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

Kartikay Prasad

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

Vijay Kumar (✉ vkumar33@amity.edu)

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

<https://orcid.org/0000-0002-3621-5025>

Systematic Review

Keywords: COVID-19, SARS-CoV-2, artificial intelligence, deep learning, structure, drug repurposing

Posted Date: March 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-370041/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Current Research in Pharmacology and Drug Discovery on January 1st, 2021. See the published version at <https://doi.org/10.1016/j.crphar.2021.100042>.

Abstract

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 caused massive chaos around the world, affecting people's lives and causing a large number of deaths. Till 15th October, 2020, over 38 million confirmed cases are reported worldwide with more than 1 million deaths reported (Coronavirus disease (COVID-2019) situation reports. 2020). The numbers of infection 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 power of artificial intelligence (AI) approaches have been attached to a wide range of applications across public health, disease prediction, and drug development. Fuchs in 2020 (Fuchs 2020) elegantly summarized the role of AI in the current COVID-19 pandemic at six major areas, including early predictions and alerts, tracking, data dashboards, diagnosis and prognosis, treatments, and social control.

Over the past decade, AI-based models have revolutionized drug discovery 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 a data and can make inferences from large number of test data. With the advent of deep learning (DL), the automatic feature extraction from raw data leads to the increase in performance compared to other computer-aided models (Chen et al. 2018; Zhavoronkov et al. 2019).

The recent applications of AI in case of COVID-19 includes virtual screening of both repurposed drugs as well as new chemical entities (**Figure 1**) (Keshavarzi Arshadi et al. 2020; Zhou et al. 2020b). Machine

learning (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 binds to different SARS-CoV-2 proteins and thus can potentially inhibit the viral replication and growth (Figure 1).

In addition, deep learning (DL) based applications have been the main focus for drug repurposing research, as their automatic feature extraction accelerates the process of drug discovery (Figure 1). The application of DL has benefitted the *de novo* drug design approaches. The current design approach utilizes the state-of-the-art DL models such as Generative Adversarial Networks (GAN) to generate data-based molecules (Lan et al. 2020). More recent approaches use generative models such as variational autoencoders (VAE) to generate sequences of atoms (Figure 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 (Baucum et al. 2020; Zhang and Lu 2019). To design a drug-generative network, the following information is critical, (a) collection of drug-like molecules, (b) a feature-representation of these molecules in silico, (c) method to increase diversity of the molecules, and (d) screening and modification of the altered molecules.

AI, in the field of computational biology and medicine, has been used to partially understand COVID-19 to discover drugs against the COVID-19 virus (Alimadadi et al. 2020; Ke et al. 2020; Mei et al. 2020; Nguyen 2020). Equipped with a strong computational power to deal with large amounts of data, AI can help scientists to expand the knowledge about the coronavirus quickly. For example, by determining the protein structures of the virus, researchers would be able to find 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 for rapid discovery of drugs against COVID-19 for several key reasons. The automatic feature extraction ability of deep learning can support models with better accuracy and reliable results. Also, the generative ability demonstrated by deep learning models can be largely exploited to create more druggable molecules and better epitope prediction, minimizing the chances of failure in trial experiments. Thus, the use of AI is essential in order to find out potential drugs against COVID-19 in timely 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, where the *de-novo* drug discovery is almost infeasible.

2 Methodology

The systematic literature search and analysis was done through Dimensions database (<https://www.dimensions.ai/> Accessed on 01 September 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 to January 1, 2020 to September 1, 2020. A staged literature search was performed for each section within this systematic review and all the relevant studies were identified and summarized in narrative manner in each relevant section. Studies were included if they reported outcomes on any results on structure biology of SARS-CoV-2 proteins or drug repurposing to COVID-19 utilizing AI. Whereas, the editorials, commentary, survey, and narrative reviews were excluded.

3 Results

A total of 2,052 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 275 articles were found (**Figure 2**). Out of these 275 articles, the number of articles has been posted more on open access online pre-print servers 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, 227 were excluded because they were not related to the drug repurposing work using the AI. The full text of the remaining 48 studies was then analysed; and finally, only 22 studies met the final inclusion criteria and were included in our study, and were summarized in **Table 1**.

3.1 AI/ML-based methods in COVID-19 drug repurposing

Recently, AI-based drug-discovery has been much publicised as an effective approach to accelerate the drug discovery process (Alvarez-Machancoses and Fernandez-Martinez 2019; Amino Acid Synthesizer ; Diaz et al. 2019; Hervey 2020; Sellwood et al. 2018; Yang et al. 2019; Zhu 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 COVID-19 (Abbasi 2020; Prasad et al. 2020; Zhou et al. 2020a; Zhou et al. 2020b). 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.

Insilico scientists, 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 3CL^{pro}. A total of 28 machine learning (ML) models including, Generative Autoencoders and Generative Adversarial Networks has generated molecular structures that were optimized with reinforcement learning (RL) approaches. Novel drug-like compounds made using these approaches are being published at www.insilico.com/ncov-sprint/ and has been continuously updated.

Ge et al. (Ge Y 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 L 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. (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 exploits 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 which do not have any 3D structures, such as proteins of SARS-CoV-2.

Recently, Ray and Kowalewski (Kowalewski and Ray 2020) have utilized their machine learning models to screen more than 10 million small molecules from Zinc database consists of 200 million chemicals. With their AI-linked drug discovery pipeline, they identified the best-in-class hits for the 65 human proteins that interacts with SARS-CoV-2. With their ML models, they prioritize the chemicals on the chemicals on the basis of toxicity and volatility (vapor pressure). The chemical features of the drugs were computed and cross-validated with the recursive feature elimination (RFE) along with the random forest and support vector machine algorithms.

Another research group uses knowledge-graph (KG) based deep learning method for drug repurposing in COVID-19, and termed as CoV-KGE (Zeng et al. 2020). The authors utilized DGL-KE, developed by Amazon supercomputing resource, AWS-AI (Wang et al. 2019) to construct a KG from 24 million Pubmed publications and DrugBank. This comprehensive KG includes 15 million edges across 39 types of relationships connecting drugs, diseases, genes/proteins. Pathways and expression profiles. Subsequently, a deep-learning approach (RotatE in DGL-KE) was used to provide high-confidence drug candidates for drug repurposing. The authors then identified 41 candidate drugs through enrichment

analysis of drug-gene signatures and SARS-CoV-induced transcriptome and proteomics data along with ongoing clinical data.

Gao et al. (Gao et al. 2020) uses structure-based drug repositioning (SBDR) machine learning models to evaluate the drug binding affinity to SARS-CoV-2 3CL^{pro}. In this work, the authors trained their 2-D fingerprint-based deep learning gradient-boosting decision tree (GBDT) model on 314 SARS-CoV-2/ SARS-CoV-3CL^{pro} 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 top 20 FDA-approved drugs along with top 20 investigational off-market drugs were selected as potent inhibitors of SARS-CoV-2 3CL protease.

Hofmarcher et al. (Hofmarcher M et al. 2020) conducted a deep ligand-based virtual screening using deep-learning network model "ChemAI," trained on more than 220 million data points across 3.6 million molecules from three public drug-discovery databases. They screen approximately 900 million compounds from the ZINC database for their inhibiting potentials to the SARS-CoV-2 3CL^{pro}, and the papain-like protease (PLP). Additionally, the authors also screened the DrugBank for drug repurposing. By ranking the compounds according to their predicted inhibitory potentials, the toxicity, and closeness to known drugs, the authors create 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-jku/sars-cov-inhibitors-chemai>.

Similar work has been reported by Ton et al. (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 3CL^{pro} inhibitors which were subsequently winnowed down to 1,000 compounds after docking simulation. Hu et al. (Hu F et al. 2020) used a multitask neural network model to predict protein-ligand binding affinities of viral proteins against a database of 4,895 drugs. They suggested 10 potential drugs with strong binding affinity to their target proteins.

Gysi and colleagues (Gysi et al. 2020) utilized AI-combined network medicine drug repurposing approach to rank 6,340 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. 2020) developed a Python-based DL toolkit, DeepPurpose that is based on encoder-decoder framework and presented a case study on SARS-CoV-2 3CL^{pro} with 13 potential repurposing candidates identified.

Batra et al. (Batra et al. 2020) train and validate a random forest algorithm on data from Smith et al. (Smith and Smith 2020). The authors then executed their models to CureFFI and DrugCentral datasets containing 1495 and 3967 drugs, respectively. They also applied their model to screen compounds from

the BindingDB dataset, and identified 19,000 additional candidates that binding strongly to either the native S-protein or the human ACE2-S protein complex.

Redka et al. (Redka et al. 2020) utilized deep learning 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, PoylpharmDB that contains over 10,224 drugs along with computed list of ~8,700 proteins predicted to interact with them. The interactions were generated with Cyclica's MatchMakerTM technology which is a deep-learning 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. (Zhang et al. 2020b) utilized the neural network algorithm trained on the PDBbind database and identify possible inhibitors of the SARS-CoV-2 3CL^{Pro}. They then using a structure model of the 3CL^{Pro}, explore the databases, ChemDiv and TargetMol to find promising compounds targeting 3CL^{Pro} protein.

Similarly, Nguyen et al. (Nguyen et al. 2020) applied the Mathematical Deep Learning (MathDL) approach to identify possible inhibitors for SARS-CoV-2 3CL^{Pro}. Their model has been trained on two datasets, ChEMBL and PDBbind database using two different CNNs. Finally, they 15 promising drug candidates for SARS-CoV-2 3CL^{Pro} using the trained CNN models on the DrugBank dataset.

Artigas et al. (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 on the basis of their association with a phenotype.

This strategy has been then used to screen the relationship between 6,605 drugs present in the DrugBank and the coronavirus-human interaction set retrieved through literature survey and consists of 122 human proteins. 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 combination of drug pirfenidone with melatonin could be a good candidate against COVID-19 and their combined mechanism of action have been identified at molecular level through the use of TPMS sampling-based models.

Bung et al. (Bung N 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 and results, 31 potential inhibitors of SARS-CoV-2 3CL^{Pro} has been identified.

In a study, Scott D. Bembenek of Denovicon Therapeutics (San Diego, CA, 92130, USA) used the Denovicon computational platform to perform molecular modeling-AI hybrid computational approach to find potential inhibitors of the SARS-CoV-2 main protease (M^{pro}, 3CL^{Pro}) (Bembenek 2020). Over 13,000

FDA-approved drugs and clinical candidates (approximately 30,000 protomers) were investigated and finally arrived at the five hits that may prove useful in the designing of future inhibitors of the main protease.

Tang et al. (Tang et al. 2020) combine AI with the structure-based drug design (SBDD) to speed-up the generation of potential candidate compounds against SARS-CoV-2. The authors generate a list of 284 molecules known to inhibit SARS-CoV-2 3CL^{Pro}, break them into 316 fragments and generate 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 favourable fragments, and the last is the existence of known pharmacophores. This AI-based approach generated a library of 4,922 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/tbwxbmu/2019-nCov>.

IBM research scientists from Singapore (Chenthamarakshan 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 samples with desired features. They used a multitask deep neural network (MT-DNN) to assess the toxicity of the generated molecules. The authors applied this framework to generate ~ 3000 novel drug candidates against the SARS-CoV-2 non-structural protein 9 (NSP9) replicase, the 3CL^{Pro}, and the receptor-binding domain (RBD) of the S protein.

Recently, Mahapatra et al. (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) are utilized for drug development and drug repurposing (Aynur Abdulla et al. 2020). It is a neural network approach built on a quadratic correlation between inputs defined by drugs and their individual doses and outputs defined by treatment efficacy and safety. The authors examined 12 drug/dose parameter 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. 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.

Very recently, Yi-Yu Ke et al. (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 3CL^{pro} inhibitors. The authors identified a total of 80 potential antiviral drugs, among them, 8 drugs were shown to inhibit a feline infectious peritonitis (FIP) virus in Fcwf-4 cells.

Many drug discovery companies are utilizing AI to accelerate the drug development and drug repurposing against the SARS-CoV-2 to confronts 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 Applications of ML in SARS-CoV-2 protein structure determination

The SARS-CoV-2, like other coronaviruses has four conserved structural proteins including, the membrane (M) protein, small envelop (E) protein, trimeric spike (S) glycoprotein, nucleocapsid (N) protein 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. There are primarily two modelling-based approaches available for the prediction of unknown protein structures. 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 machine learning methods. Senior et al. (Senior et al. 2020) of DeepMind company, UK have developed a system called AlphaFold to predict a variety of protein structures related with 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 help 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 J 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 (Zhang et al. 2020a) have used C-I-TASSER (Zheng et al. 2019) to create structural models for the SARS-CoV-2 proteins, which are available at (Zhang lab). 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 (Heo and Feig 2020) employed 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 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 (C-I-TASSER).

To get deeper insights into the molecular structures of different human coronavirus spike (S) proteins, Chen et al. (Serena H. Chen et al. 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.

3.3 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 is extensively using 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. The layered 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 (**Figure 3A**).

The overall vibrational motions of the molecules defined by the Anisotropic Network Model (Eyal et al. 2015) approach and were incorporated into an audio signal (Qin and Buehler 2019). The signal is then imported in the Max device, and sounds are generated using Ableton Live (Ableton Live Digital Audio Workstation 2020), forming 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 spike protein in twelve-tone equal temperament tuning, a total number of 3,647,770 notes are generated in the raw musical coding, reflecting the hierarchical structure of the protein in musical space (**Figure 3B**). This work overall results in a nearly two-hour piece of classical music which was uploaded to the music sharing website SoundCloud for the public to hear (SoundCloud.com).

The author also reports 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 spike protein to the human ACE2 cell receptor directly influence the audio (Buehler 2020). 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, material design by manipulating sound and popularising the physics of proteins.

These neural network-based interventions can convert protein structure to music rapidly and 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 (Amino Acid Synthesizer).

4 Conclusion And Future Challenges

COVID-19 disease has tremendously affected the lives of people around the globe. While AI technologies have been applied successfully in almost every corner of humans' lives, they have also contributed to helping humans in the extremely tough fight against COVID-19.

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.

Recently, Wynants and colleagues (Wynants et al. 2020) presented a systematic review to appraise and validate a large number of published prediction models for the diagnosis of COVID-19 through COVID-PRECISE (Precise Risk Estimation to optimise COVID-19 Care for Infected or Suspected patients in diverse settings) group in collaboration with the Cochrane Prognosis Methods Group. The authors in this

systematic review showed that the prediction models were not adequate in reporting the prediction of COVID-19 and emphasized the urgent need to develop rigorous prediction models and their validation. Thus, we believe that although a large number of studies reported the candidate drugs for repurposing, future studies aimed at developing and validating the clinical efficacy of the identified drugs against COVID-19 is urgently needed.

Although AI and its subsets have matured in recent years and are becoming an integral part of the drug development or drug repurposing. There is still a paucity of confirmatory results on any of the repurposed drugs even after ten months of the pandemic. This illustrates the uncertainties regarding the reproducibility and strong evidences of drug repurposing studies to tackle COVID-19. AI/ML technologies utilized within the drug development studies have greatly improved and could serve, in the near future, as a decision support system for policymakers, healthcare providers and society at large. The development of effective and robust in vitro and in vivo models can decrease the failure rate of repurposed drugs in the preclinical studies and clinical trials. However, challenges remain in developing these technologies, such as data and model harmonisation, 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 AI-based efficient and agile drug discovery process at an accelerated pace and at a price that every COVID-19 patient would appreciate.

Declarations

Acknowledgments

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

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Abbasi J (2020) Drug Repurposing Study Pinpoints Potential COVID-19 Antivirals JAMA 324:928 doi:10.1001/jama.2020.159482770383 [pii]

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

Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X (2020) Artificial intelligence and machine learning to fight COVID-19 Physiol Genomics 52:200-202 doi:10.1152/physiolgenomics.00029.2020

Alvarez-Machancoses O, Fernandez-Martinez JL (2019) Using artificial intelligence methods to speed up drug discovery Expert Opin Drug Discov 14:769-777 doi:10.1080/17460441.2019.1621284

Amino Acid Synthesizer <https://playgooglecom/store/apps/details?id=comsynthaminoacidplayer>

Artigas L et al. (2020) In-silico drug repurposing study predicts the combination of pirfenidone and melatonin as a promising candidate therapy to reduce SARS-CoV-2 infection progression and respiratory distress caused by cytokine storm PLoS One 15:e0240149 doi:10.1371/journal.pone.0240149PONE-D-20-13552 [pii]

Aynur Abdulla et al. (2020) Project IDentif.AI: Harnessing Artificial Intelligence to Rapidly Optimize Combination Therapy Development for Infectious Disease Intervention Advanced Therapeutics doi:<https://doi.org/10.1002/adtp.202000034>

Batra R, Chan H, Kamath G, Ramprasad R, Cherukara MJ, Sankaranarayanan S (2020) Screening of Therapeutic Agents for COVID-19 Using Machine Learning and Ensemble Docking Studies J Phys Chem Lett 11:7058-7065 doi:10.1021/acs.jpcclett.0c02278

Baucum M, Khojandi A, Vasudevan R (2020) Improving Deep Reinforcement Learning with Transitional Variational Autoencoders: A Healthcare Application IEEE J Biomed Health Inform PP doi:10.1109/JBHI.2020.3027443

Beck BR, Shin B, Choi Y, Park S, Kang K (2020) Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model Comput Struct Biotechnol J 18:784-790 doi:10.1016/j.csbj.2020.03.025S2001-0370(20)30049-0 [pii]

Bembenek SD (2020) Drug Repurposing and New Therapeutic Strategies for SARS-CoV-2 Disease Using a Novel Molecular Modeling-AI Hybrid Workflow ChemRxiv Preprint

Buehler MJ (2020) Nanomechanical sonification of the 2019-nCoV coronavirus spike protein through a materiomusical approach arXiv 2003.14258

Bung N, Krishnan SR, Bulusu G, A R (2020) De Novo Design of New Chemical Entities (NCEs) for SARS-CoV-2 Using Artificial Intelligence. ChemRxiv Preprint doi:doi:10.26434/chemrxiv.11998347.v2.

C-I-TASSER <https://zhanglabccmbmedumichedu/C-I-TASSER/2019-nCoV/>

Cantürk S, Singh A, St-Amant P, Behrmann P (2020) Machine-learning driven drug repurposing for covid-19 arXiv doi:arXiv:2006.14707v1

Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery Drug Discov Today 23:1241-1250 doi:S1359-6446(17)30359-8 [pii]10.1016/j.drudis.2018.01.039

Chenthamarakshan V et al. (2020) Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models. arXiv Preprint doi:arXiv:200401215.

Coronavirus disease (COVID-2019) situation reports. (2020)

<http://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>

Diaz O, Dalton JAR, Giraldo J (2019) Artificial Intelligence: A Novel Approach for Drug Discovery Trends Pharmacol Sci 40:550-551 doi:S0165-6147(19)30136-1 [pii]10.1016/j.tips.2019.06.005

Eyal E, Lum G, Bahar I (2015) The anisotropic network model web server at 2015 (ANM 2.0) Bioinformatics 31:1487-1489 doi:10.1093/bioinformatics/btu847btu847 [pii]

Fleming N (2018) How artificial intelligence is changing drug discovery Nature 557:S55-S57 doi:10.1038/d41586-018-05267-x10.1038/d41586-018-05267-x [pii]

Fuchs C (2020) Everyday Life and Everyday Communication in Coronavirus Capitalism. tripleC: Communication, Capitalism & Critique Journal for a Global Sustainable Information Society 18:375-399

Gao K, Nguyen DD, Chen J, Wang R, Wei GW (2020) Repositioning of 8565 Existing Drugs for COVID-19 J Phys Chem Lett 11:5373-5382 doi:10.1021/acs.jpcclett.0c01579

Ge Y et al. (2020) A Data-Driven Drug Repositioning Framework Discovered a Potential Therapeutic Agent Targeting COVID-19. bioRxiv preprint doi:<https://doi.org/10.1101/2020.03.11.986836>

Griffiths RR, Hernandez-Lobato JM (2020) Constrained Bayesian optimization for automatic chemical design using variational autoencoders Chem Sci 11:577-586 doi:10.1039/c9sc04026ac9sc04026a [pii]

Gysi DM et al. (2020) Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19 arXiv doi:arXiv:2004.07229v1 [pii] 2004.07229 [pii]

Heo L, Feig M (2020) Modeling of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Proteins by Machine Learning and Physics-Based Refinement bioRxiv doi:2020.03.25.008904 [pii] 10.1101/2020.03.25.008904

Hervey M (2020) Harnessing AI in drug discovery without losing patent protection Drug Discov Today doi:S1359-6446(20)30114-8 [pii] 10.1016/j.drudis.2020.03.007

Hofmarcher M et al. (2020) Large-scale ligand-based virtual screening for SARS-CoV-2 inhibitors using deep neural networks arXiv doi:arXiv:2004.00979

Hong L, Lin J, Tao J, J. Z (2019) BERE: An Accurate Distantly Supervised Biomedical Entity Relation Extraction Network. arXiv preprint doi:arXiv:190606916 <https://www.dimensions.ai/> (Accessed on 01 September 2020.)

Hu F, Jiang J, P. Y (2020) Prediction of Potential Commercially Inhibitors against SARS-CoV-2 by Multi-Task Deep Model. arXiv preprint doi:arXiv:200300728.

Huang K, Fu T, Glass L, Zitnik M, Xiao C, Sun J (2020) DeepPurpose: a deep learning library for drug-target interaction prediction and applications to repurposing and screening. ArXiv doi:arXiv:2004.08919v2

Jorba G et al. (2020) In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan PLoS One 15:e0228926 doi:10.1371/journal.pone.0228926 PONE-D-19-21700 [pii]

Jumper J, Tunyasuvunakool K, Kohli P, Hassabis D, team. atA (2020) Computational predictions of protein structures associated with COVID-19. DeepMind website <https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19>.

Ke YY et al. (2020) Artificial intelligence approach fighting COVID-19 with repurposing drugs Biomed J doi:10.1016/j.bj.2020.05.001 S2319-4170(20)30049-4 [pii]

Keshavarzi Arshadi A et al. (2020) Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development. Front Artif Intell 3 doi:doi: 10.3389/frai.2020.00065

Kowalewski J, Ray A (2020) Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space Heliyon 6:e04639 doi:10.1016/j.heliyon.2020.e04639 S2405-8440(20)31483-3 [pii] e04639 [pii]

Lan L et al. (2020) Generative Adversarial Networks and Its Applications in Biomedical Informatics Front Public Health 8:164 doi:10.3389/fpubh.2020.00164

Lavecchia A (2019) Deep learning in drug discovery: opportunities, challenges and future prospects Drug Discov Today 24:2017-2032 doi:S1359-6446(19)30282-X [pii] 10.1016/j.drudis.2019.07.006

Li Y, Zhang C, Bell EW, Yu DJ, Zhang Y (2019) Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13 Proteins 87:1082-1091 doi:10.1002/prot.25798

Liberati A et al. (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration PLoS Med 6:e1000100 doi:10.1371/journal.pmed.1000100

Mahapatra S, Nath P, Chatterjee M, Das N, Kalita D, Roy P, Satapathi S (2020) Repurposing Therapeutics for COVID-19: Rapid Prediction of Commercially available drugs through Machine Learning and Docking medRxiv doi:<https://doi.org/10.1101/2020.04.05.20054254>

Masters PS (2006) The molecular biology of coronaviruses Adv Virus Res 66:193-292 doi:S0065-3527(06)66005-3 [pii] 10.1016/S0065-3527(06)66005-3

Mei X et al. (2020) Artificial intelligence-enabled rapid diagnosis of patients with COVID-19 Nat Med doi:10.1038/s41591-020-0931-3 10.1038/s41591-020-0931-3 [pii]

Nguyen DD, Gao K, Chen J, Wang R, Wei GW (2020) Potentially highly potent drugs for 2019-nCoV bioRxiv doi:2020.02.05.936013 [pii] 10.1101/2020.02.05.936013

Nguyen TT (2020) Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions. Preprint doi:DOI: 10.13140/RG.2.2.36491.23846.

Prasad K et al. (2020) Targeting hub genes and pathways of innate immune response in COVID-19: A network biology perspective Int J Biol Macromol 163:1-8 doi:S0141-8130(20)33683-7 [pii] 10.1016/j.ijbiomac.2020.06.228

Qin Z, Buehler MJ (2019) Analysis of the vibrational and sound spectrum of over 100,000 protein structures and application in sonification . Extrem Mech Lett 100460 doi:https://doi.org/10.1016/j.eml.2019.100460.

Redka DS, MacKinnon SS, Landon M, Windemuth A, Kurji N, Shahani V (2020) PolypharmDB, a Deep Learning-Based Resource, Quickly Identifies Repurposed Drug Candidates for COVID-19. ChemRxiv Preprint doi:10.26434/chemrxiv.12071271.v1

Richardson P, Griffin I, Tucker C, Smith D, Oechsle O, Phelan A, Stebbing J (2020) Baricitinib as potential treatment for 2019-nCoV acute respiratory disease Lancet 395:e30-e31 doi:S0140-6736(20)30304-4 [pii] 10.1016/S0140-6736(20)30304-4

Sellwood MA, Ahmed M, Segler MH, Brown N (2018) Artificial intelligence in drug discovery Future Med Chem 10:2025-2028 doi:10.4155/fmc-2018-0212

Senior AW et al. (2019) Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13) Proteins 87:1141-1148 doi:10.1002/prot.25834

Senior AW et al. (2020) Improved protein structure prediction using potentials from deep learning Nature 577:706-710 doi:10.1038/s41586-019-1923-7 10.1038/s41586-019-1923-7 [pii]

Serena H. Chen, M. Todd Young, John Gounley, Christopher Stanley, Bhowmik aD (2020) Distinct Structural Flexibility within SARS-CoV-2 Spike Protein Reveals Potential Therapeutic Targets bioRxiv doi:https://doi.org/10.1101/2020.04.17.047548

Smith M, Smith JC (2020) Repurposing Therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface ChemRxiv Preprint

SoundCloud.com <https://soundcloud.com/user-275864738/viral-counterpoint-of-the-coronavirus-spike-protein-2019-ncov?in=user-275864738/sets/protein-counterpoint>

Tang B, He F, Liu D, Fang M, Wu Z, Xu D (2020) AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2 bioRxiv doi:2020.03.03.972133 [pii] 10.1101/2020.03.03.972133

Ton AT, Gentile F, Hsing M, Ban F, Cherkasov A (2020) Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds Mol Inform doi:10.1002/minf.202000028

Wang M et al. (2019) Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs. ArXiv doi:arXiv:1909.01315

Wynants L et al. (2020) Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal BMJ 369:m1328 doi:10.1136/bmj.m1328

Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D (2020) Improved protein structure prediction using predicted interresidue orientations Proc Natl Acad Sci U S A 117:1496-1503 doi:10.1073/pnas.1914677117 1914677117 [pii]

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction Nat Methods 12:7-8 doi:10.1038/nmeth.3213 nmeth.3213 [pii]

Yang X, Wang Y, Byrne R, Schneider G, Yang S (2019) Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery Chem Rev 119:10520-10594 doi:10.1021/acs.chemrev.8b00728

Zeng X et al. (2020) Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning J Proteome Res doi:10.1021/acs.jproteome.0c00316

Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y (2020a) Protein Structure and Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as Its Intermediate Host and the Unique Similarity between Its Spike Protein Insertions and HIV-1 J Proteome Res 19:1351-1360 doi:10.1021/acs.jproteome.0c00129

Zhang H et al. (2020b) Deep Learning Based Drug Screening for Novel Coronavirus 2019-nCov Interdiscip Sci doi:10.1007/s12539-020-00376-6 10.1007/s12539-020-00376-6 [pii]

Zhang lab <https://zhanglabccmbmedumichedu/COVID-19/>

Zhang Y, Lu Z (2019) Exploring semi-supervised variational autoencoders for biomedical relation extraction Methods 166:112-119 doi:S1046-2023(18)30330-X [pii] 10.1016/j.ymeth.2019.02.021

Zhavoronkov A et al. (2020) Potential COVID-2019 3C-like protease inhibitors designed using generative deep learning approaches. ChemRxiv Preprint doi:<https://doi.org/10.26434/chemrxiv.11829102.v2>.

Zhavoronkov A et al. (2019) Deep learning enables rapid identification of potent DDR1 kinase inhibitors Nat Biotechnol 37:1038-1040 doi:10.1038/s41587-019-0224-x 10.1038/s41587-019-0224-x [pii]

Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y (2019) Deep-learning contact-map guided protein structure prediction in CASP13 Proteins 87:1149-1164 doi:10.1002/prot.25792

Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F (2020a) Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2 *Cell Discov* 6:14 doi:10.1038/s41421-020-0153-3 153 [pii]

Zhou Y, Wang F, Tang J, Nussinov R, Cheng F (2020b) Artificial intelligence in COVID-19 drug repurposing
Lancet Digit Health doi:10.1016/S2589-7500(20)30192-8 S2589-7500(20)30192-8 [pii]

Zhu H (2020) Big Data and Artificial Intelligence Modeling for Drug Discovery *Annu Rev Pharmacol Toxicol* 60:573-589 doi:10.1146/annurev-pharmtox-010919-023324

Tables

Table 1. Summary of important AI and ML-based studies for drug repurposing in COVID-19

Author	Study keywords	Targets	No. of drugs	Reference
Kowalewski	Machine learning, SVM, REF	65 host SARS-CoV-2 targets	+10 million	(Kowalewski and Ray 2020)
Beck	Hybrid CNN and RNN model called MT-DTI	3CLpro, RdRp, helicase, 3'-to-5' exonuclease, endoRNase, and 2'-O-ribose methyltransferase	3,410	(Beck et al. 2020)
Chentamarakshan	Deep learning-based generative model, CogMol, VAE	Mpro, Spike, NSP9	3,000	(Chentamarakshan et al. 2020)
Ton	deep learning platform – Deep Docking (DD)	Mpro	1,000	(Ton et al. 2020)
Zhavoronkov	Deep learning-based generative model, AAE, GENTRL	3CLpro, Mpro	1,000	(Zhavoronkov et al. 2020)
Zhang	Dense fully CNN (DFCNN), DL, Virtual screening	3CLpro	+100 drugs & 20 peptides	(Zhang et al. 2020b)
Batra	Random forest (RF) regression algorithm, ensemble docking	S, S-ACE2 complex	187	(Batra et al. 2020)
Ke	Deep neural network	FIP	80	(Ke et al. 2020)
Gysi	Graph Neural Network	SARS-CoV-2	77	(Gysi et al. 2020)
Ge, Y	Knowledge Graph, deep learning	SARS-CoV-2	64	(Ge Y et al. 2020)
Tang	Advanced deep-Q learning network with the fragment-based drug design (ADQN-FDNN)	3CLpro	47	(Tang et al. 2020)
Zeng	Deep learning-based knowledge graph (CoV-KGE),	SARS-CoV-2	41	(Zeng et al. 2020)
Gao	2-D fingerprint, GBDT model,	3CLpro	40	(Gao et al. 2020)

	Recurrent Neural Network (RNN)			
Bung	Deep neural network, New Chemical Entities (NCE)	3CLpro	31	(Bung N et al. 2020)
Redka	Deep learning, Ligand Design, MatchMaker, PolypharmDB	3CLpro, Spike; ACE2, TMPRSS2, Cathepsin B	30	(Redka et al. 2020)
Huang	Deep Purpose, Python toolkit, CNN	3CLpro	13	(Huang et al. 2020)
Artigas	Therapeutic Performance Mapping System (TPMS), ANN, GUILDify	SARS-CoV-2 interactome	12	(Artigas et al. 2020)
Canturk	ANN, long-short term memory network (LSTM), CNN	SARS-CoV-2	12	(Cantürk et al. 2020)
Hu, F	Deep learning-based multi task models, Classification and Regression	3CLpro	10	(Hu F et al. 2020)
Hofmarcher	Deep neural network, ChemAI	3CLpro, PLP	20	(Hofmarcher M et al. 2020)
Nguyen	Mathematical Deep Learning (MathDL), CNN	3CLpro	15	(Nguyen et al. 2020)
Mahapatra	ML, Naïve Bayes classification algorithm, molecular docking	3CLpro	10	(Mahapatra et al. 2020)

Figures

Figure 1

The pipeline of AI and ML-based platforms for drug discovery and drug repurposing in COVID-19.

Figure 2

PRISMA flow chart for literature search results.

Adapted from <https://aip.scitation.org/doi/full/10.1063/1.5133026>

Figure 3

Hierarchical structure of proteins and music. (A) The three-dimensional structure of a protein can be translated into musical score through a process known as sonification which involves a deep neural network model. (B) Neural Network platform (CNN/RNN) for translating protein structure to music. Adapted from <https://towardsdatascience.com/everyprotein-is-a-song-6d30ee9add4>