Artificial intelligence-driven drug repurposing and structural biology for SARS-CoV-2

Kartikay Prasad, Vijay Kumar*

Amity Institute of Neuropsychology & Neurosciences, Amity University, Noida, UP, 201303, India

A R T I C L E   I N F O

Keywords:
COVID-19
SARS-CoV-2
Artificial intelligence
Deep learning
Structure
Drug repurposing

A B S T R A C T

It has been said that COVID-19 is a generational challenge in many ways. But, at the same time, it becomes a catalyst for collective action, innovation, and discovery. Realizing the full potential of artificial intelligence (AI) for structure determination of unknown proteins and drug discovery are some of these innovations. Potential applications of AI include predicting the structure of the infectious proteins, identifying drugs that may be effective in targeting these proteins, and proposing new chemical compounds for further testing as potential drugs. AI and machine learning (ML) allow for rapid drug development including repurposing existing drugs. Algorithms were used to search for novel or approved antiviral drugs capable of inhibiting SARS-CoV-2. This paper presents a survey of AI and ML methods being used in various biochemistry of SARS-CoV-2, from structure to drug development, in the fight against the deadly COVID-19 pandemic. It is envisioned that this study will provide AI/ML researchers and the wider community an overview of the current status of AI applications particularly in structural biology, drug repurposing, and development, and motivate researchers in harnessing AI potentials in the fight against COVID-19.

1. Introduction

The novel coronavirus disease (COVID-19) has become an unprecedented public health crisis affecting people's lives and causing a large number of deaths. Till June 2021, over 178 million confirmed cases was reported worldwide with more than 3.88 million deaths reported (https://covid19.who.int/). The numbers of infections and death are still increasing. With the continued growth of the COVID-19 pandemic, scientists and healthcare providers worldwide are working to better comprehend, alleviate, and suppress its spread. The usual symptoms of COVID-19 are pneumonia, shortness of breath, dry cough, tiredness, and fever (Huang et al., 2020a) along with several neurological complications (Khatoon et al., 2020). SARS-CoV-2 is a positive-sense single-stranded RNA virus consisting of ~30 kb genome encoding for four main structural proteins including spike (S) glycoprotein, small envelope (E) glycoprotein, membrane (M) glycoprotein, and nucleocapsid (N) protein, in addition to the sixteen non-structural proteins (NSPs) (Wu et al., 2020; Lu et al., 2020). These viral proteins have a specific role in the life cycle and pathogenicity of the virus and a complete understanding of their structure and function is essential for drug discovery.

The power of artificial intelligence (AI) approaches has been attributed to a wide range of applications across public health, disease prediction, and drug development. Fuchs in 2020 (Fuchs et al., 2020) elegantly summarized the role of AI in the current COVID-19 pandemic in general (Fleming, 2018; Lavecchia, 2019). Machine learning (ML), a subset of AI, has enabled the generation of models that can learn and study the patterns present in data and can make inferences from a large number of test data. With the advent of deep learning (DL), the automatic feature extraction from raw data leads to an increase in performance compared to other computer-aided models (Chen et al., 2018; Zhavoronkov et al., 2019). Different DL algorithms were utilized in fighting the COVID-19 pandemic including artificial neural network (ANN), convolutional neural network (CNN), and long short-term memory (LSTM). The recent applications of AI in the case of COVID-19 include the virtual screening of both repurposed drugs as well as new chemical entities (Fig. 1) (Keshavarzi ArshadiWebb et al., 2020; Zhou et al., 2020; Mottaqi et al., 2021; Piccialli et al., 2021). ML-based molecular docking has been utilized extensively for virtual screening and drug repurposing. This approach requires the following information: (a) dataset of
approved drugs or drug-like molecules, (b) three-dimensional structure of the protein target, and (c) molecular docking software or program. The molecular docking-based studies allow the identification of several chemical molecules that bind to different SARS-CoV-2 proteins and thus can potentially inhibit viral replication and growth (Fig. 1).

In addition, DL-based applications have been the main focus for drug repurposing research, as their automatic feature extraction accelerates the process of drug discovery (Fig. 1). The application of DL has benefitted the de novo drug design approaches. The current design approach utilizes state-of-the-art DL models such as Generative Autoencoders (GAE) and Generative Adversarial Networks (GAN) to generate data-based molecules (Wrapp et al., 2020). More recent approaches use generative models such as variational autoencoders (VAE) to generate sequences of atoms (Fig. 1). This approach allows the creation of unique drug molecules with greater diversity (Griffiths and Hernandez-Lobato, 2020). These autoencoders instruct molecules into a vector that captures properties such as bond order, element properties, and functional group (Zhang and Lu, 2019; Baucum et al., 2020). To design a drug-generative network, the information about (a) collection of drug-like molecules, (b) a feature-representation of these molecules in silico, (c) method to increase the diversity of the molecules, and (d) screening and modification of the altered molecules, are important (Fig. 1).

AI, in the field of computational biology and medicine, has been used to partially understand COVID-19 to discover drugs against the SARS-CoV-2 virus (Alimadadi et al., 2020; Mei et al., 2020; Ke et al., 2020; Nguyen, 2020). Equipped with a strong computational power to deal with large amounts of data, AI can help scientists to expand their knowledge about the coronavirus quickly. For example, by determining the protein structures of the virus, researchers would be able to find the machinery necessary for designing a drug or vaccine more accurately and effectively. The application of AI and its subsets are crucial for the current pandemic situation and rapid discovery of drugs against COVID-19 for several key reasons. The automatic feature extraction ability of DL can support models with better accuracy and reliable results. Also, the generative ability demonstrated by DL models can be largely exploited to create small molecule drugs and better epitope prediction, minimizing the chances of failure in trial experiments. Thus, the use of AI is essential.
to find out potential drugs against COVID-19 in time and accurately.

In this review, we provide a survey of AI-based models for COVID-19 drug discovery and structure biology of SARS-CoV-2 proteins, that will provide rapid and cost-effective therapeutic interventions in COVID-19. We propose that these AI-based technological advances are in utmost need during the COVID-19 global pandemic.

2. Methodology

The systematic literature search and analysis was done through the Dimensions database (https, 2020) in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Liberati et al., 2009). Dimensions is the most comprehensive database developed by digital science with the help of over 100 leading research organizations, worldwide. This database provides access to various contents ranging from publications to grants, funding agencies, clinical trials, patents, datasets, and policy documents.

The literatures were retrieved from the dimensions database using search keywords with “OR/AND” operators. Keywords include “COVID-19”, “SARS-CoV-2”, “drug repurposing”, “drug repositioning”, “artificial intelligence”, “machine learning”, “deep learning”, and “neural networks”. The study period was restricted from January 1, 2020, to March 31, 2021. A staged literature search was performed for the first section of this systematic review and all the relevant studies were identified based on the set of inclusion/exclusion criteria and summarized narratively. Studies were included if they reported outcomes on any results on drug repurposing or drug discovery to COVID-19 utilizing AI/ML approaches. Whereas, the editorials, commentary, survey, and narrative reviews on AI/ML were excluded. The article selection for this review followed the three-stage analysis. The first stage analysis considered only the titles and abstracts of the articles to extract relevant articles. The second stage analysis enables the selection of articles based on the introduction and conclusion to further refine the selection in the first stage. At the third and final stage of analysis, articles were explored thoroughly and selected in terms of their relevance to the review aim. The article was selected if it reported an empirical application of AI/ML in drug repurposing or drug discovery to COVID-19.

3. Results

A total of 8841 articles were identified in the Dimensions database with the search term: “COVID-19” OR “SARS-CoV-2” AND (machine learning OR artificial intelligence OR deep learning OR neural network). After adding “drug repurposing” in the search term i.e., “COVID-19” OR “SARS-CoV-2” AND (machine learning OR artificial intelligence OR deep learning OR neural network) AND (drug repurposing), a total of 516 articles were found (Fig. 2). Out of these 516 articles, most of the articles were posted on an open-access online pre-print server and the publications cover mainly the field of medical and health sciences and services. In the initial analysis where the titles and abstracts were considered, 427 were excluded because they were not related to the drug repurposing work using the AI. The full text of the remaining 89 studies was then considered.

Fig. 2. PRISMA flow diagram for the systematic review of the role of AI/ML in drug repurposing and drug development for COVID-19.
analyzed; and finally, 26 studies met the final inclusion criteria, and were included in our study. Fig. 2 shows the total number of articles obtained from the Dimensions database and the final number of articles considered for the research after applying all the inclusion/exclusion criteria. Table 1 summarizes the findings of these studies.

### 3.1. AI/ML-based methods in COVID-19 drug repurposing and drug discovery

AI-based drug discovery and drug repurposing have been much publicized as an effective approach to accelerate the drug discovery process (Zhou et al., 2020b; Prasad et al., 2020; Abbasi, 2020; Richardson et al., 2020; Ge et al., 2020; Hong et al., 2019; Beck et al., 2020). Broadly, AI can be very useful for initial drug discovery in two main ways, one is through screening millions of chemical compounds available in different databases for potential drugs in simulation tests, and the other is to identify novel drugs that can latch onto the targets, to reduce their infectivity. The strategy of drug repurposing comes out as a powerful solution for the COVID-19 (Zhou et al., 2020a, 2020b; Prasad et al., 2020; Abbasi, 2020). Within a month after the first COVID-19 report in China, two independent groups have used AI in different ways to find possible treatments for the SARS-CoV-2. Scientists from the AI drug discovery company BenevolentAI and Imperial College London utilized their in-house developed algorithms to mine the data and find the enzyme, adaptor-associated protein kinase 1 (AAK1) as a possible target for the COVID-19. The program then reported Baricitinib as one of the best inhibitors out of 378 known AAK1 inhibitors (Richardson et al., 2020). Baricitinib is an approved drug against rheumatoid arthritis.

Ge et al. (2020) identified CVL218, a promising PARP1 (Poly (ADP-Ribose) Polymerase 1) inhibitor through a Natural Language Processing (NLP) model. Here, a Biomedical Entity Relation Extraction (BERE) approach (Hong et al., 2019) was applied to the PubMed database and filtered after searching the term - candidate drug compounds, coronaviruses, or the related proteins. Beck et al. (2020) have utilized their pre-trained deep learning-based drug-target interaction model known as Molecule Transformer-Drug Target Interaction (MT-DTI) to identify FDA-approved antivirals against SARS-CoV-2 proteins (the 3CL/pro, RdRP, helicase, 3’-to-5’ exonuclease, endoRNase, and 2’-O-ribose methyltransferase). This model exploited simplified molecular-input line-entry system (SMILES) strings and amino acid sequences as 1D string inputs and thus can be easily applied to target proteins that do not have any 3D structures.

Another research group used a knowledge graph (KG) based deep learning method for drug repurposing in COVID-19 and termed as CoVKG (Zeng et al., 2020). The authors utilized a DL approach, RotatE, developed by Amazon supercomputing resource, AWS-AI (Wang et al., 2020).

| Table 1 | Summary of important AI and ML-based studies for drug repurposing and drug development in COVID-19. |
|---------|------------------------------------------------------------------------------------------|
| Author  | Study keywords | Targets | No. of drugs | Reference |
| Ke, Y. Y | Deep neural network | FIP | 80 | Ke et al. (2020) |
| Ge, Y. | Knowledge Graph, deep learning | SARS-CoV-2 | 64 | Ge et al. (2020) |
| Beck, B. R | Hybrid CNN and RNN model called MT-DTI | 3CLpro, RdRp, helicase, 3’-to-5’ exonuclease, endoRNase, and 2’-O-ribose methyltransferase | 3410 | Beck et al. (2020) |
| Zeng, X. | Deep learning-based knowledge graph (CoV-KGE), 2-D fingerprint, GBDT model, Recurrent Neural Network (RNN) | SARS-CoV-2 | 41 | Zeng et al. (2020) |
| Gao, K. | 2-D fingerprint, GBDT model, Recurrent Neural Network (RNN) | 3CLpro | 40 | Gao et al. (2020) |
| Hofmarcher, M. | Deep neural network, ChemAI | 3CLpro, PLP | 20 | Hofmarcher et al. (2020) |
| Ton, A. T | Deep learning platform – Deep Docking (DD) | Mpro | 1000 | Ton et al. (2020) |
| Hu, F. | Deep learning-based multi-task models, Classification and Regression | SARS-CoV-2 | 77 | Gysi et al. (2020) |
| Gysi, D.M | Graph Neural Network | 3CLpro | 13 | Huang et al. (2021) |
| Huang, K. | Deep Purpose, Python toolkit, CNN | S, S-ACE2 complex | 187 | Batra et al. (2020) |
| Batra, R. | Random forest (RF) regression algorithm, ensemble docking | 3CLpro, Spike; ACE2, TMPRSS2, Cathepsin B | 30 | Redka et al. (2020) |
| Redka, D.S. | Deep learning, Ligand Design, MatchMaker, PolypharmDB | 3CLpro | 100 drugs & 20 peptides | Zhang et al. (2020a) |
| Zhang, H. | Dense fully CNN (DFCNN), DL, Virtual screening | 3CLpro | 15 | Nguyen et al. (2020) |
| Nguyen, D. D | Mathematical Deep Learning (MathDL), CNN | 3CLpro | 12 | Artigas et al. (2020) |
| Artigas, L. | Therapeutic Performance Mapping System (TPMS), ANN, GUILDify | SARS-CoV-2 interactome | 10 | Mahapatra et al. (2020) |
| Mahapatra, S. | ML, Naive Bayes classification algorithm, molecular docking | 3CLpro | 12 | Blasiak et al. (2020) |
| Blasiak, A. | IDentifiAI, orthogonal array composite design (OADC), drug-dose relationships | SARS-CoV-2 | 12-drug set, over 530,000 drug combinations | Blasiak et al. (2021) |
| Chakravarty, K. | BIOiSIM, in silico simulation, and modeling | ACE2, Spike protein | >2000 | Chakravarty et al. (2021) |
| Zhavoronkov, A. | Deep learning-based generative model, AAE, GENTRIL | 3CLpro, Mpro | 1000 | Zhavoronkov et al. (2020) |
| Bung, N. | Deep neural network, New Chemical Entities (NCE) | 3CLpro | 31 | Bung et al. (2020) |
| Tang, B. | Advanced deep-Q learning network with the fragment-based drug design (ADQF-FINN) | 3CLpro | 47 | Tang et al. (2020) |
| Chenthamarakshnan, V. | Deep learning-based generative model, CogMol, VAE | Mpro, Spike, NSP9 | 3000 | Chenthamarakshnan et al. (2020) |
| Delijewski, M. | MACCS fingerprints, deep learning, gradient-boostered tree learning | 3CLpro | 290000 | Delijewski and Hancezok (2021) |
| Laponogov, I. | Corona-AI, DreamLab App platform, Machine learning, SVM, REF | 65 host SARS-CoV-2 targets | 52 | Laponogov et al. (2021) |
| Kowalewski, A. | ANN, long-short term memory network (LSTM, CNN) | SARS-CoV-2 | 12 | Cantürk et al. (2020) |
S-CoV-2/SARS-CoV-3CLpro inhibitors to predict the binding affinity of the gradient-boosting decision tree (GBDT) model on 314 SARs-CoV-2 3CLpro. In this work, the authors trained their 2-D fingerprint-based DL model on 314 SARs-CoV-2/SARS-CoV-3CLpro inhibitors to predict the binding affinities of potential protease inhibitors. A total of 8565 drugs (including 1553 FDA-approved drugs) from DrugBank were evaluated and the top 20 FDA-approved drugs along with the top 20 investigational off-market drugs were selected as potent inhibitors of SARS-CoV-2 3CL protease.

Gao et al. (2020) used structure-based drug repositioning (SBDR) ML models to evaluate the drug binding affinity to SARS-CoV-2 3CLpro. In this work, they trained their 2-D fingerprint-based DL gradient-boosting decision tree (GBDT) model on 314 SARs-CoV-2/SARS-CoV-3CLpro inhibitors to predict the binding affinities of potential protease inhibitors. A total of 8565 drugs (including 1553 FDA-approved drugs) from DrugBank were used to train the GBDT model. The model trained on the entire human proteome that combines structural information to their target proteins. Additionally, they also screened the DrugBank for drug repurposing. By ranking the compounds according to their predicted inhibitory potentials, toxicity, and closeness to known drugs, the authors created a list of 30,000 possible compounds for further screening. These top-ranked compounds were made available as a library at https://github.com/ml-iku/sars-cov-inhibitors-chemai.

Similar work has been reported by Ton et al. (2020) where they used a deep docking platform trained on a neural network to predict the results of docking simulations. From the ZINC database, the authors identified a set of 3 million candidate 3CLpro inhibitors which were subsequently winnowed down to 1000 compounds for drug repurposing after docking simulation. Hu et al. (Hu and Jiang, 2020) used a multilabel neural network model to predict protein-ligand binding affinities of viral proteins against a database of 4895 drugs. They suggested 10 potential drugs with strong binding affinity to their target proteins.

Gysi and colleagues (Gysi et al., 2020) utilized an AI-combined network medicine drug repurposing approach to rank 6340 drugs for their efficacy against SARS-CoV-2. The predictions were then validated on 918 experimentally validated drugs. They developed a multimodal approach that has combinations of different algorithms and further identified 77 potential repurposing drugs. Huang and colleagues (Huang et al., 2021) developed a Python-based DL toolkit, DeepPurpose that is based on an encoder-decoder framework and presented a case study on SARS-CoV-2 3CLpro with 13 potential repurposing candidates identified. Batra et al. (2020) trained and validated a random forest algorithm on data from Smith et al. (Smith and Smith, 2020). The authors then executed their models to CureFFI and DrugCentral databases containing 1495 and 3967 drugs, respectively. They also applied their model to screen compounds from the BindingDB database and identified 19,000 additional candidates that bind strongly to either the native S-protein or the human ACE2-S protein complex.

Redka et al. (2020) utilized DL platform, Ligand Design to identify the FDA-approved drugs and experimental medicines that have the potential to inhibit SARS-CoV-2 infection. They developed a resource, PolypharmDB contains 10,224 drugs along with the computed list of ~87000 proteins predicted to interact with them. The interactions were generated with Cytoscape’s MatchMaker TM technology which is a DL model trained on the entire human proteome that combines structural and experimental data to predict the binding of drug molecules to protein pockets.

Zhang et al. (2020a) utilized the neural network algorithm trained on the PDBbind database and identified possible inhibitors of the SARS-CoV-2 3CLpro. They then used the structural model of the 3CLpro, explored the databases, ChemDiv, and TargetMol to find promising compounds targeting 3CLpro protein. Similarly, Nguyen et al. (2020) applied the Mathematical Deep Learning (MathDL) approach to identify possible inhibitors for SARS-CoV-2 3CLpro. Their model has been trained on two datasets, ChEMBL and PDBbind database using two different CNNs. Finally, they identified 15 promising drug candidates for SARS-CoV-2 3CLpro using the trained CNN models on the DrugBank dataset. Artigas et al. (2020) utilized a systems biology and AI-based approach, the Therapeutic Performance Mapping System (TPMS) technology (Jorba et al., 2020) to repurpose drugs and drug combinations for COVID-19. TPMS method employs an ANN to measure the potential relationship between the nodes of a network (i.e., protein) grouped based on their association with a phenotype. This strategy has been then used to evaluate the effect of 6605 drugs present in the DrugBank and 122 human proteins retrieved through a literature search of the coronavirus-human interactome. A total of 12 approved drugs have been identified, out of which 4 are currently in COVID-19 clinical trials. Besides, they also identified the drug combinations using ANN of TPMS technology and suggested that a combination of drug pirfenidone with melatonin could be a good candidate against COVID-19 and their combined mechanism of action has been identified at the molecular level through the use of TPMS sampling-based models.

In a study, Scott D. Bembenek of Denovicon Therapeutics (San Diego, CA, 92130, USA) used the Denovicon computational platform to perform a molecular modeling-AI hybrid computational approach to find potential inhibitors of the SARS-CoV-2 main protease (Mpro, 3CLpro) (Bembenek et al., 2020). Over 13,000 FDA-approved drugs and clinical candidates (approximately 30,000 proteins) were investigated and finally arrived at the five hits that may prove useful in the design of future inhibitors of the main protease.

Recently, Mahapatra et al. (2020) reported the ML model based on the Naive Bayes algorithm, which predicts COVID-19 drugs with more than 70% accuracy. This approach suggested 10 FDA-approved drugs that can be repurposed to target COVID-19.

To optimize the use of drug combination therapy, an AI-based platform, Project Identif.AI (Identifying Infectious Disease Combination Therapy with Artificial Intelligence) were utilized for drug development and drug repurposing (Abdulla et al., 2020). It is a neural network approach built on a quadratic correlation between inputs defined by drugs and their doses and outputs defined by treatment efficacy and safety. The authors examined 12 drug/dose parameters and identified drug combinations that effectively inhibit vesicular stomatitis virus infection of A549 lung cells. Many of the studied drugs are currently used in COVID-19 clinical trials also. The authors also suggested the utilization of this platform for COVID-19 intervention. Blasiak et al. (2021) utilized Identif.AI to evaluate over 530,000 drug combinations against the SARS-CoV-2 live virus collected from a patient sample. Identif.AI identified the combination of remdesivir, ritonavir, and lopinavir as a potentially effective treatment against SARS-CoV-2 infection. Further experimental validation indicates that this drug combination exhibits a 6.5-fold enhanced efficacy over remdesivir alone. Also, the author showed that hydroxychloroquine and azithromycin were relatively ineffective against live SARS-CoV-2. Thus, Project Identif.AI greatly cuts the number of in vitro assays required to evaluate the drug tolerability and efficacy and can be applied along with the in vitro investigations of drug validation.

Yi-Yu Ke et al. (2020) developed an AI system trained on two different learning databases. The first one is an antiviral database against SARS-CoV, SARS-CoV-2, HIV, influenza virus, and the second database contains 210 known 3CLpro inhibitors. The authors identified a total of 80 potential antiviral drugs, among them, 8 drugs were shown to inhibit feline infectious peritonitis (FIP) virus in Fcwf-4 cells.

Bung et al. (2020) constructed a deep neural network-based generative and predictive model for SMILES input strings. The model was first trained on 1.6 million compounds from the ChEMBL database and then applied to a small dataset of protease inhibitors using transfer learning.
The authors used reinforcement learning to train the model and identified potential drug compounds. Based on the screening and docking results, 31 potential inhibitors of SARS-CoV-2 3CLpro have been identified.

Very recently, Chakravarty et al. (2021) developed an AI-integrated Bio-simulation platform for drug development and repurposing of pulmonary hypertension therapies for COVID-19. The group conducted an in-silico modeling by using AI-integrated mechanistic modeling platform BIOSIM with known preclinical in-vitro and in-vivo datasets for accurately simulating the systemic therapy disposition and site-of-action penetration of Angiotensin-Converting Enzyme (ACE) and calcium channel blockers (CCB) compounds to tissues playing role in COVID-19 pathogenesis. The group provides AI/ML-driven computational modeling for repurposing and accelerated the drug development process.

Zhavoronkov and colleagues (Zhavoronkov et al., 2020) have used an integrated AI-based drug discovery pipeline to generate novel drug compounds against SARS-CoV-2 3CLpro. A total of 28 ML models including, GAE and GAN have generated molecular structures that were optimized with reinforcement learning (RL) approaches. Novel drug-like compounds made with these approaches were published at www.insilico.com/ncov-sprint and have been continuously updated.

Tang et al. (2020) combined AI with the structure-based drug design (SBDD) to speed up the generation of potential candidate compounds against SARS-CoV-2. The authors generated a list of 284 molecules known to inhibit SARS-CoV-2 3CLpro, break them into 316 fragments, and generated potential lead compounds via an advanced deep Q-learning network with fragment-based drug design (ADQN-FBDD). This framework rewards three aspects of discovered leads, the first is a drug-likeness score, the second is the addition of pre-determined favorable fragments, and the last is the existence of known pharmacophores. This AI-based approach generated a library of 4922 covalent lead compounds with unique valid structures that are heuristically filtered, and finally, 47 lead compounds were evaluated with molecular docking and simulations. All these 47 top compounds and related derivatives based on SBOP were made available in the molecular library at https://github.com/tbwxmu/2019-nCov.

IBM research scientists from Singapore (Chentharamakshan et al., 2020) applied deep learning generative modeling framework, Controlled Generation of Molecules (CogMol) as a drug discovery approach. This framework applies a variational autoencoder trained on SMILES strings to learn molecule embeddings. The authors on these embeddings train attribute regression models to predict drug properties and protein binding affinities. The authors then went for conditional sampling using Conditional Latent (attribute) Space Sampling (CLaSS) to generate sam-

This framework rewards three aspects of discovered leads, the first is a drug-likeness score, the second is the addition of pre-determined favorable fragments, and the last is the existence of known pharmacophores. This AI-based approach generated a library of 4922 covalent lead compounds with unique valid structures that are heuristically filtered, and finally, 47 lead compounds were evaluated with molecular docking and simulations. All these 47 top compounds and related derivatives based on SBOP were made available in the molecular library at https://github.com/tbwxmu/2019-nCov.

IBM research scientists from Singapore (Chentharamakshan et al., 2020) applied deep learning generative modeling framework, Controlled Generation of Molecules (CogMol) as a drug discovery approach. This framework applies a variational autoencoder trained on SMILES strings to learn molecule embeddings. The authors on these embeddings train attribute regression models to predict drug properties and protein binding affinities. The authors then went for conditional sampling using Conditional Latent (attribute) Space Sampling (CLaSS) to generate sam-

Many drug discovery companies are utilizing AI to accelerate drug development and drug repurposing against the SARS-CoV-2 to confront the COVID-19 pandemic. At large, the success of AI platforms depends on the data that is being used to ‘train’ the algorithms. A limited data of SARS-CoV-2 can be a challenge. Some of the AI-based companies employing these intellectual approaches to find new drugs against COVID-19 are listed in supporting information, Table S1. These AI-drug discovery companies or AI-based start-ups are working to hasten a rational drug repurposing of the available drugs or discovery of novel drugs against the novel coronavirus.

3.2. AI/ML in vaccine development

Machine learning has also improved the field of vaccine design over the past two decades by creating the virtual frameworks of “Reverse Vaccinology” (RV) approaches. VaxiJen and Vaxign-ML are some of the examples of ML-based RV approaches (Doytchinova and Flower, 2007; Ong et al., 2020a). Various ML approaches like RF, SVM, RFE, and deep CNN (DCNN) have been used to identify the antigens from a given protein sequence (Bowick and Barrett, 2010; Rahman et al., 2019).

Since the outbreak of the COVID-19 pandemic, different ML-based approaches have been used to predict potential epitopes to design vaccines. Ong et al. (2020b) used Vaxin and Vaxign-ML-based RV to prioritize NSPs as vaccine candidates for SARS-CoV-2. They identified NSP3 as the most promising potential target for vaccine development after spike protein (Ong et al., 2020b). Malone et al. (2020) studied the SARS-CoV-2 proteome and offered a complete vaccine design blueprint for SARS-CoV-2 using the NEC Immune Profiler suite of tools to create an epitope map for different HLA alleles. Fast and Chen utilized neural network tools, MARIA and NetMHCPan4 to identify potential T-cell epitopes for SARS-CoV-2 spike receptor-binding domain (RBD) (Fast et al., 2020). Crossman (2020) utilized deep learning RNN and provided simulated sequences of S protein to identify possible targets for vaccine design. Rahman et al. (2020) applied immunoinformatic approaches to produce a SARS-CoV-2 anti-peptide vaccine of S, E (envelop), and M (membrane) protein. They used the ML-based ElliPro antibody epitope predictive method to predict B-specific epitopes in S-protein. Prachar et al. (2020) applied 19 joined epitope-HLA tools, including the Immune Epitope Database (IEDB), ANN (PyTorch), and position-specific weight matrices (PSSM) algorithm, to identify and validate 174 epitopes of SARS-CoV-2 binding strongly to 11 HLA alleles. In addition, Sarkar et al. (2020) applied an SVM technique to design the epitope-based vaccine of COVID-19 and predict the toxicity of designed epitopes.

We believe that an AI-based framework may quicken and improve the design and the development of the vaccine formulation which can enhance the immune response, and improve protection for prophylactic vaccines.
3.3. Applications of ML in SARS-CoV-2 protein structure determination

The SARS-CoV-2, like other coronaviruses, has four conserved structural proteins and 16 non-structural proteins, such as proteases (NSP3 and NSP5) and RdRp (NSP12) (Masters, 2006).

Computational models have been used to predict protein structures (Kuhlman and Bradley, 2019). There are primarily two modeling-based approaches available for the prediction of unknown protein structures (Haddad et al., 2020). The first one is template-based modeling, which predicts structure using similar proteins as a template, and the second one is template-free modeling, where no known related structures are available. Many of the SARS-CoV-2 proteins are close homologs to the proteins in related organisms with known structures. However, for some of the proteins, template-based modeling is not possible because of the lack of experimentally determined template structure. Recently, the prediction of structures for proteins where no template structures are available has been advanced significantly via novel ML methods. Senior et al. (2020) of DeepMind company, UK have developed a system called AlphaFold to predict a variety of protein structures related to COVID-19. AlphaFold has recently won the 

Ab initio

category of CASP13 competition. (Critical Assessment of Techniques for Protein Structure Prediction) and is based on a deep neural network, ResNet architecture (Senior et al., 2019). It is generally an unbiased model predictor and ignores similar structures when making predictions, which is indeed helpful for COVID-19, as very few similar protein structures are available. Central to AlphaFold lies three different layers of deep neural networks. The first layer is made up of a variational autoencoder weighted with an attention model which creates accurate fragments based on a single amino acid sequence. The second layer is divided into two sublayers, of which the first sublayer optimizes inter-residue distances utilizing a 1D convolutional neural network (CNN) on a contact map. The second sublayer helps in optimizing the generated substructures against a protein using a 3D CNN. The third layer then scores the generated protein against the actual model (Senior et al., 2019).

The researchers at DeepMind cross-validated their AlphaFold’s generated structure of SARS-CoV-2 spike protein with the experimentally determined spike structures by the Francis Crick Institute. Motivated with the positive results, DeepMind has applied AlphaFold to predict the other structures of SARS-CoV-2 proteins including the membrane protein, protein 3a, NSP2, NSP4, NSP6, and papain-like protease (Jumper et al., 2020). These protein structures can potentially contain druggable sites, and thus will help the drug development efforts to contain COVID-19.

Zhang et al. (2020b) have used C–I-TASSER (Zheng et al., 2019) to create structural models for the SARS-CoV-2 proteins, which are available at (Zhou et al., 2021). C–I-TASSER is an extended version of I-TASSER (Yang et al., 2015) and employs the deep convolutional neural network-based contact maps (Li et al., 2019) to guide the Monte Carlo fragment assembly simulations. C–I-TASSER, also known as “Zhang–Server”, is the top-ranked computerized server for protein structure prediction in the CASP13 challenge.

Heo and Feig (2020) employed a deep-learning neural network approach built-in as part of the transform-restrained Rosetta (trRosetta) (Yang et al., 2020) pipeline, to predict the structure of the SARS-CoV-2 proteins. The dilated ResNet-enabled trRosetta network may allow for better performance as it has various output layers for the prediction of the distances and orientation between the residues of a protein. The accuracy of the predicted structure models was further improved by applying molecular dynamics simulation-based refinement. The refined trRosetta and AlphaFold’s models in this study were further compared to the Zhang C–I-TASSER (Zheng et al., 2019) models. The authors showed significant variability among most of the predicted models, however, there is some similarity in the predicted structures of the M protein, nsp4, and papain-like protease available at (Jungnick et al., 2021).

To get deeper insights into the molecular structures of different human coronavirus spike (S) proteins, Chen et al. (Serena and Chen, 2020) have employed a combination of MD simulation and deep-learning methodology on S proteins of SARS-CoV-2, SARS-CoV-1, Middle East respiratory syndrome coronavirus (MERS-CoV), and human coronavirus HKU1. They have used unsupervised deep learning architecture based on a convolutional variational autoencoder to systematically compare S protein ensembles from MD simulations. The authors demonstrated large flexibility between the subunits of the S proteins and reveal important regions for S protein oligomerization which could be considered as potential targets for therapeutic interventions.

3.4. Deep neural network translate coronavirus protein structure into music

Based on a nanomechanical analysis of the structure and motions of atoms and molecules at different scales, MIT scientists used ML-based deep neural network models to create music to represent the structure of the SARS-CoV-2 spike protein (Buehler, 2020). The principal author, Markus Buehler specializes in developing ML models to design new proteins and has extensively used sonification to illuminate structural details that might otherwise remain elusive. Sonification is a method to translate protein structures into audible signals. In this study, they have used a unique approach of sonification termed “materiomusic”, to use the actual vibrations and structures of molecules to create music. According to the study, the hierarchical organization structure of a protein is reminiscent of the music where the primary sequence of amino acids defines the notes and the secondary structure i.e., the coil of the helix or the flatness of a sheet defines the rhythm and pitch (Fig. 3A). Also, the overall vibrational motions of the molecules were defined by the Anisotropic Network Model (Eyal et al., 2015) approach and incorporated into an audio signal (Qin and Buehler, 2019). The signal is then imported into the Max device, and sounds are generated using Ableton Live (Ableton Live Digital Audio Workstation, 2020), which forms the basis for the secondary signal. The signals from the structure and vibrations of a protein were overlaid and played together, generating a multi-dimensional image of the protein's structure. Further sonification of the SARS-CoV-2 S protein in twelve-tone equal temperament tuning, a total number of 3,647,770 notes were generated in the raw musical coding (Fig. 3B). This work overall results in a nearly 2-h piece of classical music that was uploaded to the music sharing website- SoundCloud for the public to hear (SoundCloud, 2758).

The author also reports an ML-enabled nanomechanical vibrational spectrum of five different protein structures, which provides understanding into how genetic mutations and the binding of the SARS-CoV-2 S protein to the human ACE2 cell receptor directly influence the audio (Buehler, 2020). The authors further suggested that the musical representations of proteins could also be used as a tool to design effective drug therapies, development of de novo antibodies, identification of druggable sites within the coronavirus’ structure, detecting mutations, and material design by manipulating sound.

These neural network-based interventions can convert protein structure to music rapidly and the Markus team has built up a database of over 10,000 protein songs. The team also developed a free app for the Android smartphone, called the Amino Acid Synthesizer, where users can create their own protein “compositions” from the sounds of amino acids (Kozakov et al., 2006).

4. CONCLUSION and FUTURE CHALLENGES

AI and ML are being applied in many COVID-19-related domains, two of which are accelerating the structural biology of SARS-CoV-2 proteins and structure-based drug discovery along with drug repurposing. This paper has presented a survey of literature review of AI applications in the field of computational biology and medicine. In particular, we have highlighted the role of AI in drug repurposing and the structure analysis of SARS-CoV-2 proteins.

AI/ML methods are generally relying on the application domain and the types of data. Text mining techniques and graph-based approaches
are used in drug repurposing while, in predicting drug likeliness, drug target relationship, and generation of novel drug molecules, autoencoder approaches are largely helpful. DL like Graph Convolutional Network and MT-DTI approach proved to be successful to predict available antiviral drugs that could be effective against SARS-CoV-2. According to the so far studies, AI/ML subsets, homology modeling, virtual screenings, and molecular docking are the most used SARS-CoV-2 drug repurposing approaches to identify potentially effective drugs for the treatment of COVID-19 infection. AI/ML-based drug repurposing or drug discovery in the majority of the studies were not confirmed either by experimental methods or follow-up clinical studies. This illustrates the uncertainties regarding reproducibility and strong evidence of drug repurposing studies to tackle COVID-19. However, challenges remain in developing these technologies, such as data and model harmonization, data heterogeneity and quality, data sharing and security, and biological interpretability of the models.

This is the chance to accomplish the visionary outlook of scientists to deliver an AI-based efficient and agile drug discovery process at an accelerated pace and at a price that every COVID-19 patient would appreciate.

**Author contributions**

V.K. designed the study. K.P., and V.K., performed the study and analyzed the data, and wrote the manuscript.
CRediT author contribution statement

Kartikay Prasad: Data curation, Writing – original draft. Vijay Kumar: Conceptualization, Methodology, Supervision, Writing-Reviewing and Editing.

Declaration of competing interest

No potential conflict of interest was reported by the authors.

Acknowledgements

The authors sincerely thank Amity University, Noida for providing facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cprhar.2021.100042.

References

Abdai, J. 2020. Drug repurposing study pinpoint potential COVID-19 antivirals. J. Am. Med. Assoc. 324, 928.

Ableton Live Digital Audio Workstation, https://www.ableton.com/en/live/,(2020).

Abbasi, J., 2020. Drug repurposing study pinpoints potential COVID-19 antivirals. J. Am. Med. Assoc. 324, 928.

Declaration of competing interest

No potential conflict of interest was reported by the authors.

Acknowledgements

The authors sincerely thank Amity University, Noida for providing facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cprhar.2021.100042.

References

Abdai, J. 2020. Drug repurposing study pinpoint potential COVID-19 antivirals. J. Am. Med. Assoc. 324, 928.

Ableton Live Digital Audio Workstation, https://www.ableton.com/en/live/,(2020).

Abbasi, J., 2020. Drug repurposing study pinpoints potential COVID-19 antivirals. J. Am. Med. Assoc. 324, 928.

Declaration of competing interest

No potential conflict of interest was reported by the authors.

Acknowledgements

The authors sincerely thank Amity University, Noida for providing facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cprhar.2021.100042.

References
