; Fludarabine; Gemcitabine; Emtricitabine; Tubercidin; Bestatin_hydrochloride; Panthenol; Dexpanthenol; Cladribine; Entecavir; Ubenimex	0.999 > Score ≥ 0.998
Entecavir_hydrate; Procarbazine_hydrochloride; Aztreonam; Disopyramide; Benznidazole; Clofarabine; Bucetin; Nifuroxazide; Triflupromazine_hydrochloride; Doxifluridine; Cytarabine; Cefdinir; Bupropion_hydrochloride; Fluoxetine; Tenofovir; Pentostatin; Fluoxetine_hydrochloride; Imazalil; Atenolol	0.998 > Score ≥ 0.997

Table 4 The potential drug candidates selected from the Targetmol-Bioactive Compounds

Bioactive compound	DFCNN Score
Vidarabine; Adenosine; Dulcitol; D-Sorbitol; D-Mannitol; Ganciclovir; 5′-deoxyadenosine	Score ≥ 0.999
Nelarabine; Tosedostat; Fosfomycin_Tromethamine; AICAR_(Acadesine); Mylosar; Guanosine; Inosine; Crotonoside; D-(-)-Ribose; Cordycepin; β-Glycerophosphate_disodium_salt_hydrate; Zeatin; Ethambutol_dihydrochloride; 5-Iodotubercidin; Myclobutanil; Sodium_etidronate; Atractylodin; Fludarabine; Heterophyllin_B; Gemcitabine; Emtricitabine; Disodium_clodronate_tetrahydrate; Ostarine; Tubercidin; Bestatin_hydrochloride; Panthenol; Dexpanthenol; FCCP; Cladribine; Z-VAD(OMe)-FMK; WP1066; Entecavir; Ubenimex; Batimastat; ML264; GSK4112; Degrasyn; Cefcapene_Pivoxil_Hydrochloride; Phospho(enol)pyruvic_acid_monopotassium; A-804598;SR3335;IPTG	0.999 > Score ≥ 0.998
KYA1797K; Mizoribine; 5-Hydroxy-1,7-diphenyl-6-hepten-3-one; ATPO; Entecavir_hydrate; Aztreonam; NXY-059; D-Pantothenic_acid; Bay_11-7085; Disopyramide; Benznidazole; SB_297006; Imidafenacin; Clofarabine; Bucetin; Nifuroxazide; Triflupromazine_hydrochloride; Doxifluridine; Selegiline_hydrochloride; Cytarabine; Cytidine; BGP-15; Cefdinir; Bupropion_hydrochloride; UK-371804; Fluoxetine; D-Saccharic_acid_potassium_salt; D-Glucose_6-phosphate_sodium_salt; J147; Tenofovir; *N*-Sulfo-glucosamine_sodium_salt; Pentostatin; Fluoxetine_hydrochloride; Nifurtimox; Imazalil; 5-Fluorouridine; Atenolol; Repertaxin; ACY-738	0.998 > Score ≥ 0.997

Table 5 The predicted tripeptide that has high possibility (DFCNN score ≥ 0.99) to bind with the pocket of 2019-nCov_3C-like protease by DFCNN Score

Peptide sequence	DFCNN Score
IKP; IPK; KIP; KPI; PIK; PKI	Score ≥ 0.997
GKL; LGK; LKG; KGL; KLG; GKK; KGK; KKG; AKK; KAK; KKA; KPV; KVP; PKV; PVK; VKP; VPK	0.997 > Score ≥ 0.996
GKI; IGK; IKG; KGI; KIG; LKP; LPK; KLP; KPL; PLK; PKL; LLK; LKL; KLL	0.996 > Score ≥ 0.995

3.3 Virtual Screening Against Database of Tripeptides

Peptides have the potential to exert higher binding affinity and specificity than small molecular chemical compounds; meanwhile, small peptides are easier to be synthesized compared with small molecules and antibodies. Since the known ligands of SARS_3C-like protease are compounds similar to tripeptides and the combination of 20 amino acids for tripeptide is also affordable for our method, we decide to perform virtual screening on the tripeptides. The screened tripeptides with a DFCNN score higher than 0.995 (0.997, 0.996 and 0.995) for the 2019-nCov_3C-like protease is shown in Table 5. A higher value indicates the peptide can most likely bind with the pocket of the 2019-nCov_3C-like protease. Our method found that the peptides formed by I,

K, P amino acids have the highest possibility to bind in the pocket. The combinations by G, K, L or G, K, K or K, P, V are also found to be favorable binding partners predicted by DFCNN (Table 5). The list of all tripeptides above score 0.99 is provided in Table S6. The combination of short peptides and its composition play a crucial role in affecting the overall conformation of protein [28, 29]. It was found that the tripeptide, pentapeptide and octapeptides are believed to be promising candidates for drug development of infectious diseases [30, 31]. Since these peptides are relatively easy to produce, many of the top predictions can be validated by the experimental techniques in a very fast and less expensive manner.

4 Conclusion

Designing small compound or peptide drugs to cure the 2019-nCoV is extremely urgent. Effective and safe drugs are required for treating deadly viral disease which caused an epidemic outbreak all over the globe. Researchers use different modern technologies to combat such diseases and deep learning is one among them with faster prediction and achieves greater than ~ 80% accuracy. With the extremely high speed and relatively high accuracy, our DFCNN model for 3C-like protease–ligand interaction analysis is suitable to overcome the challenge of screening tens of thousands of drugs in a short time in a certain emergency situations, such as 2019-nCov outbreak. Our deep learning model based on DFCNN is a data-driven model, which learns 3C-like protease–ligand interaction from known binding and non-binder data. The model use the binding pocket of 3C-like protease–ligand conformation instead of whole conformation of the complex; hence our model is so fast and accurate compared to all other molecular docking procedures.

The identified potential 3C-like protease-ligand pairs can be subjected to MD simulation to further check the binding stability and atomic interaction pattern, or even the binding free energy with techniques such as metadynamics to narrow down the candidate list. A variety of repurposed drugs and investigational drugs have been identified in the past. Screening National Medical products Administration (NMPA) approved drug libraries and other chemical libraries have identified novel agents. Hundreds of clinical trials involving remdesivir, chloroquine, favipiravir, chloroquine, convalescent plasma, TCM and other interventions are planned or underway. In this connection, we have performed a deep learning-based drug screening and provided potential compound and tripeptide lists for 2019-nCov_3C-like protease. Since the inhibitor candidates provided are on-market drugs, the list provided can help to facilitate the 2019-nCov_3C-like protease drug development and could be used immediately.

Funding This work was partly supported by the National Key Research and Development Program of China under Grant Nos. 2018YFB0204403 and 2016YFB0201305, National Science Foundation of China under Grant no. U1813203 and 61433012; the National Natural Youth Science Foundation of China (Grant No. 31601028), the Shenzhen Basic Research Fund under Grant Nos. JCYJ20180507182818013, GGFW2017073114031767 and JCYJ20170413093358429, the China Postdoctoral Science Foundation under Grant No. 2019M653132, CAS Key Lab under Grant No. 2011DP173015. We would also like to thank the funding support by the Shenzhen Discipline Construction Project for Urban Computing and Data Intelligence, Youth Innovation Promotion Association, CAS to Yanjie Wei.

Compliance with ethical standards

Conflict of interest The authors have declared that no competing interests exist.

References

1. Huang C, Wang Y, Li X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5
2. Lu H, Stratton CW, Tang Y (2020) Outbreak of pneumonia of unknown etiology in Wuhan China: the mystery and the miracle. J Med Virol 92(4):401–402. https://doi.org/10.1002/jmv.25678
3. Thompson R (2020) Pandemic potential of 2019-nCoV. Lancet Infect Dis 20(3):P280. https://doi.org/10.1016/s1473-3099(20)30068-2
4. Hui DS, Azhar E, Madani TA et al (2020) The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan China. Int J Infect Dis 91:264–266. https://doi.org/10.1016/j.ijid.2020.01.009
5. Xintian Xu, Chen P, Wang J, Feng J, Zhou H, Xuan Li Wu, Zhong PH (2020) Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci China Life Sci 63:457–460. https://doi.org/10.1007/s11427-020-1637-5
6. Ekins S, Puhl AC, Zorn KM et al (2019) Exploiting machine learning for end-to-end drug discovery and development. Nat Mater 18:435–441. https://doi.org/10.1038/s41563-019-0338-z
7. Zhou P, Yang X-L, Wang X-G et al (2020) Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. Nature 579:270–273. https://doi.org/10.1101/2020.01.22.914952
8. Lu R, Zhao X, Li J et al (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395(10224):565–574. https://doi.org/10.1016/S0140-6736(20)30251-8
9. Walls AC, Xiong X, Park YJ et al (2019) Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. Cell 176(5):1026–1039. https://doi.org/10.1016/j.cell.2018.12.028
10. Goetz DH, Choe Y, Hansell E et al (2007) Substrate specificity profiling and identification of a new class of inhibitor for the major protease of the SARS Coronavirus. Biochemistry 46(30):8744–8752. https://doi.org/10.1021/bi0621415
11. Kim Y, Lovell S, Tiew K-C et al (2012) Broad-spectrum antivirals against 3C or 3C-like proteases of picornaviruses, noroviruses, and coronaviruses. J Virol 86(21):11754–11762. https://doi.org/10.1128/jvi.01348-12

12. Zhang H, Liao L, Saravanan KM et al (2019) DeepBindRG: a deep learning based method for estimating effective protein–ligand affinity. PeerJ 7:e7362. https://doi.org/10.7717/peerj.7362
13. Zhang H, Liao L, Cai Y et al (2019) IVS2vec: a tool of inverse virtual screening based on word2vec and deep learning techniques. Methods 166:57–65. https://doi.org/10.1016/j.ymeth.2019.03.012
14. Fleming N (2018) How artificial intelligence is changing drug discovery. Nature 557:S55–S57. https://doi.org/10.1038/d41586-018-05267-x
15. Liu Z, Du J, Fang J et al (2019) DeepScreening: a deep learning-based screening web server for accelerating drug discovery. Database (Oxford) 2019:1–11. https://doi.org/10.1093/database/baz104
16. Chen H, Engkvist O, Wang Y et al (2018) The rise of deep learning in drug discovery. Drug Discov Today 23(6):1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039
17. Russell CD, Millar JE, Baillie JK (2020) Clinical evidence does not support corticosteroid treatment for 2019-nCoV lung injury. Lancet 395:473–475. https://doi.org/10.1016/S0140-6736(20)30317-2
18. Shu Y, McCauley J (2017) GISAID: Global initiative on sharing all influenza data—from vision to reality. Eurosurveillance 22(13):30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494
19. Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 27(1):135–145. https://doi.org/10.1002/pro.3290
20. Fiser A, Šali A (2003) MODELLER: generation and refinement of homology-based protein structure models. Methods Enzymol 374:461–491. https://doi.org/10.1016/S0076-6879(03)74020-8
21. Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15(11):2507–2524. https://doi.org/10.1110/ps.062416606
22. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291. https://doi.org/10.1107/S0021889892009944
23. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2(9):1511–1519. https://doi.org/10.1002/pro.5560020916
24. Benkert P, Tosatto SCE, Schomburg D (2008) QMEAN: a comprehensive scoring function for model quality assessment. Proteins 71:261–277. https://doi.org/10.1002/prot.21715
25. Liu Z, Li Y, Han L et al (2015) PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics 31(3):405–412. https://doi.org/10.1093/bioinformatics/btu626
26. Accelrys: Materials Studio is a Software Environment for Molecular Modeling (2009) Dassault Systèmes BIOVIA. Discovery. https://doi.org/10.1007/s10822-010-9395-8
27. Jaeger S, Fulle S, Turk S (2018) Mol2vec: unsupervised machine learning approach with chemical intuition. J Chem Inform Model 58(1):27–35. https://doi.org/10.1021/acs.jcim.7b00616
28. Santos S, Torcato I, Castanho MARB (2012) Biomedical applications of dipeptides and tripeptides. Biopolymers 98(4):288–293. https://doi.org/10.1002/bip.22067
29. Saravanan KM, Selvaraj S (2012) Search for identical octapeptides in unrelated proteins: structural plasticity revisited. Biopolymers 98(1):11–26. https://doi.org/10.1002/bip.21676
30. Wendler J, Schröder BO, Ehmann D et al (2018) Tu1860—a novel octapeptide as a promising candidate for antibiotic drug development and host derived microbiome regulation. Gastroenterology 154(6):S1040. https://doi.org/10.1016/s0016-5085(18)33486-3
31. Saravanan KM, Dunker AK, Krishnaswamy S (2017) Sequence fingerprints distinguish erroneous from correct predictions of intrinsically disordered protein regions. J Biomol Struct Dyn 36(16):4338–4351. https://doi.org/10.1080/07391102.2017.1415822