 doi:10.20944/preprints202203.0032.v1

Review

Artificial Intelligence Technologies for COVID-19 De Novo Drug Design

Giuseppe Floresta, Chiara Zagni, Davide Gentile, Vincenzo Patamia and Antonio Rescifina*

Dipartimento di Scienze del Farmaco e della Salute, Università di Catania, Viale A. Doria 6, 95125 Catania, Italy
giuseppe.floresta@unict.it, chiara.zagni@unict.it, davide.gentile@unict.it, vincenzo.patamia@unict.it.

* Correspondence: arescifina@unict.it

Abstract: The recent covid crisis has proven important lessons for academia and industry regarding digital reorganization. Among fascinating lessons from these times is the huge potential of data analytics and artificial intelligence. The crisis exponentially accelerated the adoption of analytics and artificial intelligence, and this momentum is predicted to continue into the 2020s and over. Moreover, drug development is a costly and time-consuming business, and only a minority of approved drugs return the revenue that exceeds the research and development costs. As a result, there is a huge drive to make drug discovery cheaper and faster. With modern algorithms and hardware, it is not too surprising that the new technologies of artificial intelligence and other computational simulation tools can help drug developers. In only two years of covid research, many novel molecules have been designed/identified using artificial intelligence methods with astonishing results in terms of time and effectiveness. This paper will review the most significant research on artificial intelligence in the de novo drug design for COVID-19 pharmaceutical research.

Keywords: artificial intelligence; machine learning; drug design; covid-19, structure-based drug design; ligand-based drug design.

1. Introduction

Initially noted on December 2019, in Wuhan, China, with a patient diagnosed with atypical pneumonia just a few months later, COVID-19 has announced a pandemic by the World Health Organization on 11 March 2020. Today, it is still a considerable concern for humanity [1]. A problematic aspect for the clinician to address is that disease caused by the virus can cause a broad spectrum of symptoms and disease outcomes. In most cases, the virus results in common influenza-like symptoms (cough, fever, and fatigue) or even remains asymptomatic. However, unfortunately, in 10 to 20% of the patients, the inflammation results in more complicated conditions that resulted in more than 5.5 million deaths in early 2022 [2]. Because of the COVID-19 pandemic, the research on developing new treatments/therapies and vaccines against the virus, including drug repurposing and de novo design, took over a remarkable significance and implication at a global level and time plays a fundamental aspect in this field [3, 4]. Sadly, the drug development is long and expensive process, it is estimated that during the years 2000 to 2015, the average cost of developing a new approved drug was over USD 2.5 billion per single approved molecule! Commonly, it takes up to 15 years from the design of a single drug to reach the market and only less than 15% of selected compounds that are tested on humans, are later proved as safe and effective and can be finally used.[5, 6] It sounds reasonable that humanity can't wait 15 years for the development of a new molecule for COVID-19 treatment and in this case, the computational chemistry applied in drug design is helping accelerate research and providing stunning results. In the field of computer-aided drug discovery (CADD), the scientific community worldwide is regularly developing new technology and algorithms to attain handy hit-compounds in a short time and reduce the entire cost

 doi:10.20944/preprints202203.0032.v1

[7]. Nowadays, the introduction of artificial intelligence (AI) and all the related techniques such as deep learning (DL), machine learning (ML), and other classical computational chemistry tools towards drug discovery has exhibited a significant power on the success rate and velocity of novel pharmaceutical identification [8]. AI and classical CADD tools can be used alone or combined to produce new approaches that integrate a broad range of algorithms with enhanced predictions capabilities. CADD is classically classified into two methods: structure-based drug design and ligand-based drug design. Both are faces of the same coin and massively rely on force fields, scoring functions, and algorithms to evaluate and rank the studied molecules' energy contribution in the targeted macromolecular biological system. While the computer-aided structure-based drug design (*e.g.*, docking) depend on the actual 3D structure of the targeted binding site of the targeted receptor protein to understand the stabilizing interactions at the molecular level between the studied ligand/receptor system, the ligand based-drug design (*e.g.*, 3D-QSAR modeling) approach relies on the recognition of a database of already know ligands interacting with the studied target receptor. Both structure and ligand bases technologies have several success stories and pursue a key role in the drug modern drug discovery process [9-13]. In this context, several of these approaches have been recently employed to research novel drug candidates against COVID-19. The absence of complete protection with vaccination or powerful drugs for the treatment of the infection, the mutability of the virus, and the mortality rate impel the fast discovery of novel molecules active against COVID-19, and CADD is believed to be a valuable tool to achieve the goal [14]. Also, AI and ML have been broadly exercised from the beginning of the pandemic in discovering new treatments, vaccines, and drug repurposing and helping clinicians in pharmaceutical-related big data analysis and explanation for a better understanding of the outcome of the disease. Ongoing evidence shows that AI is being exploited to find potential novel molecules, repurpose drugs, find novel drug targets, and design novel and more effective vaccines. This review describes the current state of the art and the stunning result of AI and ML in the last two years of research in the field of de novo drug design since the COVID-19 pandemic started. Despite sometimes of difficult classification because of multidimensional structure- and ligand-based approaches are used, we divided the review into three main paragraphs reflecting the application on the AI in each single paper (*i.e.* if most AI used in the structure- or ligand-based for the multidimensional approaches), a paragraph regarding de novo vaccine design is also reported.

2. Structure-based artificial intelligence methods for small molecules

SARS-CoV-2 spike protein (S protein) is the leading mediator of viral entry into the cells and infection by binding the human Angiotensin-II Converting Enzyme (ACE2) and therefore represent an attractive target for drug therapies [15]. Recently, Srinivasan et al. developed a surrogate multi-task neural network (MTNN) model that replaces docking simulation in the finding of new molecules targeting the spike protein [16]. Monte Carlo algorithms and recurrent neural network (RNN) as rollout were endorsed to explore the chemical space of millions of potential molecules using SMILES input. In this way, they discovered 97,973 new molecules not included in the existing starting databases. The molecules were docked using Vina for their ability to bind the pocket region of the SARS-CoV-2 S-protein/ACE2 complex (SARS-CoV-2 S-protein (NCBI Reference Sequence YP_009724390.1)/ACE2 receptor (PDB ID: 2AJF) [17]. The docking calculation search space was chosen as 1.2 nm × 1.2 nm × 1.2 nm and includes the binding pocket located at the S-protein/ACE2 interface [18].

Several molecules with good scores were selected and compared to FDA-approved drugs and BindingDB data set. Despite identifying compounds with great binding affinity, MTNN performance had to be improved using active learning and extending the training data to have a correlation between the SMILES string and Vina scores considering a larger domain of SMILES space. This method incremented the molecules selected (over

 doi:10.20944/preprints202203.0032.v1

300,000). However, using Vina scores selection criteria, 200 promising molecules were chosen.

This aforementioned method accelerated the exploration of the vast chemical space represented by SMILES strings to evaluate structural features and discover structural similarities between top-performing candidates.

Many of the reported computational studies have selected molecules with high affinity to the Spike protein. However, these studies are limited to the receptor ACE2/protein Spike interface [18]. This represents a limitation since it dramatically reduces the possibility of identifying potential allosteric inhibitors of ACE2-Spike complex formation, leading to identifying a limited number of potentially active compounds.

To overcome this, a new machine learning approach, namely SSnet, was used to identify new potential drugs by screening a library of approved drugs from DrugBank and ZINC databases [19] targeting two different conformations (open and closed) of the ACE2 receptor as well as ACE2 in complex with S1 domain of the S protein, that is the protein responsible for the binding with human cells [20]. After cross-validation of the hits using the Autodock Vina scoring function [21], the SSnet approach was extended to a library of 750,000 molecules in BindingDB to gain additional information regarding *de novo* drug design.

After that, to accelerate the identification of high-affinity scaffolds to test for their *in vitro* activity, a web interface, where molecules are grouped according to their similarity on a 2D map and colored based on binding affinity to the protein, was developed. This system allows selecting a certain point into the interface to explore the effect of singular scaffolds and functional groups on the binding score or affinity. Moreover, this approach can be applied to other therapeutical targets besides CoVID-19.

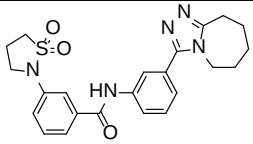
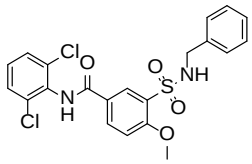
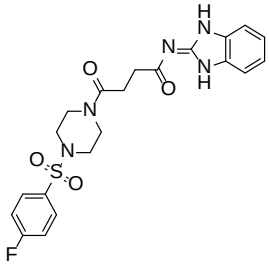
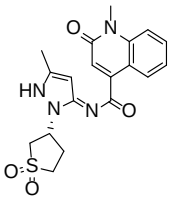
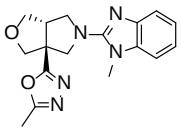
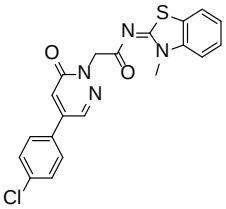
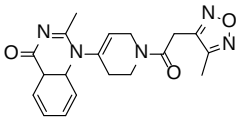
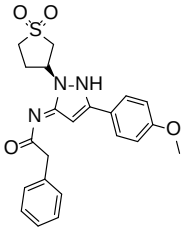
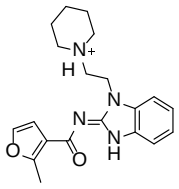
SARS-CoV-2 3CL^{pro} main protease (Picornain 3C-like protease, also referred to as M^{pro} for the main protease) is a homodimeric cysteine protease representing an attractive target for trans-variant activity since no mutations have been observed yet over this protein.

High-throughput virtual screening (HTVS) coupled with ML experiments have been performed to obtain potential virtual inhibitors against the targeted protein rather than trusting commercial "corona-focused libraries". The system was associated with an ML classification experiment where each compound is indexed into the chemical space of Mpro inhibitors, viral protease inhibitors, or a new chemical space. This approach has the advantage of taking into consideration potential drug that would otherwise have been omitted and gaining information into the possible mechanism of action of the selected compounds [22]. Initially, ZINC 15 library [23] was employed, and over 9 million compounds with a molecular weight below 200 g/mol were selected. The target SARS-CoV-2 M^{pro} crystal structure downloaded from the database has the PDB ID 6Y7M [24]. HTVS docking was performed using CmDock docking calculations considering the QuickProp, QPlogS descriptor that indicates possible soluble compounds, and 200 hits were selected. A set of M^{pro} inhibitors and viral cysteine protease inhibitors collected in the ChEMBL database with experimental IC_{50} < 100 μM were selected, and a set of 208 chemical descriptors were calculated to organize the compounds in the perspective of their representative chemical space [25].

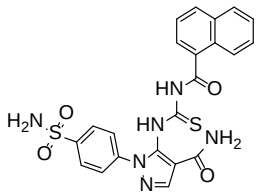
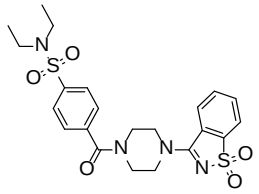
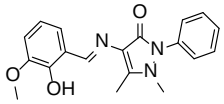
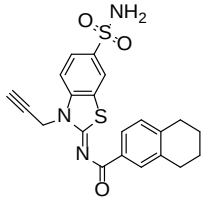
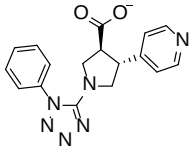
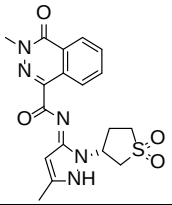
Using this HTVS method, a set of top-scoring compounds that could inhibit SARS-CoV-2 main protease has been identified for further compound prioritization in biological evaluation experiments (Table 1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0032.v1

Table 1. Selected top-scoring compounds by HTVS on the SARS-CoV-2 main protease.

n	Structure	CmDock Docking Score
1		–32.51
2		–29.02
3		–26.80
4		–25.58
5		–25.53
6		–25.05
7		–24.76
8		–24.51
9		–24.17

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0032.v1

12		–24.01
11		–23.98
12		–23.61
13		–23.53
14		–23.26
15		–23.18

A different research group screened a library of molecules for their potential ability to inhibit SARS-CoV-2 main protease (Mpro) and the receptor-binding domain (RBD) of the spike protein by using the molecular docking software AutoDock Vina.[26] Mpro is a peculiar cysteine protease of the coronavirus family and has a crucial role in mediating viral replication and transcription. The absence of a homologous human protease makes this protein an important target against Covid-19 [27].

The studies included 7675 molecules from the African Natural Product Database (AfroDB) and North African Natural Product Database (NANPDB), 43 FDA-approved antivirals, and 940 compounds derived from a machine learning study on viral Mpro [28]. AfroDB and NANPDB 470 were filtered using an ADMET predictor in order to include in the study only the compounds with low toxicity and molecular weight between 250 and 350 g/mol.

A library of 2430 compounds was selected and docked for the Mpro-ligand complex and RBD-ligand complex using Autodock Vina. 36 compounds with binding affinities ≤ –7.5 (kcal/mol) against both RBD and Mpro were selected and characterized for their binding affinity. After that, a predictor of biological activity using a Bayesian-based approach was accomplished and led to identifying 6 novel potential bioactive molecules. The leads selected NANPDB2245, NANPDB2403, fusidic acid, ZINC000095486008, ZINC0000556656943, and ZINC001645993538 (Figure 1) were subjected to molecular mechanics simulations involving molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) calculations that showed stable protein-ligand complexes with all the compounds with free binding energies < −3.58 kcal/mol with each receptor. The compounds

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0032.v1

identified showed good pharmacological profiles with little toxicity. However, in vitro studies are still needed to corroborate the findings.

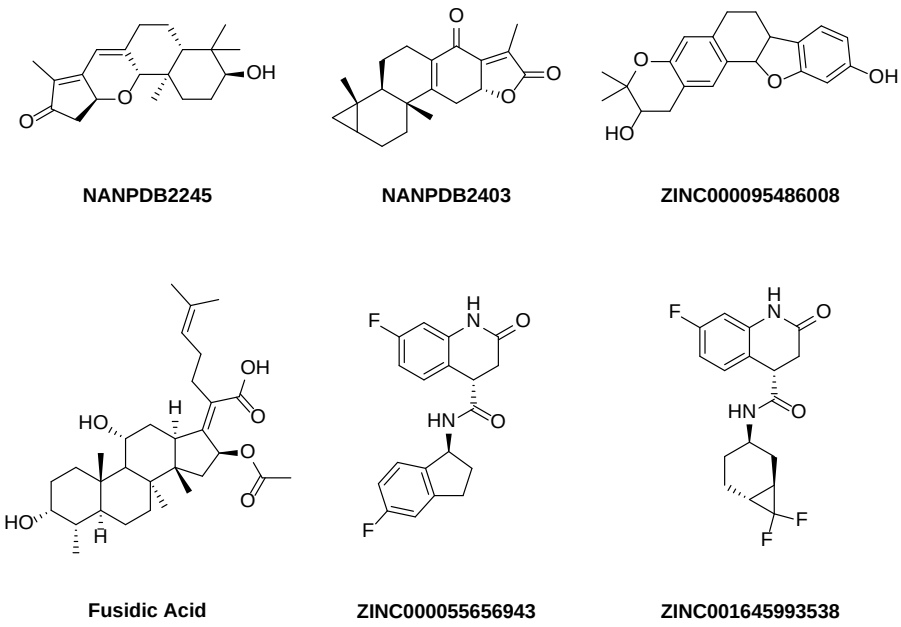


Figure 1. Ligand ID and structures of selected hit compounds.

Recently, Born et al. built up a new method for discovering and synthesizing drugs against SARS-CoV-2 [29]. This procedure that merges biology and chemistry for target-driven molecular design associated with an automatic synthesis plan generator can be virtually applied to any protein target. This approach uses deep generative models that implement a conditional molecule generator to propose drug candidates by exploring the latent space of proteins and small molecules. The promising of this approach is the possibility to generalize to unseen targets.

In this way, a conditional molecular generator can produce novel structures expressly designed to target a protein of interest [30]. In this work, 41 SARS-CoV-2 related protein targets as labeled in UniProt have been retrieved, among which Mpro and spike are the most targeted ones.

The first step of the procedure consists of encoding the selected protein sequence in a continuous and compressed latent space. The latent representation of the profile is decoded through the molecular decoder of trained variational autoencoders (VAE) generating valid molecules as drug candidates. Predictions studies of the toxicity were also performed by employing the Tox21 database [31], and the molecules were screened against 12 toxicity assays of nuclear receptor and stress response pathways. Finally, they were divided into toxic and non-toxic. The selected compounds' binding affinity was predicted using a multimodal deep learning model that classify compound-protein-interaction samples as binding or non-binding.

To assess the feasibility of synthesizing the generated compounds, the retrosynthetic pathways of a subset of candidates for each target were assessed using the interface of IBM RXN (https://rxn.res.ibm.com/) [20]. For half of the generated molecules, the synthetic route could be accomplished in one or two steps starting from commercially available materials indicating that they are attractive drug candidates.

The founders of COVID Moonshot, a non-profit, open-science consortium of scientists from around the world dedicated to discovering globally affordable and easily-manufactured antiviral drugs against COVID-19, demonstrated the utility of a de novo design using machine learning with synthesis route prediction.[32] The purpose of the study was to generate new potential drugs targeting the main protease (Mpro) of the novel

 doi:10.20944/preprints202203.0032.v1

coronavirus. In fact, while classic approaches tend to modify existing compounds exploring limited chemical space, this machine learning method searches in larger chemical space. The inconvenience of this approach is the expense of synthetic accessibility, but this can be overcome using machine learning to predict the synthetic route.

The learning-to-rank machine learning model [33] here reported consists of a classifier that predicts the activity of a molecule compared to another compound by considering the difference in pharmacophore fingerprint. The output is plotted in a curve reporting the relative activities of the considered compounds. The ligands' ranking is better than a model that directly learns IC_{50}. The model used for training was the FastAI Tabular model (J. Howardet et al., https://github.com/fastai/fastai, 2018) with the initial input features composed by Morgan, Atom Pair, and Topological Torsion fingerprints implemented in RDkit (RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org).

After a selection of reasonable chemical perturbations, a fragmentation of synthetically accessible bonds, among which amides and aromatic C–C and C–N, were performed, generating 8.8 million of molecules that have been compared to the most potent molecule in the dataset. The manifold platform was used to predict the synthetic route prediction of the identified compounds (https://postera.ai/manifold). Finally, the best five predicted molecules with no more than 4 step synthesis prediction have been synthesized and evaluated for their ability to inhibit Mpro by fluorescence assay and are reported in Table 2 together with the most potent molecules obtained from the training set. Compound 16 that showed the best IC_{50} was also tested toward OC43 coronavirus in a live virus assay showing low cytotoxicity and discrete activity toward the virus (EC_{50} = 13 uM).

Oak Ridge National Laboratory Summit supercomputer was used for in silico drug discovery using enhanced sampling molecular dynamics (MD) and assembly docking using the popular docking program Autodock Vina.[34] The Summit supercomputer is currently the fastest in the United States, hosted at the Oak Ridge Leadership Computing Facility (OLCF). Summit is an IBM AC922 system consisting of 4608 large nodes, each with six NVIDIA Volta V100 GPUs per node. The temperature replica exchange molecular dynamics (T-REMD) routine [35, 36] which was chosen here for the MD calculations (see below) uses the interconnect not only to allow for parallelization of a single simulated molecule, but also to communicate between separate replicas of the system, each carried out at a different temperature, and performs exchanges between replicas to accelerate the conformational sampling of the structures [37].

The 24 systems analyzed comprise nine protein domains. Two of these, RBD of protein S (spike) and the *N*-terminal region of protein N (nucleocapsid), are structural domains that are bound within the virion. Protein N is used for packaging the viral genome and is essential for virion assembly [38]. The remaining seven domains come from nonstructural proteins (NSPs) 3, 5, 9, 10, 15, and 16, which form the replication complex and are involved in many key tasks that create new viral particles (Table 3) [39].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0032.v1

Table 2. The 5 compounds with predicted routes < 4 steps. The top 3 compounds from the training set, with potency and cytotoxicity measurements.

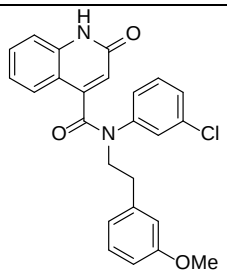
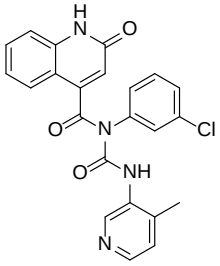
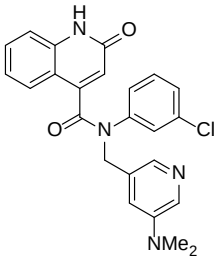
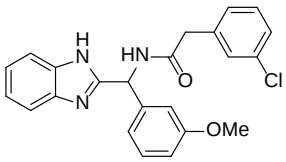
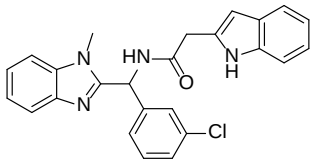
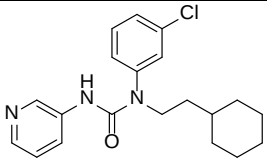
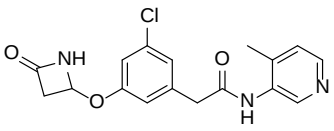
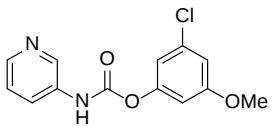
Top 5 predicted compounds		
		
16	**17**	**18**
		
19	**20**	
Top 3 training set compounds		
		
21	**22**	**23**
IC_{50} = 3.06 μM CC_{50} (Calu 3) = 18.2 μM CC_{50} (A549) = 14.1 μM	IC_{50} = 3.61 μM CC_{50} (Calu 3) = 63.5 μM CC_{50} (A549) > 100 μM	IC_{50} = 6.68 μM CC_{50} (Calu 3) > 100 μM

Table 3. 24 different model systems.

Protein/System (PDB code)		
S (Spike) Protein Receptor Binding Domain (RBD)/"Apo" (PDB ID: 6W41)	S Protein RBD/Complexed with ACE2 (PDB ID: 6W41)	MPro/monomer, CHARMM-GUI default protonation (PDB ID: 6Y2E)
MPro/dimer, CHARMM-GUI default protonation (PDB ID: 6Y2E)	MPro/dimer, "charged" protonation variant (PDB ID: 6WQF)	MPro monomer/HIE41 protonation variant (PDB ID: 6WQF)
MPro dimer/HIE protonation variant (PDB ID: 6WQF)	MPro monomer/HID41 protonation variant (PDB ID: 6WQF)	MPro dimer/HID41 protonation variant (PDB ID: 6WQF)
NSP15 (endoribonuclease)/hexamer (PDB ID: 6VWW)	NSP15 (Endoribonuclease)/monomer (PDB ID: 6VWW)	NSP10:NSP16 Complex (Methyltransferase) (PDB ID: 6W4H)
NSP10/monomer (PDB:6W4H)	NSP16/monomer (PDB:6W4H)	N (nucleocapsid) *N*-terminus phosphoprotein/monomer (PDB ID: 6M3M)
N (nucleocapsid) *N*-terminus phosphoprotein/tetramer (PDB ID: 6M3M)	N (nucleocapsid) *N*-terminus phosphoprotein/ tetramer complexed with Zn (PDB ID: 6YVO)	N (nucleocapsid) *N*-terminus phosphoprotein/monomer

 doi:10.20944/preprints202203.0032.v1

NSP9/monomer (PDB ID: 6W4B) PLPro/monomer "charged" protonation variant (PDB ID: 6W9C)	NSP9/dimer (PDB ID: 6W4B) PLPro/monomer "neutral" variant (PDB ID: 6WRH)	alternate crystal structure (PDB ID: 6YVO) NSP3 ADP ribose phosphatase/asymmetric unit (PDB ID: 6W02) NSP3 ADP ribose phosphatase (PDB ID: 6W02)

Two different docking databases were used - a potential ligand database merging the contents of SWEETLEADS [40] and the NCI-diversity database - yielding 13757 compounds - and the Enamine database using an accelerated version of Autodock (Autodock-GPU).

The authors used an experimental screening database of 2,900 chemicals tested by the National Institutes of Health, National Center for Advancing Translational Sciences (NCATS) and listed at https://opendata.ncats.nih.gov/covid19/databrowser, to compare with positives identified experimentally by NCATS.

Interestingly, all four experimentally tested compounds (*i.e.*, 100% of the tested compounds in the top 10 lists) were strongly active.

The ensemble docking performed affects database reuse limited to approximately 10,000 compounds. Many of these compounds are expected to be quite promiscuous in binding to targets. Two of the compounds identified in the richest 1% of the preliminary protein S screening have been reported in two registered clinical trials (quercetin and hypericin).

Pirolli D. et al., with the support of machine learning approaches, reported a structure-based virtual screening as an effective strategy to discover inhibitors of protein-protein interactions (PPIs) between SARS-CoV-2 RBD and human ACE2 using the ZINC database [41].

Using different ligand and structure-based approaches, a customized virtual screening (VS) strategy was set up. The first step of the VS strategy was the selection of a library focused on small molecule PPIs from a data set of 2 million compounds, using a ligand-based approach capable of recognizing chemical characteristics and scaffolds common to known modulators. For this purpose, a convolutional neural network (cNN) was trained to obtain a QSAR model capable of identifying potential PPI modulators within a virtual library of unknown molecules. The molecules classified by the cNN-based QSAR as potential PPI modulators were further filtered by the expected toxicological properties to discard compounds harmful to human health. The resulting virtual library was hooked to ACE2 to identify compounds with the best binding affinity for the spike protein interaction surface. The dataset of 30,029 ligands obtained from QSAR modeling and toxicity analysis against the Druggable Site-4 pocket region was used to screen for effective inhibitors of the protein-protein interactions of the SARS-CoV-2 RBD/ACE2 tip. Then a virtual Glide screening in SP mode and the next phase by XP docking. Based on the docking score, the first 15015 classified ligands (50%) were selected and reassessed with Prime MM-GBSA to estimate their binding free energy. The compounds were then filtered based on distance constraints by selecting only small molecules within 4.5 Å away from any atom of Tyr83 and Gln24 residues. The remaining 9730 molecules were then grouped based on their diversity, and the resulting 973 virtual hit compounds were further assembled into 66 clusters using interaction fingerprints.

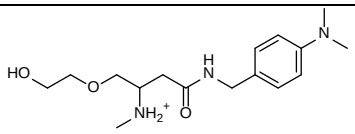
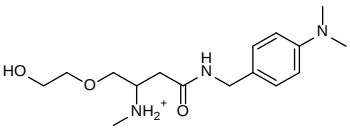
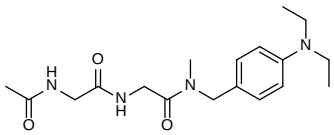
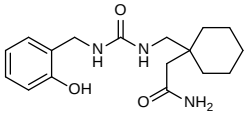
The results indicated that most of the compounds screened share a high similarity (0.6–0.7) with the training set.

Four compounds were selected as potential ACE2 surface binders capable of preventing RBD spike recognition and thus infection (Table 4). The presence of an aromatic region facilitates the interaction with ACE2 Phe28 (compounds 2 and 3) and Tyr83 (compounds 1 and 4). Furthermore, ACE2 Gln24 and Tyr83 contribute to the stabilization of ligands 1,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0032.v1

2, and 4 within the binding site by forming hydrogen bonds to their hydroxyl groups. Compound 1 is stabilized by the hydrogen bonds formed by the catechol with the oxygen atom Gln24 and the cycloheptyl fraction with Tyr83. Similarly, compound 2 is hydrogen-bonded to Tyr83 and Gln24 oxygen atoms by its hydroxyl and carbonyl oxygen atoms.

Table 4. Structures and calculated free binding energies (MM-GBSA score, in kcal/mol) of four top-ranking compounds.

Name	Structure	IUPAC Name	MM-GBSA score (kcal/mol)
Compound 1		*N*-[(1-hydroxycycloheptyl)methyl]-*N*-methyl-2,3-dihydroxybenzamide	−42.21
Compound 2		[({[4-(dimethylamino)phenyl]-methyl}carbamoyl)-methyl][2-(2-hydroxyethoxy)ethyl]-methylazanium	−42.07
Compound 3		*N*-[({[4-(diethylamino)phenyl]-methyl}(methyl)carbamoyl)-methyl]-2-acetamidoacetamide	−40.95
Compound 4		2-{1-[({[(2-hydroxyphenyl)methyl]-carbamoyl}amino)methyl]-cyclohexyl}acetamide	−39.5

3. Ligand-based artificial intelligence methods for small molecules

F. Pereira et al. succeeded in predicting five new inhibitors against SARS-CoV-2 Mpro using a CADD method based on a quantitative structure-activity relationship (QSAR) classification model that was built from 5276 organic molecules extracted from the ChEMBL database. Virtual screening was then performed using 11,162 marine natural products (MNPs) retrieved from the Reaxys® database. From the QSAR approach, 494 MNPs were selected and subsequently subjected to molecular docking against the Mpro. Among the evaluated compounds, five MNPs have been proposed as the most promising marine drugs as inhibitors of SARS-CoV-2 M^{pro}, among them a benzo[f]pyrano[4,3-b]chromene (Reaxys ID 7450892), notoamide I (Reaxys ID 19384758), hemindole SB beta-mannoside (Reaxys ID 26845562), and two derivatives of bromoindole (Reaxys IDs 10714788 and 10720065) Figure 2 [42].

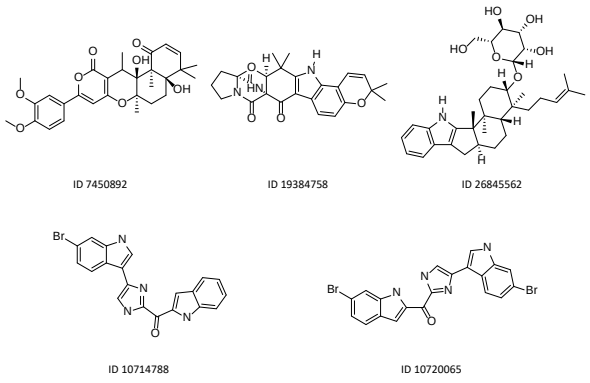


Figure 2. MNPs have been proposed as the most promising marine drugs leading as inhibitors of SARS-CoV-2 Mpro

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0032.v1

Combining a generative recurrent neural network model with transfer learning methods and active learning algorithms, R. Yassine et al. designed a novel set of small molecules capable of effectively inhibiting the 3CL protease in human cells [43]. The novelty of this work is the use of active learning methods with generative recurrent neural networks (RNNs) containing long-term memory cells (LSTM). The active learning method facilitates the selection process by focusing on the areas of the chemical space that have the best chance of success, considering structural novelties. The authors built a database consisting of multiple datasets such as FDA-approved drugs (from the ZINC database), natural products (from SuperNatural), and a manually developed database representing drug-like bioactive molecules. In the first phase of this study, by applying the RNN deep learning methodology, the LSTM-based RNN model was created to generate reliable and high-quality SMILES to design new drugs. Subsequently, molecules structurally similar to drugs with known activity against the specific SARS-CoV-2 target were generated. In this way, they were able to find a model capable of discovering new drugs using fragment-based drug discovery (FBDD) to create a library containing a series of SMILES inspired by the well-known compounds. The model generated 25,000 small molecules from the learned chemical space as described above. After removing duplicates and identical molecules from the database used for training, the remaining dataset consisted of 22,173 molecules. These molecules were then subjected to other filters such as physicochemical properties, drug similarity, and synthetic accessibility, resulting in a set of 6,962 molecules. The generated molecules were then screened for affinity to the 3CL protease. After the virtual screening, a total of 41 molecules were obtained, with a virtual screening score of less than −7.0 kcal/mol. Among these, 4 molecules resulted in a binding affinity score lower than −18 kcal/mol (Figure 3).

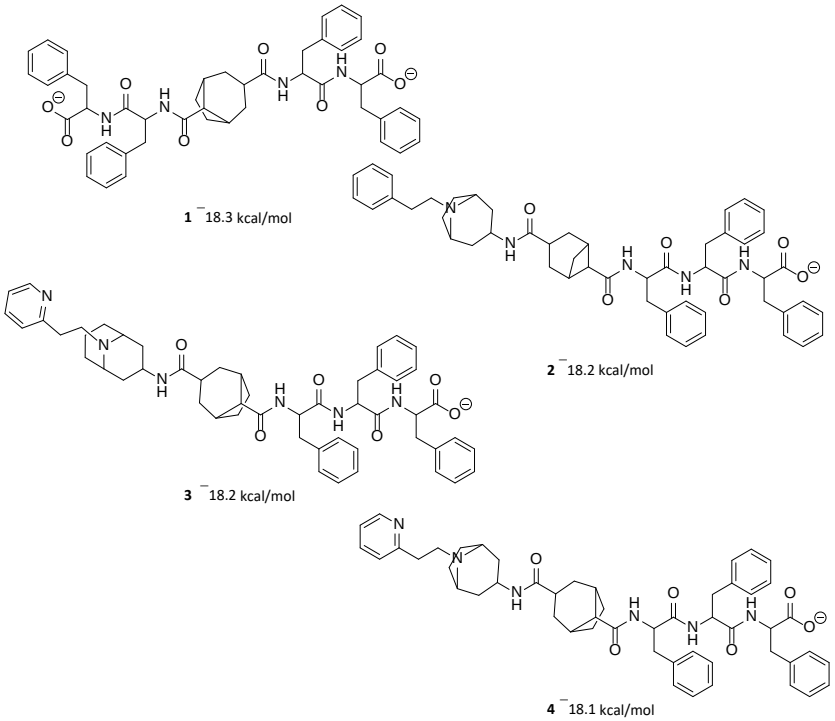


Figure 3. The generated molecules by R. Yassine et al. with the lowest binding affinity score.

The reported model developed by F. P. Silva-Jr et al. also overperformed the Chemprop model available for free on an external test set of fragments shielded against SARS-CoV-2 Mpro [44]. The method is divided into 3 main phases - formation and validation of the generative model based on general chemistry - development of the model for Mpro chemical space of SARS-CoV-inhibitors - formation of a classifier for the prediction of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0032.v1

bioactivity using transfer learning. Using the improved classifier to predict the bioactivity of the 70,000 valid SMILES, the authors classified 1,697 active molecules, and the Uniform manifold approximation and projection (UMAP) plot showed good to optimal overlap between the predicted results and actual value of inhibition for Mpro inhibitors in the studied chemical space. Among the resulting molecules, 20 compounds were classified as high-confidence hits, with probabilities ranging from 0.99 to 1.0. These molecules found to be potential inhibitors were then subjected to a docking simulation using the crystalline structure of SARS-CoV-2 Mpro (PDB: 6W79). Nine compounds showed binding poses similar to experimentally validated inhibitors in X-ray crystal complexes with Mpro. These molecules include three benzotriazoles and four benzothiazolylketones, a peptidomimetic, and an *N*-(2-pyridyl) acetamide derivative (Figure 4).

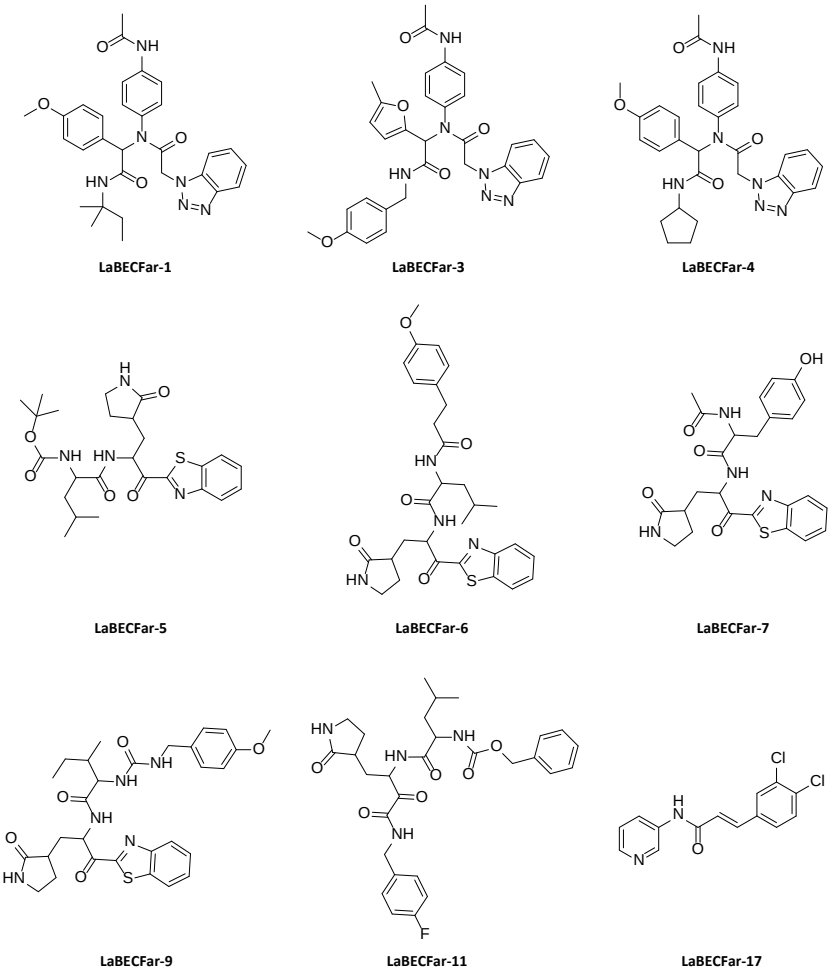


Figure 4. The nine compounds reported by F. P. Silva-Jr et al. that showed similar binding positions to the experimentally validated inhibitors in X-ray crystal complexes with Mpro.

Roy et al. recently identified new chemical entities (NCEs) starting from a dataset of approximately 1.6 million drug-like small molecules from the ChEMBL database, which were collected for pre-training of a generative model [45]. The set of molecules obtained after applying the physicochemical properties filters was screened using RDKit by applying the following four filters: Pan Assay Interference Compounds, the BRENK filter, the NIH filter, and the ZINC filter. These filters use rules to avoid toxic compounds and synthetically impractical molecules. Potential NCEs for synthesis and testing against SARS-CoV-2 were finally subjected to docking simulation and selected using a virtual screening score cutoff of −8.5 kcal/mol. The results showed that 5 of the top 15 compounds have a high virtual screening score and a more remarkable similarity to existing protease

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0032.v1

inhibitors, notably one of these NCEs possessing a higher virtual screening score of −9.1 kcal/mol (Figure 5).

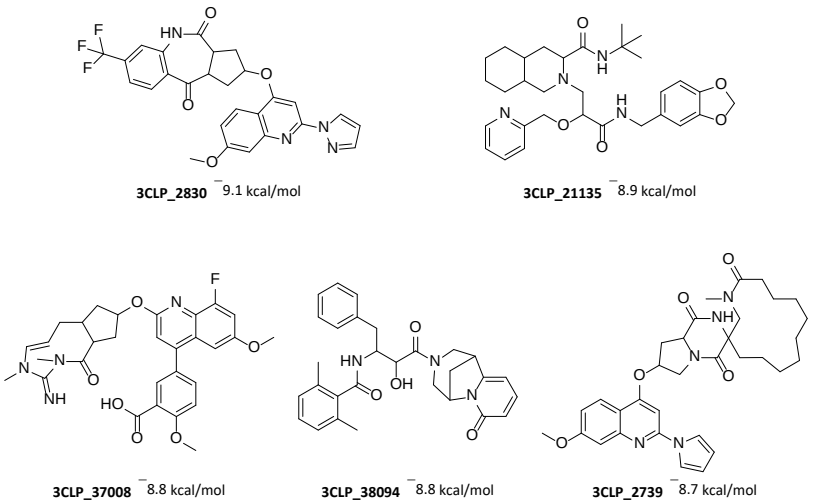


Figure 5. NCEs with the highest virtual screening score and a more remarkable similarity to existing protease inhibitors.

The unique approach to data curation coupled with random forest (RF) analysis by K. Cooper et al. produced a specific target and validated predictive fingerprints (PFFs) that have a high predictability value on multiple targets such as plasma kallikrein, HIV protease, NSP5, NSP12, AT-1, and the JAK family. This broad applicability to different biologically relevant targets (protease, RNA polymerase, G protein-coupled receptor (GPCR), tyrosine kinase, and a phenotypic assay) as well as to different chemotypes within a target is an important strength and differentiator of the applied methodology. The capability of this methodology allows for each target to create a binary decision tree for inactive or active compounds or a ternary decision tree for weakly active, moderately active, and highly active compounds also suggests that the models could be used for virtual screening of libraries of target-specific compounds to select the most active compounds for synthesis and/or clinical testing. Regarding the SARS-CoV-2 target, about 5600 FDA-approved drugs were examined in this study. Molecules that showed more than 75% inhibition were considered active, and among the wide range of FDA-approved drugs, 267 were identified as active. Leveraging the bioactivity data of ChEMBL and a subset of the data, it was trained on physicochemical characteristics from the set of 110 chemical properties, and a set of 868 compounds was then used to establish a binary classification model capable of predicting whether a molecule was active or inactive in the test; overall the model was able to identify the bioactivity with an accuracy of 65% [46].

E. Glaab et al. reported a combined virtual screening study, molecular dynamics (MD) simulation, machine learning, and in vitro experimental validation analysis, which led to the identification of small molecule inhibitors of 3CLpro with micromolar activity and to a pharmacophore model. The methodology consists of a filtering system involving screening multiple receptors and ligands in combination with a final MD simulation to confirm the binding stability for the selected compounds. The structural screening was then integrated with a machine learning-based screening for compounds selection using a molecular-descriptive data set derived from known ligands and non-ligands for 3CLPro. The best compounds selected using these in silico screening methods were then experimentally evaluated to determine the subset of stable ligands and their 3CLpro inhibitory activity using the in vitro Forster-type resonance energy transfer (FRET) assay. From the various screening analyses conducted, 95 molecules were identified and tested. 7 of the 95 tested compounds were confirmed active, and one of these showed the lowest IC_{50} value of 31 μM [47].

 doi:10.20944/preprints202203.0032.v1

The deep learning (DL) model can generate 1D or 2D sequences ligands structures. However, rational drug design requires 3D ligand structures that target the crystalline structure of proteins. To solve this problem, Q. Bai and coworkers developed MolAICal software, which allows to generate 3D drugs in the 3D pocket of protein targets by combining the merits of the deep learning model and classical algorithm. The software essentially consists of two modules. In the first one, FDA-approved drug fragments are used to train the Wasserstein-based deep learning model generative adversarial networks (WGANs). The generated fragments of the deep learning model are further used to grow the 3D ligands in the protein pocket. In the second module, drug-like molecules from the ZINC database are used to train the WGAN-based deep learning model. Then the affinities between the generated molecules and the proteins are evaluated through molecular docking experiments with AutodockVina. The membrane protein glucagon receptor (GCGR) and the non-membrane target SARS-CoV-2 Mpro were chosen to analyze the drug design capabilities of MolAICal. In this way, the software can generate various ligands that have an ever-higher 3D structural similarity with the ligand crystallized in the active site of GCGR or SARS-CoV-2 Mpro [48].

Mekni N. et al. developed a machine learning approach using support vector machine (SVM) classification, to share new knowledge for designing novel Mpro inhibitors from a data set of two million commercially available compounds. The model was able to classify two hundred new chemotypes as potentially active against the viral protease [49].

Feature selection was made by implementing a python3 script using the Sklearn libraries. The script is available in the GitHub repository (https://github.com/NedraMekni/COVID-19).

The selection of the characteristics was based on the training set, with the aim of identifying the crucial molecular descriptors able to explain the possible correlation between the activity of the Mpro inhibitors and their chemical structures, implemented the elimination of the recursive characteristics of the forest random (RF-RFE) in order to select relevant molecular descriptors [50], leading to the automatic optimization of the number of features to be selected and to the definition of the optimal number of decision trees to build the forest. Compounds labeled as active by SVM were subsequently evaluated through consensus docking studies on two PDB structures, and their binding mode was compared with known protease inhibitors. Of the 25 facilities analyzed, only five (5RF6, 5RGW, 6WCO, 5R82, and 6W79) met the criteria. On these 5 PDBs, factor B (mean of the PDB B-value) was checked to assess the quality of the protein structure.

The five best compounds selected by consensus were then subjected to molecular dynamics to investigate the stability of the binding interactions.

MD simulations at 200 ns were performed on the two best performing PDBs (6WCO and 5RGW) to verify the stability of the interactions recovered within the crystal structure.

The five best compounds selected by consensus were then subjected to molecular dynamics to investigate the stability of the binding interactions. It should be noted that the compounds selected by SVM showed all the essential interactions reported in the literature.

4. Artificial intelligence methods vaccine design

An in silico deep learning approach was proposed to predict and design a multiepitope vaccine (DeepVacPred), in combination with in silico immunoinformatics and deep neural network strategies. The DeepVacPred Computing System directly predicts 26 potential vaccine subunits from the SARS-CoV-2 tip protein sequence [51].

The DNN architecture to block 26 fragments in the SARS-CoV-2 spike protein as candidates for the vaccine subunit was the first step proposed by the authors. Subsequently, linear B cell, CTL, and HTL epitopes were used to select and construct the final vaccine.

All overlapping protein fragments with a length of 30 aa are generated by the spike protein sequence 1273 aa SARs-CoV-2. DeepVacPred first tests these proteins sequences and predicts 132 potential vaccine subunits. Following this prediction, DeepVacPred

provides 26 potential vaccine subunits for further evaluation and construction. These subunits are most likely to contain B cell epitopes and multiple T cell epitopes and have high antigenicity and low allergenicity.

In silico methods were used to study linear B cell epitopes, cytotoxic T cell epitopes (CTL), helper T cell epitopes (HTL) in the 26 candidate subunits.

The B cell epitopes, predicted on the 26 vaccine subunits, are parts of antigens that bind to immunoglobulin or antibody, capable of activating B cells to provide the immune response [52]. Linear B cell epitopes are predicted from four online servers, including BepiPred [53], SVMtrip [54], ABCPred [55], and BCPreds [56]. First, they used BepiPred for the main forecast and the other three servers to check the results of the BepiPred forecast. Additionally, the proprietary RaptorX server was used to evaluate the surface accessibility of SARS-CoV-2 to validate that the B cell epitopes in those subunits are well exposed.

CTLs recognize infected cells using class I MHCs to bind to certain CTL 26 epitopes. NetMHCpan 4.1 server [57] 43 was used to predict potential CTL epitopes. All overlapping 9aa peptide sequences in the 14 vaccine subunits are tested with the 12 most common class I alleles of human leukocyte antigen (HLA), including HLA-A1, HLA-A2, HLA-A3, HLA-A24, HLA-A26 , HLA-B7, HLA-B8, HLA-B27, HLA-B39, HLA-B44, HLA-B58, and HLA-B62 to evaluate their binding affinities and predict potential CTL epitopes [58, 59].

HTL helps other immune cells' activity and recognizes infection by using MHC class II to bind with specific HTL epitopes [60]. The NetMHCIIpan 4.0 [61] server was used to predict potential HTL epitopes. All overlapping 15aa peptide sequences in the 14 vaccine subunits are tested with the 13 most common HLA Class II alleles, including HLA-DRB1-0101, HLA-DRB1-0301, HLA-DRB1-0401, HLA-DRB1-0701, HLA-DRB1 -0801, HLA-DRB1-0901, HLA-DRB1-1001, HLA-DRB1-1101, HLA-DRB1-1201, HLA-DRB1-1301, HLA-DRB1-1401, HLA-DRB1-1501, HLA-DRB1-1601 to evaluate their binding affinities and predict potential HTL epitopes [61]. The total HLA score is calculated for each vaccine subunit.

The 3D structure of the designed vaccine was then predicted, refined, and validated by other in silico tools. The GalaxyRefine [62] server was employed to refine the 3D structure model of the final vaccine. Among the 5 refined models predicted by GalaxyRefine, the model 2 was chosen as the final vaccine model based on its quality scores with a reported RMSD of 0.58.

In conclusion, this proposed artificial intelligence (AI)-based vaccine discovery facility accelerates the vaccine design process and builds a 694 aa multiepitope vaccine containing 16 B cell epitopes, 82 CTL epitopes, and 89 HTL epitopes, which promises to fight SARS-CoV-2 viral infection showing good antigenicity, population coverage, and good physicochemical properties and structures, providing great potential for next stage COVID-19 vaccine design with actual clinical trials.

AI was used in another study to predict the rationale for designing universal vaccines against SARS-CoV-2, which contain an extensive repertoire of T-cell epitopes capable of providing coverage and protection to the global population. To achieve these goals, the authors profiled the entire SARS-CoV-2 proteome through the 100 most frequent HLA-A, HLA-B, and HLA-DR alleles in the human population, using the presentation of the cell surface antigen infected with host and immunogenicity predictors from the NEC Immune Profiler suite of tools and generated sufficiently complete epitope maps [63].

Antigen (AP) presentation was predicted by a machine learning model that integrates information from several HLA binding predictors (in this case, three distinct HLA binding predictors trained on IC50nm binding affinity data) and 13 different antigen processing predictors. The emitted AP score ranges from 0 to 1 and was used as an input to calculate immune presentation (IP) through the epitope map, penalizing those peptides that have degrees of similarity to humans compared to the human proteome and rewards peptides less similar. The resulting IP score represents those presented HLA peptides that can be recognized by circulating T cells.

 doi:10.20944/preprints202203.0032.v1

Epitope maps were created for all viral proteins and an example based on IP scores for protein S containing candidate CD8 and CD4 epitopes for the 100 most frequent human HLA-A, HLA-B, and HLA-DR alleles.

Epitope hotspots that shared significant homology with proteins in the human proteome were removed to reduce the possibility of inducing off-target autoimmune responses. In addition, the antigen presentation and immunogenic landscape of all non-synonymous mutations in 3,400 different virus sequences in the GISAID database with the AP potential of the Wuhan Genbank reference sequence were also analyzed to identify a trend by which SARS-COV-2 mutations are expected to have a reduced potential to be introduced by host-infected cells and, consequently, detected by the host's immune system.

In order to assess whether epitope hotspots are solid enough across sequenced and mutant strains of SARS-CoV-2, the AP-based Monte Carlo epitope hotspot statistical model was used, and 10 virus sequences were analyzed among the 10 most mutated viral sequences. different geographical regions.[64] Most of the hotspots were present in all sequenced viruses, however the hotspots were eliminated and/or new hotspots emerged in these divergent strains.

This is the first computational approach to generate vaccine design designs from large-scale epitope maps of SARS-CoV-2 optimized on diverse T-cell immune responses across the global population.

Finally, an HLA haplotype database of approximately 22,000 individuals was evaluated to develop a “digital twin” simulation to model the effectiveness of different combinations of hotspots in a diverse human population; the approach identified an optimal variety of epitope hotspots that could provide maximum coverage in the global population.

The CD8 epitope maps of these optimized epitope hotspots are based on AP predictions of the peptides presented on the surface of the host-infected cells and visible to the host’s CD8 T cells. Furthermore, these antigen-presented peptides are subject to IP predictions, which infer the epitopes most likely to activate a T cell.

In conclusion, the authors combined antigen presentation at the infected host cell surface and immunogenicity predictions from the NEC Immune Profiler with a robust Monte Carlo and digital twin simulation in order to delineate the entire SARS-CoV-2 proteome and identify a subset of epitope hotspots that could be exploited in rational vaccine design to provide broad coverage across the global population [65].

Kesarwani et al. examined data acquired from proteomic analyses of human cell lines infected with SARS-CoV-2 and COVID-19 patient samples to identify peptides useful for diagnostics and vaccine development. Initially, a large-scale meta-analysis of changes in 358,558 SARS-CoV-2 protein sequences detected in samples from 42 countries was performed. For sequence conservation analysis, a protein data cluster was generated for each SASR-CoV-2 protein. Hence, there were five regions and 14 regions for the nucleocapsid and spike proteins, respectively [66].

Two cell lines and four proteomes from naturally infected human patients were used for high-confidence identification of peptides using viral and human protein sequences as references. In total, 361 and 81 peptides of viral origin were identified in cell lines and patient samples, respectively. A few peptides with varying lengths were found from different parts of the same proteins, including 57 component peptides of the spike protein. Of these 57 peptides, three are components of the S1 (14–685), S2 (686–1273), and RBD (319–541) regions of the spike protein, respectively.

Therefore, the authors explored host responses to the virus in both cell lines (colon carcinoma-2 and H1299) and naturally infected COVID-19 patient samples. Many proteins are involved in biological processes related to the immune system, such as regulation of immune responses, leukocyte migration, autophagy, processing, immune system development, antigen presentation, or leukocyte-mediated cytotoxicity.

 doi:10.20944/preprints202203.0032.v1

323 and 143 human peptides were identified in the cell line and patients, respectively. Only five (MDGA1, PIK3C2A, FOXP2, DCAF5 and IVD) were detected in both sample sets. While MDGA1 plays an important role in inhibitory synapse formation [67], PIK3C2A is involved in several intracellular traffic and signalling pathways [68]. FOXP2 is a transcription factor that can regulate hundreds of genes in different tissues, including the brain [69]. DCAF5 is a receptor of the CUL4-DDB1 E3 ubiquitin-protein ligase [70], and IVD is an enzyme essential for the beta-oxidation of mitochondrial fatty acids.

Thirty-three proteins were found and matched entries in the InnateDB database. Most of these proteins are involved in immune-related functions such as protein binding (TAB1, SREBF2, HSP90AA1, RB1, STAT3, DCN, IL1R1, BNT3A2, PIK3R2, CCR6), transferase activity (TREM2 , ABL1, S100A12, C4BPB), the protein dimerization (UBE2N, CSF1R) and lipopeptide binding (EPS8, CD36) [71].

Once the proteins related to the immune system process were identified from the patients' cell line and proteomes, they were used to generate a protein interaction network. The generated protein interaction network, which includes 403 nodes and 671 edges, identified higher-ranking hubs and bottlenecks.

Therefore, multi-step filtering was applied to identify potential diagnostic peptides. Initially, to avoid cross-reactivity, the identified peptides (442) were filtered to exclude peptides from the human and human saliva microbiome (418) and subsequently peptides from a targeted group of pathogenic bacteria and viruses (129). Subsequently, using the results of the RNA-Seq data analysis of the infected cell lines, the expression of the selected peptides was verified to avoid the selection of poor peptides for diagnostic purposes. Then, four antigenic peptides were selected for attachment to known viral T-cell receptor (TCR), class I and II peptide major histocompatibility complex (pMHC), and paratopes sequences identified. They also tested the paratope binding affinity of SARS-CoV-1 T and B cell peptides that had previously been experimentally validated. The resulting antigenic peptides have a high potential for generating antibodies specific to SARS-CoV-2 and the peptides of the paratopes can be used directly to develop a COVID-19 diagnostic assay

The antigenic peptides found in this study have a high potential for generating antibodies specific to SARS-CoV-2, and the paratopes peptides can be used directly to develop a COVID-19 diagnostic assay. In addition, the paratope binding affinity of SARS-CoV-1 T and B cell peptides that had previously been experimentally validated was also tested. The assembled paratopes showed a greater binding affinity for SARS-CoV-2 antigens and proteins than SARS-CoV-1.

In conclusion, the authors in this study explored both the cell line and proteomes of naturally infected COVID-19 patients via in silico methods and identified four SARS-CoV-2 antigens and three antigen-binding peptides that could be used to develop diagnostic assays. The proposed antigenic peptides can be used for antibody generation, and the paratopes sequences can be used directly for COVID-19 diagnostic test and vaccine development.

5. Conclusions and perspective

AI technologies have shown impressive ability in COVID-19 research from real-time tracking of the virus spread to the developing of novel drugs and vaccines faster than never seen before. Unfortunately, the still high number of positive cases and deaths from COVID-19 infection and the absence of effective treatments and complete cover by vaccination persist in influencing global health, resulting in colossal concern that requires the necessary discovery of novel molecules for the cure and prevention of the infection. The structural definition of targetable proteins for the therapy and prevention of the infection has recently boosted the structure-based virtual identification of small molecules. The discovery of some promising compounds has driven the optimization of those scaffolds through ligand-based drug design.

From ligand-based-AI technologies, where only the ligands structures are considered to train the models, to structure-based AI technologies where the protein target is also

taken into account in the process of drug design and vaccine design aided by AI. In this review, we reported and discussed the more cutting-edge technologies in the field of COVID-19 de novo drug design. ML on its own and other AI-based technologies are being of critical importance in responding to problems for COVID-19 research. They showed to be efficient tools to quickly analyze large amounts of data, estimate drug repurposing against COVID-2019, identify association of these repurposed drugs, or estimate dosage adjustments and other clinical related issues such as early diagnosis, identifying people at risk, and predicting disease evolution [72, 73].

The field of de novo drug design is also part of the AI leaded research in the COVID-19 field. Several molecules have been already identified from impressive huge databases of billions of compounds as reported in this review. Additionally, pairing different approaches considering all the information produced by omics sciences should lead to developing personalized strategies, on top due to the mutability of this RNA virus and the emergence of drug resistance problems, it is mandatory to start considering targeting multiple targets that will be more effective and might help in overcoming future drug resistance.

Interestingly, although none of the de novo identified molecules has entered clinical trials, some of the repurposed molecules identified by AI technologies have entered this phase. These are mainly already used, and approved antibiotics, anti-inflammatory, antivirals, anticancer, and ACE2 drugs, and some already are in clinical trials according to ClinicalTrials.gov (https://clinicaltrials.gov/, accessed on 10 February 2022) [74]. In conclusion, as the pandemic crisis has exponentially accelerated the adoption of analytics and AI is not surprising to say that AI will lead the prevention and the research not for only COVID-19 related issue but also many other diseases in the following decade, speeding up the drug design process making AI at the forefront for the fighting of public health problems.

Author Contributions: Conceptualization, G.F. and A.R.; formal analysis, C.Z., V.P., G.F., D.G. and A.R.; investigation, C.Z., V.P., G.F., D.G. and A.R.; resources, C.Z., V.P., G.F., D.G. and A.R.; writing—original draft preparation, C.Z. V.P. G.F. and D.G.; writing—review and editing, C.Z. V.P. G.F. and A.R.; supervision, A.R. and G.F.; project administration, A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cucinotta, D.; Vanelli, M., WHO Declares COVID-19 a Pandemic. *Acta Biomed* **2020,** 91, (1), 157-160.
2. Pastorino, R.; Pezzullo, A. M.; Villani, L.; Causio, F. A.; Axfors, C.; Contopoulos-Ioannidis, D. G.; Boccia, S.; Ioannidis, J. P. A., Change in age distribution of COVID-19 deaths with the introduction of COVID-19 vaccination. *Environ Res* **2022,** 204, (Pt C), 112342.
3. Gupta, R. K.; Nwachuku, E. L.; Zusman, B. E.; Jha, R. M.; Puccio, A. M., Drug repurposing for COVID-19 based on an integrative meta-analysis of SARS-CoV-2 induced gene signature in human airway epithelium. *PLoS One* **2021,** 16, (9), e0257784.
4. Sultana, J.; Crisafulli, S.; Gabbay, F.; Lynn, E.; Shakir, S.; Trifiro, G., Challenges for Drug Repurposing in the COVID-19 Pandemic Era. *Front Pharmacol* **2020,** 11, 588654.
5. Brown, D. G.; Wobst, H. J.; Kapoor, A.; Kenna, L. A.; Southall, N., Clinical development times for innovative drugs. *Nat Rev Drug Discov* **2021**.
6. Wouters, O. J.; McKee, M.; Luyten, J., Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **2020,** 323, (9), 844-853.
7. Yu, W.; MacKerell, A. D., Jr., Computer-Aided Drug Design Methods. *Methods Mol Biol* **2017,** 1520, 85-106.
8. Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A., Jr.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G., Rethinking drug design in the artificial intelligence era. *Nat Rev Drug Discov* **2020,** 19, (5), 353-364.
9. Floresta, G.; Apirakkan, O.; Rescifina, A.; Abbate, V., Discovery of High-Affinity Cannabinoid Receptors Ligands through a 3D-QSAR Ushered by Scaffold-Hopping Analysis. *Molecules* **2018,** 23, (9).

10. Floresta, G.; Abbate, V., Machine learning vs. field 3D-QSAR models for serotonin 2A receptor psychoactive substances identification. *RSC Advances* **2021,** 11, (24), 14587-14595.
11. Floresta, G.; Amata, E.; Barbaraci, C.; Gentile, D.; Turnaturi, R.; Marrazzo, A.; Rescifina, A., A Structure- and Ligand-Based Virtual Screening of a Database of "Small" Marine Natural Products for the Identification of "Blue" Sigma-2 Receptor Ligands. *Mar Drugs* **2018,** 16, (10).
12. Floresta, G.; Cilibrizzi, A.; Abbate, V.; Spampinato, A.; Zagni, C.; Rescifina, A., 3D-QSAR assisted identification of FABP4 inhibitors: An effective scaffold hopping analysis/QSAR evaluation. *Bioorg Chem* **2019,** 84, 276-284.
13. Floresta, G.; Gentile, D.; Perrini, G.; Patamia, V.; Rescifina, A., Computational Tools in the Discovery of FABP4 Ligands: A Statistical and Molecular Modeling Approach. *Mar Drugs* **2019,** 17, (11).
14. Francis, A. I.; Ghany, S.; Gilkes, T.; Umakanthan, S., Review of COVID-19 vaccine subtypes, efficacy and geographical distributions. *Postgrad Med J* **2021**.
15. Gallagher, T. M.; Buchmeier, M. J., Coronavirus spike proteins in viral entry and pathogenesis. *Virology* **2001,** 279, (2), 371-4.
16. Srinivasan, S.; Batra, R.; Chan, H.; Kamath, G.; Cherukara, M. J.; Sankaranarayanan, S., Artificial Intelligence-Guided De Novo Molecular Design Targeting COVID-19. *ACS Omega* **2021,** 6, (19), 12557-12566.
17. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **2010,** 31, (2), 455-61.
18. Mohapatra, S.; Nath, P.; Chatterjee, M.; Das, N.; Kalita, D.; Roy, P.; Satapathi, S., Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking. *PLoS One* **2020,** 15, (11), e0241543.
19. Verma, N.; Qu, X.; Trozzi, F.; Elsaied, M.; Karki, N.; Tao, Y.; Zoltowski, B.; Larson, E. C.; Kraka, E., SSnet: A Deep Learning Approach for Protein-Ligand Interaction Prediction. *bioRxiv* **2021**, 2019.12.20.884841.
20. Karki, N.; Verma, N.; Trozzi, F.; Tao, P.; Kraka, E.; Zoltowski, B., Predicting Potential SARS-COV-2 Drugs-In Depth Drug Database Screening Using Deep Neural Network Framework SSnet, Classical Virtual Screening and Docking. *Int J Mol Sci* **2021,** 22, (4).
21. Koes, D. R.; Baumgartner, M. P.; Camacho, C. J., Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model* **2013,** 53, (8), 1893-904.
22. Jukic, M.; Skrlj, B.; Tomsic, G.; Plesko, S.; Podlipnik, C.; Bren, U., Prioritisation of Compounds for 3CL(pro) Inhibitor Development on SARS-CoV-2 Variants. *Molecules* **2021,** 26, (10).
23. Sterling, T.; Irwin, J. J., ZINC 15--Ligand Discovery for Everyone. *J Chem Inf Model* **2015,** 55, (11), 2324-37.
24. Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R., Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. *Science* **2020,** 368, (6489), 409-412.
25. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012,** 40, (Database issue), D1100-7.
26. Kwofie, S. K.; Broni, E.; Asiedu, S. O.; Kwarko, G. B.; Dankwa, B.; Enninful, K. S.; Tiburu, E. K.; Wilson, M. D., Cheminformatics-Based Identification of Potential Novel Anti-SARS-CoV-2 Natural Compounds of African Origin. *Molecules* **2021,** 26, (2).
27. Citarella, A.; Scala, A.; Piperno, A.; Micale, N., SARS-CoV-2 M(pro): A Potential Target for Peptidomimetics and Small-Molecule Inhibitors. *Biomolecules* **2021,** 11, (4).
28. Ton, A. T.; Gentile, F.; Hsing, M.; Ban, F.; Cherkasov, A., Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Mol Inform* **2020,** 39, (8), e2000028.
29. Born, J.; Manica, M.; Cadow, J.; Markert, G.; Mill, N. A.; Filipavicius, M.; Janakarajan, N.; Cardinale, A.; Laino, T.; Martinez, M. R., Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Mach Learn-Sci Techn* **2021,** 2, (2).
30. Meyers, J.; Fabian, B.; Brown, N., De novo molecular design and generative models. *Drug Discov Today* **2021,** 26, (11), 2707-2715.
31. Huang, R.; Xia, M.; Nguyen, D.-T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Rossoshek, A.; Simeonov, A., Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Frontiers in Environmental Science* **2016,** 3.
32. Morris, A.; McCorkindale, W.; Consortium, T. C. M.; Drayman, N.; Chodera, J. D.; Tay, S.; London, N.; Lee, A. A., Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model. *Chem Commun (Camb)* **2021,** 57, (48), 5909-5912.
33. Agarwal, S.; Dugar, D.; Sengupta, S., Ranking chemical structures for drug discovery: a new machine learning approach. *J Chem Inf Model* **2010,** 50, (5), 716-31.
34. Acharya, A.; Agarwal, R.; Baker, M. B.; Baudry, J.; Bhowmik, D.; Boehm, S.; Byler, K. G.; Chen, S. Y.; Coates, L.; Cooper, C. J.; Demerdash, O.; Daidone, I.; Eblen, J. D.; Ellingson, S.; Forli, S.; Glaser, J.; Gumbart, J. C.; Gunnels, J.; Hernandez, O.; Irle, S.; Kneller, D. W.; Kovalevsky, A.; Larkin, J.; Lawrence, T. J.; LeGrand, S.; Liu, S. H.; Mitchell, J. C.; Park, G.; Parks, J. M.; Pavlova, A.; Petridis, L.; Poole, D.; Pouchard, L.; Ramanathan, A.; Rogers, D. M.; Santos-Martins, D.; Scheinberg, A.; Sedova, A.; Shen, Y.; Smith, J. C.; Smith, M. D.; Soto, C.; Tsaris, A.; Thavappiragasam, M.; Tillack, A. F.; Vermaas, J. V.; Vuong, V. Q.; Yin, J.; Yoo, S.; Zahran, M.; Zanetti-Polzi, L., Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *J Chem Inf Model* **2020,** 60, (12), 5832-5852.
35. Earl, D. J.; Deem, M. W., Parallel tempering: theory, applications, and new perspectives. *Phys Chem Chem Phys* **2005,** 7, (23), 3910-6.

36. Sugita, Y.; Kitao, A.; Okamoto, Y., Multidimensional replica-exchange method for free-energy calculations. *The Journal of Chemical Physics* **2000,** 113, (15), 6042-6051.
37. Bernardi, R. C.; Melo, M. C. R.; Schulten, K., Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim Biophys Acta* **2015,** 1850, (5), 872-877.
38. Chang, C. K.; Hou, M. H.; Chang, C. F.; Hsiao, C. D.; Huang, T. H., The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res* **2014,** 103, 39-50.
39. Astuti, I.; Ysrafil, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. *Diabetes Metab Syndr* **2020,** 14, (4), 407-412.
40. Novick, P. A.; Ortiz, O. F.; Poelman, J.; Abdulhay, A. Y.; Pande, V. S., SWEETLEAD: an in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. *PLoS One* **2013,** 8, (11), e79568.
41. Pirolli, D.; Righino, B.; De Rosa, M. C., Targeting SARS-CoV-2 Spike Protein/ACE2 Protein-Protein Interactions: a Computational Study. *Mol Inform* **2021,** 40, (6), e2060080.
42. Gaudencio, S. P.; Pereira, F., A Computer-Aided Drug Design Approach to Predict Marine Drug-Like Leads for SARS-CoV-2 Main Protease Inhibition. *Mar Drugs* **2020,** 18, (12).
43. Yassine, R.; Makrem, M.; Farhat, F., Active Learning and the Potential of Neural Networks Accelerate Molecular Screening for the Design of a New Molecule Effective against SARS-CoV-2. *Biomed Res Int* **2021,** 2021, 6696012.
44. Santana, M. V. S.; Silva-Jr, F. P., De novo design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning. *BMC Chem* **2021,** 15, (1), 8.
45. Bung, N.; Krishnan, S. R.; Bulusu, G.; Roy, A., De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future Med Chem* **2021,** 13, (6), 575-585.
46. Cooper, K.; Baddeley, C.; French, B.; Gibson, K.; Golden, J.; Lee, T.; Pierre, S.; Weiss, B.; Yang, J., Novel Development of Predictive Feature Fingerprints to Identify Chemistry-Based Features for the Effective Drug Design of SARS-CoV-2 Target Antagonists and Inhibitors Using Machine Learning. *ACS Omega* **2021,** 6, (7), 4857-4877.
47. Glaab, E.; Manoharan, G. B.; Abankwa, D., Pharmacophore Model for SARS-CoV-2 3CLpro Small-Molecule Inhibitors and in Vitro Experimental Validation of Computationally Screened Inhibitors. *J Chem Inf Model* **2021,** 61, (8), 4082-4096.
48. Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X., MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Brief Bioinform* **2021,** 22, (3).
49. Mekni, N.; Coronnello, C.; Langer, T.; Rosa, M.; Perricone, U., Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors. *Int J Mol Sci* **2021,** 22, (14).
50. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P., Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* **2003,** 43, (6), 1947-58.
51. Yang, Z.; Bogdan, P.; Nazarian, S., An in silico deep learning approach to multi-epitope vaccine design: a SARS-CoV-2 case study. *Sci Rep* **2021,** 11, (1), 3238.
52. Sanchez-Trincado, J. L.; Gomez-Perosanz, M.; Reche, P. A., Fundamentals and Methods for T- and B-Cell Epitope Prediction. *J Immunol Res* **2017,** 2017, 2680160.
53. Jespersen, M. C.; Peters, B.; Nielsen, M.; Marcatili, P., BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res* **2017,** 45, (W1), W24-W29.
54. Yao, B.; Zhang, L.; Liang, S.; Zhang, C., SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. *PLoS One* **2012,** 7, (9), e45152.
55. Saha, S.; Raghava, G. P., Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* **2006,** 65, (1), 40-8.
56. El-Manzalawy, Y.; Dobbs, D.; Honavar, V., Predicting linear B-cell epitopes using string kernels. *J Mol Recognit* **2008,** 21, (4), 243-55.
57. Li, M.; Jiang, Y.; Gong, T.; Zhang, Z.; Sun, X., Intranasal Vaccination against HIV-1 with Adenoviral Vector-Based Nanocomplex Using Synthetic TLR-4 Agonist Peptide as Adjuvant. *Mol Pharm* **2016,** 13, (3), 885-94.
58. Nielsen, M.; Lundegaard, C.; Blicher, T.; Lamberth, K.; Harndahl, M.; Justesen, S.; Roder, G.; Peters, B.; Sette, A.; Lund, O.; Buus, S., NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* **2007,** 2, (8), e796.
59. Emini, E. A.; Hughes, J. V.; Perlow, D. S.; Boger, J., Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol* **1985,** 55, (3), 836-9.
60. Nielsen, M.; Lundegaard, C.; Blicher, T.; Peters, B.; Sette, A.; Justesen, S.; Buus, S.; Lund, O., Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS Comput Biol* **2008,** 4, (7), e1000107.
61. Reynisson, B.; Barra, C.; Kaabinejadian, S.; Hildebrand, W. H.; Peters, B.; Nielsen, M., Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. *J Proteome Res* **2020,** 19, (6), 2304-2315.
62. Heo, L.; Park, H.; Seok, C., GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res* **2013,** 41, (Web Server issue), W384-8.
63. Dimitrov, I.; Flower, D. R.; Doytchinova, I., AllerTOP--a server for in silico prediction of allergens. *BMC Bioinformatics* **2013,** 14 Suppl 6, S4.

64. Ali, M.; Pandey, R. K.; Khatoon, N.; Narula, A.; Mishra, A.; Prajapati, V. K., Exploring dengue genome to construct a multi-epitope based subunit vaccine by utilizing immunoinformatics approach to battle against dengue infection. *Sci Rep* **2017,** 7, (1), 9232.
65. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; Cheng, Z.; Yu, T.; Xia, J.; Wei, Y.; Wu, W.; Xie, X.; Yin, W.; Li, H.; Liu, M.; Xiao, Y.; Gao, H.; Guo, L.; Xie, J.; Wang, G.; Jiang, R.; Gao, Z.; Jin, Q.; Wang, J.; Cao, B., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **2020,** 395, (10223), 497-506.
66. Kesarwani, V.; Gupta, R.; Vetukuri, R. R.; Kushwaha, S. K.; Gandhi, S., Identification of Unique Peptides for SARS-CoV-2 Diagnostics and Vaccine Development by an In Silico Proteomics Approach. *Front Immunol* **2021,** 12, 725240.
67. Takeuchi, A.; O'Leary, D. D., Radial migration of superficial layer cortical neurons controlled by novel Ig cell adhesion molecule MDGA1. *J Neurosci* **2006,** 26, (17), 4460-4.
68. Domin, J.; Pages, F.; Volinia, S.; Rittenhouse, S. E.; Zvelebil, M. J.; Stein, R. C.; Waterfield, M. D., Cloning of a human phosphoinositide 3-kinase with a C2 domain that displays reduced sensitivity to the inhibitor wortmannin. *Biochem J* **1997,** 326 (Pt 1), 139-47.
69. Lai, C. S.; Fisher, S. E.; Hurst, J. A.; Vargha-Khadem, F.; Monaco, A. P., A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **2001,** 413, (6855), 519-23.
70. Jin, J.; Arias, E. E.; Chen, J.; Harper, J. W.; Walter, J. C., A family of diverse Cul4-Ddb1-interacting proteins includes Cdt2, which is required for S phase destruction of the replication factor Cdt1. *Mol Cell* **2006,** 23, (5), 709-21.
71. Zou, Z.; Xie, X.; Li, W.; Song, X.; Tan, Y.; Wu, H.; Xiao, J.; Feng, H., Black carp TAB1 up-regulates TAK1/IRF7/IFN signaling during the antiviral innate immune activation. *Fish Shellfish Immunol* **2019,** 89, 736-744.
72. Russo, G.; Reche, P.; Pennisi, M.; Pappalardo, F., The combination of artificial intelligence and systems biology for intelligent vaccine design. *Expert Opin Drug Discov* **2020,** 15, (11), 1267-1281.
73. Malik, Y. S.; Sircar, S.; Bhat, S.; Ansari, M. I.; Pande, T.; Kumar, P.; Mathapati, B.; Balasubramanian, G.; Kaushik, R.; Natesan, S.; Ezzikouri, S.; El Zowalaty, M. E.; Dhama, K., How artificial intelligence may help the Covid-19 pandemic: Pitfalls and lessons for the future. *Rev Med Virol* **2021,** 31, (5), 1-11.
74. Pires, C., A Systematic Review on the Contribution of Artificial Intelligence in the Development of Medicines for COVID-2019. *J Pers Med* **2021,** 11, (9).