Research Article ISSN 2639-9458

Microbiology & Infectious Diseases

Calming the Storm: Identifying Multi-Cytokine Inhibiting Drugs with Machine Learning for COVID-19 Induced Cytokine Storms

James Hou[1], Valentina L. Kouznetsova[2,3], Igor F. Tsigelny[2,3,4]*

[1]*REHS Program, San Diego Supercomputer Center, UC San Diego, California, USA.*

[2]*San Diego Supercomputer Center, UC San Diego, California, USA.*

[3]*BiAna, La Jolla, California, USA.*

[4]*Dept. of Neurosciences, UC San Diego, California, USA.*

***Correspondence:**
Igor F. Tsigelny, San Diego Supercomputer Center, UC San Diego, California, USA.

Received: 02 Jan 2022; **Accepted:** 28 Jan 2022; **Published:** 04 Feb 2022

Citation: Hou J, Kouznetsova VL, Tsigelny IF. Calming the Storm: Identifying Multi-Cytokine Inhibiting Drugs with Machine Learning for COVID-19 Induced Cytokine Storms. Microbiol Infect Dis. 2022; 6(1): 1-7.

ABSTRACT

In COVID-19, patients in severe condition often suffer from a major complication that leads to lung injury, ARDS and possibly death: the cytokine storm. The cytokine storm is composed of many cytokines, including IL-6, IL-2, and TNF-a for COVID-19. To combat such effects, a cocktail of cytokine-inhibiting drugs are administered. However, a combination of drugs can be overly taxing upon the patient, thus creating the demand for a drug that targets multiple cytokines. This project identifies multi-cytokine inhibiting compounds from FDA-approved drugs with machine-learning methods. Many machine-learning algorithms were applied to the task and Support Vector Machines proved best with strong performances across all cytokines. Under the constraints of limited data (30–60 samples) for some cytokines, we significantly boosted modeling power and accuracy with the application of data dimension reduction technique, Principle Component Analysis. After exhaustive exploration, the FDA-approved hepatitis-C drug—glecaprevir—was identified with confidences of 80.52% for TNF-a, 99.04% for IL-2, and 98.23% for IL-6.

Keywords

Cytokine, Cytokine storm, COVID-19, Machine learning, Repurposing.

Acronyms and Abbreviations

ARDS, acute respiratory distress syndrome; CDK, Chemistry Development Kit; COVID-19, coronavirus disease 2019; CSF, colony stimulating factor; GB, gradient boosted tree; GUI, graphical user interface; IC_{50}, half maximal inhibitory concentration; IFN, interferon; IL, interleukin; LR, logistic regression; MLP, multilayer perceptron; PCA, principal component analysis; RBF, radial basis function; ReLU, rectified linear unit; SARS, severe acute respiratory syndrome; SMILES, simplified molecular-input line-entry system; SVM, support vector machine; TNF, tumor necrosis factor.

Introduction

Cytokines

Cytokines are small water-soluble messenger proteins secreted by multiple types of cells. Cytokines have a large role in immunoregulation and inflammation control, managing much of the immune response. They are often secreted by cells and used either to activate its parent cell or neighboring cells in processes called autocrine and paracrine action. Some major producers and affected cells are B Cells, T Cells, Macrophages, and Neutrophils. The cytokines invoke change or activate cells by binding to their respective receptors on a cell and thus, changing the cell's behavior. Major groups of cytokines include Interleukins (ILs), Interferons (IFNs), Tumor Necrosis Factor (TNF), Chemokines and Colony Stimulating Factors (CSFs). Cytokines form a complex network of interactions, as they can induce cells to produce other cytokines [1,2].

Cytokine Storm

The cytokine storm is defined as an uncontrolled release of cytokines. They occur in both viral diseases and non-infectious diseases. As mentioned, the overlapping and redundant nature of cytokines can cause a snowballing effect. Since cytokines are effective at low levels, such a reaction can cause a large influx of inflammatory cytokines. This also comes with the flood of immune cells activated by the cytokines into the local area. This causes great damage to the organs of the local area and the repeated entry of these immune cells can cause significant damage to membranes. In severe cases, the cytokine storm affects multiple organs, resulting in multiple organ failure [3-5,1].

Cytokine Storm in COVID-19

Coronavirus disease 2019 (COVID-19), an upper respiratory disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has become a worldwide pandemic, in total infecting 230 million people and killing nearly five million people (September 24th, 2021) [6]. It was observed that there was a high occurrence of cytokine storms in severe cases of COVID-19. The cytokine storm can cause significant injury to a patient. The rapid and concentrated influx of immune cells damage capillaries, and vascular barriers. Furthermore, the cytokines induced activity can cause apoptosis in the epithelial cells. As a result, patients suffer Acute Respiratory Distress Syndrome (ARDS), lung injury and multiple organ failure. Thus, the cytokine storm can be detrimental to an already weakened patient. It is observed that severe cases of COVID-19 are associated with the rise of IL-1β, IL-6, IL-2, IL-8, and TNF-α levels in serum. Furthermore, IL-6 is associated with poor prognosis [7-12]. For the purposes of this study, we will be targeting cytokines IL-6, IL-2, and TNF-α due to the availability of relevant data.

Treatment

To combat the effects of the cytokine storm in COVID-19, multiple approaches have been explored. Corticosteroids, hydroxychloroquine, and chloroquine have been tested for their suppression of inflammation. However, they have been shown to have problematic side effects. Tocilizumab has also been tested for IL-6 inhibition but neglects the other cytokines [4,8,9,11,13]. A cocktail of cytokines inhibitors could be used to counteract the storm but would likely be overly taxing upon a patient's body and potentially injured liver. Thus, there is an urgent need for a drug that can target multiple cytokines. To combat the problem, this study proposes to use machine-learning (ML) methods to identify potential candidates for multi-cytokine inhibition from FDA-approved drugs. In repurposing already FDA-approved drugs, we avoid the many years and hundreds of millions of dollars required for drug development and testing. Furthermore, the current drugs have known side effects and dosages that can ensure safety.

Methods

Throughout the research process, a number of software and databases were utilized to conduct in-silico experiments. The main language used to conduct the theoretical experiments was Python and a collection of its libraries including NumPy, Scikit Learn, PyTorch, and Matplotlib. In gathering data, web databases such as PubChem, ZINC15, and DrugBank were used in conjunction with PaDEL-Descriptor for data preparation.

To find an inhibitor of multiple types of cytokines, we formulate the task as an inhibitor prediction or classification problem: known inhibitors (targeted towards a type of cytokine) labeled as positive samples; random molecules or drugs labeled as negative cases.

Three models are trained, each to classify inhibitors for their own cytokine. The learned models were then applied to a preserved set of FDA-approved drugs and the intersection of the three models formed our candidate pool that applies to all three cytokines. As a note, this method could be extended to more combinations of different cytokines.

Data Collection and Processing

As shown in Figure 1, we first collected the tested compounds and proven inhibitors of TNF-α (1525 compounds), IL-6 (19 compounds), and IL-2 (29 compounds) from PubChem, and retrieved their Simplified Molecular-Input Line-Entry System (SMILES), which describes the 2D structures of each compound. For each tested compound, we removed those that have a reported inhibiting activity value (IC_{50}) greater than 5 μM, leaving the more effective compounds in the dataset. Then we download a collection of FDA-Approved Drugs from ZINC15. Two-hundred FDA-approved drugs were partitioned from the original set to form our preserved set. We submitted the SMILES of all downloaded data into PaDEL-Descriptor program developed using the Chemistry Development Kit (CDK)—to generate 1875 distinct chemical and physical descriptions for each compound [14]. After filtering nominal and missing entries (not useful to training a machine learning model), we retained 1200 descriptors for each compound. With the necessary features prepared, the data was then divided into a training set and a testing set for each type of cytokine Figure 2. For each type, its inhibitors (positive class) were paired with an equal amount of random FDA-approved drugs from the remaining FDA-approved drugs list (negative class). Then both classes' data for each of cytokines were combined in common datasets, which were then split into training sets and testing sets in an 80–20 training-test set. The aforementioned process left the TNF-α dataset with 2260 training samples and 566 testing samples, IL-6 dataset with 30 training samples and 8 testing samples, IL-2 dataset with 42 training samples and 12 testing samples.

Data Cleaning and Feature Selection

Due to the nature of the dataset and a possibility of overfitting—a small number of samples (30–60 for IL-6 and IL-2) and a large number of features—it was very difficult to achieve consistent and strong performance when training. Thus, we conducted feature selection through means of principal component analysis (PCA). The correlation between all features is calculated and the resulting matrix is decomposed into major components or eigenvectors (Figure 3).

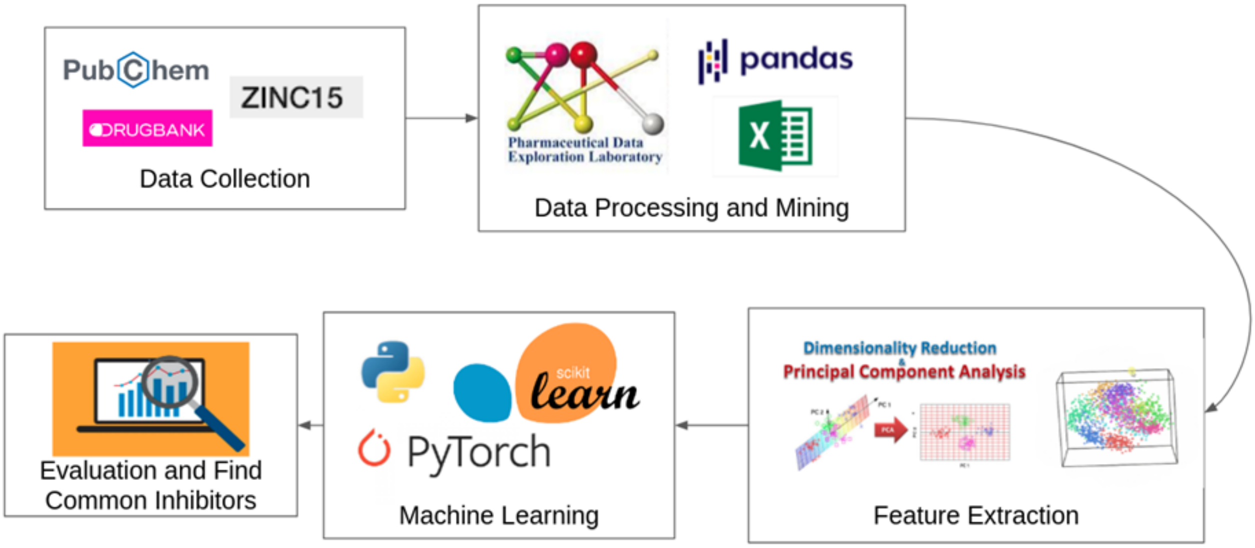
PubChem
DRUGBANK
ZINC15
Data Collection
Pharmaceutical Data Exploration Laboratory
pandas
Data Processing and Mining
Dimensionality Reduction & Principal Component Analysis
Feature Extraction
scikit learn
PyTorch
Machine Learning
Evaluation and Find Common Inhibitors

Figure 1: The research workflow and resources. Data sources used in this project include PubChem, DrugBank, and Zinc15. The data is then fed into the data processing pipeline, which generate features/descriptors of the molecules. Through further feature extraction, the data is condensed down for better efficiency and performance during the machine-learning phase. Finally, the inhibitor predicting models are deployed to identify potential inhibitors to TNF-α, IL-6, and IL-2.

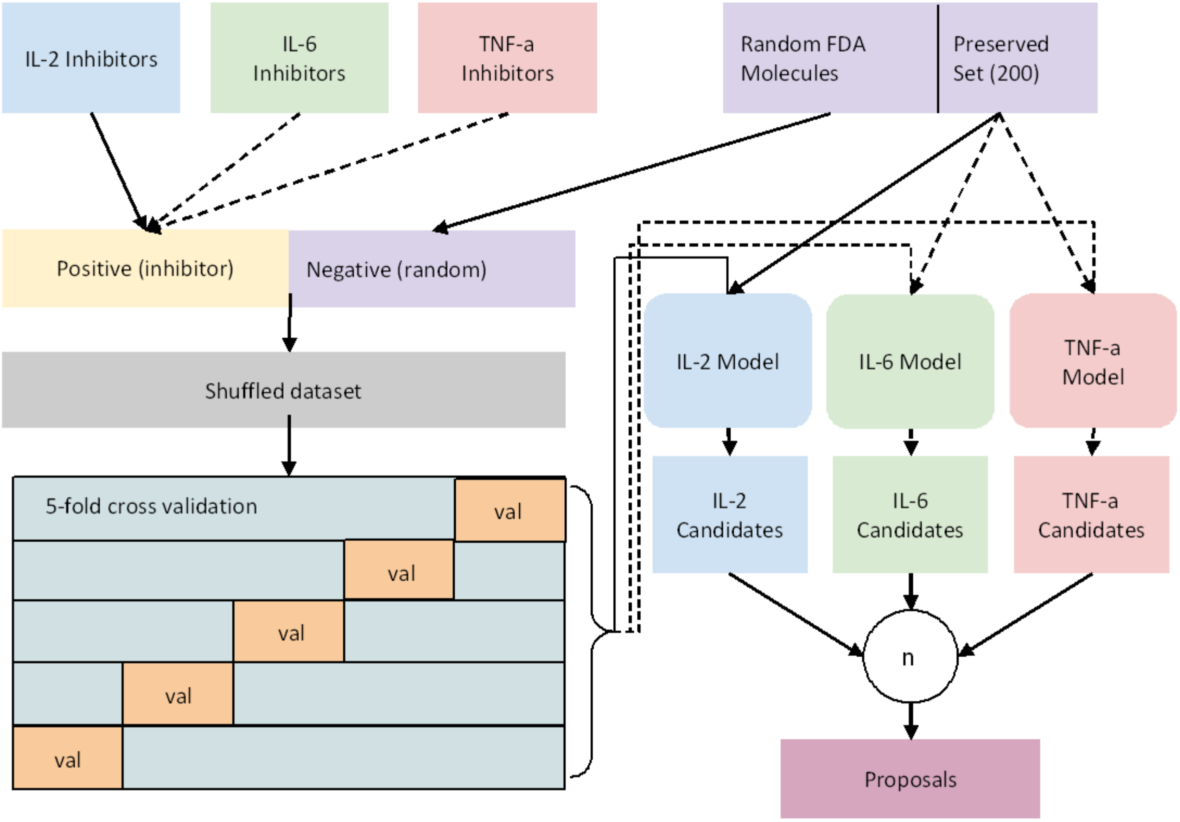

Figure 2: A flowchart of data organization and process that details how cytokine-specific models are trained and then used to make a multi-targeted inference. Following the solid line, known IL-2 inhibitors form the true samples of the dataset, while randomly sampled molecules are used as false inhibitors. The dataset is then shuffled and divided into 5-folds, used in evaluating the trained model. The data is taken and used to train the inhibitor-predicting model. Following training, the preserved set of FDA molecules (unseen) is run through the model to identify potential inhibitors. We repeat the same process for the other cytokines following the dotted lines. The candidates are aggregated together to form the multi-target inhibitor predictions.

Note: Dotted lines indicate that the process done to IL-2 is repeated for other cytokine types.

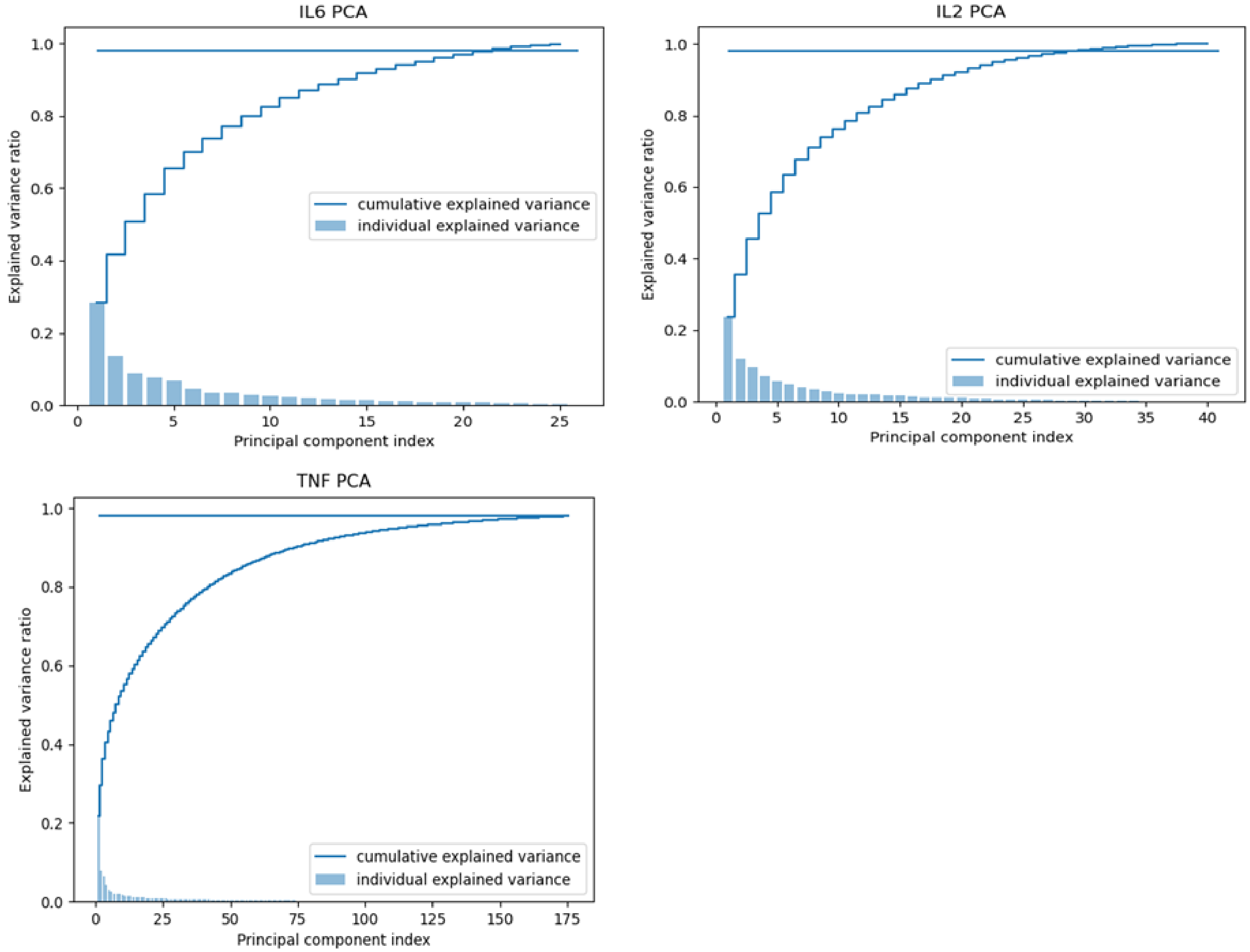

Figure 3: Energy graphs of Principle Component Analysis applied to (**A**) IL-6, (**B**) IL-2, and (**C**) TNF-α data. The energy graphs visualize the amount of "information" obtained from each successive eigenvector. The horizontal line is the boundary of 98%, which we use as the threshold to the number of eigenvectors needed.

We selected the smallest amount of eigenvectors that preserves 98% of the information and project the data onto the subspace spanned by the chosen eigenvectors, which left us with condensed features for training: IL-6 reduced to 25 features, IL-2 reduced to 40 features, and TNF-α reduced to 175 features. As to be discussed in Supplementary Materials, the addition of PCA significantly improved the performance of the models (Supplementary Tables S1–S3).

Machine Learning

For each type of cytokine, we trained a model with its respective data. We selected Support Vector Machines (SVM) as the model to predict if a compound is an inhibitor to its cytokine type. SVM are a class of supervised learning methods that are very effective in high-dimensional space. They are also fit for the problem, as they are robust to limited data with large amounts of features.

During the training, the processed samples are submitted into the model and predictions are matched against labels to assess performance. The SVM model is trained with the Scikit Learn solver and the performance of the models are evaluated by final test accuracy.

Due to limitations of data, a thorough hyper parameter search was done to achieve the best performance (Table 1). In SVM, a number of kernel types can be applied to achieve optimal performance: Linear, Polynomial, Gaussian Radial Basis Function, Sigmoid. Another parameter is the gamma, which is used to weight the kernels (Auto or Scale). Through some grid search, the optimal combinations for each cytokine model were for TNF-α: Polynomial + Auto, IL-2: Polynomial + Scale, and IL-6: Sigmoid + Auto. To ensure that the model generalizes well, we ran 5-fold cross validation and averaged the accuracies to obtain the final one. In addition, L2 regularization (ridge regression) was applied to the models.

Table 1: Average percent accuracies over 5-fold cross validation for each type of SVM kernel. The row names are in the following format: cytokine target, kernel-scaling method. The column heads are the different types of SVM kernels. RBF—radial basis function.

	Linear	Polynomial	Gaussian RBF	Sigmoid
TNF-α, Auto	93.0	95.4	94.0	79.4
TNF-α, Scale	93.0	92.6	94.8	90.4
IL-6, Auto	68.6	70.7	54.3	81.1
IL-6, Scale	68.6	42.5	68.6	86.4
IL-2, Auto	81.2	86.2	58.5	70.6
IL-2, Scale	81.2	60.9	81.1	82.7

Last, we applied the trained models to the preserved set of FDA-approved drugs to identify candidates. The model outputs are +1 (inhibitor) and −1 (random molecule), and we collected those predicted as inhibitors. For better understanding, we also used Platt's Scaling, a learned function that maps real values to probability ranges, to obtain a probability. The predicted inhibitors for each type of cytokine were compared to find common drugs among them that inhibited all three types.

Results

The predicted inhibitors simultaneously affecting all three targets: TNF-α, IL-2, and IL-6 are presented in Table 2; affecting just two targets are presented in Table 3. The SVM model identified a promising candidate: glecaprevir, It is an FDA-approved drug that serves as an antiviral treatment against hepatitis-C virus. The drug is shown to be relatively safe with minimal to no genotoxicity. The predicted probabilities of inhibition for this drug for each type of model were for TNF-α—80.5216%, IL-2—99.0453%, and IL-6—98.2354%. All probabilities reported are created through Platt's Scaling, as mentioned in Section *Methods*, Subsection *Machine Learning*. Other potential candidates were identified in with the SVM model (efinaconazole, alvimopan, and olodaterol) though with a lower probability threshold, 35% (default is 50%).

Table 2: Identified potential simultaneous inhibitors of three cytokines IL-6, IL-2, and TNF-α. Indicated by asterisk (*) are compounds that have ≥ 50% probability. SVM—support vector machines, LR—logistic regression, GBT—gradient boosted trees.

SVM ≥ 35%	LR ≥ 35%	GBT ≥ 35%
efinaconazole	venetoclax*	acarbose
alvimopan		delafloxacin
glecaprevir*		doxycycline
olodaterol		iohexol
		mitoxantrone
		osimertinib*
		teniposide

Table 3: Inhibitors affecting simultaneously at least two cytokines.

TNF-α–IL2	TNF-α–IL6	IL6–IL2
methohexital ribociclib enzacamene triclocarban	N/A	ceftaroline fosamil valrubicin erythromycin ceftriaxone Synribo (omacetaxine) canqrelor vinblastine gadofosveset ioxilan dalfopristin ceftolozane simeprevir isavuconazonium tetracycline ceftotetan fluticasone furoate ecteinascidin spinosyn D CDTR-PI (cefditoren pivoxil) verteporfin acarbose idarubicin daunorubicin ombitasvir

Through other models (explored in Section *Results,* Subsection *Study of Different ML Algorithms*) we also found promising candidates venetoclax (through LR), a medication for lymphocytic leukemia, and osimertinib (through GBTs), and a drug to treat non-small-cell lung carcinomas. Venetoclax and osimertinib both interestingly reduce neutrophil count, which might affect cytokine production and immune cell recruitment. Some lower prediction score (35–50% probability) candidates identified were include acarbose, delafloxacin, doxycycline, iohexol, and mitoxantrone.

Study of Different ML Algorithms

Beyond the combination of PCA with SVM used for the results, other means of classification were explored. First, the standard MultiLayer Perceptron (MLP) was utilized (Figure 4).

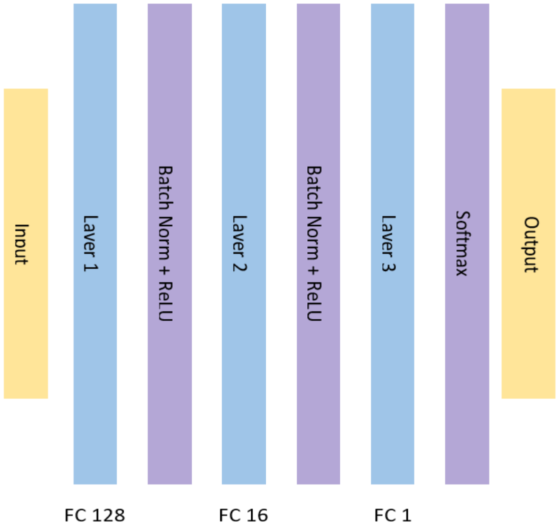

Figure 4: Model architecture of proposed MultiLayer Perceptron (MLP). FC—fully connected layer. ReLU—Rectified Linear Unit, BatchNorm—batch normalization The MLP model is composed of 3 feedforward blocks and the final probability is computed through a softmax function.

The MLP is constructed with three feed-forward layers and rectified linear unit (ReLU) as its activation function. The input of the model was not processed with PCA due to the model's property of implied feature extraction. The model was trained for 250 epochs with a learning rate of 0.01.

The second approach taken was a logistic regression model, a standard machine-learning algorithm used for classification and regression. Due to its limited modeling power, the logistic regression was done on PCA cleaned data. Using Scikit Learn, the model was trained for 2000 iterations and L1 regularization to prevent overfitting.

The last method was Gradient-Boosted Trees (GBTs), where trees are congregated based on gradients to make a collective decision. Similarly, the Scikit Learn solver was used to train the model at a learning rate of 0.1 and the input of the model was PCA-processed. After thorough evaluation with 5-fold cross validation, none of the models performed as well as the SVM in combination with PCA-processed data (Table 4).

Table 4: Model accuracies (%) for each type of cytokine. Measured across 5-fold cross validation.

	MLP (no PCA)	Logistic Regression	GB Trees	SVM
TNF-α	89.1	93.3	94.0	95.4
IL-6	86.0	76.4	81.4	86.4
IL-2	83.2	82.9	70.8	86.2

Discussion

Through the proposed technique, we identified a promising FDA-approved candidate for multi-cytokine inhibition (TNF-α, IL-6, IL-2) in glecaprevir with predicted probabilities of TNF-α—80.52%, IL-2—99.04%, and IL-6—98.23%. With the combination of PCA and SVMs, optimal accuracies of TNF-α—95.4%, IL-6—86.4%, IL-2—86.2% over 5-fold cross-validation to avoid overfitting. Our study suggests that clinical trials be tested on the efficacy of glecaprevir on cytokine storm inhibition. If successful, the candidate can significantly reduce the complications in severe COVID-19 cases without the immense side effects of a cocktail of drugs to combat multiple types of cytokines. Furthermore, this drug is not restricted to only COVID-19 induced cytokine storms but also for other cytokine storms involving TNF-α, IL-6, and IL-2.

A significant work of this project is its efficacy despite the limitations of data. Pulling from existing databases, only TNF-α had a significant number of valid samples (2260): IL-6 had 31 samples and IL-2 had 47. The lack of data posed a significant challenge to learning an effective machine-learning model to identify inhibitors. The challenge was overcome with the application of PCA to reduce the noise in the data and made the data easier to learn. This was further improved by the model selection, as SVMs are very good with small datasets. Overall, we suggest that for future similar experiments/projects, researchers can follow these proposed techniques for strong performance on small datasets.

What differentiates this work from previous attempts at cytokine inhibition is that it is targeted towards multiple types of cytokines. Through multiple criteria, we were able to identify compounds that are versatile enough for the range of cytokines. Previously, to inhibit a cytokine storm, a mix of single target drugs was administered, putting further stress on a weakened patient. Thus, condensing the treatments to one drug has significant upside for both recovery and inhibition.

This project is advantageous in that it can identify preapproved FDA drugs and apply them to different use cases. This controls risk of the proposed treatment as it has been proved relatively safe in vivo and has known side effects. The utilization of in-silico research for new treatments will continue to develop and provide more insightful options with faster iterations.

We recommend that future research and experimental testing the listed compounds and their efficacy against cytokine storms. Furthermore, we suggest further experimentation with the proposed technique for other disease-induced cytokine storms.

Conclusion

We developed a set of machine-learning models to predict the possible FDA-approved drugs that would simultaneously inhibit at least two of proteins related to immune response in COVIF-19: TNF-α, IL-2, and IL-6. One of the selected drugs—glecaprevir—is predicted to be a simultaneous strong inhibitor of all three cytokines.

References

1. Tisoncik JR, Korth MJ, Simmons CP, et al. Into the eye of the cytokine storm. Microbiol Mol Biol Rev. 2012; 76: 16-32.
2. Zhang JM, An J. Cytokines, inflammation, and pain. Int Anesthesiol Clin. 2007; 45: 27-37.
3. Costela Ruiz VJ, Illescas Montes R, PuertaPuerta JM, et al. SARS-CoV-2 infection: The role of cytokines in COVID-19 disease. Cytokine Growth Factor Rev. 2020; 54: 62-75.
4. Mustafa MI, Abdelmoneim AH, Mahmoud EM, et al. Cytokine storm in COVID 19 patients, its impact on organs and potential treatment by QTY code-designed detergent-free chemokine receptors. Mediators Inflamm. 2020; 8198963.
5. Ragab D, Eldin HS, Taeimah M, et al. The COVID-19 cytokine storm; What we know so far. Front Immunol. 2020; 11: 1446.
6. https://covid19.who.int/
7. Chen G, Wu D, Guo W, et al. Clinical and immunological features of severe and moderate coronavirus disease 2019. J Clin Invest. 2020; 130: 2620-2629.
8. Hu B, Huang S, Yin L. The cytokine storm and COVID 19. J Med Virol. 2021; 93: 250-256.
9. Pedersen SF, Ho YC. SARS-CoV-2: A storm is raging. J Clin Invest. 2020; 130: 2202-2205.
10. Qin C, Zhou L, Hu Z, et al. Dysregulation of immune response in patients with Coronavirus 2019 (COVID-19) in Wuhan, China. Clin Infect Dis. 2020; 71: 762-768.

11. Rokni M, Hamblin MR, Rezaei N. Cytokines and COVID-19: Friends or foes? Hum Vaccin Immunother. 2020; 16: 2363-2365.

12. Sinha P, Matthay MA, Calfee CS. Is a "Cytokine Storm" relevant to COVID-19? JAMA Intern Med. 2020; 180: 1152-1154.

13. Sun X, Wang T, Cai D, et al. Cytokine storm intervention in the early stages of COVID-19 pneumonia. Cytokine Growth Factor Rev. 2020; 53: 38-42.

14. Yap CW. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011; 32: 1466-1474.