Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 6668985, 13 pages
https://doi.org/10.1155/2021/6668985

Research Article

Deep Learning Approach for Discovery of In Silico Drugs for Combating COVID-19

Nishant Jha,[1] **Deepak Prashar**,[1] **Mamoon Rashid**,[2] **Mohammad Shafiq**,[3] **Razaullah Khan**,[4] **Catalin I. Pruncu**,[5,6] **Shams Tabrez Siddiqui**,[7] **and M. Saravana Kumar**[8]

[1]*School of Computer Science & Engineering, Lovely Professional University, Phagwara, India*
[2]*Department of Computer Engineering, Faculty of Science and Technology, Vishwakarma University, Pune, India*
[3]*Cyberspace Institute of Advanced Technology, GuangZhou University, Guangzhou, China*
[4]*Department of Engineering Management, University of Engineering and Applied Sciences, Swat 19060, Pakistan*
[5]*Design,Manufacturing & Engineering Management, University of Strathclyde, Glasgow G1 1XJ, UK*
[6]*Mechanical Engineering, Imperial College London, Exhibition Road South Kensington, London, UK*
[7]*College of Computer Science and Information Technology, Jazan University, Jazan, Saudi Arabia*
[8]*Department of Mechanical Engineering, Mount Zion College of Engineering and Technology, Pudukkottai, India*

Correspondence should be addressed to Catalin I. Pruncu; c.pruncu@imperial.ac.uk

Received 30 December 2020; Accepted 8 July 2021; Published 23 July 2021

Academic Editor: Daniel Espino

Copyright © 2021 Nishant Jha et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Early diagnosis of pandemic diseases such as COVID-19 can prove beneficial in dealing with difficult situations and helping radiologists and other experts manage staffing more effectively. The application of deep learning techniques for genetics, microscopy, and drug discovery has created a global impact. It can enhance and speed up the process of medical research and development of vaccines, which is required for pandemics such as COVID-19. However, current drugs such as remdesivir and clinical trials of other chemical compounds have not shown many impressive results. Therefore, it can take more time to provide effective treatment or drugs. In this paper, a deep learning approach based on logistic regression, SVM, Random Forest, and QSAR modeling is suggested. QSAR modeling is done to find the drug targets with protein interaction along with the calculation of binding affinities. Then deep learning models were used for training the molecular descriptor dataset for the robust discovery of drugs and feature extraction for combating COVID-19. Results have shown more significant binding affinities (greater than −18) for many molecules that can be used to block the multiplication of SARS-CoV-2, responsible for COVID-19.

1. Introduction

The first case of COVID-19 was detected in December 2019, and from then, it has overgrown, affecting millions of people around the globe. More than 2 million cases have been confirmed, with over 0.15 million deaths globally [1, 2]. Drug repurposing is defined as discovering and identifying newer applications for existing drugs in the treatment of various diseases [3]. Recent advancements in drug discovery using deep learning have made it possible to speed up identifying and developing new pharmaceuticals [4]. Various drugs, such as Arbidol, remdesivir, and favipiravir, have been tested to cure COVID-19 patients and many others are in the testing phase [4]. Biomedical researchers are investigating drugs for treating the patients, with an attempt to develop a vaccine for preventing the virus [5]. On the other hand, computer scientists have developed early detection models for COVID-19 from CT scans and X-ray images [5]. These techniques are a subset of deep learning and have been applied successfully in various fields [5]. Over the past few years, a significant increase in the quantity of biomedical data has resulted in the emergence of new technologies such as parallel synthesis and HTS (high-

throughput screening), to mining large-scale chemical data [6]. Since COVID-19 is transmitted from person to person, electronic devices based on artificial intelligence may play a crucial role in preventing the spread of this virus. With the expansion of the role of health epidemiologists, the pervasiveness of electronic health data has also increased [7]. The increasing availability of electronic health data provides a massive opportunity for healthcare to enhance healthcare for both discoveries and practical applications [7]. For training machine learning algorithms, these data can be used to improve their decision-making in terms of disease prediction [7].

As the increase in the number of cases infected by coronavirus rapidly outnumbered the medical services available in hospitals, a significant burden on healthcare systems was imposed [7]. Because of the limited supply of hospital services and the delay in time for diagnostic test results, it is common for health professionals to provide patients with sufficient medical care. However, since the number of cases tested for coronavirus is growing increasingly day by day, testing is not feasible due to time and cost factors [7]. This paper aims at suggesting a technique based on deep learning which would be helpful in rapidly finding the drugs for combating the pandemic. Deep learning is currently an area that is quickly emerging and constantly expanding. To optimize its performance, it programs computers using data. Using the training data or its previous encounters, it learns the parameters to optimize the computer programs. It can also forecast the future using the data. Deep learning also lets us operate the statistics of the data to construct a mathematical model. The main goal of deep learning is that it learns without any human intervention from the feed data, and it automatically learns from the data (experience) provided and gives us the desired output where it searches the data trends/patterns [8]. Deep learning techniques have achieved greater efficiency in various tasks, including drug development, prediction of properties, and drug target forecasting. As drug development is a complex task, the deep learning approach makes this process faster and cheaper.

The challenges with COVID-19 at present make it necessary to look for some alternatives in medicine or drugs to combat the rise of cases due to COVID-19 infection. One of the significant challenges is the processing delay for the finalization of the drugs for vaccine formulation. However, many pharmaceuticals companies have achieved success to some extent after passing through different trials. Hence, predicting the most probable drugs for the vaccination formulation can speed up vaccine formulation and thus save many human lives. Another challenge is that most of the testing for vaccine formulation is done on a clinical basis where all the drug combinations are tried to get the desired selection of drugs. Still, there is less utilization of computational techniques for the same at present. Thus, there is an hour to look after some alternatives using some machine intelligence techniques to provide some solutions with more accuracy and at a faster note.

Based on the above challenges, the main contributions of the paper are as follows:

(1) Deep learning approach based on logistic regression, SVM, and Random Forest along with QSAR modeling is proposed to discover some drugs for the treatment of COVID-19

(2) QSAR modeling is done to find the drug targets with protein interaction along with the calculation of binding affinities

(3) Deep learning models are used for training the molecular descriptors dataset for the robust discovery of drugs and feature extraction for combating COVID-19

The rest of the article is organized as follows. Section 2 deals with the literature reviewed. Section 3 deals with the significance of work. Section 4 deals with the suggested methodology followed by Section 5, dealing with results, and the paper is concluded in Section 6.

2. Literature Review

Artificial intelligence techniques have been utilized in various areas of drug and vaccine development [9]. This utilization and further advancements are essential for immediately discovering a cure for the current pandemic. Many studies have been done previously, and many are ongoing to find a less complex and easy-to-use technique that would speed up the drug discovery process. In [10], the authors have trained a model based on LSTM (long short-term memory) for reading the SMILE fingerprints of a molecule for predicting IC50, binding to RdRp. The authors in [11] have suggested a B5G framework, which supports the diagnosis of COVID-19 through low latency and 5G. Choi et al. [12] proposed the MT-DTI model for predicting the drugs approved by FDA having solid affinities for the ACE2 receptor with TMPRSS2. The authors in [13] have reviewed all state-of-the-art research studies related to medical imaging and deep learning. Deep learning techniques and feature engineering were compared in order to efficiently diagnose COVID-19 from CT images [14]. Various neural network architectures and generative models such as RNN, autoencoders with adversarial learning, and reinforcement learning are suggested for ligand-based drug discovery [15]. Classification performance of DNN on imbalance compound datasets is explored by applying data balancing techniques in [16]. A novel approach for deep docking large numbers of molecular structures accurately is suggested in [17]. The effects of deep learning in drug design and complimentary tools were reviewed [18].

In [19], a systematic review of the application of deep learning techniques for predicting drug response in cancer cell lines has been done. A QSAR model (quantitative structure-activity relationship) is developed [20], which implements deep learning to predict antiplasmodial activity and cytotoxicity of untested compounds for screening malaria. In [21], the authors have built a multitask DNN model and compared the results with a single-task DNN model. In [22], various machine learning and deep learning algorithms used for drug discovery are reviewed, and their applications were discussed. However, various studies

suggest deep learning for drug discovery or detecting COVID-19 lacks proper practical implementation with results. Most studies have just reviewed different deep learning techniques to be used for the development of drugs. This paper will give a practical implementation on various datasets available online with efficient results. Upon analyzing various studies, we found that various studies claim HCS (high content screening) as an efficient technique for screening chemical compounds for discovering drugs. At present, deep learning techniques have been producing faster and efficient results.

The basic idea of the screening process is that the cells are exposed to various compounds, and automated optical microscopy is done to see what happens, creating thriving images of cells. A quantitative and qualitative analysis of the result can be done by using an automated HCS pipeline. HCS branches out from microscopy, and Giuliano et al. first coined the terminology in the 1990s [23]. HCS research can cover several fields, such as discovering drugs that can be defined as a form of cell phenotypic screen. It includes methods of analysis that produce simultaneous readouts of multiple parameters considering cells or cell compounds. In this phase, the screening aspect is an early discovery stage in a series of various steps needed to identify new drugs. It acts as a filter to target potential applicants that can be used for further development. Small molecules classified as a low molecular weight organic compound, e.g., proteins, peptides, or antibodies, can be the substances used for this purpose [24].

3. Significance of the Work

Hospitals are using trial and error techniques for COVID-19 drug discovery [9]. It results in an emergence of virtual screening to discover chemical compounds due to the inefficiency of the lab-based HTS technique (high-throughput screening) [9]. Also, drug discovery and development is a complex and time-consuming process [25]. It is estimated that the preapproval cost of production of new drugs has increased at the rate of 8.5% annually from 802 million USD to 2870 million USD [26, 27]. Finding molecules with the required characteristics is one of the significant challenges in drug discovery. A practical and quality drug needs to be balanced regarding safety and potency against its target and other properties such as ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) and physicochemical properties [25]. This paper aims to increase the speed of discovering new molecules using deep learning, thereby reducing the cost of producing new drugs. Deep learning techniques will help us navigate large chemical spaces to find new chemical compounds [25]. The significance of using deep learning techniques for combating COVID-19 [1] is summarized in Table 1.

4. Suggested Methodology

This section includes a description of the proposed methodology.

4.1. Dataset Preparation and Preprocessing. We have used the combination of the datasets from the sources [29–31]. Each of the datasets contains a set of chemical compounds with respective binding activity to a target protein calculated by $\mathrm{pIC50} = -\log_{10}(\mathrm{IC}_{50})$ [32]. Preprocessing is done for removing the invalid and replicated compounds. The entries with IC50 measurements with filtered out compounds having suspicious measures are depicted by the "DATA VALIDITY COMMENT" column. For repeated records groups, if the standard deviation (SD) of the activity is found more significant than 1 log unit, then these datasets are deleted from the dataset, and a single entry is kept with the median of the activity [32]. Data preprocessing is one of the significant phases in data mining as it helps in achieving data integrity. Before preprocessing, data cleaning needs to be done as raw data contain abnormalities and errors affecting the results [33]. After preprocessing, conversion of SMILES [34] representations to molecular representations takes place. These are open datasets that contain the binding, ADMET, and functional information for various drugs like bioactive compounds [35]. The database containing the datasets has over 5 million bioactivity measurements for over 1 million compounds and over 5000 target proteins [35].

A minor challenge may occur in data mining algorithms due to variation in range and distribution of every variable in the large datasets due to distance measurements; also, these may contain noisy variables, which makes the learning of the algorithms more difficult [33]. These challenges can be handled by min-max normalization where the value of each variable is adjusted in a uniform range of 0 to 1 [33]. It is given in the following equation:

$$Y_{\text{normalised}} = \frac{Y_x - Y_{\text{minimum}}}{Y_{\text{maximum}} - Y_{\text{minimum}}}, \tag{1}$$

where $Y_{\text{normalised}}$ is the normalised value, Y_x is the value of interest, Y_{minimum} is the minimum value, and Y_{maximum} is the maximum value.

Apart from the dataset, the system used for performing the experiments has UBUNTU 20.04 LTS OS installed with 16 GB RAM and Intel Core i7-8700 processor. The language used for building the model is Python 3.7 with NumPy, pandas, TensorFlow, Bunch, tqdm, Matplotlib, scikit-learn, NVIDIA GPU, CUDA 9.0, Pytorch 0.4.1, Mordred, and RDkit. For evaluating the binding affinities, PyRx is used. We have used the regression model and QSAR techniques as regression models help us define relationships between dependent and independent variables and show the strength of the impact of various independent variables on dependent variables. QSAR helps in maintaining the quantitative structural relationships in molecular predictions.

4.2. Model Development and Evaluation Parameters. As mentioned above, developing a QSAR model can help us in defining the relationship between the chemical

Table 1: Summary of applications of deep learning for combating COVID-19.

S. no.	Application	Explanation
1	Pandemic tracking [1]	(i) Bidirectional GRU along with attentional techniques are used for analyzing patterns in respiratory images for mass scale screening of COVID-19 (ii) Application of deep learning (DL) techniques for identification of geographical hazards and spreading at the community level
2	Predicting the structure of proteins [2]	(i) CNN, DNN, and deep ResNet architecture are utilized for the identification of characteristics of proteins (ii) Virus-host prediction and early prevention of virus infectivity can be done using DL architectures
3.	Drug discovery [25]	(i) GAN and reinforcement learning techniques should be implemented for discovering the chemical compounds inhibiting COVID-19
4.	Medical imaging[28]	(i) DL architecture should be used for extraction of features and prediction of possible cases of COVID-19 from CT scan or chest X-ray images

structures and their endpoints by using various statistical methods for the construction of predictive models for revealing the origin of bioactivity [36]. Generally, a QSAR model is depicted by the equation of the form $X = m(X) +$ Err that can be utilized or prediction of endpoints or new compounds in terms of time-consuming and cost approaches. In order to derive the global molecular features for the SMILES, some notations are there [36], which are given in the following equation:

$$\begin{aligned} pqrstu &\rightarrow p + q + r + s + t + u\left(X_m\right), \\ pqrstu &\rightarrow pq + qr + rs + st + tu\left(XX_m\right), \\ pqrst &\rightarrow pqr + qrs + rst + stu\left(XXX_m\right). \end{aligned} \tag{2}$$

Also, these global descriptors are described as follows [36]:

(1) BOND is defined as the presence or absence of double (=), triple (#), and stereochemical (@) bond in SMILES

(2) PAIR is defined as the coincidence of I, N, O, P, S, Br, Cl, F, #, @, and =

(3) NOSP is defined as the presence or absence of P, S, O, and N

(4) HALO is defined as the presence and absence of halogens

The optimal attributes for the SMILES are calculated by the following equation [36]:

$$\begin{aligned} W\left(X_{\text{epoch}}, \text{Threshold}\right) == &\sum TW\left(Xm\right) + \sum TW\left(XXm\right) \\ &+ \sum TW\left(XXXm\right) + \sum TW\left(\text{NOSP}\right) \\ &+ \sum TW\left(\text{BOND}\right) + \sum TW\left(\text{HALO}\right) \\ &\cdot \sum TW\left(\text{PAIR}\right). \end{aligned} \tag{3}$$

The chemical endpoints [36] can be given in the following equation:

$$\text{End} = T_0 + T_1 \times W\left(X_{\text{epoch}}, \text{Threshold}\right), \tag{4}$$

where T_0 is the intercept and T_1 is the correlation coefficient.

The development of the QSAR model consists of two significant steps: (i) describing the molecular structure and (ii) the multivariate analysis for correlation of molecular descriptors with observable characteristics [33]. Successful development of the model also includes data preprocessing and statistical evaluations. For evaluating the performance of the QSAR model, the statistical method suggested in [33] is used in the following equation:

$$\begin{gathered} x^2 > 0.5, \\ Y^2 > 0.6, \\ \frac{Y^2 - Y_0^2}{Y^2} < 0.1, \\ \text{or } \frac{Y^2 - Y_0''^2}{Y^2} < 0.1, \\ 0.85 \le z \le 1.15, \\ \text{or } 0.85 \le z'' a \le 1.15, \end{gathered} \tag{5}$$

where x^2 is the cross-validated explained variance, Y^2 is the coefficient of determination, Y_0^2 and $Y_0''^2$ are the predicted vs. observed activities and vice versa, respectively, and x^2 is calculated by the following equation:

$$X^2 = \frac{\sum_{j=1}^{\text{training}}\left(P_j - \widehat{P}_j\right)2}{\sum_{j=1}^{\text{training}}\left(P_j - \widehat{P}\right)2}, \tag{6}$$

where P_j are the measured values, $\widehat{P}_j$ are the predicted values, and P_j is the mean value of the entire dataset. This equation is also used for the calculation of external x^2, i.e., the compounds that are not used in the QSAR model development earlier and are given in the following equation:

$$X_{\text{external}^2} = 1 - \frac{\sum_{j=1}^{\text{training}}\left(P_j - \widehat{P}_j\right)2}{\sum_{j=1}^{\text{training}}\left(P_j - \widehat{P}_j\right)2}. \tag{7}$$

For measuring the internal chemical diversity [28], let x and y be two molecules having Z_X and Z_Y as their Morgan fingerprints [28]. The number of common fingerprints is

defined as $Z_x \cap Z_y \vee$ and the total number of fingerprints is defined as $Z_x \cup Z_y \vee$. The Tanimoto similarity [28] between x and y is defined in the following equation:

$$S(x, y) = \left| \frac{Z_x \cap Z_y}{Z_x \cup Z_y} \right|. \tag{8}$$

And the Tanimoto distance [28] is given by

$$S_d(x, y) = 1 - S(x, y). \tag{9}$$

We have used RDKit [28] for the implementation of Tanimoto distance. In earlier studies, the QSAR models were developed for small compounds that used limited quantitative characteristics [32]. Various algorithms were suggested for covering significant features, including hundreds or thousands of molecular descriptors. We have used the OPLRAreg algorithm suggested in [32] to illustrate the flexibility of mathematical modeling and show how the division of characteristics and regions helps enhance the features of OSAR datasets. The OPLRAreg is given in Algorithm 1.

Due to advancements in deep learning techniques, there has been an increase in the use of neural networks in a variety of applications including healthcare [25]. A neural network can be defined as a group of layers consisting of perceptrons called multilayer perceptron (MLP) or simply a neuron [25]. The perceptrons are the main building blocks of a perceptron and consist of three parts, weights, $v = [v_1, v_2 \ldots\ldots, v_n], v_j \in R$, biases, $b \in R$, and an activation function, $f(n)$ [25]. Let the input vector given to a perceptron be defined as, $x = [x_1, x_2 \ldots\ldots, x_n]^{Q}$. Then, the output is given in the following equation:

$$f(vx + b) = f. \tag{10}$$

Both v and x should be in the same direction. Furthermore, for enabling the matrix multiplication, b and x_1 should be appended to the weight and input vector, respectively [25] so that $v = [v_1 v_2 \ldots v_n b]$ and $x = [x_1 x_2 \ldots x_n 1]^{Q}$.

And the output is given by

$$f(vx) = f\left(v_1 x_1 + v_2 x_2 + \cdots + v_n x_n + b\right). \tag{11}$$

Due to an increase in the efficiency of computation, matrix multiplication is required for training larger networks with forward passing and backpropagation for optimizing the network parameters [25]. The different types of classification methods are given in the following sections.

4.2.1. Logistic Regression. Logistic regression is the most used method of modeling for the prediction of risk [37]. A logistic regression model uses a role to render the model range output between zero and one and should therefore be used for classification. The logistic function is defined in [37] as follows:

$$Y(x = 1) = \frac{1}{1 + \exp - (\alpha r + s)}, \tag{12}$$

where r is the input and α and s are called as model parameters. The output given is the modeled probability of the input belonging to a class [37]. For interpreting the meaning of the weights, rearrange the above equation as follows [37]:

$$\log)\log)\ \alpha r + s. \tag{13}$$

$Y(x = 1)/Y(x = 0)$ is called as the odds. The modeling of odds is done through a linear equation [37]. Like most of the ML (machine learning) models, optimization of the parameters is done w.r.t. loss function [37]. Consider a given set of data points $\left\{(p_j, q_j)\right\}_j$, where p_j is defined as the input and q_j is the true output. Let $\widehat{q_j}$ denote the output of the logistic regressor. Then $\alpha \wedge s$ are selected according to [37] in the following equation:

$$\begin{gathered} \alpha^*, \\ s^* \operatorname{argmin}_{\alpha s}. \end{gathered} \tag{14}$$

This is also known as the log-loss function. The problem of minimization is solved iteratively until the convergence of parameters, using a coordinate descent algorithm [37].

4.2.2. Random Forest. Random Forest is an ensemble approach that combines several decision trees to make predictions. More reliable and precise predictions can be made by combining several poor learners. In addition, ensemble techniques decrease variance and are less vulnerable to overfitting [37]. The Random Forest algorithm [38] is given in Algorithm 2.

As a sequence of questions, a decision tree is best defined. The principle is that questions are asked, and new questions are asked based on the responses, thus creating a tree. Data points are identified using the leaf nodes in the tree [37] by following the trajectory of the questions and answers. The tree is designed by determining which question to ask at each node and determined based on the information obtained from each possible query or the degree to which the uncertainty in the dataset [37] is reduced. The uncertainty in the dataset [37] is defined in the following equation:

$$\text{Entropy}(X) = -\sum_{|XzX|} y(X) \log_2\ y(X). \tag{15}$$

The information is acquired by knowing the value of certain feature F and is given in the following equation:

$$\text{Gain}(F) = \text{Entropy}(X) - \sum_{z \in \text{values}} \frac{|Xz|}{|X|} \text{Entropy}\left(X_z\right), \tag{16}$$

where X_z is defined as the subset where the feature F takes z value. Therefore, during the construction of a decision tree, a feature is to decide each node as explained in [37]. Here, the construction is either terminated once the entropy of the subset has reached zero or the tree has reached its maximum depth [37]. Upon evaluation of a sample, the tree's trajectory is decided until the leaf node is reached. An approximate probability can also be given as output by comparing the class sizes found in the leaf node [37].

(1) OPLRAreg is evaluated for $P = 1$//linear regression
$x \longrightarrow$ currenterror
olderror $\longrightarrow \infty$
temperror $\longrightarrow \infty$
$Y_{\text{best}} \longrightarrow$
(2) $Z \longrightarrow \{Y \in y \vee F_{q_1,Y} \neq 0\}$//choosing implicit features
$P \longrightarrow 2$
(3) **For** $j \longrightarrow 1; j \longrightarrow j+1; j \leq Z$ **do**//choose the best partition feature in the region
(4) Evaluate OPLRAreg having two regions and partition feature y_f
(5) **If** $k <$ temperror **then**
(6) $k \longrightarrow$ temperror
(7) $Y_{\text{best}} \longrightarrow y_f$
(8) **End if**
(9) End for
olderror$\leftarrow$currenterror
temperror $\longrightarrow$ currenterror
(10) $Y^* \leftarrow Y_{\text{best}}$
(11) **While** currenterror $< (1-\alpha)$olderror **do**//increase the number of regions
$P + 1 \leftarrow P$
(12) Evaluate OPLRAreg with P regions and partition feature y_f
(13) olderror$\leftarrow$currenterror
(14) $k \longrightarrow$ currenterror
(15) End While
(16) **Return** y_f, Breakpoints as $B_{\text{q y}}$, and regression coefficients as $F_{q_1,Y}$

Algorithm 1: OPLRAreg algorithm.

for $j = 1$ to X **do**
generation of random samples Φ
while stopping criteria $\neq$ true **do**
select randomly f of all features
training of tree on f
end
$f_{\text{RF}}(n) = (1/X)\sum_{j=1}^{X} f_{\text{Tree}}(n \vee \Phi)$

Algorithm 2: Random Forest algorithm.

4.2.3. Support Vector Machine (SVM). The support vector machine (SVM) is an algorithm for classification that involves creating a hyperplane. A set of features is used in order to classify an object. Thus, the hyperplane will lie in p-dimensional space if there are p features [39]. The hyperplane is generated through SVM optimization, which in turn maximizes the distance from the nearest points, also known as support vectors [39]. Let $y_j = [y_{j1}, \ldots, y_{jm}]^N$ be an arbitrary observation feature vector in the training set, $\underline{x_j}$ corresponding label to y_j, with a weight vector $v = [v_1, \ldots\ldots, v_q]^N$ with $\vee v \vee^2 = 1$ and T be the threshold. The constraints defined for the classification problem [39] are given in equations (17) to (20):

$$vNy_j + T > 0 \text{ for } \underline{x_j} = +1, \tag{17}$$

$$vNy_j + T < 0 \text{ for } \underline{x_j} = -1. \tag{18}$$

Let $f(y_j) = V^N Y_j + T$, then the output of the model $\widehat{x_j}$ can be given as follows:

$$\widehat{x_j} = \{1 \text{ for } f(y_j) \geq 0, \{0 \text{ for } f(y_j) < 0. \tag{19}$$

Instead of using $\|v\|^2 = 1$, for margin maximization, the lower bound on the margin along with the optimization problem can be defined for minimization of $\|v\|^2$ [39]. The constraints for the optimization problem can be derived from equations (17) and (18), respectively, [39] as follows:

$$\widehat{x_j} vNy_j + T \geq 1. \tag{20}$$

In some of the cases, it is required to implement a soft margin, allowing some points to lie on the wrong side of the hyperplane [39] in order to provide an efficient model. A cost parameter M is introduced, which plays a major role in the assignment of penalties to errors, where $M > 0$ [39]. Then, the minimized objective function [39] is defined as follows:

$$\vee v \vee 2 + M \sum_j \beta_j, \tag{21}$$

where β_j = slack variable. The constraints to the optimization problems [39] are now modified in the following equation:

$$\underline{x_j} Ny_j + T \geq 1 - \beta_j, \quad \beta_j \geq 0. \tag{22}$$

Most of the datasets are not linearly separable. But through a nonlinear transformation into a high-dimensional space, a dataset is more likely to be linearly separable [37]. Therefore, each sample is transformed using a nonlinear function [37] so that

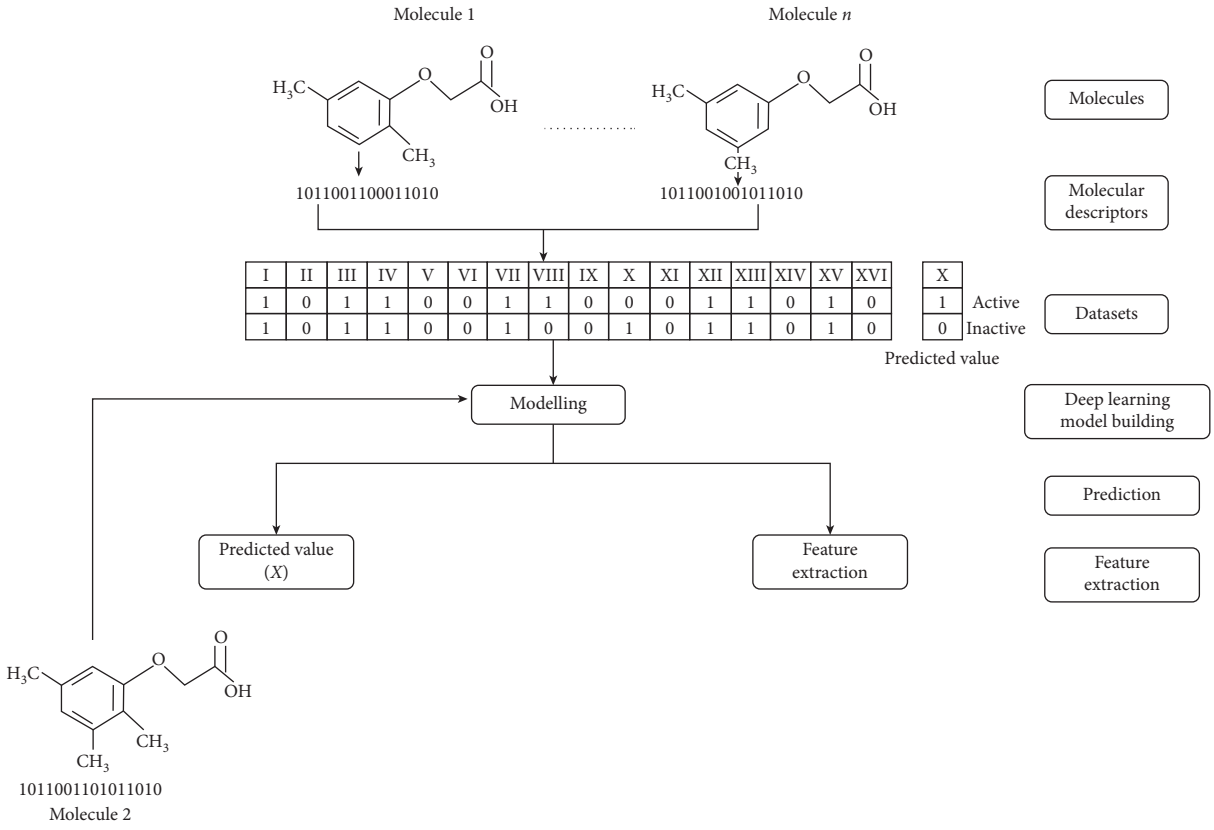

Figure 1: Overall workflow of the suggested methodology.

$$f\colon R^X \longrightarrow R^Y, \quad x > y. \tag{23}$$

And then the problem is considered using $m_j = f(y_j)$ [37]. Furthermore, using Lagrange optimization, the dual problem of maximizing [37] is defined as follows:

$$\sum_j \left[\delta_j - \frac{1}{2}\sum_i \delta_j \delta_i T_j T_i \lambda y_i\right], \quad \lambda = yjt, \tag{24}$$

subject to the condition

$$\sum_j \delta_j T_j = 0, \quad \delta_j \geq 0 \forall j. \tag{25}$$

The overall structure of the workflow and QSAR modeling [36, 40] is explained in Figure 1. First, we have to select the number of molecules. It can be of any number. Each molecule has its molecular descriptors that describe the molecules' physical and chemical properties that help us differentiate between the molecules. Here, 1 and 0 are the binary descriptors that show the presence/absence of the molecular descriptors. A collection of these descriptors constitutes the dataset. Values of X (active/inactive) show the biological activity we want to predict. This dataset is now used for training the deep learning model, which therefore gives our results. The working of the proposed approach is represented in a flowchart, as depicted in Figure 2.

5. Results

Our goal is to develop a deep learning model to suggest novel and effective drugs for combating SARS-CoV-2 or combating COVID-19. Our regression-based models and Random Forest model were trained on a dataset of approximately 1.5 million drug-like molecules from the data sources [29–31]. The molecules were represented in Simplified Molecular Input Line Entry System (SMILES) format helping our model learn the required features for designing novel drug-like molecules. SMILES are defined as the character strings for representing drug molecules. For example, an atom of carbon can be represented as C, oxygen atom as O, double bond as =, and CO_2 molecule can be represented as C(=O)=O. The maximum length of the string can be taken as 25 [41]. SMILES grammar's learning problem and reproducing it for generating novel small molecules is considered a classification problem [42]. The SMILES strings should be considered a time series, where every symbol is considered a time point. At a given point, the model was trained for predicting the class of the next symbols in the time series.

We will only retrieve the coronavirus proteinase during preprocessing of the bioactivity data that can be reported as IC50 values in nM (nanomolar) units [43]. The data for bioactivity is in the IC50 unit. Compounds with less than 1000 nM values will be considered active, whereas compounds with values greater than 10,000 nM will be considered inactive. As for such values, the intermediate value is

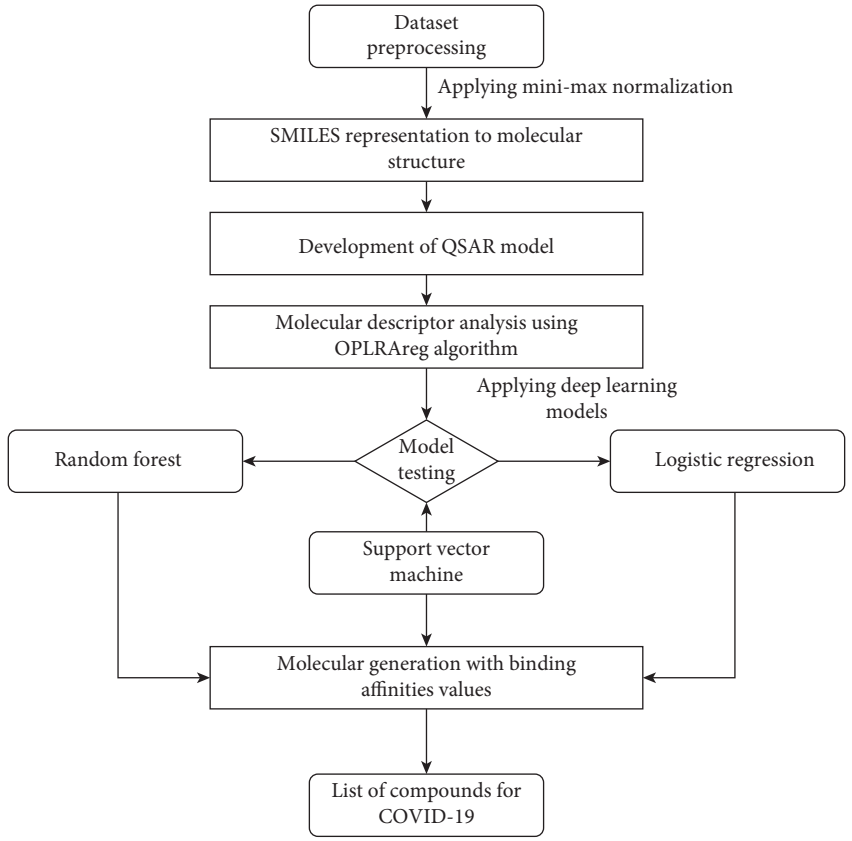

Figure 2: Flowchart depicting the complete working of the proposed approach.

Table 2: Calculated values of Lipinski descriptors.

MW	Log*P*	NumH donors	NumH acceptors
281.3	1.90	0.0	5.0
416.5	3.82	0.0	2.0
422.2	2.67	0.0	3.0
294.3	3.63	0.0	4.0
339.3	3.54	0.0	5.0
338.4	3.41	0.0	5.0
297.0	3.45	0.0	3.0
277.2	4.10	0.0	3.0
278.3	3.30	0.0	3.0
282.4	4.11	0.0	2.0

between 1,000 and 10,000 nM [43]. To evaluate the model, Lipinski descriptors [43] were used as given in Table 2.

Upon analyzing the pIC50 values, the actives and inactives have shown a significant difference, which is expected as the values of IC < 1000nM = active, IC50 > 10000nM = inactive, corresponding to pIC50 >6 = active and pIC50 <5 = inactive. Out of the 4 Lipinski descriptors [43], only log*P* showed no difference between the actives and inactives, while the other three descriptors showed significant differences between the actives and inactives. This can be better understood by Figures 3–7 , respectively. A scatter plot has also been drawn to show that the two bioactivity classes (active/inactive) are spanning similar chemical spaces.

Figures 3–7 show that our model can explore the chemical spaces that are further adapted for generating the smaller molecules specific to a target of interest. The SARS-CoV-2 contains the proteins responsible for the cation and replication of the virus [44]. The functioning of the proteins can be stopped by introducing the drug molecules capable of blocking the protein. Therefore, we have to find the molecules with a high binding affinity to bind the protein effectively. Various drugs/compounds have been tested for finding a high binding relationship, but the results are not very good. We have created novel molecules for binding with the coronavirus, using deep learning and QSAR modeling. After the generation of the molecules, PyRx was used for evaluating the binding affinities. We have also build a regression model using a Random Forest algorithm for acetylcholinesterase inhibitors, as shown in Figure 8. The binding affinities for

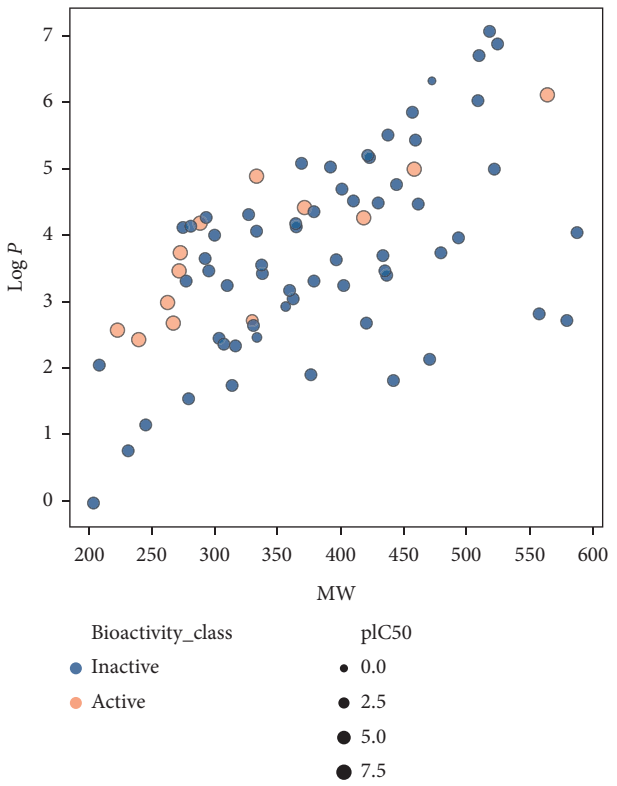

Figure 3: Scatter plot of MW vs. log*P*.

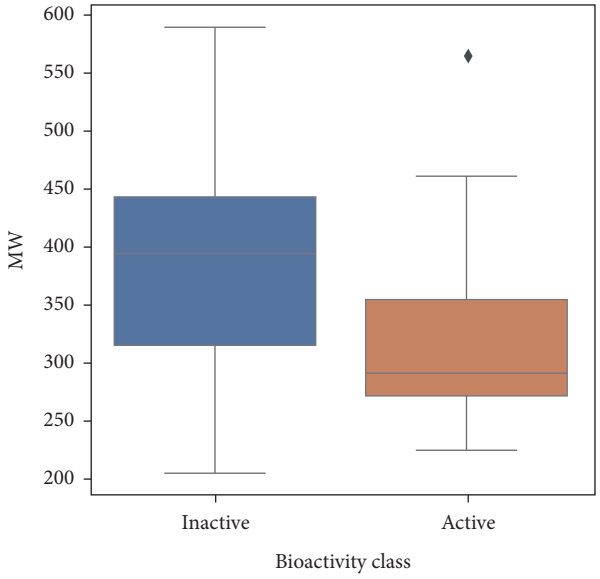

Figure 4: Box plot of MW.

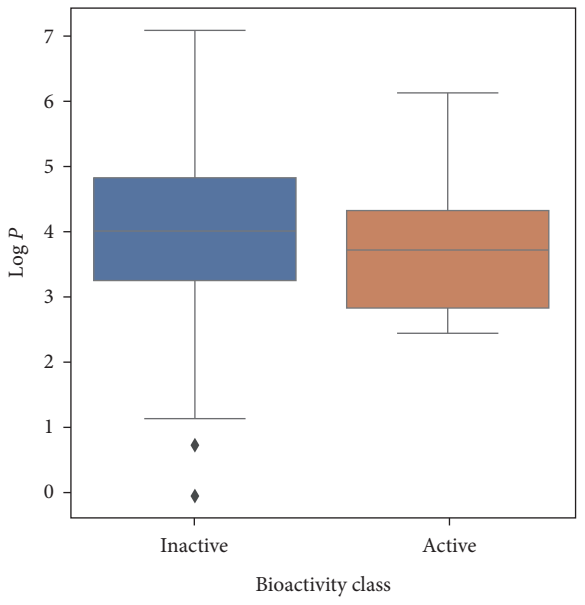

Figure 5: Box plot of log*P*.

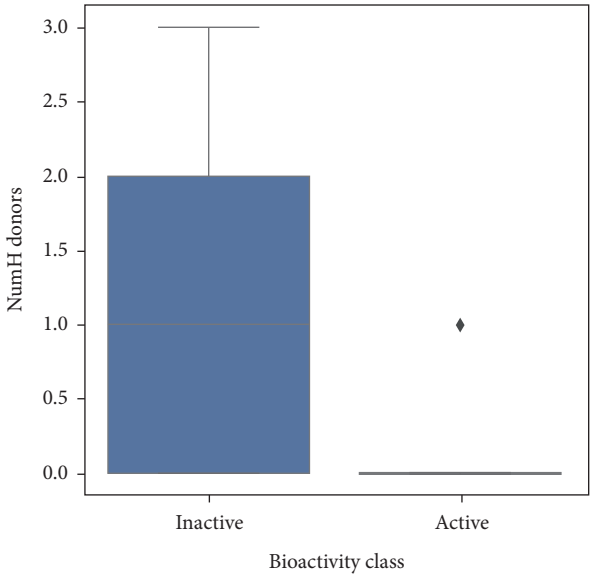

Figure 6: Box plot of NumH donors.

leading drugs for other diseases such as HIV inhibitors range from −10 to −11. Also, the most recent drug remdesivir, which is clinically tested, has the binding affinity of −13. By convention, the more negative the scores are, the more effective the drugs would be. QSAR modeling, docking analysis, and use of regression model generate a list of bioactive compounds from which top 100 compounds were selected, which may have the potential to be effective against SARS-CoV-2. The methodology suggested in this paper is easy to use and can be a possible technique for the discovery of anti-COVID-19 drugs and also shortening the clinical development period required for drug repositioning. Our proposed methodology can give the binding affinity more than the present drugs being tested, making our approach efficient. The proposed list of top 100 chemical structures or molecules generated using our proposed approach through SMILES software is shown in Table 3.

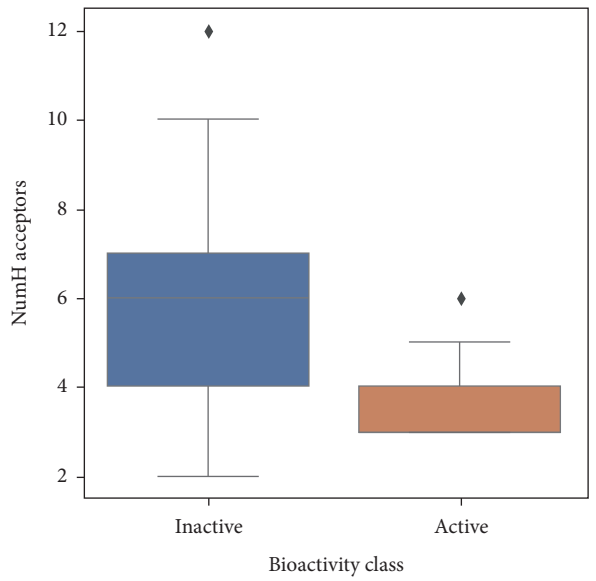

Figure 7: Box plot of NumH acceptors.

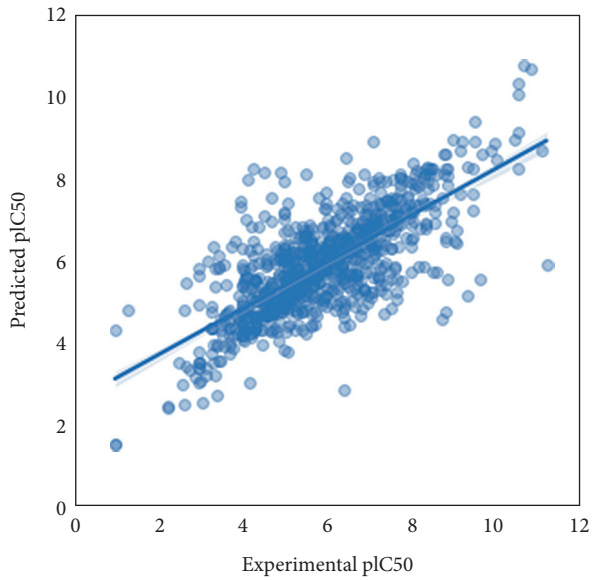

Figure 8: Scatter plot for experimental vs. predicted values of pIC50 for regression model developed for acetylcholinesterase inhibitors.

Table 3: Top 100 compounds generated using the proposed approach.

Serial no. of the chemical structure generated	SMILES generated chemical structure generated through the proposed approach	Binding affinity value (kcal/mol)
1	Cc1ccc(C2CNCCN2C)cc1	−23.1
2	CCOC(CO)c1ccccc1	−15.2
3	CC(=O)Nc1cnn(C)n1	−24.6
4	CCC(C)NCc1ncccn1	−21.5
5	CC(C)=C1CC(N)C1	−20.4
6	CN1CCCc2cc(CON)ccc21	−18.9
7	CC12CNCC1CN(CC(N)=O)C2	−28.9
8	CCNC(C)C(C)c1cnccc1C	−19.5
9	CCN(Cc1ccccc1)C(C)CCCNC	−18.1

Table 3: Continued.

Serial no. of the chemical structure generated	SMILES generated chemical structure generated through the proposed approach	Binding affinity value (kcal/mol)
10	CCC(=O)c1cc(C)ccn1	−18.3
11	C=CC(O)c1cc(C)ccn1	−21.5
12	C#CCCOc1cnccc1C	−16.8
13	Cn1nc2ccccc2c1S(N)(=O)=O	−19.8
14	Cn1cnn(CC(N)=O)c1=O	−23.1
15	CC(NCCSc1ccccc1)c1ccncc1	−21.6
16	Cc1ccsc1-c1ccc(O)nc1	−21.9
17	N#Cc1ncccc1N1CC2CC1CN2	−19.6
18	N#Cc1cnccc1SCC(N)=O	−23.6
19	N#Cc1ccc(C2NCCCCC2=O)cn1	−23.5
20	CC(C)C(C)Sc1ccc(C#N)cn1	−18.6
21	Cc1ccnc(C=CCCN)c1	−24.2
22	CCOC(CC)C(=O)c1cnccc1C	−15.9
23	Cc1ccncc1C(O)CNCC(C)C	−22.2
24	CS(=O)(=O)c1ncc(N)cn1	−21.1
25	OCC(O)CCSCc1ccccc1	−19.8
26	COC(=O)CNCc1cc(C)ccn1	−19.5
27	CCOC(c1ccccc1)C(CC)NN	−18.0
28	Cc1ccncc1C(=O)CCCN(C)C	−19.3
29	C=CCCSCCNc1cc(C)ccn1	−21.2
30	CCNC(=S)NNC(=O)Cc1ccccc1	−23.6
31	OC(CCCc1ccccc1)c1cccnc1	−20.4
32	CC(=O)CC(C)c1cnccc1C	−17.3
33	CN1CCC(O)(c2ccoc2)CC1	−18.1
34	Cc1ccnc(NC(=O)C#CCN)c1	−24.1
35	N#Cc1cnccc1NCCCO	−21.0
36	CCSCc1cncc(C#N)c1	−19.4
37	NC1=CCOC1=O	−16.4
38	CNC(CSC1CCCCC1)Cc1cccnc1	−18.7
39	COC(=O)c1ccc(C(C)C=O)cc1	−14.3
40	CC(=O)CC(O)c1cnccc1C	−21.0
41	CCCNCc1ccccc1S(N)(=O)=O	−20.8
42	N#Cc1nccnc1N1CCCOCC1	−22.0
43	CCC(CC)Oc1ncccc1C#N	−16.8
44	CC(C)(C)C(C)(N)c1ccccc1	−17.0
45	CN(C)NCc1ccccc1	−20.0
46	NC12CCCC1CNC2	−24.3
47	C(=Cc1ccccc1)CNCc1cccnc1	−23.8
48	CCNCCNc1ncccc1C#N	−26.6
49	CC(C)OCc1ccc(C#N)cn1	−18.3
50	NC1Cc2csnc2C1	−26.5
51	Cc1ccsc1C1NCCCCC1O	−21.3
52	N#CCCNCc1cncnc1	−20.8
53	COC(=O)c1ccccc1C#CCO	−15.5
54	N#CC1CN(CCN)C(=O)O1	−19.4
55	CC(CCO)Nc1ccc(C#N)cn1	−22.6
56	NC1CC2(CCNC2=O)C1	−21.8
57	C#CC(CO)NCc1cnccc1C	−22.6
58	CN1CCCc2cccc(OCC#N)c21	−16.2
59	NNC(c1ccncc1)C1CCCCC1	−23.8
60	C#CCCSc1ncccn1	−17.2
61	Cc1ccncc1C(C)(N)C(C)C	−22.6
62	NS(=O) (=O)c1ccc(SCCO)cc1	−21.0
63	Cc1ccnc(CC(=O)C(=O)O)c1	−18.8
64	CN1CC2CCN(CC(N)=O)C2C1	−25.9
65	O=C=NCc1ccncn1	−20.9
66	Cc1cscc1C1CC(O)CN1	−19.2
67	O=C(CC1CCCCC1)NC1CCCNCC1	−22.4

Table 3: Continued.

Serial no. of the chemical structure generated	SMILES generated chemical structure generated through the proposed approach	Binding affinity value (kcal/mol)
68	CC(O)Cc1cncnc1	−20.8
69	CCC(CC)Oc1ccc(C#N)cn1	−16.1
70	Cc1ccnc(NN=CC(C)C)c1	−19.7
71	COC(CNCCCOCc1ccccc1)OC	−12.3
72	N#Cc1ncccc1C1CCCCC1	−18.3
73	NC1COC2COCC12	−19.9
74	COC(=O)c1ccccc1C=CCCO	−18.6
75	CCCC(C)Sc1ncccn1	−16.9
76	CC(C)CC(=O)NCCCc1ccccc1	−16.8
77	CCC(CC#N)Nc1ccc(C#N)cn1	−21.8
78	CCCC(C)C(=O)c1cc(C)ccn1	−19.0
79	CCOc1cncnc1	−18.6
80	NCCCCC(O)c1ccccc1	−21.0
81	N#CCNc1ccncc1C#N	−21.6
82	N#Cc1cnccc1NCC=CCN	−27.2
83	CCCOCC(NC)c1cc(C)ccn1	−18.6
84	Nc1ccc(S(N)(=O)=O)cc1	−22.4
85	c1cncc(OCCNC2CCCCC2)c1	−20.8
86	CSCC(C)CNc1ncccc1C#N	−21.1
87	CC(N)CNc1cncnc1	−26.8
88	CC(C)(N)CNC(=O)Cc1ccccc1	−22.2
89	NC(CO)c1ccncn1	−26.9
90	CC(=O)OCSc1ncccn1	−19.3
91	CN1CCCc2cccc(C=O)c21	−16.4
92	CCNc1cc(NCC(C)(C)O)ccn1	−25.6
93	CCC(CC)CC(=O)COCc1ccccc1	−13.0
94	C=CCCC(=O)OCc1ccccc1	−13.9
95	CN(CCCO)C(=O)Oc1ccccc1	−18.8
96	CSCCC(=O)c1cncnc1	−19.6
97	CC(C)CCCC(O)CCOCc1ccccc1	−13.9
98	COc1ccncc1C#N	−18.1
99	CNc1nc(N)ncc1N	−28.4
100	c1ccc(CONCCNc2ccncc2)cc1	−25.4

6. Conclusion

Drug development is a time-consuming and expensive process. Deep learning has achieved excellent performance in a lot of tasks. Drug discovery is one of the areas that can be benefitted from this. The use of deep learning techniques has made the process of drug development more manageable and cheaper. Deep learning-based models can learn the feature representations based on present drugs that can be used to explore the chemical spaces in search of more drug-like molecules. The available data for automating the processes and better predictions are what deep learning techniques promise for efficient drug discovery. These techniques have proven effective in scanning peptides or detecting COVID-19 from the CT scan or X-ray images. These techniques can speed up the drug development process but require clinical testing for more validation and accuracy [45].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] A. Asraf, M. Z. Islam, M. R. Haque, and M. M. Islam, "Deep learning applications to combat novel coronavirus (COVID-19) pandemic," *SN Computer Science*, vol. 1, no. 6, pp. 363–367, 2020.

[2] X. Zeng, X. Song, T. Ma et al., "Repurpose open data to discover therapeutics for COVID-19 using deep learning," *Journal of Proteome Research*, vol. 19, no. 11, pp. 4624–4636, 2020.

[3] S. Pushpakom, F. Iorio, P. A. Eyers et al., "Drug repurposing: progress, challenges and recommendations," *Nature Reviews Drug Discovery*, vol. 18, no. 1, pp. 41–58, 2019.

[4] K. Arora and A. S. Bist, "Artificial intelligence based drug discovery techniques for covid-19 detection," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 2, no. 2, pp. 120–126, 2020.

[5] T. T. Nguyen, "Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions," 2020, https://arxiv.org/abs/2008.07343.

[6] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discovery Today*, vol. 23, 2018.

[7] D. M. Matta and M. K. Saraf, "Prediction of COVID-sing machine learning techniques," Dissertation, Blekinge Institute of Technology, Karlskrona, Sweden, 2020, http://urn.kb.se/resolve?urn=urn:nbn:se:bth-20232.

[8] X. Yang, S. Nazir, H. U. Khan, M. Shafiq, and N. Mukhtar, "Parallel computing for efficient and intelligent industrial internet of health things: an overview," *Complexity*, vol. 2021, Article ID 6636898, 11 pages, 2021.

[9] A. Keshavarzi Arshadi, J. Webb, M. Salem et al., "Artificial intelligence for COVID-19 drug discovery and vaccine development," *Frontiers in Artificial Intelligence*, vol. 3, p. 65, 2020.

[10] S. Patankar, *Deep Learning-Based Computational Drug Discovery to Inhibit the RNA Dependent RNA Polymerase: Application to SARS-CoV and COVID-19*, Science Open, Berlin, Germany, 2020.

[11] M. A. Rahman, M. S. Hossain, N. A. Alrajeh, and N. Guizani, "B5G and explainable deep learning assisted healthcare vertical at the edge: COVID-I9 perspective," *IEEE Network*, vol. 34, no. 4, pp. 98–105, 2020.

[12] Y. Choi, B. Shin, K. Kang, S. Park, and B. R. Beck, "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*, vol. 12, no. 11, p. 1325, 2020.

[13] S. Bhattacharya, P. K. R. Maddikunta, Q. V. Pham et al., "Deep learning and medical image processing for coronavirus (COVID-19) pandemic: a survey," *Sustainable cities and society*, vol. 65, Article ID 102589, 2020.

[14] H. Wang, L. Wang, E. H. Lee et al., "Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 48, 2020.

[15] I. I. Baskin, "The power of deep learning to ligand-based novel drug discovery," *Expert Opinion on Drug Discovery*, vol. 15, pp. 1–10, 2020.

[16] S. Korkmaz, "Deep learning-based imbalanced data classification for drug discovery," *Journal of Chemical Information and Modeling*, vol. 60, no. 9, pp. 4180–4190, 2020.

[17] F. Gentile, V. Agrawal, M. Hsing et al., "Deep docking: a deep learning platform for augmentation of structure based drug discovery," *ACS Central Science*, vol. 6, 2020.

[18] F. Piroozmand, F. Mohammadipanah, and H. Sajedi, "Spectrum of deep learning algorithms in drug discovery," *Chemical Biology & Drug Design*, vol. 96, no. 3, pp. 886–901, 2020.

[19] D. Baptista, P. G. Ferreira, and M. Rocha, "Deep learning for drug response prediction in cancer," *Briefings in Bioinformatics*, vol. 22, 2020.

[20] B. J. Neves, R. C. Braga, V. M. Alves et al., "Deep learning-driven research for drug discovery: tackling malaria," *PLoS Computational Biology*, vol. 16, no. 2, Article ID e1007025, 2020.

[21] B. Ramsundar, B. Liu, Z. Wu et al., "Is multitask deep learning practical for pharma?" *Journal of Chemical Information and Modeling*, vol. 57, no. 8, pp. 2068–2076, 2017.

[22] L. Patel, T. Shukla, X. Huang, D. W. Ussery, and S. Wang, "Machine learning methods in drug discovery," *Molecules*, vol. 25, no. 22, p. 5277, 2020.

[23] K. A. Giuliano, R. L. DeBiasio, R. T. Dunlay et al., "High-content screening: a new approach to easing key bottlenecks in the drug discovery process," *Journal of Biomolecular Screening*, vol. 2, no. 4, pp. 249–259, 1997.

[24] S. Bergström and O. Ivarsson, *Automation of a Data Analysis Pipeline for High-Content Screening Data*, Linköping University, Linköping, Sweden, 2015.

[25] E. Sandström, *Molecular Optimization Using Graph-To-Graph Translation*, Umeå University, Umeå, Sweden, 2020.

[26] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of Health Economics*, vol. 22, no. 2, pp. 151–185, 2003.

[27] M. Shafiq, X. Yu, A. K. Bashir, H. N. Chaudhry, and D. Wang, "A machine learning approach for feature selection traffic classification using security analysis," *The Journal of Supercomputing*, vol. 74, no. 10, pp. 4867–4892, 2018.

[28] M. Benhenda, "ChemGAN challenge for drug discovery: can ai reproduce natural chemical diversity?" 2017, https://arxiv.org/abs/1708.08227.

[29] ChEMBL Database, 2020, https://www.ebi.ac.uk/chembl/g/#search_results/targets/query=coronavirus.

[30] Molecularsets/Moses, 2020, https://github.com/molecularsets/moses.

[31] https://datascience.nih.gov/covid-19-open-access-resources%20.

[32] J. Cardoso-Silva, G. Papadatos, L. G. Papageorgiou, and S. Tsoka, "Optimal piecewise linear regression algorithm for QSAR modelling," *Molecular informatics*, vol. 38, no. 3, Article ID 1800028, 2019.

[33] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, and V. Prachayasittikul, "A practical overview of quantitative structure-activity relationship," *EXCLI Journal*, vol. 8, 2009.

[34] https://pubchem.ncbi.nlm.nih.gov/%20.

[35] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, and A. Hersey, "Yvonne light, shaun McGlinchey, david michalovich, bissan Al-lazikani, john P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, 2012.

[36] W. Shoombuatong, P. Prathipati, W. Owasirikul et al., "Towards the revival of interpretable QSAR models," in *Advances in QSAR Modeling*Springer, Berlin, Germany, 2017.

[37] L. Abrahamsson Kwetczer, *Hospital Readmission Risk Prediction Using Machine Learning*, KTH, Stockholm, Sweden, 2020.

[38] S. Tober, *Tree-based Machine Learning Models with Applications in Insurance Frequency Modelling*, KTH, Stockholm, Sweden, 2020.

[39] A. Tahir, F. Chen, H. U. Khan et al., "A systematic review on cloud storage mechanisms concerning e-healthcare systems," *Sensors*, vol. 20, no. 18, p. 5392, 2020.

[40] C. Nantasenamat, "Best practices for constructing reproducible QSAR models," in *Ecotoxicological QSARs*, pp. 55–75, Humana, New York, NY, USA, 2020.

[41] Wikipedia contributors, "Simplified molecular-input line-entry system," 2020.

[42] AI Speeds Drug Discovery to Fight COVID-19, 2020, https://towardsdatascience.com/ai-speeds-drug-discovery-to-fight-covid-19-b853a3f93e82.

[43] "Computational drug discovery," 2020, https://github.com/dataprofessor%20.

[44] "COVID-drug discovery for COVID-19," 2020, https://github.com/AshishKempwad/%20.

[45] "SARS-CoV-2 drug discovery using genetic algorithm and deep learning," 2020, https://github.com/Skyquek/fch-drug-discovery%20.