ML-DTD: Machine Learning-Based Drug Target Discovery for the Potential Treatment of COVID-19

Sovan Saha 1, Piyali Chatterjee 2, Anup Kumar Halder 3,4,*, Mita Nasipuri 5, Subhadip Basu 5,*, and Dariusz Plewczynski 3,4,*

1 Department of Computer Science & Engineering, Institute of Engineering & Management, Salt Lake Electronics Complex, Kolkata 700091, India
2 Department of Computer Science & Engineering, Netaji Subhash Engineering College, Techno City, Panchpota, Garia, Kolkata 700152, India
3 Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland
4 Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Banacha 2c Street, 02-097 Warsaw, Poland
5 Department of Computer Science & Engineering, Jadavpur University, 188, Raja S.C. Mallick Road, Kolkata 700032, India
* Correspondence: subhadip.basu@jadavpuruniversity.in (S.B.); dariusz.plewczynski@pw.edu.pl (D.P.)

Abstract: Recent research has highlighted that a large section of druggable protein targets in the Human interactome remains unexplored for various diseases. It might lead to the drug repurposing study and help in the in-silico prediction of new drug-human protein target interactions. The same applies to the current pandemic of COVID-19 disease in global health issues. It is highly desirable to identify potential human drug targets for COVID-19 using a machine learning approach since it saves time and labor compared to traditional experimental methods. Structure-based drug discovery where druggability is determined by molecular docking is only appropriate for the protein whose three-dimensional structures are available. With machine learning algorithms, differentiating relevant features for predicting targets and non-targets can be used for the proteins whose 3-D structures are unavailable. In this research, a Machine Learning-based Drug Target Discovery (ML-DTD) approach is proposed where a machine learning model is initially built up and tested on the curated dataset consisting of COVID-19 human drug targets and non-targets formed by using the Therapeutic Target Database (TTD) and human interactome using several classifiers like XGBBoost Classifier, AdaBoost Classifier, Logistic Regression, Support Vector Classification, Decision Tree Classifier, Random Forest Classifier, Naive Bayes Classifier, and K-Nearest Neighbour Classifier (KNN). In this method, protein features include Gene Set Enrichment Analysis (GSEA) ranking, properties derived from the protein sequence, and encoded protein network centrality-based measures. Among all these, XGBBoost, KNN, and Random Forest models are satisfactory and consistent. This model is further used to predict novel COVID-19 human drug targets, which are further validated by target pathway analysis, the emergence of allied repurposed drugs, and their subsequent docking study.

Keywords: COVID-19 human drug targets; COVID-19 human drug non-targets; machine learning; gene set enrichment analysis; network centrality; protein sequence; docking

1. Introduction

Discovering Drugs or vaccines for the pandemic outbreak of COVID-19 is a time taking process. If it can be done following an exact process, it may take a couple of months to be ready to use as no such effective therapeutic drugs are discovered to combat COVID-19. Moreover, it has several variants produced by continuous variations, and some variants, Delta and Omicron, spread the infection very quickly. In that case, Drug repurposing comes into play, and it may give alternative solutions for therapeutics or medical practitioners.
to face challenges. As we know, most known drugs are used for repositioning, and they produce promising results. For both cases, druggable target selection through appropriate ranking is significant because selecting appropriate drug targets helps us to determine the proper drug which can be used further in the clinical development and trials for COVID-19 treatment. Many antiviral, antibiotics or other types like Flavipiravir, Remdesivir, Lopinavir, Ritonavir, Azithromycin, and Hydroxy-chloroquine were tried but not able to produce promising results [1]. Drug repurposing requires more understanding of the interaction between pathogens of COVID-19 and host proteins. Identifying host proteins and potential drug targets will greatly facilitate drug repurposing or identification.

Experimental techniques to find druggable targets are affinity pull-downs, pooled RNAi Data [2], and, more recently, genome –CAS9 screens [3], which are always time-consuming and labor-intensive. Comparatively, to deal with the growing number of large-scale human genomics and proteomics datasets, the computational approach is appropriate in terms of cost and time. Structure-based drug discovery [4–6] is a kind of computational approach in which protein’s druggability is determined by molecular docking methods [4–6] to reveal ideas on macromolecule’s 3D conformation [4–6], protein-ligand interactions [4–6], and binding affinity of the target proteins [4–6]. However, this is limited to proteins whose 3-D Structure is known but will not apply to proteins without having information.

According to Behery et al. [7], computational methods are categorized into ligand, docking, and chemogenomic. Chemogenomic methods are superior to the other two methods in terms of time cost or unavailability of 3-D structure proteins. Chemogenomic methods are classified into machine learning-based, graph-based, and network-based, where machine learning methods have turned out to produce reliable results. This paper used a combination of Physico-chemical features, sequence information, and features derived from Smiles (Simplified Molecular Input Line). They adopted a methodical prediction scheme exploiting ensemble classifiers like LightBoost, XGBoost, ExtraTree, deep learning (DBN), and traditional machine learning methods (random forest), support vector machine.

Another work Wang et al. [8] presented is based on deep learning. They represented drug-target pairs using a scheme where drug molecules are encoded as fingerprint features, and target proteins are represented using their evolutionary information in the form of Legendre moments. Features set thus formed are again reduced by sparse Principal Component Analysis (SPCA). The Deep Long Short-Term Memory (DeepLSTM) model predicts Drug Target Interaction (DTI).

The research work [9], based on Machine learning techniques, proposed a work that can predict druggability scores of protein targets by exploiting distinguishable features from non-targets. In this approach, protein sequence-derived features, protein interaction information, and gene ontology terms are extracted as features, and the model is built on a training dataset. A very recent work based on neural networks on DTI is proposed by Li et al. [10], where 2D paired distance maps of target proteins and molecular graphs of drugs are used. A mutual interaction neural network is designed to capture the interactive impacts between drugs and targets and combine transformed modules by constructing a communicative Message Passing Neural Network (CMPNN).

Another recent work by Barman et al. [1] proposed a network biology approach to identify host targets for discovering drugs by constructing a human protein-protein interaction network (PPIN) with the above SARS-CoV-2 targeted proteins. This work combines essential network centrality measures and functional properties of the human proteins to identify the critical human targets of SARS-CoV-2, namely PRKACA, RHOA, CDK5RAP2, and CEP250.

Adhami et al. [11] presented repurposing novel candidate drugs for COVID-19. This work proposes a network-based drug repurposing strategy where seven potential drugs are identified. First, the PPIN is constructed for 524 human proteins. Then the target miRNAs of the mentioned module genes were separately obtained from the mirWalking 2.0 database because of the vital role of miRNAs in biological processes. Thus, they specified
seven clusters of proteins as the complexes of proteins associated with the SARS-CoV-2 virus. Moreover, seven therapeutic candidate drugs were identified to control gene regulation in COVID-19. They are Paclitaxel, Bortezomib, Carboplatin, Crizotinib, Cytarabine, Daunorubicin, and Vorinostat.

Inspired by these works described and considering the current need of the present time, we have hypothesized that druggable COVID-19 human targets can be predicted using GSEA, protein sequence-derived features, and network properties from the PPIN. The host targets for COVID-19 drugs are fetched from TTD (Therapeutic Target Database) [12]. A positive-negative target dataset is constructed by considering approved, clinical and investigational drugs with corresponding indications and their respective targets. Next, ML-DTD is applied to this dataset by a set of Machine learning classifiers to predict COVID-19 human drug targets. Gradually, some novel COVID-19 human drug targets are also identified, from which five COVID-19 repurposed drugs are highlighted.

2. Material & Methods

2.1. Dataset

A COVID-19 drug target is usually considered a protein indispensably linked with the process of COVID-19 disease progression. Several drugs could handle these targets to produce the desired consequences for the recovery from the pandemic of COVID-19. The Therapeutic Target Database (TTD) [12] is used in this research work on COVID-19 since it involves relevant information for implementing the drug discovery process, which demands both drugs and their associated therapeutic targets [13]. Besides this, 204,961 reviewed human proteins; their respective sequences are also extracted from UniProt [14]. Positive samples comprise ninety COVID-19 human drug targets enlisted in the TTD database. Negative samples are formed of ninety COVID-19 human drug non-targets randomly selected from the 204,961 reviewed human proteins after removing the positive samples.

2.2. Workflow of ML-DTD

The proposed methodology of ML-DTD consists of the following stages (please see Figure 1): (a) Feature Extraction, (b) Feature Selection, (c) Data Cleaning and Preprocessing, (d) Classification Algorithms, (e) Prediction of novel COVID-19 human drug targets, (f) Drug-Repurposing Study of novel COVID-19 human drug targets and (g) Docking study of the repurposed drugs detected through novel COVID-19 human drug targets. In feature extraction, several characteristics or attributes are fetched for the COVID-19 target and non-target human proteins. Once fetched, feature selection is applied to select the most significant features by filtering out the less important ones. Then the data obtained is preprocessed in the data cleaning and preprocessing stage to filter out the null values and make the data available for processing further. The refined data is inputted into the various machine learning models for learning and predicting by splitting the data into training and test set. These models are discussed in brief in the classification algorithms section. Once the accuracy of the models is fetched, the top performing models are used to predict the novel drug targets in the prediction of novel COVID-19 human drug targets stage. Finally, the related novel drug targets are mapped to their corresponding existing drugs by COVID-19Db in the drug-Repurposing study of novel COVID-19 human drug targets stage. These drugs are validated by docking studies with the other available COVID-19 proteases by DockCoV2 in the last stage, i.e., Docking study of the repurposed drugs detected through novel COVID-19 human drug targets.

(a) Feature Extraction

To execute Machine Learning algorithms, a pre-defined set of features are needed from the training set to generate the output for the test set. This research initially computed forty feature values for ninety positive and negative samples derived from Protein-Protein Interaction Network (PPIN), protein sequence, and Gene Set Enrichment Analysis (GSEA). More in-depth details of these forty features are highlighted in Figure 2. Network-based features
like Closeness Centrality [15], Betweenness Centrality [16], etc., are computed by using the CytoNCA app [17] of the Cytoscape [18], where proteins are given as the input. Their corresponding interactions are fetched from the String Db [19]. Protein Sequence-based features like Polarity, Non-Polarity, etc., are computed for each amino acid sequence of proteins obtained from UniProt [14] using the web server Pfeature [20]. GSEA of proteins have been computed on KEGG [21], Molecular Function (MF) [22], Cellular Component (CC) [22], Biological Process (BP) [22], and Reactome Pathway [23] by using python modules, and the top two ranks in each of them have been included in the feature list.

Figure 1. Computational workflow of ML-DTD.

Figure 2. PPIN features Protein Sequence features, and GSEA Ranking features.
(b) Feature Selection

Feature selection is an automatic or manual selection of a subset of the most relevant and appropriate features from a set of pre-defined features that can be used to build machine learning models. Out of the initial selection of forty features, the top twenty features are selected based on the ranking provided by the SelectKBest function of the python module sklearn [24]. SelectKBest can select k best features based on the type of scoring function the user will provide. In this case, the \textit{chi2 scoring function} is used for this purpose. The detailed ranking and the score of the top twenty selected features are highlighted in Table 1.

**Table 1.** Top twenty features selected by the SelectKBest module. SelectKBest module assigns a score to each feature which is then used for ranking. Features with lower rank are eliminated while higher ranking features are considered the most significant ones.

| Sl. No. | Specifications/Features                                      | SelectKBest Score |
|--------|-------------------------------------------------------------|-------------------|
| 1      | GO (BP) Rank2                                               | 7.62              |
| 2      | GO (MF) Rank1                                               | 6.73              |
| 3      | GO (BP) Rank1                                               | 5.25              |
| 4      | GO (MF) Rank2                                               | 3.94              |
| 5      | GO (CC) Rank1                                               | 3.34              |
| 6      | Reactome Rank2                                              | 2.89              |
| 7      | GO (CC) Rank2                                               | 2.82              |
| 8      | Reactome Rank1                                              | 1.61              |
| 9      | KEGG Rank1                                                  | 1.59              |
| 10     | KEGG Rank2                                                  | 1.27              |
| 11     | Betweenness                                                | 1.14              |
| 12     | Tiny (PCP_TN)                                               | 0.72              |
| 13     | Degree                                                      | 0.6               |
| 14     | Network                                                     | 0.56              |
| 15     | Sulphur Content (PCP_SC)                                    | 0.35              |
| 16     | Solvent Accessibility (Intermediate) (PCP_SA_IN)             | 0.32              |
| 17     | Cyclic (PCP_CY)                                             | 0.25              |
| 18     | Small (PCP_SM)                                              | 0.21              |
| 19     | Subgraph                                                    | 0.212             |
| 20     | Aromaticity (PCP_AR)                                        | 0.18              |

(c) Data Cleaning and Preprocessing

Data cleaning and preprocessing is the initial step to transforming the raw data into an understandable format that can be used for further scientific analysis. It can be categorized into the following stages: (1) Gathering the data along with feature values (2) import of the data and inclusion of python library files (3) handling missing (null) values (4) Detection of dependent and independent variables (5) handling of categorical values (if present) (6) Train and Test splitting of data (7) Scaling of feature to normalize them with a fixed range of values. All these steps are followed, and finally, the entire curated dataset with features is split into 80% train and 20% test data to execute machine learning algorithms. This research problem is categorized as a binary classification problem with two labels: Zero and One. Zero stands for \textit{COVID-19 human drug non-targets}, and One stands for \textit{COVID-19 human drug targets}. The entire working mechanism is highlighted in Figure 1.

(d) Classification Algorithms

Out of the several algorithms, the following eight algorithms are used to predict COVID-19 human drug targets and non-targets: (1) XGBoost Classifier [25], (2) AdaBoost Classifier [26], (3) Logistic Regression [27], (4) Support Vector Classification [28], (5) Decision Tree Classifier [29], (6) Random Forest Classifier [30], (7) Naive Bayes Classifier [31], and (8) K-Nearest Neighbour Classifier [32].
Extreme Gradient Boosting (XGBBoost) is one of the most popular classifiers in recent times. Here the decision trees are sequentially created. Weights have a significant impact on this classifier. All independent variables are allocated some weights, which are then transmitted to the decision trees to predict results. Variable weights wrongly predicted by the decision tree are increased and serve as an input to the second decision tree. Thus, a robust model is ensembled through each type of predictor.

AdaBoost Classifier is another methodology of ensemble learning like XGBBoost. It is primarily created to enhance binary classifier efficiency. In this classifier, weak classifiers are turned into stronger ones using weights through an iterative process similar to the XGBBoost classifier.

Logistic Regression (LR) is mainly used as a classification algorithm. It leads to estimating the probability of an event occurring based on the independent variables as provided in the dataset. The dependent variable thus gets confined between 0 and 1 since the result is always a probability.

Support Vector Machine (SVM) is a popular machine learning algorithm for regression and classification-based problems. The data here are mapped explicitly in a vector space to generate a hyperplane so that $n$-dimensional space can be segregated into classes which can be used to plot a new data point in an appropriate categorization later.

Decision Tree comes under the classification of supervised machine learning algorithms. It is mainly known for rule-based decisions. It is a tree-based model where each node represents a test to be performed on the feature, its associated edges signify the decision rules, whereas the leaf nodes denote the possible results of the test.

Random Forest Classifier (RF) is one of the easy and flexible machine learning algorithms used in machine learning. Trees form the forest. Forests having significantly more trees are generally more robust. Decision trees are created on random data samples by the random forest. Prediction from each tree is obtained, and the best solution is picked up by voting. It also gives an idea about the importance of the features.

The naive Bayes Classifier is known as the probability-based classifier. It will give the probability of whether the test data belongs to a class or not rather than specifying the label of the class to which it belongs. It is based on the Bayes theorem.

The K-Nearest Neighbour Classifier (KNN) is another classification algorithm that is often called a lazy learner machine learning algorithm since it does not consider training data for learning; instead, it uses them for computing the similarity with the new data points, and it places the new data points into the category which is most similar to the available ones.

All these eight machine learning algorithms have been implemented with the help of sklearn [24] on the previous curated training and test data as stated in the earlier section, and the performance accuracy of the best models is estimated.

(e) Prediction of novel COVID-19 human drug targets

The top three machine learning model predictors out of eight are now used to predict the COVID-19 human drug targets from the remaining 204,871 human proteins. The predicted COVID-19 human drug targets are further validated with the other COVID-19 human drug targets detected through several in-silico and in-vitro methodologies as specified in various works of literature [1,33–36]. A significant overlap of the predicted targets with the others has been observed, which motivates us to analyze those COVID-19 human drug targets further, which are not overlapped/matched. These might evolve as novel potential COVID-19 human drug targets if they relate to COVID-19 drugs or drug target pathways.

(f) Drug-Repurposing Study of novel COVID-19 human drug targets

The unmatched COVID-19 human drug targets now serve as an input to COVID19Db [37] to detect any possible connection with existing COVID-19 repurposed drugs. COVID19Db has an inbuilt drug discovery web server. This tool [37] initially led to the integration of the drug–target–pathway interactions from two central resources: (1) KEGG Pathway [21] resources and (2) DrugCentral [38]. Once the integration is done, the drug discovery tool is implemented with detailed results. The tool also generates PubMed database links to
support the mapped drugs with corresponding drug targets [37]. This tool fetches disease
pathway analysis, i.e., the number of targets, pathways, and target-pathway interactions,
along with COVID-19-supported PubMed shreds of evidence for these unmatched COVID-
19 human drug targets. Once the information is fetched, these novel COVID-19 human
drug targets are further validated by docking studies in the next section.

(g) Docking study of the repurposed drugs detected through novel COVID-19 human
drug targets

Molecular docking is an intrinsic requirement for any drug discovery [39]. With the
help of DockCoV2 [40], a docking study is done on the repurposed drugs generated from
the COVID19Db [37] for the novel COVID-19 human drug targets, as discussed earlier.
DockCoV2 uses AutoDock Vina (version 1.1.2) [41] as its core docking utility. Docking anal-
yses of the repurposed drugs have been performed with six COVID-19 proteins, including
spike protein [42], 3CLpro [43], PLpro [44], RdRp [45], N protein [46], and ACE2 [47]. The
study shows that the repurposed drugs might be the potential contenders for COVID-19.

3. Results & Discussion

This section can be broadly classified into two categories: 1) Performance Analysis
of the Machine Learning Models used in ML-DTD: In this category, the results of the
performance level of the models of ML-DTD will be discussed 2) Detection and Validation
of novel COVID-19 Human Drug Targets: In this second category, the procedure of detecting
novel human drug targets and their corresponding validation will be highlighted.

3.1. Performance Analysis of the Machine Learning Models Used in ML-DTD

In this section, the empirical results of the proposed machine learning methodology
have been discussed. The positive and negative samples are formed from the TTD and
human datasets. Eight machine learning models have been implemented on this curated
dataset after dividing it into 80% training and 20% test set. The efficiency of the models has
been estimated in terms of Accuracy, Precision, Recall, F1-Score, Cohens Kappa Score [48],
and Area Under Curve [49], which is reported in Table 2 and Figure 3. XGBBoost Classifier
obtained the best accuracy score of 0.80. The KNN and Random Forest models stand second
and third by obtaining a score of 0.77 and 0.72, respectively.

Table 2. The efficiency of the models used in ML-DTD. ML-DTD uses six parameters: Accuracy,
Precision, Recall, F1-Score, Cohens Kappa Score, and Area Under Curve (AUC) to estimate the overall
performance.

| Sl. No. | Models        | Accuracy | Precision | Recall | F1-Score | Cohens Kappa Score | Area under Curve |
|---------|---------------|----------|-----------|--------|----------|--------------------|------------------|
| 1       | XGBBoost      | 0.80     | 0.73      | 0.87   | 0.79     | 0.61               | 0.83             |
| 2       | AdaBoost      | 0.75     | 0.65      | 0.93   | 0.76     | 0.51               | 0.79             |
| 3       | LR            | 0.66     | 0.60      | 0.75   | 0.66     | 0.34               | 0.75             |
| 4       | SVM           | 0.66     | 0.59      | 0.81   | 0.68     | 0.34               | 0.74             |
| 5       | Decision Tree | 0.58     | 0.52      | 0.75   | 0.61     | 0.19               | 0.60             |
| 6       | Random Forest | 0.72     | 0.63      | 0.87   | 0.73     | 0.45               | 0.83             |
| 7       | Naive Bayes   | 0.63     | 0.57      | 0.75   | 0.64     | 0.29               | 0.74             |
| 8       | KNN           | 0.77     | 0.83      | 0.62   | 0.71     | 0.53               | 0.81             |

In contrast, comparing the corresponding F1-Score, the top three model rankers remain
the same, but their order changes slightly. XGBBoost Classifier holds its first position, while
KNN and Random Forest model interchanges their position by obtaining an F1 score of
0.79, 0.71, and 0.73, respectively. The area under the curve (AUC) visually gives a more
accurate representation for the prediction of COVID-19 human drug targets since it is
calculated based on each model’s Receiver Operating Characteristic curve (ROC curve) to
highlight the work’s quality. Figure 3 represents the eight machine learning models’ ROC
and Area Under Curve (AUC). XGBBoost and Random Forest models predict the highest value in the AUC = 0.83 for the curated dataset, while KNN stands second with AUC = 0.81, respectively.

**Figure 3.** Comparison of Accuracy, Precision, Recall, F1-Score, Cohens Kappa Score, and AUC of the proposed ML-DTD.

### 3.2. Detection and Validation of Novel COVID-19 Human Drug Targets

The three top ranking models, i.e., XGBBoost, KNN, and Random Forest models, as detected in the earlier stage, are selected and used to predict the novel COVID-19 targets from the human proteome consisting of 204,871 human proteins, which is obtained after removing the proteins involved in our curated dataset from the initial interactome consisting of 204,961 reviewed human proteins. XGBBoost, KNN, and Random Forest models successfully predict 3814, 1687, and 4144 COVID-19 novel human drug targets in the human proteome.

These results are validated with the COVID-19 human drug targets detected through the in-vitro methodologies of Gordon et al. [35]. Significant overlap is observed in this validation, highlighted through the Venn diagram and the minimal triangular matrix in Figure 4. All the Venn diagrams and minimal triangular matrix are generated through the multi-list comparator tool of molbiotool [50]. Further analyses have been carried out, and the prediction of each model has been compared with the other in-silico methodologies of Tehrani et al. [36], Barman et al. [1], Chen et al. [33], and Saha et al. [34]. The results are highlighted in Figures 5–7.

**Figure 4.** Overlapping of ML-DTD and Gordon et al. predicted COVID-19 human drug targets.
The above results highlighted that some COVID-19 human drug targets are not overlapped with the existing targets, like 1070 in the case of KNN, 2680 in the case of RF, and 2468 in the case of the XGBoost model. These unmatched COVID-19 human drug targets for each model are then uploaded separately in COVID19Db [37], which gives two detailed results of the drugs associated with these targets: (1) Drug Target Pathway Interactions and (2) Number of target-pathway interactions of the potential actionable drugs along with the PubMed link of supporting shreds of evidence for COVID-19 (please see Supplementary Table S1 for XGBoost, Table S2 for RF and Table S3 for KNN). These potential drugs from each of the three models are collected, and it has been observed that potential drugs from each of the three models are: (1) Bosutinib, (2) Crizotinib, (3) Midostaurin, (4) Nintedanib, and Sunitinib (please see Table 3).
each of the three models are collected, and it has been observed that there is an overlap of 209 drugs (please see Figure 8). Information for only these 209 drugs is retained for each of the three models from the above two results. These are then sorted in descending order based on the number of target-pathway interactions. The top ten drugs are selected from each model, and analyses have been made to see the possibility of overlap. The result is highlighted in Figure 9, which shows a common overlap of 5 repurposed COVID-19 drugs, and they are: (1) Bosutinib, (2) Crizotinib, (3) Midostaurin, (4) Nintedanib, and Sunitinib (please see Table 3).

![Figure 8](image_url)

**Figure 8.** Overlap of COVID-19Db-derived potential COVID-19 drugs for the unmatched/novel targets predicted by the three top-performing models.

![Figure 9](image_url)

**Figure 9.** Overlap of top ten COVID-19Db-derived potential COVID-19 drugs for the unmatched/novel targets predicted by the three top-performing models.

All these 5 drugs are also docked with six COVID-19 proteins, including spike protein [42], 3CLpro [43], PLpro [44], RdRp [45], N protein [46], and ACE2 [47] by DockCoV2 [40]. DockCoV2 generates the docking score for each of them, highlighted in Table 4, where more negative scores indicate a strong binding of a protein-ligand complex. The generated scores are quite satisfactory for these five drugs. The best poses for each of them are displayed in Table 5. Few samples of best poses of Nintedanib with the other COVID-19 proteins are also highlighted in Figure 10. So, both COVID19Db and DockCoV2 prove that the unmatched targets (1070 in the case of KNN, 2680 in the case of RF, and 2468 in the case of the XGBoost model) can be treated as the novel COVID-19 human drug targets (please see Supplementary Table S4) since they have a good docking score with the COVID-19...
proteins as well as they are linked with the COVID-19 repurposed drugs through target pathway analysis as discussed earlier.

Table 3. Number of target-pathway interactions of the top five potential actionable COVID-19 repurposed drugs. COVID19Db maps these detected top 5 drugs with the disease pathway and generates the number of targets, pathways, and target-pathway interactions. It also highlights the PubMed literature links as evidence that these top 5 drugs are already associated with or recommended for the possible treatment of COVID-19 in other in-silico or in-vitro research methodologies.

| Drug Name | The Number of Target-Pathway Interactions | The Number of Targets | The Number of Pathways | Evidence for COVID-19 |
|-----------|------------------------------------------|-----------------------|------------------------|-----------------------|
| Sunitinib | 91                                       | 24                    | 44                     | Link to PubMed        |
| Bosutinib | 66                                       | 20                    | 43                     | Link to PubMed        |
| Crizotinib| 35                                       | 15                    | 25                     | Link to PubMed        |
| Midostaurin| 47                                      | 6                     | 39                     | Link to PubMed        |
| Nintedanib| 24                                       | 6                     | 22                     | Link to PubMed        |

Table 4. Docking scores of top five potential actionable COVID-19 repurposed drugs with COVID-19 protein structures. The top five potential actionable COVID-19 repurposed drugs are also docked with the available COVID-19 proteases like 3CLpro, RdRp, PLpro, N, ACE2, and Spike RBD by DockCoV2. The docking score represents how well these drugs can be docked with COVID-19 proteases. More negative value highlights a good docking that represents that the drug and protease can dock well with each other by using less energy.

| COVID-19 Structure Details | Predicted Drugs Docking Score by DockCoV2 |
|---------------------------|-------------------------------------------|
| Structure Name            | Bosutinib | Crizotinib | Midostaurin | Nintedanib | Sunitinib |
| 3C-like protease          | 3CLpro    | −7         | −7.4        | −8.5        | −7.6       | −7.1       |
| RNA-dependent RNA polymerase | RdRp     | −8.4       | −7.9        | −8.9        | −9.1       | −7.6       |
| Papain-like protease      | PLpro     | −7.7       | −7.3        | −8.8        | −8.1       | −6.9       |
| Nucleocapsid protein      | N         | −9         | −9.5        | −10.1       | −9.2       | −8.8       |
| ACE2 Receptor Protein     | ACE2      | −8.8       | −9.2        | −14.5       | −10.9      | −8         |
| Spike Protein             | Spike RBD | −7         | −7          | −8.8        | −8.6       | −8.3       |

Table 5. Best pose Links of the docking of top five potential actionable COVID-19 repurposed drugs with COVID-19 protein structures. DockCoV2 also generates the best docking pose for each of the top five detected potential COVID-19 drugs and COVID-19 proteases, the link of which is hyperlinked in the table below.

| Predicted Drugs Best Pose by DockCoV2 |
|---------------------------------------|
| Structure ID                          | Bosutinib | Crizotinib | Midostaurin | Nintedanib | Sunitinib |
| 3CLpro                                | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link |
| RdRp                                  | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link |
| PLpro                                 | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link |
| N                                     | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link |
| ACE2                                  | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link |
| Spike RBD                             | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link | Best Pose Link |
Figure 10. Best poses results of Docking of Nintedanib (CID: 135423438) with spike protein, 3CLpro, PLpro, RdRp, N protein, and ACE2.

4. Conclusions

In this paper, ML-DTD is proposed for identifying novel COVID-19 human drug targets, which are initially tested on a curated dataset of positive samples (obtained from the TTD dataset) and negative samples (formed from reviewed human proteome). PPIN features, protein sequence features, and GSEA ranking are used for this purpose. Then the best performing models are used to predict COVID-19 human drug targets in the remaining human proteome other than the curated dataset. A significant overlap has been found between the predicted and existing targets obtained from various existing works of literature. The non-overlapping targets are further analyzed and validated through COVID19Db and DockCoV2, highlighting that these unmatched targets can be the novel COVID-19 human drug targets. The accuracy of the proposed methodology can be further improved by including the protein domain and other allied features. Besides, UniProt and TTD do not include any strain specific annotations. So, other data sources to extract strain specific protein variants needs to be explored to prioritize drugs against evolving strains of SARS-CoV-2. Currently, this methodology is only used for COVID-19, which can be extended to other diseases in our future works.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/vaccines10101643/s1, Table S1: Drug Target Pathway Interaction of 2468 unmatched targets by XGBoost (generated by Covid-19Db); Table S2: Drug Target Pathway Interaction of 2680 unmatched targets by RF (generated by Covid-19Db); Table S3: Drug Target Pathway Interaction of 1070 unmatched targets by KNN (generated by Covid-19Db); Table S4: NOVEL COVID-19 Human Targets predicted by XGBoost, KNN and RF.

Author Contributions: Data curation, S.S. and S.B.; formal analysis, S.S., P.C., A.K.H., S.B. and D.P.; funding acquisition, P.C., M.N., S.B. and D.P.; investigation, S.B. and D.P.; methodology, P.C., M.N., S.B. and D.P.; project administration, S.S., P.C. and A.K.H.; resources, P.C. and S.B.; software, S.S.; supervision, S.S. and A.K.H.; validation, M.N. and D.P.; visualization, S.S. and M.N.; writing-original draft, S.S.; writing-review & editing, S.S., A.K.H. and M.N.; All authors have read and agreed to the published version of the manuscript.

Funding: UGC Research Award (F.30-31/2016(SA-II)) from UGC, Government of India, and DBT project (No.BT/PR16356/BID/7/596/2016), Ministry of Science and Technology, Government of India; Excellence Initiative: Research University (IDUB) IDUB grant BOB-IDUB-622-197/2021; Pol-
ish National Science Center (2019/35/O/ST6/02484 and 2020/37/B/NN2/03757), Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund; European Commission Horizon 2020 Marie Skłodowska-Curie ITN Enpathy grant ‘Molecular Basis of Human enhanceropathies;’ and National Institute of Health USA 4D/Nucleome grant 1U54DK107967-01 ‘Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation’; Marie Skłodowska-Curie Action: co-supported as RENOIR Project by the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 691152 and by Ministry of Science and Higher Education (Poland), grant Nos. W34/H2020/2016, 329025/PnH/2016; Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme; Polish Ministry of Science and Higher Education (decision no. 7054/IA/SP/2020 of 28 August 2020).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Authors are thankful to the CMATER research laboratory of the Computer Science Department, Jadavpur University, India, for providing infrastructure facilities during the progress of the work.

Conflicts of Interest: The authors declare that they have no competing interest.

References
1. Barman, R.K.; Mukhopadhyay, A.; Maulik, U.; Das, S. A network biology approach to identify crucial host targets for COVID-19. Methods 2022, 203, 108–115. [CrossRef]

2. McFarland, J.M.; Ho, Z.V.; Kugener, G.; Dempster, J.M.; Montgomery, P.G.; Bryan, J.G.; Krill-Burger, J.M.; Green, T.M.; Vazquez, F.; Boehm, J.S.; et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat. Commun. 2018, 9, 4610. [CrossRef]

3. Behan, F.M.; Iorio, F.; Picco, G.; Gonçalves, E.; Beaver, C.M.; Migliardi, G.; Santos, R.; Rao, Y.; Sassi, F.; Pinnelli, M.; et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature 2019, 568, 511–516. [CrossRef]

4. Saha, S.; Halder, A.K.; Bandypadhyay, S.S.; Chatterjee, P.; Nipanpadri, M.; Bose, D.; Basu, S. Drug repurposing for COVID-19 using computational screening: Is Fostamatinib/R406 a potential candidate? Methods 2022, 203, 564–574. [CrossRef]

5. Rutwick Surya, U.; Praveen, N. A molecular docking study of SARS-CoV-2 main protease against phytochemicals of Boerhavia diffusa Linn. for novel COVID-19 drug discovery. Virusdisease 2021, 32, 46–54. [CrossRef]

6. Hosseini, M.; Chen, W.; Xiao, D.; Wang, C. Computational molecular docking and virtual screening revealed promising SARS-CoV-2 drugs. Precis. Clin. Med. 2021, 4, 1–16. [CrossRef]

7. El-Behery, H.; Attia, A.F.; El-Fishawy, N.; Torkey, H. Efficient machine learning model for predicting drug-target interactions with case study for Covid-19. Comput. Biol. Chem. 2021, 93, 107536. [CrossRef]

8. Wang, W.; Lv, H.; Zhao, Y.; Liu, D.; Wang, Y.; Zhang, Y. DLS: A Link Prediction Method Based on Network Local Structure for Predicting Drug-Protein Interactions. Front. Bioeng. Biotechnol. 2020, 8, 330. [CrossRef]

9. Dezső, Z.; Ceccarelli, M. Machine learning prediction of oncology drug targets based on protein and network properties. BMC Bioinformatics 2020, 21, 104. [CrossRef]

10. Li, F.; Zhang, Z.; Guan, J.; Zhou, S. Effective drug–target interaction prediction with mutual interaction neural network. Bioinformatics 2022, 38, 3582–3589. [CrossRef]

11. Adhami, M.; Sadeghi, B.; Rezapour, A.; Haghdost, A.A.; MotieGhader, H. Repurposing novel therapeutic candidate drugs for coronavirus disease-19 based on protein-protein interaction network analysis. BMC Biotechnol. 2021, 21, 22. [CrossRef]

12. Chen, X.; Ji, Z.L.; Chen, Y.Z. TTD: Therapeutic Target Database. Nucleic Acids Res. 2002, 30, 412–415. [CrossRef]

13. Drews, J. Drug Discovery: A Historical Perspective. Science 2000, 287, 1960–1964. [CrossRef]

14. The UniProt Consortium. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 2017, 45, D158–D169. [CrossRef]

15. Wuchty, S.; Stadler, P.F. Centers of complex networks. J. Theor. Biol. 2003, 223, 45–53. [CrossRef]

16. Joy, M.P.; Brock, A.; Ingber, D.E.; Huang, S. High-Betweenness Proteins in the Yeast Protein Interaction Network. J. Biomed. Biotechnol. 2005, 2005, 594674. [CrossRef]

17. Tang, Y.; Li, M.; Wang, J.; Pan, Y.; Wu, F-X. CytoNCA: A cytoscape plug-in for centrality analysis and evaluation of protein interaction networks. Biosystems 2015, 127, 67–72. [CrossRef]

18. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003, 13, 2498–2504. [CrossRef]

19. Szklarczyk, D.; Gable, A.L.; Nastou, K.C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N.T.; Legeay, M.; Fang, T.; Bork, P.; et al. The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids 2021, 49, D605–D612. [CrossRef]
20. Pande, A.; Patiyal, S.; Lathwal, A.; Arora, C.; Kaur, D.; Dhall, A.; Mishra, G.; Kaur, H.; Sharma, N.; Jain, S.; et al. Computing wide range of protein/peptide features from their sequence and Structure. *BioRxiv* 2019. [CrossRef]

21. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017, 45, D353–D361. [CrossRef]

22. Consortium, T.G.O. The Gene Ontology project in 2008. *Nucleic Acids Res.* 2007, 36, D440–D444. [CrossRef]

23. Fabregat, A.; Sidireopoulos, K.; Viteri, G.; Forner, O.; Marin-Garcia, P.; Arnau, V.; D’Eustachio, P.; Stein, L.; Hermjakob, H. Reactome pathway analysis: A high-performance in-memory approach. *BMC Bioinform.* 2017, 18, 142. [CrossRef]

24. 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, 2825–2830.

25. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

26. Schapire, R.E. A brief introduction to boosting. In Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 31 July–6 August 1999; Volume 2, pp. 1401–1406.

27. Bacaër, N. Verhulst and the logistic equation (1838). In *A Short History of Mathematical Population Dynamics*; Bacaër, N., Ed.; Springer London: London, UK, 2011; pp. 35–39.

28. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* 1995, 20, 273–297. [CrossRef]

29. Kamirski, B.; Jakubczyk, M.; Szufel, P. A framework for sensitivity analysis of decision trees. *Cent. Eur. J. Oper. Res.* 2018, 26, 135–159. [CrossRef]

30. Tin Kam, H. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 271, pp. 278–282.

31. Hand, D.J.; Yu, K. Idiot’s Bayes: Not So Stupid after All? *Int. Stat. Rev. Rev. Int. De Stat.* 1989, 57, 238–247. [CrossRef]

32. Fix, E.; Hodges, J.L. Discriminatory Analysis. *Nonparametric Discrimination: Consistency Properties.* 1951; 20.

33. Chen, Z.; Wang, C.; Feng, X.; Nie, L.; Tang, M.; Zhang, H.; Xiong, Y.; Swisher, S.K.; Srivastava, M.; Chen, J. Interactomes of SARS-CoV-2 and human coronaviruses reveal host factors potentially affecting pathogenesis. *EMBO J.* 2021, 40, e107776. [CrossRef]

34. Saha, S.; Chatterjee, P.; Nasipuri, M.; Basu, S. Detection of spreader nodes in human-SARS-CoV protein-protein interaction network. *PeerJ* 2021, 9, e12117. [CrossRef]

35. Gordon, D.E.; Jang, G.M.; Bouhaddou, M.; Xu, J.; Obernier, K.; White, K.M.; O’Meara, M.J.; Rezelj, V.V.; Guo, J.Z.; Swaney, D.L.; et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020, 583, 459–468. [CrossRef]

36. Samavarchi-Tehrani, P.; Abdouni, H.; Knight, J.; Astori, A.; Samson, R.; Lin, Z.-Y.; Kim, D.-K.; Knapp, J.; St-Germain, J.; Go, C.; et al. A SARS-CoV-2-host proximity interactome. *BioRxiv* 2020.

37. Zhang, W.; Wang, C.; Feng, X.; Nie, L.; Tang, M.; Zhang, H.; Xiong, Y.; Swisher, S.K.; Srivastava, M.; Chen, J. Interactomes of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Acta Pharm. Sin. B* 2020, D353–D361. [CrossRef]

38. Avram, S.; Bologa, C.G.; Holmes, J.; Bocci, G.; Wilson, T.B.; Nguyen, D.-T.; Curpan, R.; Halip, L.; Bora, A.; Yang, J.J.; et al. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.* 2021, 49, D1160–D1169. [CrossRef]

39. Pinzí, L.; Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. *Int. J. Mol. Sci.* 2019, 20, 4331. [CrossRef]

40. Chen, T.-F.; Chang, Y.-C.; Hsiao, Y.; Lee, K.-H.; Lin, Y.-H.; Tu, Y.-C.E.; Huang, H.-C.; Chen, C.-Y.; Juan, H.-F. DockCoV2: A drug database against SARS-CoV-2. *Nucleic Acids Res.* 2021, 49, D1152–D1159. [CrossRef]

41. Trotter, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 2010, 31, 455–461. [CrossRef]

42. Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature 2020, 581*, 215–220. [CrossRef]

43. Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 2020, 582, 289–293. [CrossRef]

44. Rut, W.; Lv, Z.; Zmudzinski, M.; Patchett, S.; Nayak, D.; Snipas, S.J.; El Oualid, F.; Huang, T.T.; Bekes, M.; Drag, M.; et al. Activity profiling and crystal structures of inhibitor-bound SARS-CoV-2 papain-like protease: A framework for anti-COVID-19 drug design. *Sci. Adv.* 2020, 6, eabd4596. [CrossRef]

45. Yin, W.; Mao, C.; Luan, X.; Shen, D.-D.; Shen, Q.; Su, H.; Wang, X.; Zhou, F.; Zhao, W.; Gao, M.; et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science 2020*, 368, 1499–1504. [CrossRef] [PubMed]

46. Kang, S.; Yang, M.; Hong, Z.; Zhang, L.; Huang, Z.; Chen, X.; He, S.; Zhou, Z.; Zhou, Z.; Chen, Q.; et al. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm. Sin. B* 2020, 10, 1228–1238. [CrossRef] [PubMed]

47. Towler, P.; Staker, B.; Prasad, S.G.; Menon, S.; Tang, J.; Parsons, T.; Ryan, D.; Fisher, M.; Williams, D.; Dales, N.A.; et al. ACE2 X-Ray Structures Reveal a Large Hinge-bending Motion Important for Inhibitor Binding and Catalysis. *J. Biol. Chem.* 2004, 279, 17996–18007. [CrossRef] [PubMed]

48. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med.* 2012, 22, 276–282. [CrossRef]
49. Kumar, R.; Indrayan, A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics* **2011**, *48*, 277–287. [CrossRef]

50. Available online: [https://molbiotools.com/](https://molbiotools.com/) (accessed on 1 May 2022).