BioDrugs (2023) 37:649–674
https://doi.org/10.1007/s40259-023-00611-8

REVIEW ARTICLE

Staying Ahead of the Game: How SARS-CoV-2 has Accelerated the Application of Machine Learning in Pandemic Management

Alexander H. Williams[1,2,3] · **Chang-Guo Zhan**[1,2]

Accepted: 28 May 2023 / Published online: 18 July 2023

Abstract

In recent years, machine learning (ML) techniques have garnered considerable interest for their potential use in accelerating the rate of drug discovery. With the emergence of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, the utilization of ML has become even more crucial in the search for effective antiviral medications. The pandemic has presented the scientific community with a unique challenge, and the rapid identification of potential treatments has become an urgent priority. Researchers have been able to accelerate the process of identifying drug candidates, repurposing existing drugs, and designing new compounds with desirable properties using machine learning in drug discovery. To train predictive models, ML techniques in drug discovery rely on the analysis of large datasets, including both experimental and clinical data. These models can be used to predict the biological activities, potential side effects, and interactions with specific target proteins of drug candidates. This strategy has proven to be an effective method for identifying potential coronavirus disease 2019 (COVID-19) and other disease treatments. This paper offers a thorough analysis of the various ML techniques implemented to combat COVID-19, including supervised and unsupervised learning, deep learning, and natural language processing. The paper discusses the impact of these techniques on pandemic drug development, including the identification of potential treatments, the understanding of the disease mechanism, and the creation of effective and safe therapeutics. The lessons learned can be applied to future outbreaks and drug discovery initiatives.

Key Points

Machine learning (ML) and artificial intelligence (AI) methodologies have risen in prominence since the beginning of the coronavirus pandemic.

Machine learning techniques have been utilized within the pharmaceutical sciences for both drug repurposing and for novel drug discovery against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus.

Additional resources for the validation of these repurposed and newly discovered compounds are required, as many lack sufficient data concerning their in vitro and in vivo potency against their purported target.

Extended author information available on the last page of the article

1 Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [a.k.a. coronavirus disease 2019 (COVID-19)] pandemic has over 600 million confirmed cases, with over 6.5 million deaths caused by this disease [1, 2]. The SARS-CoV-2 virus is the descendent of the SARS-CoV-1 (colloquially known as SARS), which caused a global outbreak of respiratory illness from 2002 to 2004 [3, 4]. These viruses are members of the *Coronaviridae* family and are viruses that releases a single-stranded RNA into the infected cells. The released strand encodes the four structural proteins: spike (S), envelope (E), membrane (M), and nucleocapsid (N), along with sixteen nonstructural proteins, and nine accessory proteins, which come together to form a complete copy of the virion [5–7]. The spike protein is responsible for recognizing receptors on a host cell's surface, primarily the angiotensin converting enzyme 2 (ACE2) located on the surface of human cells [8, 9]. Once the virus has recognized this protein, it begins a process of attachment and fusion with the host membrane. The virus's single-stranded RNA is then released into the host cell, where the viral genome is translated

into their protein products by the host's ribosomes [6]. These genes encode large polyproteins, which are later cleaved by the translated papain like protease (PLpro) to release both non-structural (i.e., RNA-dependent RNA polymerase (RdRp) and helicase) and structural proteins (e.g., spike, membrane, etc.) The RdRp protein synthesizes new copies of the viral RNA, and final assembly of the new SARS-CoV-2 virion occurs at the endoplasmic reticulum and Golgi apparatus interface. The overproduction of these new virions eventually leads the lysis and subsequent death of the host cell [10].

In comparison with the original SARS outbreak, the SARS-CoV-2 pandemic has been exacerbated by several factors. The first factor is the virus's communicability, with an estimated reproductive number (R0) of 2.2– 3.58 [11–13]. This means that, on average, every infected person will infect two to three people. Second, the virus has a long incubation period of up to 2 weeks, during which an infected person can spread the disease to others without showing any symptoms. The third factor is the virus lethality, with a case fatality rate of 2.3%. While the introduction of SARS-CoV-2 vaccines in late 2020 introduced an additional protection from SARS-CoV-2 infection [providing an 86% vaccine effectiveness (VE) against infection, hospitalization, intensive care unit (ICU) admission, and death] [14], breakthrough infections are still possible, and thus antiviral treatments are still being developed.

In contrast to the pressing nature of the coronavirus pandemic, the development of new drugs (i.e., de novo drug discovery) is a complex and challenging process that is often hindered by prohibitive costs and long timelines. However, researchers have recently begun implementing artificial intelligence (AI) and machine learning (ML) into the drug discovery process [15–23]. By analyzing large volumes of data, AI can identify potential drug candidates that might have been missed using conventional screening methods, in a fraction of the time [24–26]. In this review, we provide a brief overview of machine learning and its applications to drug discovery and how these methodologies have accelerated drug discovery for therapeutics targeted towards SARS-CoV-2, either through identifying drugs that could be repurposed, or by designing new therapeutics de novo tailored for the SARS-CoV-2 virus. We also examine the challenges and limitations of these techniques and their potential impact on the future of drug development.

2 An Introduction to Artificial Intelligence and Machine Learning

Since the start of the COVID-19 pandemic, the use of artificial intelligence (AI) and machine learning (ML) methodologies within healthcare has increased rapidly. Since 2020, over 30,000 articles per year mentioning machine learning have been listed on PubMed, a 400% increase over the per year average of 2010–2019. Drug discovery has not been unaffected by this trend, with over 3500 articles submitted utilizing machine learning to advance the discovery rates of new chemical entities since 2019 [27, 28].

Machine learning has the potential to improve our ability to both discover new compounds, repurpose existing drugs, and accelerate the drug discovery pipeline. The ability for these algorithms to analyze vast sets of existing chemical structures, analyze biological and preclinical data, and detect patterns and relationships that may not be immediately apparent to humans has the potential to significantly speed up the drug discovery process. Machine learning can also help identify new targets for drug development, predict the toxicity and pharmacokinetics of potential drug candidates, and optimize lead compounds for potency and selectivity [15, 16, 19, 29–31].

2.1 Machine Learning Overview

Machine learning is a quickly growing field, focused on creating algorithms that can learn from the data they are provided (i.e., training data), identify patterns within that data, and then proceed to use what it has "learned" to make decisions and predictions when presented with novel data. These types of algorithms, which require significant amounts of computational time due to their iterative training nature, have become more accessible to numerous fields in conjunction with the arrival of large datasets and the widespread availability of cheap computational power. Machine learning has been implemented in many areas of healthcare and drug development, from identifying cancer to predicting the toxicity of drug candidates [18, 26, 32–35].

Unlike typical algorithms, whose actions are determined at the time of their writing, machine learning algorithms must first train on their input data to make those decisions. This learning can either be supervised or unsupervised, which refers to whether the algorithm has access to the correct output to the input training data. For example, for a machine learning algorithm created for the identification of antiviral drugs, the training data would likely be composed of a set of thousands of drugs, each compound containing its molecular descriptors (e.g., LogP, molecular weight, etc.) along with its endpoint label (i.e., active versus inactive). A supervised training algorithm may have access to the correct endpoint of a drug (e.g., the drugs' antiviral classification or potency against an antiviral target), while the unsupervised algorithm would only have the compounds' descriptors, clustering the compounds into discrete clusters. Unlike supervised learning, unsupervised learning cannot be used in regression problems (i.e., drug activity) due to its lack of access to these endpoint values.

The power of these machine learning algorithms lies in their capacity to enhance their predictions over time. The more data and the longer they are trained, the better they become. With additional data, the model can learn new patterns and generalize better to new, unseen data. Similarly, longer training times allow the model to better understand the relationships within the data, leading to more accurate predictions. This capability of continuous improvement is a key advantage of machine learning algorithms, and it allows them to adapt to changing environments and to provide more precise results.

2.2 Commonly Used Learning Models

When building a machine learning model, the output desired can usually be classified into two broad categories: regression or categorization. For example, when working on drug discovery, a regression model could be used to predict the binding affinity of a compound to a particular protein target, using previously tested compounds as the training data. On the other hand, a classification model might be useful to determine whether a compound can permeate through the blood–brain barrier. Both methodologies have their strengths and limitations that need to be considered before selecting the appropriate method. While regression models can offer a continuous output, making them suitable for predicting a wide range of numerical values, they may struggle with extreme outliers or non-linear relationships. Conversely, classification models can provide a binary or categorical output, making them ideal for tasks such as determining whether a compound is active or inactive. However, they may be less precise in predicting values that fall between discrete categories.

2.2.1 Models For Categorization

Categorization can be roughly split into two different types of problems: classification and clustering. Classification is a supervised learning process, in which the goal is to predict the class of unseen data on the basis of a labeled dataset, on which a model has been previously trained. Conversely, clustering is an unsupervised learning process in which a model has been trained to group data points based on their features, without a priori knowledge of what the endpoints of the training data are. These models have been previously applied to numerous fields, including machine vision (e.g., object identification), medical diagnoses and financial transaction categorization.

2.2.1.1 Logistic Regression Logistic regression is one of the simplest methods of classification, as its very nature is binary and can only take in two endpoint classes, usually represented as either a 0 or a 1. These models use a sigmoidal function to separate these two classes and each input variable is given its own weight, these values are then summed to determine the log likelihood of the current datapoint using the following equations (Eqs. 1, 2):

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \tag{1}$$

or…

$$p = \frac{1}{1 + e^{-(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n)}}. \tag{2}$$

In this scenario, p represents the likelihood of the current datapoint being the endpoint class of 1 (e.g., if using this model were to be used as a central nervous system (CNS) drug classifier, our endpoint of 1 may be that a drug is CNS active). While these models can give outputs of anywhere from 0 to 1, these outputs are commonly rounded to the nearest integer to give a binary classifier. After the initial round of weights is used, the accuracy of the model is tested, and through an optimization algorithm (e.g., gradient descent, steepest descent, etc.), these weights are incrementally changed to optimize the accuracy of the model (Fig. 1).

Logistic regression is a commonly used classification algorithm due to its simplicity and robustness; however, its assumption that the input and output variable share a linear relationship limits it in applications where the relation may have a logarithmic, quadratic, or exponential relationship.

2.2.1.2 *K*-Means Clustering *K*-Means clustering is an unsupervised learning method that aims to categorize data into an integer number of clusters. This method starts by randomly choosing a K number of centroids to represent the center of each cluster within the dataset, where K is provided by the user (Fig. 2) [37–41]. Using these centroid positions, each point from the dataset has its distance to each centroid calculated and is given a cluster designation based on what the nearest centroid is. Iteratively, the centroid of each cluster is moved closer to the mean location of the datapoints within each cluster, reassigning the datapoints' cluster designation each time the centroids are recalculated. This process continues until the centroid positions converge, or when a maximum number of iterations are reached.

This methodology is commonly used across numerous fields, ranging from computer vision, where it is used to separate the feature space of images based on colors and texture, to biology, where *K*-means clustering can help to cluster genes based on their expression patterns within tissues. While *K*-means is a versatile methodology, it can

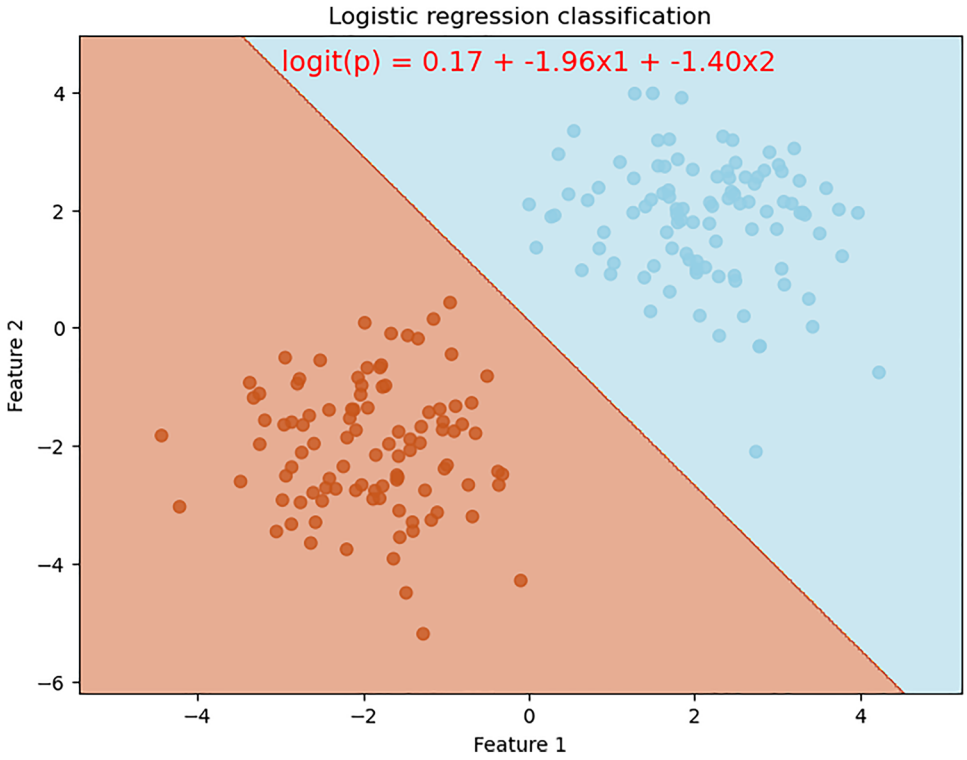

Fig. 1 Scatter plot of a two-cluster dataset (blue and brown points), each cluster centered around [− 2, − 2] and [2, 2], respectively. The decision boundary was obtained using a logistic regression classifier, represented as the dark brown line bifurcating the two sets. The dataset consists of two clusters of points, where each cluster corresponds to a specified different class. A logistic regression classifier forms the scikit-learn python library [36] was trained on the dataset to classify new datapoints into one of the two classes based on their features. The plot shows that the decision boundary (defined by the equation in red) separates the two clusters reasonably well, capturing the underlying relationship between the variables

struggle in situations where the clusters may not be spherical in shape, have outliers, or when the clusters are unevenly sized. Due to the random starting points of the centroids, they may encounter scenarios where they converge in a local minimum instead of the global minimum (i.e., the true cluster center).

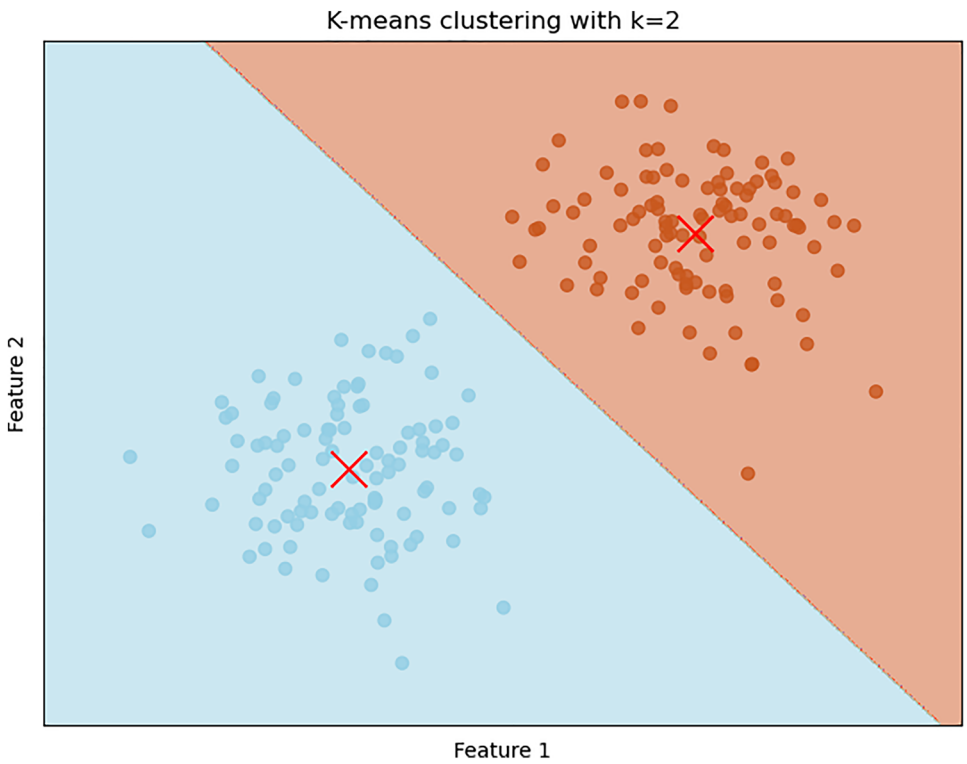

Fig. 2 Scatter plot of a two-cluster dataset (blue and brown points); cluster centers shown as red stars and the decision boundary of the *K*-means classifier model using the scikit-learn library is shown. The model was trained on the dataset to classify new datapoints into one of the two classes based on their features

2.2.2 Models For Regression

Regression modeling is simply attempting to predict a continuous numerical value based on the input features for a certain datapoint. These models are widely used in finance, economics and the natural sciences.

Linear regression is one of the simplest possible models for regression; using the equation of a linear line to fit the data (i.e., $Y = mX + B$), multiple variables can be used to establish an equation such as:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n + \varepsilon. \quad (3)$$

These b values can be tuned over an iterative process known as least squares, which will minimize the sum of the squared errors between the predicted and actual values. Linear regression (Fig. 3) is one of the most interpretable models, due to the coefficients being clearly represented for each of the variables used to establish the relationship; however, much like logistic regression, there is a base assumption that the relationship between the input and output is indeed linear and can perform poorly with variables that have a non-linear relationship to the output response.

Polynomial regression allows for these types of relations by changing the equation to an nth degree polynomial, as shown below.

$$Y = b_0 + b_1X_1 + b_2X^2 + \cdots + b_nX_n{}^n + \varepsilon \quad (4)$$

While this model does allow for non-linear relationships to be represented, it can also add the danger of overfitting the data, as the use of unlimited polynomial features allows for the complexity of the fitting line to increase, perfectly fitting the data.

2.2.3 Models for Both Regression and Classification

2.2.3.1 Support Vector Machines Support vector machines (SVM) are a machine learning algorithm that employs a hyperplane (i.e., a line in two-dimensional space, or a plane in three-dimensional space) to separate data into distinct classes. This methodology is known for its ability to manage both high dimensionality and non-linear datasets. The algorithm achieves this group separation by finding the hyperplane that has the highest margin (i.e., is the farthest away) from the group members on either side of the plane (Fig. 4). For regression problems, instead of separating the points, the algorithm looks to find a hyperplane that best fits the data, while maintaining the maximum margin on either side of the plane. SVM models have been used in a variety of chemical applications, including the prediction of drug blood/brain partitioning behavior to predicting protein/ligand binding affinities [42–47].

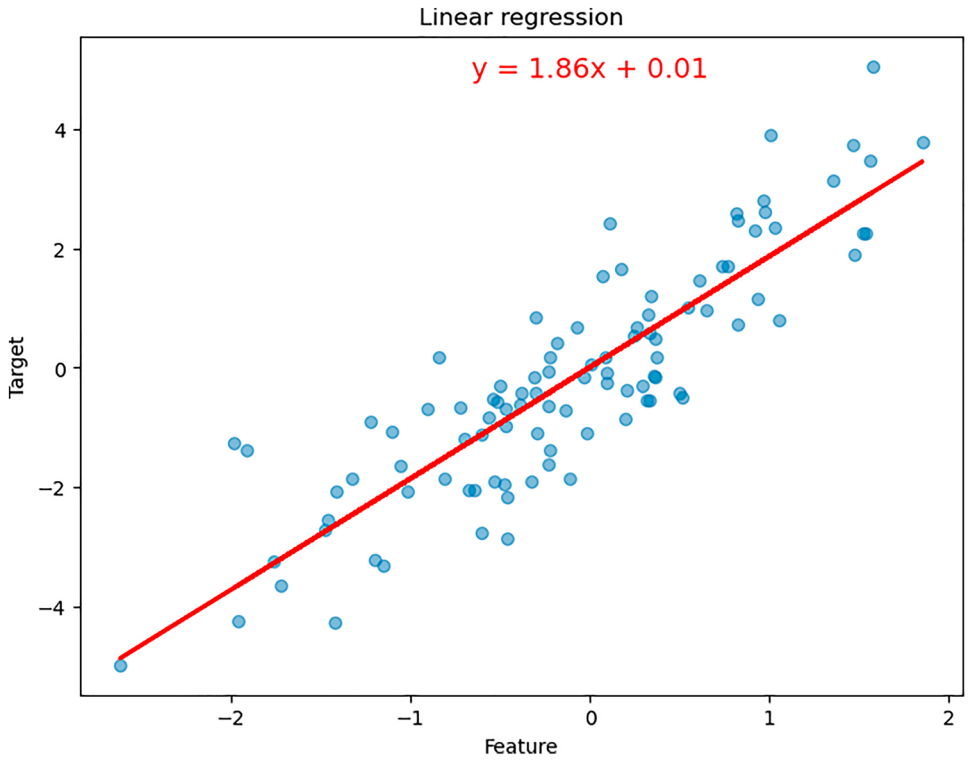

Fig. 3 Scatter plot of a random dataset (blue points) and the linear regression line (red line) obtained using a linear regression model from the scikit-learn library. The dataset consists of 100 samples generated from a normal distribution with a linear relationship between the feature and target variables. The linear regression model was trained on the dataset to predict the target variable based on the feature variable. The regression equation is represented in red text

2.2.3.2 Decision Trees and Random Forest Decision trees are a machine learning model that utilizes cascading partitions in the dataset to make a final prediction, which can either be a numerical value for regression problems, or a class label for classification problems [48–50]. The root node of the tree represents the entire dataset, with each branch being partitioned on a specific input feature. For each attempted split along a particular feature, the algorithm measures whether the newly created nodes contain an abundance of one class over another (Fig. 6). If the two nodes have a preponderance of one class over another, then that split provides additional information and is kept. Conversely, if the dataset split resulted in two nodes with an even distribution of the two groups, then the split did not provide any additional information. Like with other machine learning methodologies, a balance must be struck between how well the model does on the training set and its overall generalizability. Creating too many nodes will increase the performance on the training set but will overall destroy the generalizability of new, unseen data due to being overtrained.

Random forest models proceed to take many decision trees, each trained on a random subset of the training data and features (Fig. 5). Once these individual trees are trained, their independent evaluations are aggregated, and a final prediction is made based on the preponderance of the votes. These models have been used to assess many chemical properties, from overall function to the bitterness of their taste [43, 51].

2.2.3.3 Neural Networks Artificial neural networks (ANNs), also known as neural networks, are a type of machine learning model inspired by the structure and function of the human brain [52, 53]. A neural network is made up of several layers of interconnected nodes, or "neurons," that process and transmit data. Each neuron receives input from other neurons, applies a non-linear function to the weighted sum of those inputs, and then sends the result of this function as its output to neurons in the next layer. These applied functions, known as activation functions, can be changed on the basis of the problem the model is constructed for [54–56]. For example, the sigmoid function, much like when used in logistic regression, can be used in binary classification problems, where the output of the model is being used to classify two separate endpoints. The non-linearity of these activation functions' output leads to the model being able to identify patterns between the input variables and final output that are similarly non-linear.

In contrast to most activation functions, which are non-linear in nature (e.g., sigmoid, rectified linear unit (ReLU), etc.) the "identify function" or linear activation function, can be used to simply output the weighted sum of the inputs without transformation, leaving the final output of the neuron directly proportional to the weighted inputs. This function is useful in applications such as linear regression, but falters in applications where the relationship between the data and output is non-linear.

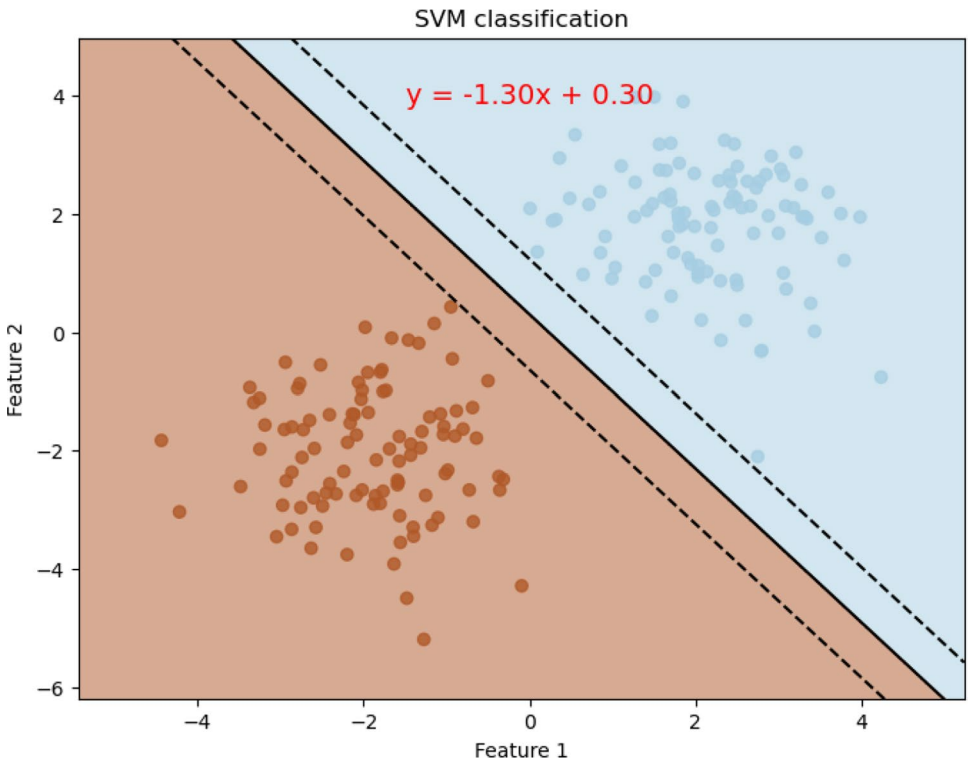

Fig. 4 Scatter plot of a two-cluster dataset (brown/blue points) and the decision boundary obtained using a support vector machine (SVM) classifier (black line) with a margin (dashed lines) using the scikit-learn library. The dataset consists of two clusters of points, where each cluster corresponds to a different class. The SVM classifier was trained on the dataset to classify new datapoints into one of the two classes based on their features. The plot shows that the SVM decision boundary is a linear boundary that maximizes the margin between the two classes. The dashed lines represent the margin, which is the perpendicular distance between the decision boundary and the closest points from each class. The dividing line equation for the SVM model is represented in red

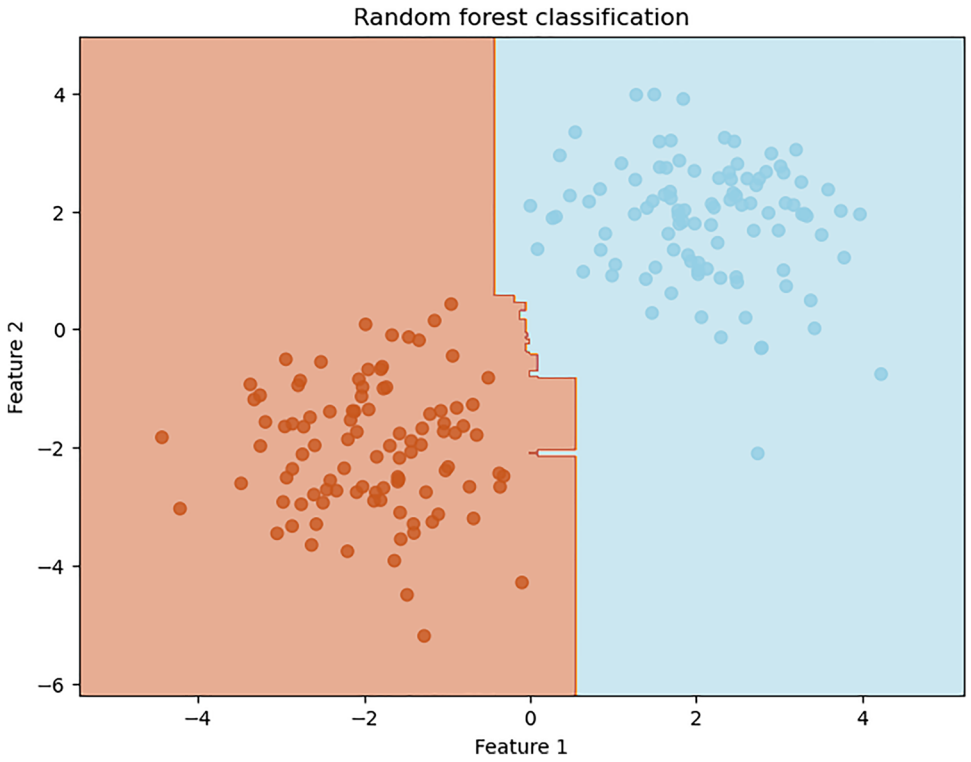

Fig. 5 Visualization of the decisions made by a random forest classifier on a two-cluster dataset. The plot shows the binary predictions made by each decision tree in the random forest for each datapoint in the dataset. The dataset consists of two clusters of points, where each cluster corresponds to a different class. The random forest classifier was trained on the dataset to classify new data points into one of the two classes based on their features. The blue/brown overlay on the graph represents the prediction made by the model at the given (*X*, *Y*) coordinates. The boundary line between the brown and blue positions shows the ability for random forest classifiers to give boundaries to the clusters that are not linear in nature, unlike the logistic regression model shown above

The ability of these models to be used in both regression and classification tasks, as well as their ability to interpret complicated, non-linear relationships between the input variables and output endpoints, have made them popular for tasks that require many variables, including disease diagnosis and financial modeling (Fig. 7). However, the inherent complexity of these models leaves them in a "black box" space [57–59], where even those who constructed the model have limited understanding of why the model outputs a certain response. Additionally, neural networks require a large amount of data and computational resources to be effective, which may be cost prohibitive when working with experimental data. Neural networks have seen extensive use in recent years in the field of drug discovery, these models can be applied to a wide variety of endpoints, including activity towards targets, potential toxicity, and potential drug/drug interactions [57, 60–66].

2.3 Representing Molecular Structures for Machine Learning

Applying machine learning to tasks involving molecular structures has been a growing field, even before the coronavirus pandemic. Quantitative structure–activity relationship (QSAR) modeling is a methodology that has sought to predict the properties of chemical compounds based upon their physiochemical features.

Molecular descriptors are numerical or categorical factors that characterize a molecule's physical qualities, such as its size, shape, and electrical properties. These descriptors may be estimated using computational approaches, such as quantum mechanics or molecular mechanics, or from experimental measurements (e.g., melting point, logP, logD, etc.) Simple one-dimensional molecular descriptors include molecular weight and lipophilicity, whereas more complicated three-dimensional descriptors include molecule shape and electrostatic potential [67–69]. Using experimental data and the generated molecular descriptors for each compound, a QSAR model can be generated using the above-described learning models to predict the endpoint of interest. QSAR models have been used extensively to predict endpoints such as liver toxicity, cardiotoxicity, and carcinogenicity [70–76].

In addition to physiochemical properties, molecular fingerprints are a method to represent molecular structures for machine learning. Molecular fingerprints are a way to encode the structure of a molecule as a binary vector or bitstring (Fig. 8), where each bit represents the presence or absence of a particular structural feature or substructure. There are many different types of molecular fingerprints, each with its own set of rules for encoding structural information. Some of the most popular types of fingerprints include ECFP (extended-connectivity fingerprints) [77], MACCS (Molecular ACCess System) [78], and Morgan [77, 79] fingerprints. The most used of these sets is the Morgan fingerprints (MFP), also known as extended-connectivity

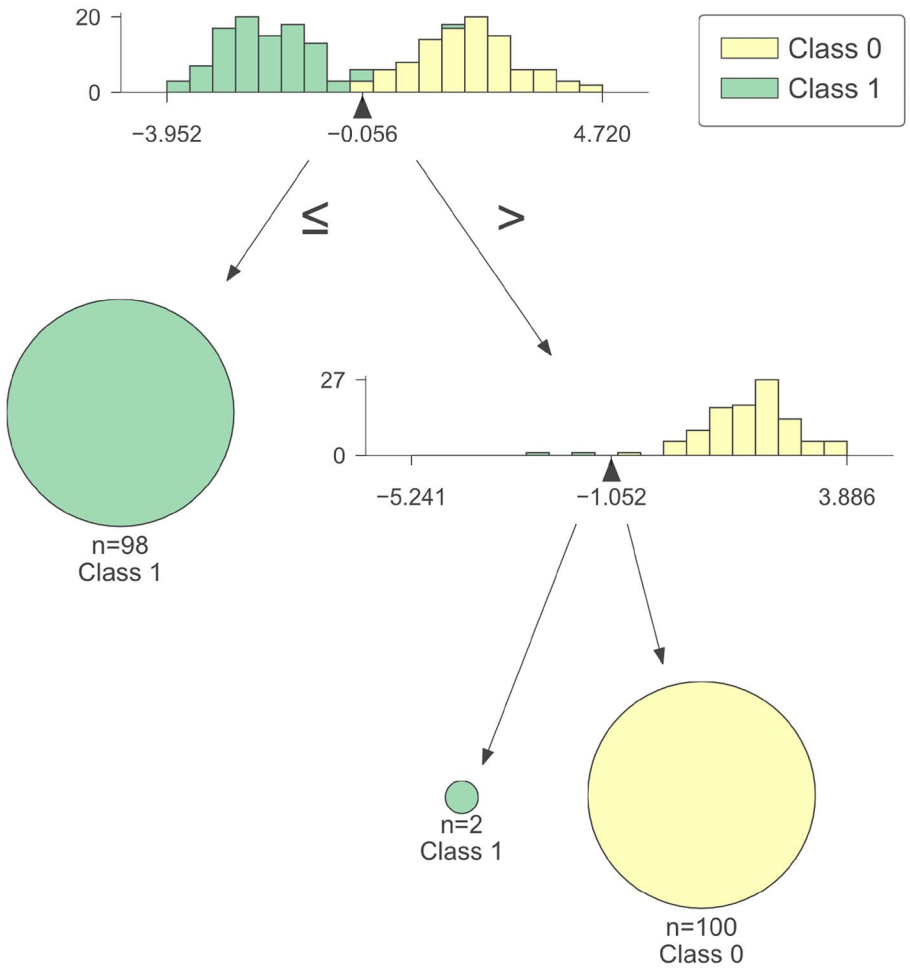

Fig. 6 Example decision tree using the same dataset as in Fig. 5 with the dtreeviz python library. The first decision within the tree represents the value on the *X*-axis, which splits the clusters (class 1 representing the brown points within Fig. 5) into two sets, with set 1 containing 98% of the points within cluster 1. The other set contains 102 points, which are further separated based upon the *X*-axis value again, which further split the points into two class 1 points and 100 points for class 0, represented by the blue points in Fig. 5. This model, along with other decision trees, would be used within the random forest model to make a consensus position for the final prediction of a given point

fingerprints, and encode the local molecular environment around each atom in a molecule as a series of circular substructures or "rings." The substructures within each ring are hashed to produce a unique identifier for each ring, and the resulting set of ring identifiers is concatenated into a binary bitstring, which represents the Morgan fingerprint for the molecule. Multiple sets of Morgan fingerprints are commonly used, with MF1024 (i.e., the resulting bitstring being 1024 bits long) being the most popular [80].

3 Machine Learning Methods for Drug Repurposing

Due to the severity of COVID-19, some of the earliest uses of machine learning focused on the adept power of these models to function as classifiers to identify already approved chemical entities to be repurposed against the virus. Before the beginning of the COVID-19 pandemic, multiple studies had already been performed to train and deploy machine learning models that could accurately repurpose compounds for use in other disease states [81–86].

To identify relevant papers using similar methodologies for the COVID-19 pandemic, we used the following search query:

> ("machine learning" OR "artificial intelligence" OR "deep learning" OR "neural networks") AND ("drug repurposing" OR "drug repositioning") AND ("COVID-19" OR "SARS-CoV-2")

The most relevant papers based on these keywords can be separated into two categories, those models relying on large sets of experimental data and known associations between drugs, proteins, and disease states, and those models that relied on the structural features of compounds. We will refer to these two types of models as knowledge-based and QSAR-based methodologies (Table 1).

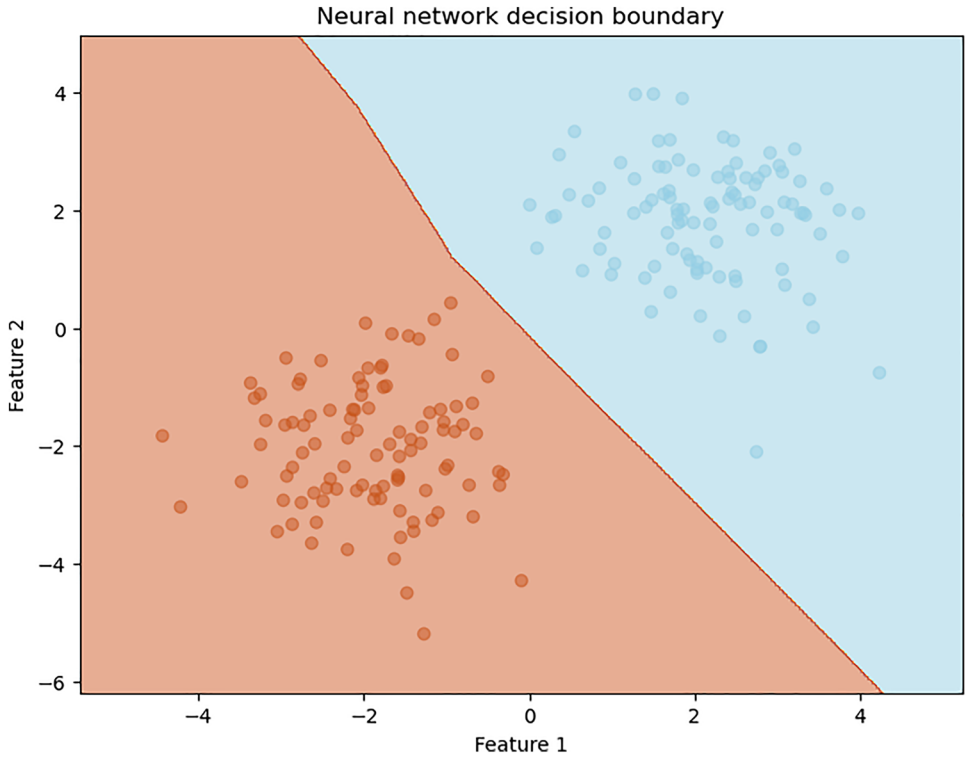

Fig. 7 Visualization of the decisions made by a neural network classifier on a two-cluster dataset using the scikit-learn MLPClassifer. The plot shows the binary predictions made by the neural network for each data point in the dataset. The dataset consists of two clusters of points, where each cluster corresponds to a different class. The neural network classifier was trained on the dataset to classify new data points into one of the two classes based on their features. The blue/brown overlay on the graph represents the prediction made by the neural network model at the given (*X*, *Y*) coordinates

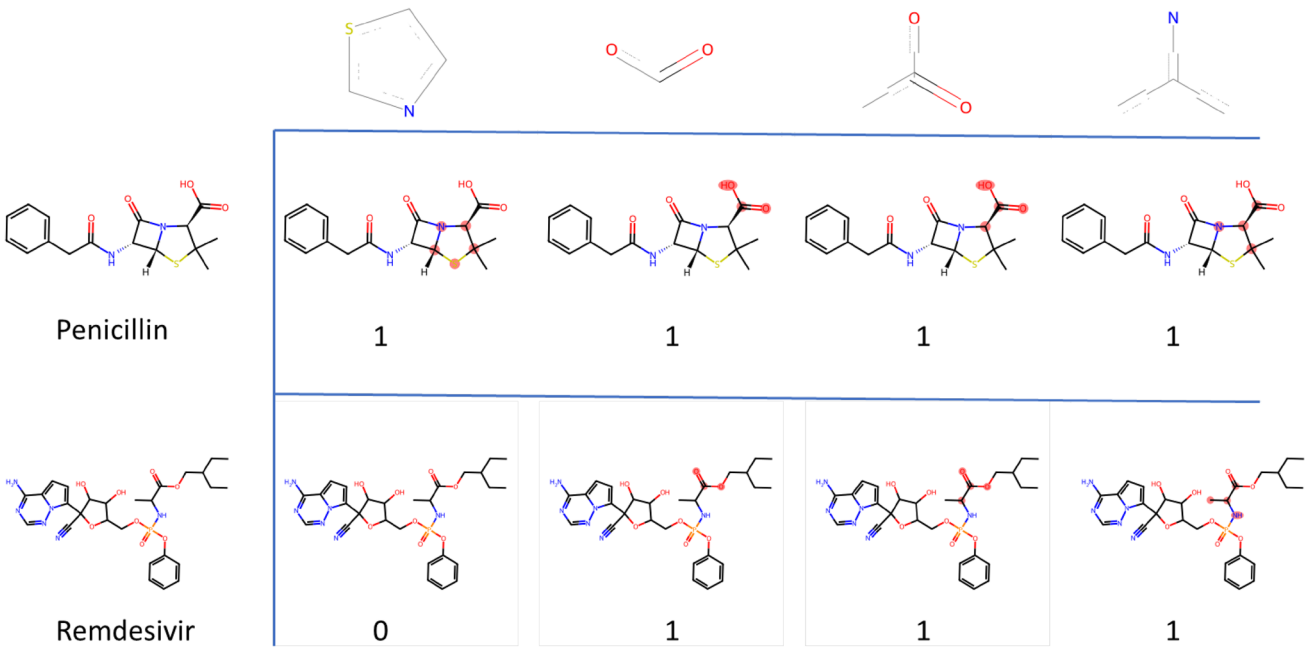

Fig. 8 Simplified representation of two compounds and how they may be represented as a string of bits. For each compound (penicillin and remdesivir in this example) the compounds are described in a bit string by the presence or absence of a set of fingerprints (thiazolidine, acyl, carboxyl, and amino groups, shown as red circles if they are present in this case). These two compounds differ by the presence of a thiazolidine group, which sets the first bit of remdesivir at 0. This process can be expanded to an unlimited number of potential substructures to differentiate the compounds within a given dataset.

3.1 Using Knowledge-Based Models for Drug Repurposing

As the body of research involving the pharmaceutical sciences continues to grow, the internet has allowed for the near-instant access to millions of studies concerning the interactions between drugs, proteins, genes, and disease states. This deluge of large, high-dimensional datasets, brought on by the ever-lowering cost of genetic sequencing, was once seen as a threat to interpretability due to the sheer volume that was being created, the so called "curse of dimensionality" [96–98]. However, deep learning neural networks (DNNs), a subset of machine learning & ANNs that utilize large, multilayer architectures to extract features from knowledge bases (i.e., knowledge-based/deep learning neural networks) that would otherwise be impractical or impossible for humans to identify, have been seen as a method of sifting through this ever-expanding set of data. DNNs have been recently used in several fields concerning the biological sciences, including cancer diagnosis [99–102], radiology [64, 103, 104], and protein three-dimensional (3D) structure prediction [105–112]. One of the seminal studies concerning the use of deep learning architecture for drug property prediction and repurposing came from Aliper et al. [81], which used a set of transcriptomic profiles from the LINCS Project [113, 114] that were obtained from a set of drug perturbations within several cell lines. Combining this transcriptome perturbation data with the MeSH database to classify the therapeutic use of these drugs, the authors sought to develop a model that could predict the therapeutic use of novel compounds using similar transcriptome profiles. Using a neural network consisting of three hidden layers, with 200 nodes within each layer, the DNN model was able to learn the associations between therapeutic uses and the transcriptional profiles that had been provided within the training set.

The availability of SARS-CoV-2-related datasets for use in these machine learning applications has been bolstered by both the decisions of major journals to make research relating to the pandemic open to the wider public and the increasing popularity of prepublished paper archives such as BioRxiv, where datasets related to the pandemic can be released before the editing process for their related manuscripts is complete. Numerous attempts were made to ingest these datasets into DNNs to predict potential drugs that could be repositioned for COVID-19 [21, 22, 65, 87, 93, 115–139]. Zeng et al. [87] utilized this deep learning approach by combining a known set of drug–gene, drug–disease, gene–gene, and gene–disease interactions from the Global Network of Biomedical Relationships (GNBR) [140], drugs from the DrugBank database [141], and finally a set of genes and proteins associated with SARS-CoV-2. This data mining involved 24 million research articles to compose this relational database [140] and created an overall network of over 15 million edges. Using a deep learning model, which was previously developed by the Amazon Web Services (AWS) AI laboratory for use with large knowledge based graphs [142], the authors trained a model that was validated using a known set of COVID-19 active compounds. From this model, the model predicted over 40 compounds that could be effective against COVID-19 including tetrandrine, nadide, estradiol, and rifampicin, none of which have been subsequently approved for COVID-19 treatment. Notably absent from the predictions are drugs such as remdesivir and tocilizumab, drugs that obtained emergency use authorization for use in patients with COVID-19. Additionally, drugs that were known to not be effective clinically (e.g.,

Table 1 List of curated studies using the above search query

Author	Paper	Model Type	Citation
Zeng et al.	Repurpose open data to discover therapeutics for COVID-19 using deep learning	Knowledge-based	[87]
Gysi et al.	Network medicine framework for identifying drug-repurposing opportunities for COVID-19	Knowledge-based	[88]
Santos et al.	Machine learning and network medicine approaches for drug repositioning for COVID-19	Knowledge-based	[89]
Ge et al.	An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19	Knowledge-based	[90]
Pham et al.	A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing	Knowledge-based	[21]
Smith et al.	Expert-augmented computational drug repurposing identified baricitinib as a treatment for COVID-19	Knowledge-based	[91]
Kumar et al.	Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2	QSAR-based	[92]
Beck et al.	Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model	QSAR-based	[93]
Gawriljuk et al.	Machine learning models identify inhibitors of SARS-CoV-2	QSAR-based	[94]
Kadioglu et al.	Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning	QSAR-based	[95]

hydroxychloroquine, chloroquine, ivermectin) were recommended as potential antiviral agents against COVID-19 [143–145]. The authors point out this flaw, recommending that further development of the model attempt to filter out drugs that may perform well in vitro but do not validate within the clinic.

Gysi et al. [88] took a similar approach by using a large dataset containing human protein–protein interactions along with SARS-CoV-2/human proteins interactions, as well as drug–target interactions from the DrugBank database [141]. Using this model, the authors predicted the efficacy of over 6000 compounds within DrugBank; concurrently the authors also assessed the efficacy of 918 compounds within VeroE6 cells from the African green monkey. Of these tested compounds, only 37 compounds had what the authors classified as a "strong effect" (i.e., viral reduction > 80% within VeroE6 cells) on the overall infection rate. Of the authors' models, model P1, which included both direct drug–protein interactions, as well as metabolic drug–protein interactions within the knowledge graph, showed the greatest predictive ability in selecting compounds from DrugBank validation set, which showed the so-called strong effect against COVID-19 infection. Notably, two drugs that have obtained emergency use authorizations, namely ritonavir and dexamethasone, both appear in the model's top ten predictions, with another commonly used anti-COVID-19 agent fluconazole being within the top 50. Similar interactome methodologies, such as those taken on by Santos et al. [89], identified favipiravir as a top candidate against SARS-CoV-2, a drug which has had approval for use against COVID-19 in Italy, Russia, and India [146].

Pham et al. [21] expanded on these methodologies, using the L1000 database of gene expression [114], the STRING database of protein–protein interactions, the DrugBank database for drug–target interactions, and the gene expression profile of 8 SARS-CoV-2 infected patients in comparison with 12 healthy patients. Using the molecular fingerprints of the compounds within DrugBank that target proteins that are within both the STRING and L1000 databases, the authors sought to train a model that could identify how certain compounds would impact the expression of certain genes based upon the proteins that those drugs interact with (i.e., the phenotypic response to the drug dose). By including the gene expression data of both healthy and COVID-19 presenting patients, compounds could be screened for their effectiveness by comparing the impact to gene expression each compound induces to the genes that are most associated with infection. Using multiple different machine learning methodologies (e.g., KNN, neural networks, linear regression, etc.) and different sets of information (i.e., chemical descriptors, drug–target interactions, drug–gene interactions, etc.), the authors finally found that their graph-based neural network performed the best when predicting the gene expression values for the training set of compounds tested. Using this model on the DrugBank database of compounds, the authors identified several macrocyclic, antifungal, and antiviral drugs (e.g., faldaprevir, alisporivir, and anidulafungin) predicted to have positive effects against COVID-19. Of these compounds predicted, alisporivir has received the most attention, and has been the subject of clinical trials [147–149].

Finally, efforts by BenevolentAI, a London-based biotech company, utilized a similar knowledge-based approach, utilizing a knowledge graph containing nearly 30 million PubMed papers and numerous structured databases containing the relationships between drugs, drug–targets, genes, disease states, and the biological mechanisms underlying those disease states. Smith et al. focused on identifying potential compounds that could counteract the cytokine storm induced by the SARS-CoV-2 infection, as well as the replication of the virus via the clathrin mediated endocytosis (CME) pathway, with specific focus placed on the protein AAK1 due to its association with both endocytosis and membrane trafficking. By focusing on these SARS-CoV-2-related pathways, Smith et al.'s knowledge graph identified several US Food and Drug Administration (FDA) approved compounds that were predicted to inhibit the endocytosis pathway, including sunitinib, baricitinib, and fedratinib, all of which possessed $pK_d > 7$ binding affinity for the AAK1 protein. After publication of this finding, trials for baricitinib were conducted to determine the effectiveness of this compound on COVID-19 infections, and it was found that baricitinib in combination with remdesivir improved inpatient outcomes and lowered overall mortality [150]. Barcitinib would later obtain an emergency use authorization from the FDA for use against COVID-19 [151].

3.2 QSAR-Based Methodologies for Drug Repurposing

Using chemical features to predict the properties of compounds has been an established methodology for over 30 years [152–159]. With the introduction of different fingerprint classification systems as described above (e.g., MACCS, daylight, SMILES, Morgan, etc.) [160–162], molecules can be represented in such a way as to be understandable and enterable into machine learning models. Using these fingerprints, machine learning models can identify patterns in the active and inactive compound sets and learn to discriminate between the sets. These trained models can then be used with new unseen compound fingerprints to generate a prediction on whether they are active or inactive.

Kumar et al. [92], developed a machine learning model that was able to discriminate between antiviral compounds and those that would be inactive against COVID-19. Using a dataset of known compounds (DrugRepV) that were active in inhibiting coronaviruses, over 1100 in total, the authors

generated over 17,000 chemical and structural descriptors, ranging from 1D to 3D descriptors, using the open-source PaDel descriptor software [163]. These molecular features, along with the known pIC_{50} values against each of their respective viral targets, were used to train several models, making sure to remove redundant features from the models to prevent overtraining. Several machine learning models, including SVM and random forest models, were trained on this experimental and descriptor dataset, using ten-fold cross validation to prevent overtraining on any one chemotype. The authors were able to achieve models that had good predictive pIC_{50} correlation with the experimental results with their SVM model performing the best with R^2 values ranging from 0.53 in all coronaviruses to 0.81 when filtered down to the drugs that showed inhibition against SARS-CoV-2. Applying this model to a database of approved drugs, the authors identified verteporfin, alatrofloxacin, metergoline, rescinnamine, leuprolide, and telotristat ethyl as potential candidates for inhibitors of the SARS-CoV-2, going further to incorporate in silico docking methodologies to show their potential as spike inhibitors; however, they were unable to perform any in vitro or in vivo to test their repurposing hypothesis (Fig. 9).

Gawriljuk et al. [94] took a similar approach, utilizing the extended connectivity fingerprint (ECFP6), a set of molecular descriptors developed by ChemAxon, which represents a compound as a fixed length binary representation, which can then be subsequently used as the input features for numerous machine learning models [164, 165]. The authors then collected available in vitro inhibition data from several drug repurposing studies for use in COVID-19, collecting over 60 compounds in total. These activities values, along with the set of molecular descriptors, were fed into numerous machine learning regression models, including random forest, support vector machines, decision trees, and a deep learning neural network. Out of these models, the Bayesian method created by the external Assay Central [166, 167] software performed the best at predicting the inhibition within the training set of compounds. Using this trained model, the authors then proceeded to predict the inhibition of a subset of FDA-approved compounds that were available to them; these top scoring compounds were

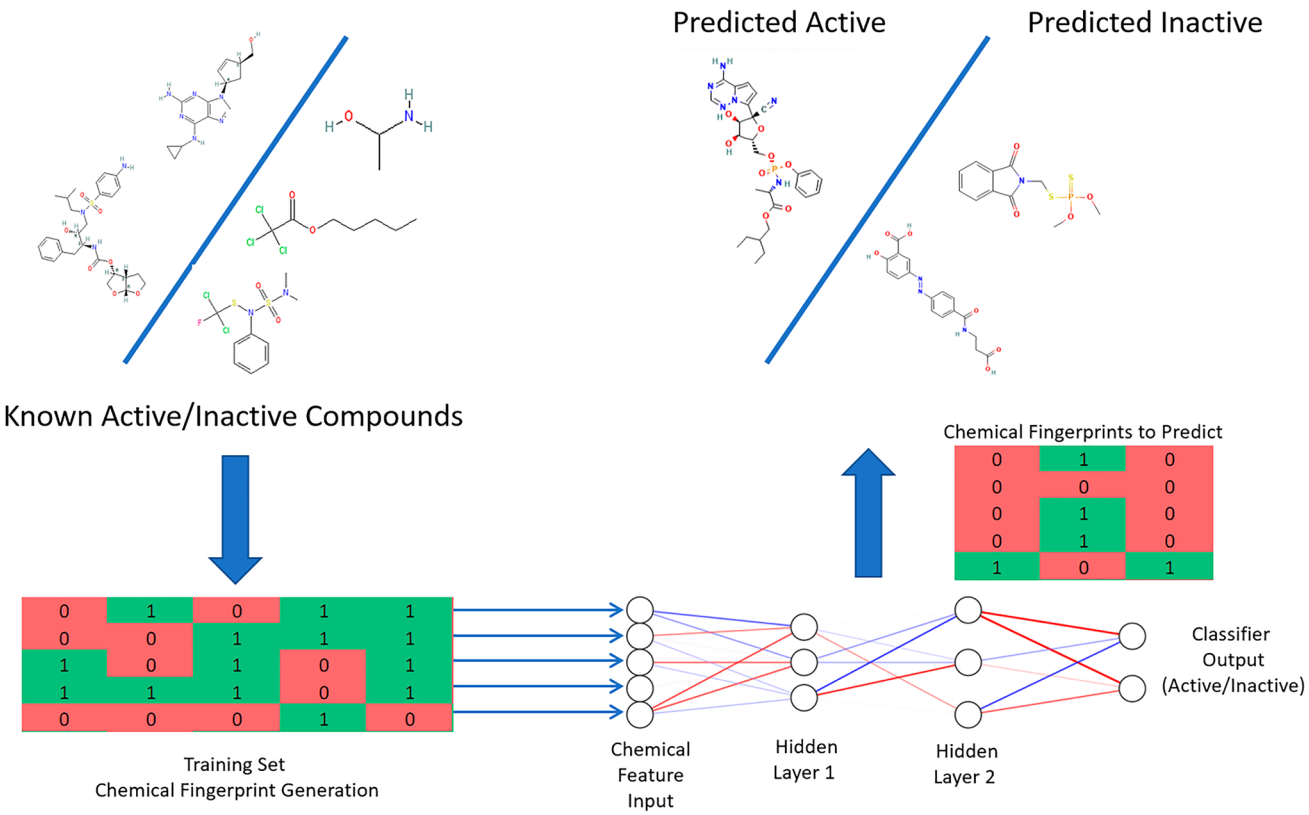

Fig. 9 A generalized machine learning neural network for the classification of molecules as active or inactive, in this case as antiviral compounds. To create such a classifier, known active and inactive compounds for the desired class are needed as a training/validation set for the model to assess its predictions selected. Chemical fingerprints (e.g., MACCS, Morgan Fingerprints etc.) for these compounds are generated and used as the features the neural network trains upon to learn the patterns that identify a compound as active or inactive. Once the model is trained, new compounds may have these same features generated and passed into the neural network to have their activity predicted. This allows for quick filtering of compounds and prioritization in testing their activities

then used within HeLa–ACE2 cells, where the compounds showed high inhibitory potential, with IC_{50} values in the sub millimolar range (540 nM–8.4 μM). Strangely, the authors of this study have chosen to obfuscate the names of these top performing compounds, and to date no follow-up studies using the chosen monikers (CPI1062 and CPI1155) have been performed. Using the ECFP6 fingerprints of the active and inactive compounds, the authors also provided the most populous fragments from each set. While the authors were able to develop a model that could accurately predict the binding affinity based on the curated database of in vitro data, the dataset used is rather limited with only 60 compounds, which contains compounds such as hydroxychloroquine and chloroquine, which are now known to not provide any benefit within clinical trials [145, 168]. The authors also show that the most populous fragment from the ECFP6 fingerprint set is a tertiary amine, which appears in both chloroquine and hydroxychloroquine, showing that the presence of these compounds is indeed biasing the dataset. With large datasets of known COVID-19 inhibitors now published [24, 169], this method should be reapplied to see if the observed accuracies still hold; additionally, increasing the overall chemical space within the training dataset would make this methodology more generalizable.

Kadioglu et al. [95] utilized a combination of clinical data, molecular docking, and machine learning to identify potential drugs for COVID-19 from the FDA-approved drugs and a set of zinc natural compounds. Using AutoDock Vina, the authors used these two libraries to dock each compound against the SARS-CoV-2 spike protein and nucleocapsid protein, and a methyltransferase to rank order compounds for the next step in their in silico pipeline. Using clinical data, the authors additionally trained a supervised learning-based classifier model on the basis of a set of known active and inactive SARS-CoV-2 compounds to further validate the screened compounds that showed high affinity against targets during the docking study against the COVID-19 targets. Rather than using molecular descriptors such as ECFP6 or MACCS, the authors decided to use simpler property values of the compounds (i.e., hydrogen bond donors, rotatable bonds, PSA, total surface area, etc.) After the neural network model was trained, and was externally validated, showing a 1.0 precision, the docked compounds were then screened using this same model. The top scoring compound simeprevir has been externally validated by other labs to inhibit COVID-19 in vitro [170, 171]; however, the authors' chosen target of the spike protein for this compound does not match with the rest of the literature, which focuses on simeprevir inhibiting the main protease of SARS-CoV-2 [170, 172–174].

Beck et al. took a natural language processing approach to determining molecular features, rather than using the more established fingerprinting methods. Previously published methodology [175] was used that can represent linear SMILES strings of molecular compounds [176] and the FASTA [177] sequence of protein targets as matrices capable of being used as the input for a neural network, along with a curated database of drug/protein interactions and binding affinity data from BindingDB [178, 179]. The authors trained a model that could accurately predict the binding affinity of compounds, with a Pearson correlation score of 0.9. Using this trained model, they then proceeded to predict the binding affinity of each FDA-approved drug against a set of target viral proteins, focusing on those compounds that were predicted to have a binding affinity lower than 1000 nM. These compounds were subsequently docked into their respective protein targets using AutoDock Vina [180] to determine their validity. For both the antiviral subset and whole subset of the library of FDA-approved drugs, the predicted binding affinity of the compounds were correlated with the binding affinity predicted by AutoDock Vina, potentially making this methodology a suitable replacement for molecular docking when working with large datasets of compounds. Notable top predictions from this study include ritonavir, which was initially seen as a potential therapeutic until further clinical trials showed that the compound had negligible effect on hospitalization [181]. Remdesivir was ranked in the top 5% of compounds when tested, however, suggesting that this prediction model should be used not as an absolute source of binding affinity, but as a method of prioritization for large sets of predictive compounds.

3.3 Potential Improvements in Drug Repurposing Methodologies

While many of these networks were able to make predictions of compounds that would later be validated, many predicted compounds would later go on to be removed from emergency use authorization once further clinical trials were conducted (e.g., hydroxychloroquine, ivermectin, chloroquine) (Table 2). The appearance of these known inactive compounds suggests that these machine learning methodologies for drug repurposing, while novel, suffer from the same issues that other in silico and in vitro methodologies have when attempting to translate into in vivo and clinical results within patients.

Notably, when using patient endpoint data, such as within Pham et al.'s study utilizing the gene expression data of infected and healthy patients, these failed compounds did not appear within the top results reported by the authors. Using such clinical endpoint data may increase the predictive power over these other models and should be included in combination with the in vitro data that many of these models utilize. Many of these studies were performed within the first two years of the pandemic and can benefit from the updates in knowledge that have been subsequently

Table 2 Papers implementing knowledge-based graphs to predict active compounds against COVID-19 and which predicted known inactive compounds

COVID-19 ineffective drug	Papers predicting
Ivermectin	Zeng et al. [87], Gysi et al. [88]
Chloroquine	Zeng et al. [87], Gysi et al. [88], Santos et al. [89], Ge et al. [90]
Hydroxychloroquine	Zeng et al. [87], Gysi et al. [88]

generated. Utilization of these results may elucidate new chemical space for antiviral compounds for COVID-19, which may have been previously overlooked.

4 Machine Learning Methodologies for De Novo Drug Design

While drug repurposing for COVID-19 has been extensively pursued due to the potentially lower hurdles for the discovered candidates to progress through clinical trials, de novo drug discovery still has a role to play in helping to manage future pandemics. As seen over the past three years, the SARS-CoV-2 virus is not merely one strain but has consistently mutated over the course of the pandemic. By the time the first wave of SARS-CoV-2 infections hit the USA, the D614G variant of the initial Wuhan strain had already become dominant and subsequent strains (e.g., B.1.617.2 or Delta, and B.1.1.529 or Omicron) have caused additional waves [182–188]. While initial efforts towards developing vaccines and monoclonal antibody treatments (mAbs) were successful, the mutations contained within these variants have been effective in providing resistance to both types of treatments [189–201]. In addition to the threats caused by the ongoing mutation of this virus, these repurposed drugs are not optimized to inhibit the proteins used by these viruses. Remdesivir, one of the standout antivirals that was given emergency use authorization against SARS-CoV-2, was originally targeted against the Hepacivirus C RNA-dependent RNA polymerase (RdRp), and exhibits an average EC_{50} 700 nM versus SARS-CoV-2 infection; in contrast, remdesivir has an EC_{50} of 3 to 90 nM against the Ebola virus, and 70 nM within the original SARS-CoV-1 [202, 203]. This loss in activity can potentially be attributed to the differences in the structure of RdRp between SARS-CoV-1 and SARS-CoV-2, which have 80% similarity [202, 203]; however, additional in silico and in vitro studies need to be performed to identify the exact residue change that confers this selectivity to the SARS-CoV-1 RdRp. De novo drug discovery offers the opportunity to develop novel treatments that specifically target the SARS-CoV-2 virus's protein structures, thereby providing a more effective and tailored approach to treating COVID-19 versus utilizing the limited chemical space explored by the currently existing set of FDA approved compounds. For this section, we will focus on the efforts made in de novo drug design for both small molecules (e.g., remdesivir) and biological (e.g., tixagevimab) entities against SARS-CoV-2, which utilize machine learning to accelerate their efforts.

To identify studies that fit this description, the following search query was utilized:

> [("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network" OR "random forest" OR "support vector machine" OR "convolutional neural network" OR "generative adversarial network" OR "autoencoder") AND ("COVID-19" OR "SARS-CoV-2" OR "coronavirus") AND ("de novo compound" OR "novel compound" OR "drug discovery" OR "virtual screening")]

Through this search, the following studies were identified (Table 3).

4.1 Small-Molecule Drug Design

4.1.1 Utilization of Machine Learning to Accelerate Virtual Screening (VS)

Traditional virtual screening methods are time consuming and computationally demanding, limiting their efficiency in screening vast chemical databases, such as the ZINC or Enamine libraries, which can potentially contain billions of compounds. By utilizing known active and inactive compounds against a given target, a machine learning model can be trained and subsequently used to filter out compounds within these large sets of compounds, saving on both computational time and eventual in vitro testing resources.

Kumari and Subbarao [204] took a similar approach to the described drug repurposing studies. Using a set of known compounds that have assay response data against the 3-chymotrypsin like protease (3CLPro, also known as the main protease [211]) of SARS-CoV-2, the authors proceeded to train a machine learning model that could classify compounds as being active or inactive against 3CLPro based on a set of over 100 calculated two-dimensional descriptors. The author's best model, a convolutional neural network, showed an accuracy of over 85% based on the designated set of test compounds. This model was then applied to several compound sets, including natural compounds from the ZINC database, the NCI IV divest, which contains primarily natural products, along with the FDA-approved compound set. After applying a Lipinski's rule of five filter, the authors identified nine flavonoid compounds from the phytochemical dataset, but did not go further and test these compounds in vitro.

Table 3 List of curated studies using the above search query

Author	Paper	Method	Citation
Kumari et al.	Deep learning model for virtual screening of novel 3C-like protease enzyme inhibitors against SARS coronavirus diseases	Machine learning assisted virtual screening	[204]
Srinivasan et al.	Artificial intelligence-guided de novo molecular design targeting COVID-19	Machine learning assisted virtual screening	[205]
Bung et al.	De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence	Generational network/virtual screening	[206]
Arshia et al.	De novo design of novel protease inhibitor candidates in the treatment of SARS-CoV-2 using deep learning, docking, and molecular dynamic simulations	Generational network/virtual screening	[207]
Magar et al.	Potential neutralizing antibodies discovered for novel coronavirus using machine learning	Machine learning assisted antibody screening	[208]
Williams et al.	Fast prediction of binding affinities of SARS-CoV-2 spike protein and its mutants with antibodies through intermolecular interaction modeling-based machine learning	Machine learning assisted antibody screening	[209]
Xu et al.	Discovery of potential flavonoid inhibitors against COVID-19 3CL proteinase based on virtual screening strategy	Machine learning assisted virtual screening	[210]

Flavonoid compounds were also identified by Xu et al. [210] as potential inhibitors through a combination of machine learning and molecular docking. Using a set of 66 active and 66 inactive compounds against the SARS-Cov-1 3CLPro protein, the authors trained a classifier model using several machine learning methodologies, with their logistic regression model performing the best. After training this model, 2030 known natural compounds were passed through the model to determine if they were active or inactive against 3CLPro. These results were used in combination with the publicly available crystal structure of 3CLPro from SARS-CoV-1, which shares nearly perfect identity with the SARS-CoV-2 3CLPro. Using a combination of molecular docking and molecular mechanics Poisson–Boltzmann surface area via the Schrödinger Maestro suite, the authors identified six flavonoid compounds with rutin scoring the highest among the other potential inhibitors.

While flavonoid-based compounds have been tested for their efficacy against COVID-19 [212], with some showing nanomolar efficacy in terms of inhibition, most of the in vitro evidence for these compounds inhibitory activity comes from kinase targets, rather than this protease target. With no in vitro testing performed on the compounds from Kumasi et al., it is impossible to validate the efficacy of these compounds as SARS-CoV-2 inhibitors. However, rutin, the compound identified as the top scorer in the study by Xu et al. [210], has had its in vitro activity determined, and was found to be inactive against SARS-CoV-2's PLpro protein [213].

Srinivasan et al. implemented a Monte Carlo tree search machine learning model to reduce the amount of docking required to screen compounds against COVID-19 [205]. This model was created by using molecular features (i.e., building blocks) to identify moieties that confer strong binding energy as predicted by AutoDock Vina [180]. The authors were successful in creating a model that could predict the AutoDock Vina [180] binding scores of compounds binding against the spike protein/ACE2 complex. This model can drastically reduce the computational cost for each tested compound, as each compound can have its binding affinity estimated through its structure alone, rather than having to use computationally expensive methodologies such as docking. The authors then used this model to score over 97,000 molecules against the spike/ACE2 complex.

4.1.2 Using Machine Learning to Generate New Compounds for COVID-19

Generative models are a type of machine learning algorithm that can be used to create new molecules within a particular chemical space. These models are trained on existing datasets of molecules, such as those from the PubChem database, to learn the underlying patterns and rules governing the molecular structure of the compounds. Once trained, the generative model can be used to design new molecular entities that have not yet been explored. These molecules can be generated with specific properties, such as a high binding affinity to a target protein or a low toxicity profile. This is done by using the learned patterns and rules to iteratively generate new molecules and assess their properties using computational models or in vitro and in vivo assays [214–221].

Bung et al. [206] utilized machine learning to generate new sets of compounds using a generative model trained on drug-like SMILES strings. Generative models are a type of machine learning model that learns to generate new data that is similar to a given set of training data [222–224]. They are used to model the underlying distribution of the training

data and can be used to generate new samples that follow the same distribution. Using a set of 1.6 million drug-like small molecules, the authors trained a new generative model to generate new SMILES strings based upon the given set. The model was trained in combination with RDKit, which was used to determine whether the SMILES strings generated were chemically feasible (e.g., correct number of bonds, valid atom types, SMILES grammar, etc.). After training, 97% of SMILES strings generated by the generative model were valid molecules. After learning the structure of valid chemical entities along with the grammar of SMILES, this model was subsequently trained on the subset of compounds that could inhibit 3CLPro, along with a set of known antiviral compounds (7665 compounds in total). By identifying the underlying features within this set of antiviral compounds, the model then proceeded to generate over 40,000 compounds that contained similar features to the initial dataset; these compounds were then docked into the structure of 3CLPro to determine their viability against the protein structure. Over 1200 of these compounds were found to have a VINA docking score of less than − 7.0 kcal/mol. Interestingly, the highest scored compounds from this generative model had high Tanimoto similarity to known human immunodeficiency virus (HIV) protease inhibitors, such as darunavir, indinavir, and saqunavir, showing that the model was able to learn the features of the provided antiviral compounds. Other studies of anti-HIV antivirals have shown limited efficacy against COVID-19 with IC_{50} values in the micromolar range. Indinavir specifically was shown to have an $IC_{50} > 200$ μM to inhibit the main protease of SARS-CoV-2. Additional in vitro validation will be required for these compounds to determine their efficacy against SARS-CoV-2.

Similarly, Arshia et al. [207] utilized a generative network based on the ChEMBL[225] and ZINC [226, 227] SMILES strings to generate new potential inhibitors of the main protease of SARS-CoV-2. The long short-term memory [228] (LSTM) chem network is a generative recurrent neural network, which has learned to generate de novo designs of compounds based on these training sets [25, 225–227]. After compound generation, these compounds are then prepared and subsequently docked into the crystal structure of the main protease using AutoDock Vina [180]. The docking scores of these compounds are then used to fine tune the neural network for future generations, each subsequent generation being more finely tuned for the docking receptor. The top generated compounds were then simulated in complex with Mpro utilizing GROMACS. Compared with the crystal structures of remdesivir and N3 with the Mpro protein, the docking score of the generated compounds showed a notable improvement over the known inhibitors. However, the authors provide no additional studies concerning in vitro efficacy of these compounds, leaving their actual potency unknown. Additionally, the generative network appears to have no weighting concerning typical ADMET properties (i.e., LogP, molecular weight, etc.) as the produced compounds are extremely large, lipophilic, and contain numerous aromatic rings. These molecules easily break Lipinski's rule of five for drug-like compounds, raising concerns about their overall applicability in vitro and in vivo (Fig. 10).

4.2 Antibody Design

Efforts to implement machine learning to accelerate de novo drug discovery are not limited to small molecules, but also extend to the design of biological compounds, such as antibodies. Much like with their small molecule counterparts, antibodies can be described through a set of descriptors (e.g., their primary sequence) that can be used as an input to many machine learning methods. Even before the pandemic began, several groups were studying the use of machine learning to modify existing antibody designs to improve their affinity with their antigens. Using a set of training data consisting of Fab fragment binding affinities with a chosen antigen, Liu et al. [231] showed that even with only the one dimensional primary sequence data of the antibody, a trained neural network would be able to suggest new sequences that encode antibodies, with increased EC_{50}

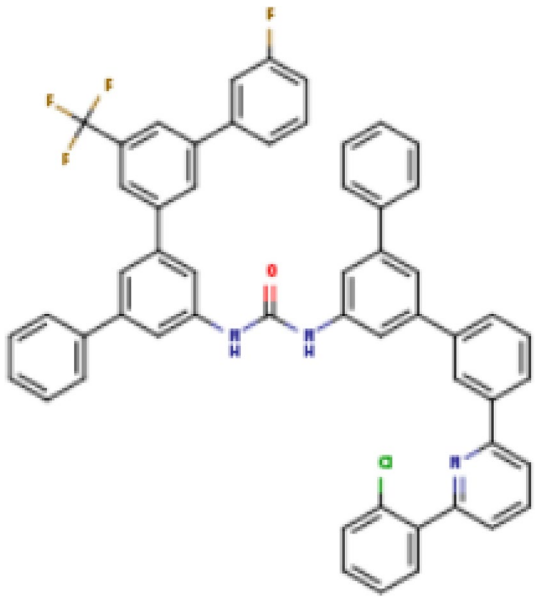

Fig. 10 Ligand A from Arshia et al.'s [207] generative model submitted to the SWISSADME [229] service. Ligand A, while showing enhanced binding affinity to the main protease, exhibits undesirable qualities, such as a large number of rotatable bonds (12), high lipophilicity (6.97 logP), and large size [866 molecular weight (MW)]. This compound breaks two of Lipinski's rule of five [230], [MW > 500 and Moriguchi octanol-water partition coefficient (MLOGP) > 4.15]

values compared with the original training set. This method of training neural networks to design new antibodies based on the one-dimensional primary sequences and experimental binding affinity data was further expanded on by Akbar et al. [232], who designed a similar deep learning neural network trained on 11,000 HER2-binder and non-binder CDR-H3 sequences. However, this methodology did not go into the in vitro stage to evaluate these antibodies, thus relying on in silico methodologies to determine the binding affinity between antibody and antigen.

Magar et al. [208] utilized a similar approach to designing, utilizing the known IC_{50} values of numerous antibodies with their respective viruses, including human immunodeficiency virus (HIV), influenza, Dengue, etc. as training data for their machine learning model [208]. For each of these pairs of antibody/antigen structures, the known interactions for each were extracted and used to train a classifying machine learning model to determine whether the antibody will be able to recognize and neutralize the antigen. Unlike the previous examples of designing antibodies with machine learning, which used the primary sequence of the antibody as the input layer for the ML model, the authors used a graph model that treated each atom within the antibody and antigen as its own input node; each node including specific features concerning the atom (e.g., aromaticity, atom type, residue type, etc.). With these two sets of input nodes concatenated into one total input layer, several machine learning algorithms were implemented to determine which could most accurately predict whether the binding ability of the test set of antibodies could be found. The authors' XG-Boost algorithm, an implementation of gradient boosting that combines several decision trees to create an overall strong predictive model, showed the best accuracy concerning the training data, with exceptional performance displayed in predicting the binding efficacy of antigens concerning SARS-CoV-1. Looking at feature importance, the authors noted that the inclusion of a methionine residue upon the antibody surface was a crucial feature to increase antibody/antigen recognition, noting the importance of surface sulfur atoms in protein–protein interactions [233]. Using this trained model, Magar et al. proceeded to screen thousands of hypothetical antibody candidates for the SARS-CoV-2 virus using molecular dynamics simulations of the designed antibodies to determine their overall stability [208].

Free energy calculations can be a powerful tool to determine the relative binding affinities of both ligands and antibodies to their protein targets [234–238]. Using free energy calculations, such as molecular mechanics with generalized Born and surface area solvation (MM-GBSA), a reliable methodology was developed early in the pandemic that allowed for the quick estimation of the experimental binding affinity (K_d) of the spike protein to the ACE2 receptor [236, 239]. This methodology was able to successfully predict the binding affinity of both the Alpha and Omicron variants of the SARS-CoV-2 spike protein before experimentally validated assays were published [236, 239]. However, these methods are not without their drawbacks; methods such as MM-GBSA are liable to vastly overestimate the binding energy between two entities [240–246], requiring methods such as a linear regression model to return the predicted values to a reasonable value in line with experimentally determined results [236–239, 247]. This overestimation is especially evident when comparing entities that have differing charge states (e.g., different protonation states of a ligand causing a large change in the predicted ΔG value, while the experimentally determined ΔG value would remain constant between the two states). With these limitations in mind, implementing a free energy-based methodology to predict the binding affinity of a wide range of known SARS-CoV-2 targeting antibodies, each with their own unique charge states, would be a significant challenge. However, groups such as Dong et al., have shown that the introduction of additional energy and interaction features (e.g., hydrogen bonding interactions, rotatable bonds, ligand charges, element count, etc.) via a random forest machine learning model can vastly improve the experimental/predicted binding affinity correlation than with just free energy methods alone [248].

Williams et al., utilized this combination of machine learning and free energy methods to predict the experimentally determined binding affinity of numerous antibodies with both the wild type spike protein and its known Kappa (B.1.617.1) and Delta (B.1.617.2) variants [235]. Along with the decomposed energy terms of MM-GBSA (i.e., Van der Waals (vdW), electrostatic (EEL), solvation), several additional intermolecular interaction terms were added to characterize the complex between each antibody and the SARS-CoV-2 spike protein (e.g., hydrogen bonds, charges of the spike and antibody within the crystal structure, surface area of the spike, and antibody proteins, etc.,) Using this set of features, the available spike/antibody binding affinity data was split into three separate K-folds to allow for each data point to be used at least once within the validation set to avoid overtraining on any of the few data points that were available at that time. Williams et al., proceeded to train a two hidden layer perceptron neural network model, which accepted 11 different features of the antibody/spike complex as its input, these input nodes then forward fed into the two hidden layers, which would finally feed into a single output node, providing the binding free energy (ΔG) prediction of the antibody/spike complex in question. The model's three K-folds were able to successfully predict the binding affinity with high accuracy to the known experimental values, with an RMSE of 0.4 kcal/mol averaged across the three folds, improving on the overestimation of binding affinity that would have resulted with MM-GBSA alone. With this trained model, Williams et al. proceeded to predict the

binding affinity of over 20 antibodies with 11 variants of the SARS-CoV-2 virus. This model successfully predicted that several antibodies would have diminished binding affinity with the B.1.617.2 (Delta) variant spike protein, including Ly-CoV555, without any previous training data concerning the Omicron variant [193]. The authors propose that this methodology could be used to scan hundreds of potential SARS-CoV-2 variants, to determine which mutation or combination of mutations could potentially threaten antibody efficacy as the pandemic continues.

5 Discussion

5.1 Successes and Limitations

Machine learning methods have rapidly matured since the SARS-CoV-2 virus first appeared. Multiple studies for drug repurposing and de novo design of new compounds for COVID-19 have been able to identify both new and existing compounds for use in inhibiting infection by the virus.

While these models do show great promise, it is essential to note the limitations inherent to these machine learning models. Unlike simpler methods, such as multiparameter optimization (MPO) and linear regression, machine learning models are intrinsically opaquer in terms of their understandability, due to their more complex structures in comparison (e.g., each parameter within a linear regression will have a coefficient roughly equivalent to their overall importance if the input values are standardized). This is especially true within neural networks that have many input nodes and hidden layers; though the weights between these layers may be known, changing a certain input value while keeping the other variables constant will not necessarily produce a linear correlative output as seen within models such as a linear regression.

Despite the considerable amount of effort and resources invested in training machine learning models, many studies that employ these models fail to validate the generated designs in vitro or in vivo. This is a significant limitation that restricts the overall impact that these models may have within the scientific community. Without proper in vitro validation of these predictions, it is impossible to determine whether the designs generated by the machine learning models are accurate or whether they merely reflect chance correlations. This lack of validation makes it challenging to determine the usefulness of these models in practice and may result in wasted resources and false leads.

This lack of in vitro or clinical validation is especially concerning within drug repurposing, as the recommended compounds are currently on the market and are potentially obtainable by those without medical training. Off label and unauthorized use of ivermectin contained within livestock products and hydroxychloroquine within aquarium supplies led to the poisoning of numerous people who thought that these compounds were viable treatments for COVID-19, despite the lack of clinical trials showing efficacy [249–252]. The validation of these predictions with experimental in vitro and clinical results will not only provide appropriate recommendations to the healthcare community, but will simultaneously generate additional data to use within the next generation of trained models.

5.2 Lessons for the Future

In terms of utilizing these machine learning methods in preparation for the next pandemic, emphasis must be placed on the methodologies that provide information that can be quickly implemented for healthcare decisions. While machine learning methods for de novo drug design are essential for exploring chemical space that would otherwise go ignored, developing new antiviral medications is a long and expensive process that does not fit within the short timeframe that an emerging pandemic provides. Methodologies seeking to identify existing approved pharmaceuticals can provide chemical entities with known safety characteristics and can be prescribed for off-label use until a decision can be made later for use against the emerging pandemic.

In conclusion, the emergence of the COVID-19 pandemic has highlighted the crucial role that machine learning (ML) techniques can play in accelerating the rate of drug discovery. With the ability to analyze large datasets, including experimental and clinical data, ML techniques have proven to be effective in identifying drug candidates, repurposing existing drugs, and designing new compounds for COVID-19. Overall, the utilization of ML techniques in drug discovery has the potential to transform the drug development process and improve public health outcomes.

Declarations

Author contributions Conceptualization, CG Zhan; Data curation and analysis, AH Williams; Writing – original draft preparation, AH Williams; Writing – review and editing, AH Williams and CG Zhan; Funding acquisition, CG Zhan. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported in part by the funding of the Molecular Modeling and Biopharmaceutical Center at the University of Kentucky College of Pharmacy and the National Science Foundation (Directorate for Mathematical and Physical Sciences, NSF Grant DMS-2245903 under funding opportunity NSF 22-600—Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences).

Conflicts of interest Authors declare no conflicting interests concerning the information presented within this review.

Ethics approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication Not Applicable.

Availability of data and material Not Applicable.

Code availability Not Applicable.

References

1. Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. Lancet Infect Dis. 2022;22(9):1293–302.
2. Richards F, Kodjamanova P, Chen X, Li N, Atanasov P, Bennetts L, Patterson BJ, Yektashenas B, Mesa-Frias M, Tronczynski K. Economic burden of COVID-19: a systematic review. Clinicoecon Outcomes Res CEOR. 2022;14:293.
3. Giannis D, Ziogas IA, Gianni P. Coagulation disorders in coronavirus infected patients: COVID-19, SARS-CoV-1, MERS-CoV and lessons from the past. J Clin Virol. 2020;127: 104362.
4. Nicholls J, Dong XP, Jiang G, Peiris M. SARS: clinical virology and pathogenesis. Respirology. 2003;8:S6–8.
5. Troyano-Hernáez P, Reinosa R, Holguín Á. Evolution of SARS-CoV-2 envelope, membrane, nucleocapsid, and spike structural proteins from the beginning of the pandemic to September 2020: a global and regional approach by epidemiological week. Viruses. 2021;13(2):243.
6. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol. 2021;19(3):155–70.
7. Bai C, Zhong Q, Gao GF. Overview of SARS-CoV-2 genome-encoded proteins. Sci China Life Sci. 2022;65(2):280–94.
8. Zhou T, Tsybovsky Y, Gorman J, Rapp M, Cerutti G, Chuang G-Y, Katsamba PS, Sampson JM, Schön A, Bimela J. Cryo-EM structures of SARS-CoV-2 spike without and with ACE2 reveal a pH-dependent switch to mediate endosomal positioning of receptor-binding domains. Cell Host Microbe. 2020;28(6):867.e865-879.e865.
9. Nguyen HL, Lan PD, Thai NQ, Nissley DA, O'Brien EP, Li MS. Does SARS-CoV-2 bind to human ACE2 more strongly than does SARS-CoV? J Phys Chem B. 2020;124(34):7336–47.
10. Domingo P, Mur I, Pomar V, Corominas H, Casademont J, de Benito N. The four horsemen of a viral Apocalypse: the pathogenesis of SARS-CoV-2 infection (COVID-19). EBioMedicine. 2020;58: 102887.
11. Viceconte G, Petrosillo N. COVID-19 R0: magic number or conundrum? Infect Dis Rep. 2020;12(1):8516.
12. Bulut C, Kato Y. Epidemiology of COVID-19. Turk J Med Sci. 2020;50(9):563–70.
13. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J: The reproductive number of COVID-19 is higher compared to SARS coronavirus. J Travel Med. 2020;20(3):318–31.
14. Zheng C, Shao W, Chen X, Zhang B, Wang G, Zhang W. Real-world effectiveness of COVID-19 vaccines: a literature review and meta-analysis. Int J Infect Dis. 2022;114:252–60.
15. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today. 2015;20(3):318–31.
16. Lo Y-C, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discov Today. 2018;23(8):1538–46.
17. Kim E, Choi A-S, Nam H. Drug repositioning of herbal compounds via a machine-learning approach. BMC Bioinform. 2019;20(10):33–43.
18. Vo AH, Van Vleet TR, Gupta RR, Liguori MJ, Rao MS. An overview of machine learning and big data for drug toxicity evaluation. Chem Res Toxicol. 2019;33(1):20–37.
19. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77.
20. Kowalewski J, Ray A. Predicting novel drugs for SARS-CoV-2 using machine learning from a > 10 million chemical space. Heliyon. 2020;6(8):e04639.
21. Pham TH, Qiu Y, Zeng JC, Xie L, Zhang P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. Nat Mach Intell. 2021;3(3):247–57.
22. El-Behery H, Attia AF, El-Feshawy N, Torkey H. Efficient machine learning model for predicting drug-target interactions with case study for Covid-19. Comput Biol Chem. 2021;93:107536.
23. Lv H, Shi L, Berkenpas JW, Dao FY, Zulfiqar H, Ding H, Zhang Y, Yang LM, Cao RZ. Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. Brief Bioinform. 2021;1–10:bbab320. https://doi.org/10.1093/bib/bbab320.
24. Liu Y, Gan J, Wang R, Yang X, Xiao Z, Cao Y. DrugDev-Covid19: an atlas of anti-COVID-19 compounds derived by computer-aided drug design. Molecules. 2022;27(3):683.
25. Gupta A, Müller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G. Generative recurrent networks for de novo drug design. Mol Inf. 2018;37(1–2):1700111.
26. Zhang L, Zhang H, Ai H, Hu H, Li S, Zhao J, Liu H. Applications of machine learning methods in drug toxicity prediction. Curr Top Med Chem. 2018;18(12):987–97.
27. White J. PubMed 2.0. Med Ref Serv Q. 2020;39(4):382–7.
28. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50(D1):D20–6. https://doi.org/10.1093/nar/gkab1112.
29. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today. 2017;22(11):1680–5.
30. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, Hickey AJ, Clark AM. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18(5):435–41.
31. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine learning methods in drug discovery. Molecules. 2020;25(22):5277.
32. Castillo TJM, Arif M, Niessen WJ, Schoots IG, Veenland JF. Automated classification of significant prostate cancer on MRI: a systematic review on the performance of machine learning applications. Cancers (Basel). 2020;12(6):1606.
33. Cuocolo R, Cipullo MB, Stanzione A, Romeo V, Green R, Cantoni V, Ponsiglione A, Ugga L, Imbriaco M. Machine learning for the identification of clinically significant prostate cancer on MRI: a meta-analysis. Eur Radiol. 2020;30:6877–87.
34. Cuocolo R, Cipullo MB, Stanzione A, Ugga L, Romeo V, Radice L, Brunetti A, Imbriaco M. Machine learning applications in

prostate cancer magnetic resonance imaging. Eur Radiol Exp. 2019;3(1):1–8.
35. Yang H, Sun L, Li W, Liu G, Tang Y. In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front Chem. 2018;6:30.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
37. Sinaga KP, Yang M-S. Unsupervised K-means clustering algorithm. IEEE Access. 2020;8:80716–27.
38. Kodinariya TM, Makwana PR. Review on determining number of cluster in K-means clustering. Int J. 2013;1(6):90–5.
39. Pham DT, Dimov SS, Nguyen CD. Selection of K in K-means clustering. Proc Inst Mech Eng C J Mech Eng Sci. 2005;219(1):103–19.
40. Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. Pattern Recogn. 2003;36(2):451–61.
41. Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat). 1979;28(1):100–8.
42. Golmohammadi H, Dashtbozorgi Z, Acree WE Jr. Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. Eur J Pharm Sci. 2012;47(2):421–9.
43. Shar PA, Tao W, Gao S, Huang C, Li B, Zhang W, Shahen M, Zheng C, Bai Y, Wang Y. Pred-binding: large-scale protein–ligand binding affinity prediction. J Enzyme Inhib Med Chem. 2016;31(6):1443–50.
44. Zsila F, Bikadi Z, Malik D, Hari P, Pechan I, Berces A, Hazai E. Evaluation of drug–human serum albumin binding interactions with support vector machine aided online automated docking. Bioinformatics. 2011;27(13):1806–13.
45. Passerini A, Punta M, Ceroni A, Rost B, Frasconi P. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. Proteins Struct Funct Bioinform. 2006;65(2):305–16.
46. Cai Y-D, Lin SL. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochim Biophys Acta (BBA) Proteins Proteom. 2003;1648(1–2):127–33.
47. Bradford JR, Westhead DR. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics. 2005;21(8):1487–94.
48. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. Int J Comput Sci Issues (IJCSI). 2012;9(5):272.
49. Denisko D, Hoffman MM. Classification and interaction in random forests. Proc Natl Acad Sci. 2018;115(8):1690–2.
50. Fratello M, Tagliaferri R. Decision trees and random forests. Encyclopedia of bioinformatics and computational biology. 2019;1:374–83.
51. Shi H, Liu S, Chen J, Li X, Ma Q, Yu B. Predicting drug–target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics. 2019;111(6):1839–52.
52. Bishop CM. Neural networks and their applications. Rev Sci Instrum. 1994;65(6):1803–32.
53. Abdi H. A neural network primer. J Biol Syst. 1994;2(03):247–81.
54. Agostinelli F, Hoffman M, Sadowski P, Baldi P. Learning activation functions to improve deep neural networks. arXiv preprint arXiv:14126830 2014.
55. Rasamoelina AD, Adjailia F, Sinčák P. A review of activation function for artificial neural network. In: 2020 IEEE 18th World symposium on applied machine intelligence and informatics (SAMI). IEEE. 2020. p. 281–6.
56. Yuen B, Hoang MT, Dong X, Lu T. Universal activation function for machine learning. Sci Rep. 2021;11(1):18757.
57. Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Ann Transl Med. 2018;6(11):216.
58. Dayhoff JE, DeLeo JM. Artificial neural networks: opening the black box. Cancer Interdiscipl Int J Am Cancer Soc. 2001;91(S8):1615–35.
59. Benítez JM, Castro JL, Requena I. Are artificial neural networks black boxes? IEEE Trans Neural Netw. 1997;8(5):1156–64.
60. Temurtas H, Yumusak N, Temurtas F. A comparative study on diabetes disease diagnosis using neural networks. Expert Syst Appl. 2009;36(4):8610–5.
61. Khemphila A, Boonjing V. Heart disease classification using neural network and feature selection. In: 2011 21st international conference on systems engineering: 16–18 Aug 2011. 2011. p. 406–9.
62. Sadighpour L, Rezaei S, Paknejad M, Jafary F, Aslani P. The application of an artificial neural network to support decision making in edentulous maxillary implant prostheses. J Res Pract Dent. 2014:2014:369025. https://doi.org/10.5171/2014.369025.
63. Jung S-K, Kim T-W. New approach for the diagnosis of extractions with neural network machine learning. Am J Orthod Dentofac Orthop. 2016;149(1):127–33.
64. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol. 2018;36:257–72.
65. de Souza JG, Fernandes MAC, Barbosa RD. A novel deep neural network technique for drug–target interaction. Pharmaceutics. 2022;14(3):625.
66. Karnati M, Seal A, Sahu G, Yazidi A, Krejcar O. A novel multi-scale based deep convolutional neural network for detecting COVID-19 from X-rays. Appl Soft Comput. 2022;125:109109.
67. Andrade CH, Pasqualoto KF, Ferreira EI, Hopfinger AJ. 4D-QSAR: perspectives in drug design. Molecules. 2010;15(5):3281–94.
68. Hopfinger A, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, Duraiswami C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. J Am Chem Soc. 1997;119(43):10509–24.
69. Potemkin V, Grishina M. Principles for 3D/4D QSAR classification of drugs. Drug Discov Today. 2008;13(21–22):952–9.
70. Myshkin E, Brennan R, Khasanova T, Sitnik T, Serebriyskaya T, Litvinova E, Guryanov A, Nikolsky Y, Nikolskaya T, Bureeva S. Prediction of organ toxicity endpoints by qsar modeling based on precise chemical-histopathology annotations. Chem Bioogyl Drug Des. 2012;80(3):406–16.
71. Yang L, Wang Y, Chang J, Pan Y, Wei R, Li J, Wang H. QSAR modeling the toxicity of pesticides against *Americamysis bahia*. Chemosphere. 2020;258: 127217.
72. Klüver N, Vogs C, Altenburger R, Escher BI, Scholz S. Development of a general baseline toxicity QSAR model for the fish embryo acute toxicity test. Chemosphere. 2016;164:164–73.
73. Pavan M, Netzeva T, Worth A. Validation of a QSAR model for acute toxicity. SAR QSAR Environ Res. 2006;17(02):147–71.
74. Huang S-H, Tung C-W, Fülöp F, Li J-H. Developing a QSAR model for hepatotoxicity screening of the active compounds in traditional Chinese medicines. Food Chem Toxicol. 2015;78:71–7.
75. Rodgers AD, Zhu H, Fourches D, Rusyn I, Tropsha A. Modeling liver-related adverse effects of drugs using k nearest neighbor quantitative structure–activity relationship method. Chem Res Toxicol. 2010;23(4):724–32.
76. Fjodorova N, Vračko M, Novič M, Roncaglioni A, Benfenati E. New public QSAR model for carcinogenicity. In: Chemistry central journal: 2010. Springer. p. 1–15.

77. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54.
78. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015;71:58–63.
79. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. 2016;30:595–608.
80. Pattanaik L, Coley CW. Molecular representation: going long on fingerprints. Chemistry. 2020;6(6):1204–7.
81. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Mol Pharm. 2016;13(7):2524–30.
82. March-Vila E, Pinzi L, Sturm N, Tinivella A, Engkvist O, Chen HM, Rastelli G: On the integration of in silico drug design methods for drug repurposing. Front Pharmacol. 2017;8:298. https://doi.org/10.3389/fphar.2017.00298.
83. Wen M, Zhang ZM, Niu SY, Sha HZ, Yang RH, Yun YH, Lu HM. Deep-learning-based drug–target interaction prediction. J Proteome Res. 2017;16(4):1401–9.
84. Chang Y, Park H, Yang HJ, Lee S, Lee KY, Kim TS, Jung J, Shin JM. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. Sci Rep. 2018;8(1):8857. https://doi.org/10.1038/s41598-018-27214-6.
85. Wan FP, Hong LX, Xiao A, Jiang T, Zeng JY. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. Bioinformatics. 2019;35(1):104–11.
86. Zeng XX, Zhu SY, Liu XR, Zhou YD, Nussinov R, Cheng FX. DeepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics. 2019;35(24):5191–8.
87. Zeng XX, Song X, Ma TF, Pan XQ, Zhou YD, Hou Y, Zhang Z, Li KL, Karypis G, Cheng FX. Repurpose open data to discover therapeutics for COVID-19 using deep learning. J Proteome Res. 2020;19(11):4624–36.
88. Gysi DM, do Valle I, Zitnik M, Ameli A, Gan X, Varol O, Ghiassian SD, Patten JJ, Davey RA, Loscalzo J et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proced Natl Acad Sci USA 2021;118(19):e2025581118.
89. Santos SD, Torres M, Galeano D, Sanchez MD, Cernuzzi L, Paccanaro A. Machine learning and network medicine approaches for drug repositioning for COVID-19. Patterns. 2022;3(1):100396.
90. Ge YY, Tian TZ, Huang SL, Wan FP, Li JX, Li SY, Wang XT, Yang H, Hong LX, Wu N et al: An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. Signal Transduct Target Ther. 2021;6(1):165.
91. Smith DP, Oechsle O, Rawling MJ, Savory E, Lacoste A, Richardson PJ. Expert-augmented computational drug repurposing identified baricitinib as a treatment for COVID-19. Front Pharmacol. 2021;12:1699.
92. Kumar A, Loharch S, Kumar S, Ringe RP, Parkesh R. Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2. Comput Struct Biotechnol J. 2021;19:424–38.
93. Beck BR, Shin B, Choi Y, Park S, Kang K. 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. 2020;18:784–90.
94. Gawriljuk VO, Zin PPK, Puhl AC, Zorn KM, Foil DH, Lane TR, Hurst B, Tavella TA, Costa FTM, Lakshmanane P, et al. Machine learning models identify inhibitors of SARS-CoV-2. J Chem Inf Model. 2021;61(9):4224–35.
95. Kadioglu O, Saeed M, Greten HJ, Efferth T. Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning. Comput Biol Med. 2021;133:104359.
96. Jiang X, Neapolitan RE. Mining pure, strict epistatic interactions from high-dimensional datasets: ameliorating the curse of dimensionality. PLoS One. 2012; 7(10):e467712012.
97. Chattopadhyay A, Lu T-P. Gene-gene interaction: the curse of dimensionality. Ann Transl Med. 2019;7(24):813.
98. Carse B, Fogarty TC: Tackling the "curse of dimensionality" of radial basis functional neural networks using a genetic algorithm. In: Parallel Problem solving from nature—PPSN IV: international conference on evolutionary computation—the 4th international conference on parallel problem solving from Nature Berlin, Germany, September 22–26, 1996 Proceedings 4: 1996. Springer. p. 707–19.
99. Fakoor R, Ladhak F, Nazi A, Huber M. using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the ICML workshop on the role of machine learning in transforming healthcare (WHEALTH). Atlanta, GA. 2013.
100. Levine AB, Schlosser C, Grewal J, Coope R, Jones SJ, Yip S. Rise of the machines: advances in deep learning for cancer diagnosis. Trends Cancer. 2019;5(3):157–69.
101. Sun W, Zheng B, Qian W. Computer aided lung cancer diagnosis with deep learning algorithms. In: Medical imaging 2016: computer-aided diagnosis: 2016. SPIE: 241-248.
102. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. 2021;13(1):1–17.
103. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, Langlotz CP, Amrhein TJ, Lungren MP. Deep learning to classify radiology free-text reports. Radiology. 2018;286(3):845–52.
104. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. J Magn Reson Imaging. 2019;49(4):939–54.
105. Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, Su C, Wu Z, Xie Q, Berger B. High-resolution de novo structure prediction from primary sequence. BioRxiv 2022:2022.2007. 2021.500999.
106. Wei L, Zou Q. Recent progress in machine learning-based methods for protein fold recognition. Int J Mol Sci. 2016;17(12):2118.
107. Wei G-W. Protein structure prediction beyond AlphaFold. Nat Mach Intell. 2019;1(8):336–7.
108. Noé F, De Fabritiis G, Clementi C. Machine learning for protein folding and dynamics. Curr Opin Struct Biol. 2020;60:77–84.
109. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
110. Heo L, Feig M. High-accuracy protein structures by combining machine-learning with physics-based refinement. Proteins Struct Funct Bioinform. 2020;88(5):637–42.
111. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics. 2006;22(12):1456–63.
112. AlQuraishi M. Machine learning in protein structure prediction. Curr Opin Chem Biol. 2021;65:1–8.
113. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics. 2016;32(15):2338–45.
114. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, Rouillard AD, Tan CM, Chen EY, Golub TR. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014;42(W1):W449–60.

115. Zhang PL, Wei ZQ, Che C, Jin B. DeepMGT-DTI: transformer network incorporating multilayer graph information for drug–target interaction prediction. Comput Biol Med. 2022;142:105214.
116. Yazdani-Jahromi M, Yousefi N, Tayebi A, Kolanthai E, Neal CJ, Seal S, Garibay OO. AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. Brief Bioinform. 2022;23(4):bbac272. https://doi.org/10.1093/bib/bbac272.
117. Yang YQ, Zhou DS, Zhang XB, Shi YL, Han JX, Zhou LP, Wu LY, Ma MF, Li JT, Peng SL et al. D3AI-CoV: a deep learning platform for predicting drug targets and for virtual screening against COVID-19. Brief Bioinform. 2022;23(3):bbac147. https://doi.org/10.1093/bib/bbac147.
118. Wei BM, Zhang Y, Gong X. DeepLPI: a novel deep learning-based model for protein–ligand interaction prediction for drug repurposing. Sci Rep. 2022;12(1):18200.
119. Wang SW, Sun Q, Xu YJ, Pei JF, Lai LH. A transferable deep learning approach to fast screen potential antiviral drugs against SARS-CoV-2. Brief Bioinform. 2021;22(6):bbab211.
120. Wang JJ, Wen NF, Wang CY, Zhao LL, Cheng L. ELECTRA-DTA: a new compound-protein binding affinity prediction model based on the contextualized sequence encoding. J Cheminform. 2022;14(1):1–14.
121. Timmons JA, Anighoro A, Brogan RJ, Stahl J, Wahlestedt C, Farquhar DG, Taylor-King J, Volmar CH, Kraus WE, Phillips SM. A human-based multi-gene signature enables quantitative drug repurposing for metabolic disease. Elife 2022;11:e68832.
122. Surianarayanan C, Chelliah PR. Leveraging artificial intelligence (AI) capabilities for COVID-19 containment. New Gener Comput. 2021;39(3–4):717–41.
123. Su XR, Hu L, You ZH, Hu PW, Wang L, Zhao BW. A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2. Brief Bioinform. 2022;23(1):bbab526.
124. Silveira EC. Screening anti-inflammatory, anticoagulant, and respiratory agents for SARS-CoV-2 3CL(Pro) inhibition from chemical fingerprints through a deep learning approach. Clin Transl Investig. 2022;74(1):31–9.
125. Shorten C, Khoshgoftaar TM, Furht B. Deep learning applications for COVID-19. J Big Data. 2021;8:18.
126. Ray S, Lall S, Bandyopadhyay S. A deep integrated framework for predicting SARS-CoV2-human protein–protein interaction. IEEE Trans Emerg Top Comput Intell. 2022;6(6):1463–72.
127. Rajput A, Thakur A, Mukhopadhyay A, Kamboj S, Rastogi A, Gautam S, Jassal H, Kumar M. Prediction of repurposed drugs for Coronaviruses using artificial intelligence and machine learning. Comput Struct Biotechnol J. 2021;19:3133–48.
128. Pan XQ, Lin X, Cao DS, Zeng XX, Yu PS, He LF, Nussinov R, Cheng FX. Deep learning for drug repurposing: methods, databases, and applications. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(4):e1597.
129. Moovarkumudalvan B, Geethakumari AM, Ramadoss R, Biswas KH, Mifsud B. Structure-based virtual screening and functional validation of potential hit molecules targeting the SARS-CoV-2 main protease. Biomolecules. 2022;12(12):1754.
130. Li ZR, Zhong Q, Yang J, Duan YJ, Wang WJ, Wu CK, He KL. DeepKG: an end-to-end deep learning-based workflow for biomedical knowledge graph extraction, optimization and applications. Bioinformatics. 2022;38(5):1477–9.
131. Lee CY, Chen YPP. New insights into drug repurposing for COVID-19 using deep learning. IEEE Trans Neural Netw Learn Syst. 2021;32(11):4770–80.
132. Kanapeckaite A, Mazeikiene A, Geris L, Burokiene N, Cottrell GS, Widera D. Computational pharmacology: new avenues for COVID-19 therapeutics search and better preparedness for future pandemic crises. Biophys Chem. 2022;290:106891.
133. Joshi T, Sharma P, Mathpal S, Joshi T, Maiti P, Nand M, Pande V, Chandra S. Computational investigation of drug bank compounds against 3C-like protease (3CL(pro)) of SARS-CoV-2 using deep learning and molecular dynamics simulation. Mol Divers. 2022;26(4):2243–56.
134. Hooshmand SA, Ghobadi MZ, Hooshmand SE, Jamalkandi SA, Alavi SM, Masoudi-Nejad A. A multimodal deep learning-based drug repurposing approach for treatment of COVID-19. Mol Divers. 2021;25(3):1717–30.
135. Harigua-Souiai E, Heinhane MM, Abdelkrim YZ, Souiai O, Abdeljaoued-Tej I, Guizani I. Deep learning algorithms achieved satisfactory predictions when trained on a novel collection of anticoronavirus molecules. Front Genet. 2021;12:744170.
136. Deepthi K, Jereesh AS, Liu YS. A deep learning ensemble approach to prioritize antiviral drugs against novel coronavirus SARS-CoV-2 for COVID-19 drug repurposing. Appl Soft Comput. 2021;113:107945.
137. Choi Y, Shin B, Kang K, Park S, Beck BR. Target-centered drug repurposing predictions of human angiotensin-converting enzyme 2 (ACE2) and transmembrane protease serine subtype 2 (TMPRSS2) interacting approved drugs for coronavirus disease 2019 (COVID-19) treatment through a drug–target interaction deep learning model. Viruses-Basel 2020;12(11):1325.
138. Anwaar MUU, Adnan F, Abro A, Khan RAA, Rehman AUU, Osama M, Rainville C, Kumar S, Sterner DEE, Javed S, et al. Combined deep learning and molecular docking simulations approach identifies potentially effective FDA approved drugs for repurposing against SARS-CoV-2. Comput Biol Med. 2022;141:105049.
139. Abdel-Basset M, Hawash H, Elhoseny M, Chakrabortty RK, Ryan M. DeepH-DTA: deep learning for predicting drug-target interactions: a case study of COVID-19 drug repurposing. IEEE Access. 2020;8:170433–51.
140. Percha B, Altman RB. A global network of biomedical relationships derived from text. Bioinformatics. 2018;34(15):2614–24.
141. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014;42(1):D1091-1097.
142. Wang MY. Deep graph library: towards efficient and scalable deep learning on graphs. In: ICLR workshop on representation learning on graphs and manifolds. 2019.
143. Reis G, Silva EA, Silva DC, Thabane L, Milagres AC, Ferreira TS, Dos Santos CV, Campos VH, Nogueira AM, de Almeida AP. Effect of early treatment with ivermectin among patients with Covid-19. N Engl J Med. 2022;386(18):1721–31.
144. Popp M, Stegemann M, Metzendorf M-I, Gould S, Kranke P, Meybohm P, Skoetz N, Weibel S. Ivermectin for preventing and treating COVID-19. Cochrane Database Syst Rev. 2021;2021(7):CD015017. https://doi.org/10.1002/14651858.CD015017.pub2.
145. Torjesen I: Covid-19. Hydroxychloroquine does not benefit hospitalised patients, UK trial finds. BMJ Br Med J (Online). 2020;369:m2263.
146. Hassanipour S, Arab-Zozani M, Amani B, Heidarzad F, Fathalipour M, Martinez-de-Hoyo R. The efficacy and safety of Favipiravir in treatment of COVID-19: a systematic review and meta-analysis of clinical trials. Sci Rep. 2021;11(1):11022.
147. De Wilde AH, Falzarano D, Zevenhoven-Dobbe JC, Beugeling C, Fett C, Martellaro C, Posthuma CC, Feldmann H, Perlman S, Snijder EJ. Alisporivir inhibits MERS-and SARS-coronavirus replication in cell culture, but not SARS-coronavirus infection in a mouse model. Virus Res. 2017;228:7–13.

148. Softic L, Brillet R, Berry F, Ahnou N, Nevers Q, Morin-Dewaele M, Hamadat S, Bruscella P, Fourati S, Pawlotsky J-M. Inhibition of SARS-CoV-2 infection by the cyclophilin inhibitor alisporivir (Debio 025). Antimicrob Agents Chemother. 2020;64(7):e00876-e1820.
149. Pawlotsky J-M. COVID-19 pandemic: time to revive the cyclophilin inhibitor alisporivir. Clin Infect Dis. 2020;71(16):2191–4.
150. Kalil AC, Patterson TF, Mehta AK, Tomashek KM, Wolfe CR, Ghazaryan V, Marconi VC, Ruiz-Palacios GM, Hsieh L, Kline S. Baricitinib plus remdesivir for hospitalized adults with Covid-19. N Engl J Med. 2021;384(9):795–807.
151. Rubin R. Baricitinib is first approved COVID-19 immunomodulatory treatment. JAMA. 2022;327(23):2281–2281.
152. Dunnivant FM, Elzerman AW, Jurs PC, Hasan MN. Quantitative structure property relationships for aqueous solubilities and henrys law constants of polychlorinated-biphenyls. Environ Sci Technol. 1992;26(8):1567–73.
153. Fisher SW, Lydy MJ, Barger J, Landrum PF. Quantitative structure–activity-relationships for predicting the toxicity of pesticides in aquatic systems with sediment. Environ Toxicol Chem. 1993;12(7):1307–18.
154. Karickhoff SW, McDaniel VK, Melton C, Vellino AN, Nute DE, Carreira LA. Predicting chemical-reactivity by computer. Environ Toxicol Chem. 1991;10(11):1405–16.
155. Moriguchi I, Hirono S, Liu Q, Matsushita Y, Nakagawa T. Fuzzy adaptive least-squares and its use in quantitative structure-activity-relationships. Chem Pharm Bull. 1990;38(12):3373–9.
156. Nabivach VM, Dmitrikov VP. Use of the correlation equations for the prediction of the retention data in gas–liquid-chromatography. Usp Khim. 1993;62(1):27–38.
157. Narayszabo G, Balogh T. The average molecular electrostatic-field as a QSAR descriptor. 4. hydrophobicity scales for amino-acid residues-alpha. J Mol Struct Theochem. 1993;103(3):243–8.
158. Stanton DT, Egolf LM, Jurs PC, Hicks MG. Computer-assisted prediction of normal boiling points of pyrans and pyrroles. J Chem Inf Comput Sci. 1992;32(4):306–16.
159. Suzuki T, Ohtaguchi K, Koide K. Correlation and prediction of autoignition temperatures of hydrocarbons using molecular-properties. J Chem Eng Jpn. 1992;25(5):606–8.
160. Cereto-Massague A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallve S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015;71:58–63.
161. Dong J, Cao DS, Miao HY, Liu S, Deng BC, Yun YH, Wang NN, Lu AP, Zeng WB, Chen AF. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminform. 2015;7:1–10.
162. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. 2016;30(8):595–608.
163. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74.
164. Russo DP, Zorn KM, Clark AM, Zhu H, Ekins S. Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. Mol Pharm. 2018;15(10):4361–70.
165. Chen B, Sheridan RP, Hornak V, Voigt JH. Comparison of random forest and pipeline pilot Naive Bayes in prospective QSAR predictions. J Chem Inf Model. 2012;52(3):792–803.
166. Lane TR, Foil DH, Minerali E, Urbina F, Zorn KM, Ekins S. Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery. Mol Pharm. 2020;18(1):403–15.
167. Minerali E, Foil DH, Zorn KM, Ekins S. Evaluation of assay central machine learning models for rat acute oral toxicity prediction. ACS Sustain Chem Eng. 2020;8(42):16020–7.
168. Rosenke K, Jarvis MA, Feldmann F, Schwarz B, Okumura A, Lovaglio J, Saturday G, Hanley PW, Meade-White K, Williamson BN. Hydroxychloroquine proves ineffective in hamsters and macaques infected with SARS-CoV-2. BioRxiv. 2020.
169. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, Funk K, Kinney R, Liu Z, Merrill W. Cord-19: the covid-19 open research dataset. arXiv. 2020.
170. Lo HS, Hui KPY, Lai H-M, He X, Khan KS, Kaur S, Huang J, Li Z, Chan AK, Cheung HH-Y. Simeprevir potently suppresses SARS-CoV-2 replication and synergizes with remdesivir. ACS Cent Sci. 2021;7(5):792–802.
171. Muturi E, Hong W, Li J, Yang W, He J, Wei H, Yang H. Effects of simeprevir on the replication of SARS-CoV-2 in vitro and in transgenic hACE2 mice. Int J Antimicrob Agents. 2022;59(1):106499.
172. Rahman MM, Saha T, Islam KJ, Suman RH, Biswas S, Rahat EU, Hossen MR, Islam R, Hossain MN, Mamun AA. Virtual screening, molecular dynamics and structure–activity relationship studies to identify potent approved drugs for Covid-19 treatment. J Biomol Struct Dyn. 2021;39(16):6231–41.
173. Abhithaj J, Dileep F, Sharanya C, Arun K, Sadasivan C, Jayadevi V. Repurposing simeprevir, calpain inhibitor IV and a cathepsin F inhibitor against SARS-CoV-2 and insights into their interactions with Mpro. J Biomol Struct Dyn. 2020;1:23–35.
174. Ahmed S, Mahtarin R, Ahmed SS, Akter S, Islam MS, Mamun AA, Islam R, Hossain MN, Ali MA, Sultana MU. Investigating the binding affinity, interaction, and structure–activity-relationship of 76 prescription antiviral drugs targeting RdRp and Mpro of SARS-CoV-2. J Biomol Struct Dyn. 2021;39(16):6290–305.
175. Shin B, Park S, Kang K, Ho JC: Self-attention based molecule representation for predicting drug-target interaction. In: Machine learning for healthcare conference: 2019. PMLR. p. 230–48.
176. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6.
177. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. Genomics. 1991;11(3):635–50.
178. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res. 2007;35(suppl_1):D198–201.
179. Tanoli Z, Alam Z, Vähä-Koskela M, Ravikumar B, Malyutina A, Jaiswal A, Tang J, Wennerberg K, Aittokallio T: Drug Target Commons 2.0: a community platform for systematic analysis of drug–target interaction profiles. Database. 2018;2018:bay083. https://doi.org/10.1093/database/bay083.
180. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–61.
181. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, Ruan L, Song B, Cai Y, Wei M. A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19. N Engl J Med. 2020;382(19):1787–99.
182. Kupferschmidt K, Wadman M: Delta variant triggers new phase in the pandemic. In: American Association for the Advancement of Science; 2021.
183. Christensen PA, Olsen RJ, Long SW, Subedi S, Davis JJ, Hodjat P, Walley DR, Kinskey JC, Saavedra MO, Pruitt L. Delta variants of SARS-CoV-2 cause significantly increased vaccine breakthrough COVID-19 cases in Houston, Texas. Am J Pathol. 2021;192(2):320–31.

184. CDC Statement on B. 1.1. 529 (Omicron variant), Media Statement Release on November 26, 2021. https://www.cdc.gov/media/releases/2021/s1126-B11-529-omicron.html.
185. Torjesen I. Covid-19: Omicron may be more transmissible than other variants and partly resistant to existing vaccines, scientists fear. In. British Medical Journal Publishing Group; 2021.
186. Rao S, Singh M. The Newly Detected B. 1.1. 529 (Omicron) variant of SARS-CoV-2 With multiple mutations: implications for transmission, diagnostics, therapeutics, and immune evasion. DHR Proc. 2021;1(S5):7–10.
187. Sahoo JP, Samal KC. World on alert: WHO designated South African new COVID strain (Omicron/B. 1.1. 529) as a variant of concern. Biotica Res Today. 2021;3(11):1086–8.
188. Zhang X, Wu S, Wu B, Yang Q, Chen A, Li Y, Zhang Y, Pan T, Zhang H, He X. SARS-CoV-2 Omicron strain exhibits potent capabilities for immune evasion and viral entrance. Signal Transduct Target Ther. 2021;6(1):1–3.
189. Jahanshahlu L, Rezaei N. Monoclonal antibody as a potential anti-COVID-19. Biomed Pharmacother. 2020;129: 110337.
190. Clark SA, Clark LE, Pan J, Coscia A, McKay LG, Shankar S, Johnson RI, Griffiths A, Abraham J. Molecular basis for a germline-biased neutralizing antibody response to SARS-CoV-2. bioRxiv 2020.
191. Chen RE, Winkler ES, Case JB, Aziati ID, Bricker TL, Joshi A, Darling TL, Ying B, Errico JM, Shrihari S, VanBlargan LA. In vivo monoclonal antibody efficacy against SARS-CoV-2 variant strains. Nature. 2021;596(7870):103–8.
192. Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y, Wang M, Yu J, Zhang B, Kwong PD. Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7. Nature. 2021;593(7857):130–5.
193. Starr TN, Greaney AJ, Dingens AS, Bloom JD. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. Cell Rep Med. 2021;2(4): 100255.
194. Bertoglio F, Fühner V, Ruschig M, Heine PA, Abassi L, Klünemann T, Rand U, Meier D, Langreder N, Steinke S. A SARS-CoV-2 neutralizing antibody selected from COVID-19 patients binds to the ACE2-RBD interface and is tolerant to most known RBD mutations. Cell Rep. 2021;36(4): 109433.
195. Williams MA, Hall DR, Hulswit RJ, Bowden TA, Fry EE. Antibody evasion by the P. 1 strain of SARS-CoV-2. Cell. 2021;184:1–16.
196. Planas D, Saunders N, Maes P, Guivel-Benhassine F, Planchais C, Buchrieser J, Bolland W-H, Porrot F, Staropoli I, Lemoine F. Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. Nature. 2022;602(7898):671–5.
197. Chen J, Wang R, Gilby NB, Wei G-W. Omicron variant (B. 1.1. 529): infectivity, vaccine breakthrough, and antibody resistance. J Chem Inf Model 2022;62(2):412–22.
198. Huo J, Dijokaite-Guraliuc A, Liu C, Zhou D, Ginn HM, Das R, Supasa P, Selvaraj M, Nutalai R, Tuekprakhon A. A delicate balance between antibody evasion and ACE2 affinity for Omicron BA. 2.75. Cell Rep. 2022;42:111903.
199. Cao Y, Song W, Wang L, Liu P, Yue C, Jian F, Yu Y, Yisimayi A, Wang P, Wang Y. Characterization of the enhanced infectivity and antibody evasion of Omicron BA. 2.75. Cell Host Microbe. 2022;30(11):1527–39.
200. Chakraborty C, Bhattacharya M, Sharma AR. Emerging mutations in the SARS-CoV-2 variants and their role in antibody escape to small molecule-based therapeutic resistance. Curr Opin Pharmacol. 2022;62:64–73.
201. Yue C, Song W, Wang L, Jian F, Chen X, Gao F, Shen Z, Wang Y, Wang X, Cao Y. Enhanced transmissibility of XBB. 1.5 is contributed by both strong ACE2 binding and antibody evasion. bioRxiv 2023:2023.2001. 2003.522427.
202. Pardo J, Shukla AM, Chamarthi G, Gupte A. The journey of remdesivir: from Ebola to COVID-19. Drugs Context. 2020;9:1–9.
203. Frediansyah A, Nainu F, Dhama K, Mudatsir M, Harapan H. Remdesivir and its antiviral activity against COVID-19: a systematic review. Clin Epidemiol Glob Health. 2021;9:123–7.
204. Kumari M, Subbarao N. Deep learning model for virtual screening of novel 3C-like protease enzyme inhibitors against SARS coronavirus diseases. Comput Biol Med. 2021;132:104317.
205. Srinivasan S, Batra R, Chan H, Kamath G, Cherukara MJ, Sankaranarayanan SK. Artificial intelligence-guided De novo molecular design targeting COVID-19. ACS Omega. 2021;6(19):12557–66.
206. Bung N, Krishnan SR, Bulusu G, Roy A. De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. Future Med Chem. 2021;13(6):575–85.
207. Arshia AH, Shadravan S, Solhjoo A, Sakhteman A, Sami A. De novo design of novel protease inhibitor candidates in the treatment of SARS-CoV-2 using deep learning, docking, and molecular dynamic simulations. Comput Biol Med. 2021;139: 104967.
208. Magar R, Yadav P, Barati Farimani A. Potential neutralizing antibodies discovered for novel corona virus using machine learning. Sci Rep. 2021;11(1):1–11.
209. Williams AH, Zhan C-G. Fast prediction of binding affinities of SARS-CoV-2 spike protein and its mutants with antibodies through intermolecular interaction modeling-based machine learning. J Phys Chem B. 2022;126(28):5192–206.
210. Xu Z, Yang L, Zhang X, Zhang Q, Yang Z, Liu Y, Wei S, Liu W. Discovery of potential flavonoid inhibitors against COVID-19 3CL proteinase based on virtual screening strategy. Front Mol Biosci. 2020;7: 556481.
211. Li Z, Lin Y, Huang Y-Y, Liu R, Zhan C-G, Wang X, Luo H-B. Reply to Ma and Wang: Reliability of various in vitro activity assays on SARS-CoV-2 main protease inhibitors. Proc Natl Acad Sci USA. 2021;118(8):e2024937118. https://doi.org/10.1073/pnas.2024937118.
212. Ngwa W, Kumar R, Thompson D, Lyerly W, Moore R, Reid T-E, Lowe H, Toyang N. Potential of flavonoid-inspired phytomedicines against COVID-19. Molecules. 2020;25(11):2707.
213. Pitsillou E, Liang J, Ververis K, Lim KW, Hung A, Karagiannis TC. Identification of small molecule inhibitors of the deubiquitinating activity of the SARS-CoV-2 papain-like protease: in silico molecular docking studies and in vitro enzymatic activity assay. Front Chem. 2020;8:623971.
214. Choi J, Yun JS, Song H, Kim NH, Kim HS, Yook JI. Exploring the chemical space of protein–protein interaction inhibitors through machine learning. Sci Rep. 2021;11(1):1–10.
215. Christensen AS, Faber FA, von Lilienfeld OA. Operators in quantum machine learning: response properties in chemical space. J Chem Phys. 2019;150(6):064105.
216. Coley CW. Defining and exploring chemical spaces. Trends Chem. 2021;3(2):133–45.
217. Deng Z-L, Du C-X, Li X, Hu B, Kuang Z-K, Wang R, Feng S-Y, Zhang H-Y, Kong D-X. Exploring the biologically relevant chemical space for drug discovery. J Chem Inf Model. 2013;53(11):2820–8.
218. Öztürk H, Özgür A, Schwaller P, Laino T, Ozkirimli E. Exploring chemical space using natural language processing methodologies for drug discovery. Drug Discov Today. 2020;25(4):689–705.
219. Ramakrishnan R, von Lilienfeld OA. Machine learning, quantum chemistry, and chemical space. Rev Comput Chem. 2017;30:225–56.
220. Reymond J-L. The chemical space project. Acc Chem Res. 2015;48(3):722–30.

221. Sperandio O, Reynès CH, Camproux A-C, Villoutreix BO. Rationalizing the chemical space of protein–protein interaction inhibitors. Drug Discov Today. 2010;15(5–6):220–9.
222. Kingma DP, Mohamed S, Jimenez Rezende D, Welling M. Semi-supervised learning with deep generative models. Adv Neural Inf Process Syst. 2014;27:1–16.
223. Harshvardhan G, Gourisaria MK, Pandey M, Rautaray SS. A comprehensive survey and analysis of generative models in machine learning. Comput Sci Rev. 2020;38: 100285.
224. Salakhutdinov R. Learning deep generative models. Annu Rev Stat Appl. 2015;2:361–85.
225. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res. 2015;43(W1):W612–20.
226. Irwin JJ, Shoichet BK. ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model. 2005;45(1):177–82.
227. Sterling T, Irwin JJ. ZINC 15—ligand discovery for everyone. J Chem Inf Model. 2015;55(11):2324–37.
228. Santana MV, Silva-Jr FP. De novo design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning. BMC Chem. 2021;15(1):8.
229. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017;7:42717.
230. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 1997;23(1–3):3–25.
231. Liu G, Zeng H, Mueller J, Carter B, Wang Z, Schilz J, Horny G, Birnbaum ME, Ewert S, Gifford DK. Antibody complementarity determining region design using high-capacity machine learning. Bioinformatics. 2020;36(7):2126–33.
232. Akbar R, Robert PA, Weber CR, Widrich M, Frank R, Pavlović M, Scheffer L, Chernigovskaya M, Snapkov I, Slabodkin A. In silico proof of principle of machine learning-based antibody design at unconstrained scale. In: MAbs: 2022. Taylor & Francis. p. 2031482.
233. Morgan RS, McAdon JM. Predictor for sulfur-aromatic interactions in globular proteins. Int J Peptide Protein Res. 1980;15(2):177–80.
234. Williams AH, Zhan C-G. Generalized methodology for the quick prediction of variant SARS-CoV-2 spike protein binding affinities with human angiotensin-converting enzyme II. J Phys Chem B. 2022;126(12):2353–60.
235. Williams AH, Zhan C-G. Fast prediction of binding affinities of SARS-CoV-2 spike protein and its mutants with antibodies through intermolecular interaction modeling-based machine learning. J Phys Chem B. 2022;126(28):5194–206.
236. Williams AH, Zhan C-G. Fast prediction of binding affinities of the SARS-CoV-2 spike protein mutant N501Y (UK variant) with ACE2 and miniprotein drug candidates. J Phys Chem B. 2021;125(17):4330–6.
237. Guan D, Rahman MT, Gay EA, Vasukuttan V, Mathews KM, Decker AM, Williams AH, Zhan C-G, Jin C. Indole-containing amidinohydrazones as nonpeptide, dual RXFP3/4 agonists: synthesis, structure–activity relationship, and molecular modeling studies. J Med Chem. 2021;64(24):17866–86.
238. Yang J-F, Williams AH, Penthala NR, Prather PL, Crooks PA, Zhan C-G. Binding Modes and selectivity of cannabinoid 1 (CB1) and cannabinoid 2 (CB2) receptor ligands. ACS Chem Neurosci. 2020;11(20):3455–63.
239. Williams AH, Zhan C-G. Generalized methodology for the quick prediction of variant SARS-CoV-2 spike protein binding affinities with human angiotensin-converting enzyme II. J Phys Chem B. 2022;126(12):2353–60.
240. Srivastava HK, Sastry GN. Molecular dynamics investigation on a series of HIV protease inhibitors: assessing the performance of MM-PBSA and MM-GBSA approaches. J Chem Inf Model. 2012;52(11):3088–98.
241. Rastelli G, Del Rio A, Degliesposti G, Sgobba M. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. J Comput Chem. 2010;31(4):797–810.
242. Weng G, Wang E, Chen F, Sun H, Wang Z, Hou T. Assessing the performance of MM/PBSA and MM/GBSA methods. 9. Prediction reliability of binding affinities and binding poses for protein–peptide complexes. Phys Chem Chem Phys. 2019;21(19):10135–45.
243. Sun H, Duan L, Chen F, Liu H, Wang Z, Pan P, Zhu F, Zhang JZ, Hou T. Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches. Phys Chem Chem Phys. 2018;20(21):14450–60.
244. Chen F, Sun H, Wang J, Zhu F, Liu H, Wang Z, Lei T, Li Y, Hou T. Assessing the performance of MM/PBSA and MM/GBSA methods. 8. Predicting binding free energies and poses of protein–RNA complexes. RNA. 2018;24(9):1183–94.
245. Chen F, Liu H, Sun H, Pan P, Li Y, Li D, Hou T. Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein–protein binding free energies and re-rank binding poses generated by protein–protein docking. Phys Chem Chem Phys. 2016;18(32):22129–39.
246. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov. 2015;10(5):449–61.
247. Williams A, Zhou S, Zhan C-G. Discovery of potent and selective butyrylcholinesterase inhibitors through the use of pharmacophore-based screening. Bioorg Med Chem Lett. 2019;29(24): 126754.
248. Dong L, Qu X, Zhao Y, Wang B. Prediction of binding free energy of protein–ligand complexes with a hybrid molecular mechanics/generalized born surface area and machine learning method. ACS Omega. 2021;6(48):32938–47.
249. DeJong C, Wachter RM. The risks of prescribing hydroxychloroquine for treatment of COVID-19—first, do no harm. JAMA Intern Med. 2020;180(8):1118–9.
250. Chai PR, Ferro EG, Kirshenbaum JM, Hayes BD, Culbreth SE, Boyer EW, Erickson TB. Intentional hydroxychloroquine overdose treated with high-dose diazepam: an increasing concern in the COVID-19 pandemic. J Med Toxicol. 2020;16:314–20.
251. Liu J, Cao R, Xu M, Wang X, Zhang H, Hu H, Li Y, Hu Z, Zhong W, Wang M. Hydroxychloroquine, a less toxic derivative of chloroquine, is effective in inhibiting SARS-CoV-2 infection in vitro. Cell discovery. 2020;6(1):16.
252. Temple C, Hoang R, Hendrickson RG. Toxic effects from ivermectin use associated with prevention and treatment of Covid-19. N Engl J Med. 2021;385(23):2197–8.

Authors and Affiliations

Alexander H. Williams[1,2,3] · Chang-Guo Zhan[1,2]

✉ Chang-Guo Zhan
zhan@uky.edu

[1] Molecular Modeling and Biopharmaceutical Center, University of Kentucky, 789 South Limestone Street, Lexington, KY 40536, USA

[2] Department of Pharmaceutical Sciences, College of Pharmacy, University of Kentucky, 789 South Limestone Street, Lexington, KY 40536, USA

[3] GSK Upper Providence, 1250 S. Collegeville Road, Collegeville, PA 19426, USA