Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associate cytokine storm

Anjali Dhall
IIIT-Delhi: Indraprastha Institute of Information Technology Delhi
Sumeet Patiyal
IIIT-Delhi: Indraprastha Institute of Information Technology Delhi
Neelam Sharma
IIIT-Delhi: Indraprastha Institute of Information Technology Delhi
Naorem Leimarembi Devi
IIIT-Delhi: Indraprastha Institute of Information Technology Delhi
Gajendra P. S. Raghava (raghava@iiitd.ac.in)
Indraprastha Institute of Information Technology Delhi https://orcid.org/0000-0002-8902-2876

Research Article

Keywords: STAT3 inhibitor, Machine Learning Techniques, COVID-19, FDA-approved drugs, cytokine storm

Posted Date: June 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-495671/v1

Abstract

It has been shown in the past that levels of cytokines, including interleukin 6 (IL6), is highly correlated with the disease severity of COVID-19 patients. IL6 mediated activation of STAT3 is responsible to proliferate proinflammatory responses that leads to promotion of cytokine storm. Thus, STAT3 inhibitors may play a crucial role in managing pathogenesis of COVID-19. This paper describes a method developed for predicting inhibitors against the IL6-mediated STAT3 signaling pathway. The dataset used for training, testing, and evaluation of models contains small-molecule based 1564 STAT3 inhibitors and 1671 non-inhibitors. Analysis of data indicates that rings and aromatic groups are significantly abundant in STAT3 inhibitors compared to non-inhibitors. In order to build models, we generate a wide range of descriptors for each chemical compound. Firstly, we developed models using 2-D and 3-D descriptors and achieved maximum AUC 0.84 and 0.73, respectively. Secondly, fingerprints (FP) are used to build prediction models and achieved 0.86 AUC and accuracy of 78.70% on validation dataset. Finally, models were developed using hybrid features or descriptors, achieve a maximum of 0.87 AUC on the validation dataset. We used our best model to identify STAT3 inhibitors in FDA-approved drugs and found few drugs (e.g., Tamoxifen, and Perindopril) that can be used to manage COVID-19 associated cytokine storm. A webserver “STAT3In” (https://webs.iiitd.edu.in/raghava/stat3in/) has been developed to predict and design STAT3 inhibitors.

Key Messages

- IL6 medicated activation of STAT3 is responsible for cytokine storm.
- STAT3 inhibition is an important strategy to mitigate COVID-19 severity.
- In silico models have been developed for predicting STAT3 inhibitors.
- Repurposing of FDA-approved drug for managing COVID-19 via STAT3 inhibition.
- A webserver, “STAT3In” for designing and predicting STAT3 inhibitors.

Introduction

Several studies have been shown that the elevated level of pro-inflammatory cytokines, hyper-inflammatory responses, immune dysfunctions, and cytokine storm is associated with high mortality rate in COVID-19 patients. The activation of interleukin-6 (IL6) mediated STAT3 signaling pathway is correlated with the cytokine storm in COVID-19. The Janus kinase/signal transducer and activator of transcription (JAK/STAT) pathway plays a vital role in mediating signals to various cytokines, hormones, and growth factors [1]. In mammalian, STATs (STAT1, 2, 3, 4, 5a, 5b and 6) are cytoplasmic transcription factors (TF) which participate in normal cellular events, including differentiation, proliferation, and angiogenesis [2]. Out of the STATs family, STAT3 is a pleiotropic TF encoded by the STAT3 gene. STAT3 is activated in response to various cytokines including IL6, IL10, IL12, etc, chemokines and various growth factors [3, 4]. The binding of these factors to the cell surface receptor results in phosphorylation of JAKs by adding the phosphate group. Further phosphorylated STAT3 monomers form a homodimer molecule

which translocate into the nucleus and binds to the specific target gene promoters to regulate the gene transcription process [4], as shown in **Fig. 1**. However, upregulation of the STAT3 is linked with various pathological events contributing to cancer progression, proliferation, invasion, migration, angiogenesis, cytokine storm in COVID-19 and other diseases [5, 6]. They play a vital role in the pathogenesis of distinct human cancers, making a promising potential therapeutic target that is supported by various preclinical and clinical studies [7]. Aberration of STAT3 is participated in oncogenesis by increasing the mRNA levels of several genes (Bcl-xL, Fas, Fas-L, CASP3) involved in apoptosis, cell growth, and angiogenesis [8–10]. Increasing evidences indicated that mutations in the STAT3 gene have been associated with various auto-immune disorders like Type 1 diabetes (T1D) and other inflammatory diseases such as pulmonary fibrosis and acute lung injury [11–13].

During coronavirus infection, hyper-activation of STAT3 promotes COVID-19 pathogenesis, cytokine storm production, viral replication, immunopathological responses, development of M2-like macrophages which causes lung fibrosis and lymphopenia [1, 14–16]. Thus, it is necessary to target IL6 mediated STAT3 activation. Currently, several STAT3 inhibitors are in clinical trials and researchers have attempted to develop drugs against STAT3. One of the STAT3 inhibitor, pyrrolidinesulphonylaryl molecules (6a), has shown promising activity against IL6/STAT3 signaling in breast cancer [17]. STAT3 direct inhibitors which are under clinical trials in cancer immunotherapy include FDA-approved drugs such as Celecoxib, BBI608, Pyrimethamine, etc [18]. However, a number of STAT3 inhibitors is steadily increasing, and finding of novel STAT3 inhibitor remains a major scientific challenge even in current COVID-19 scenario. Thus, it is essential to target IL6/STAT3 signaling pathway in order to get a better therapeutic candidates against COVID-19. To the best of our knowledge, there is no computational tool that can accurately predict STAT3 inhibitors.

In the current study, an attempt has been made to develop computational model for predicting STAT3 inhibitors. We have used 3236 chemical compounds (STAT3 inhibitors and non-inhibitors) and 16,112 (2-D, 3-D, and FP) descriptors, to generate prediction models (such as RF, DT, LR, XGB, SVM and GBM). Additionally, to serve the scientific community, we provide a computation tool “STAT3In” (https://webs.iiitd.edu.in/raghava/stat3in/) for the prediction and designing of potential STAT3 inhibitor candidates.

Materials And Methods

Dataset Collection

In this study, we have obtained STAT3 inhibitors and non-inhibitors from the PubChem bioassay record (AID 862) (https://pubchem.ncbi.nlm.nih.gov/bioassay/862). In this bioassay, a total of 194,698 compounds were tested to check their ability to prevent or reduce IL6-mediated STAT3 transcription. We collected a total of 194,698 chemical compounds with STAT3 inhibition and non-inhibition activity from this bioassay, which comprises 1724 inhibitors and 192974 non-inhibitors chemical compounds. We [redacted] 192974 non-inhibitors chemicals. 1724 chemical

compounds with IL6-mediated STAT3 inhibition property were considered positive datasets and called inhibitors, and 1724 chemical compounds that do not have the inhibition property were considered negative datasets and known as non-inhibitors.

After that, PubChem substance IDs and compound IDs were used to download the 2-D and 3-D structure files for the 1724 positive and negative dataset. However, out of 1724 compounds, only 1565 inhibitors and 1671 non-inhibitors compound structures were available. So, the final dataset contains 1565 positive and 1671 negative chemical compounds. To evaluate the performance of the model, we have divided the whole dataset into an 80:20 ratio. 80% of the data has been taken as a training set, which comprises 1265 positive and 1323 negative chemical compounds. The remaining 20% of data has been taken as a validation set consisting of 300 positive and 348 negative chemical compounds.

Descriptors of molecules

Chemical descriptors are the representative features of chemical molecules that are responsible for their activity. In this study, we have used PaDEL software [19] to calculate the molecules' descriptors. This software can compute a number of molecular descriptors for a single chemical compound. It computes several 1-D/2-D/3-D and binary fingerprints (FP) (e.g., Fingerprinter, Extended, KlekotaRoth count, SubStructure, MACCS keys, etc). We have computed 1444 2-D descriptors, 136 3-D descriptors, and 14532 FP descriptors for the 1564 positive and 1671 negative dataset in the current study. This 2-D, 3-D, and FP descriptors were used to develop various machine learning models.

Pre-processing of Data

The calculated descriptors were lying in a varying range, so to pre-process the dataset, we have normalized each descriptor file using a standard scaler package of scikit learn. sklearn.preprocessing.StandardScaler is a method that uses a z-score algorithm for normalizing the data. After normalizing the data, we have also removed the null values from each descriptor file. 2-D and FP descriptor files do not have any null values; only a few null values were found in 3-D descriptor file. After removing those null values, we were left with 1444 2-D, 116 3-D, and 14532 FP descriptors/features for the whole dataset.

Selection and ranking of significant descriptors

Previous studies have shown that all the descriptors calculated using PaDEL are not relevant [20, 21]. Therefore, selecting the most significant descriptors is a vital step to develop any prediction model. In this study, we have used three major feature selection techniques (i) VarianceThreshold-based method, (ii) correlation-based method, and (iii) SVC-L1-based method. We have used the VarianceThreshold package of scikit (sklearn.feature_selection) to remove the low-variance features from all the descriptors. Initially, there were 1444 2-D, 116 3-D, and 14532 FP descriptors, and after removing low variance features, we were left with 622 2-D, 66 3-D, and 2251 FP descriptors. After that, a correlation-based feature selection method was used to select those features which correlate less than 0.6 (< 0.6) with each other. Thus, we have removed the features which have a correlation of greater than or equal to 0.6 (> = 0.6). After that, we

were left with 74 2-D, nine 3-D, and 1622 FP out of 622 2-D, 66 3-D, and 2251 FP descriptors, respectively. Finally, we have used the SVC-L1 feature selection technique to get the most significant feature set. This is a popular method often used to minimize the feature vector size. Using the SVC-L1 approach, we get the most important 41 2-D, five 3-D, and 116 FP descriptors. Further, we used the combination of 2D + 3D + FP descriptors to develop a hybrid model. These 162 features were ranked based on their importance to classify inhibitor/non-inhibitors using the feature-selector program. This program utilizes Gradient Boosting Decision Tree (GBDT), also known as LightGBM, a popular machine learning algorithm, to rank the features. It estimates the rank of the feature by calculating that how many times a feature is used to split the data across all trees [22]. The features selected and ranked by this method were used to develop different machine learning models. The performance of the models was computed on top 10,20,30,…..,116 features, respectively.

Five-fold cross-validation

In this study, we have used the standard 5-fold CV technique to evaluate our prediction model. The whole dataset was partitioned into 80:20 proportion, forming 80% training dataset and 20% external validation dataset. This technique was applied to the training dataset, where the 80% training dataset was divided into five sets of the same size. Among these five sets, four sets are used for training, and the remaining fifth set will be used for testing purposes. The same process is iterated five times so that each of the five sets will be used at least once for testing the model. These five training and testing sets were used to develop the prediction models. Then the overall performance of the model was evaluated on the 20% external validation dataset. It is a standard method that has been used extensively in the past [23–25].

Machine Learning based classifiers

In this study, we have used several machine learning techniques to develop the prediction models to classify STAT3 inhibitors/non-inhibitors. We have implemented Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Support Vector Classifier (SVC), Gaussian Naive Bayes (GNB), K-nearest neighbor (KNN), and XGBoost (XGB) to develop classification models. These machine learning methods have been implemented using Scikit's sklearn package [26].

Performance evaluation parameters

In order to evaluate the performance of different prediction models, we have used the standard evaluation parameters. In this study, we have used both threshold-dependent and independent parameters. The model's performance was measured using threshold-dependent parameters such as sensitivity (Sens), specificity (Spec), accuracy (Acc), and Matthews correlation coefficient (MCC). On the other hand, the threshold-independent parameter, i.e., the area under receiver operating characteristic curve (AUC), was also used to evaluate the performance of the model. These parameters are well established in the literature and are extensively used in many studies for evaluating the model's performance evaluation [27, 28].

$$\text{Sensitivity}(Sens) = \frac{TP}{TP + FN} \times 100$$

1

$$Specificity(Spec) = \frac{TN}{TN + FP} \times 100$$

2

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FN + FP} \times 100$$

3

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100$$

4

Where FP, FN, TP, and TN are false positive, false negative, true positive, and true negative predictions, respectively.

Results

Functional groups analysis

To calculate the frequency of different functional groups of IL6-mediated STAT3 inhibitors (positive dataset) and non-inhibitors (negative dataset), we utilized the ChemmineR package [29]. We analysed average frequency values and found that in the positive dataset, the abundance of rings, and aromatic groups is significantly higher as compared to non-inhibitors. On the other hand, the frequency of secondary amines (R2NH), tertiary amines (R3N), and ester (ROR) groups is significantly elevated in non-inhibitors compounds, i.e., STAT3 non-inhibitors, as represented in **Figure 2**.

Additionally, we also observed that the occurrence of the rings and aromatic groups in few existing STAT3 inhibitors such as Napabucasin (BBI608), an FDA-approved drug used for treating advanced malignancies, and STAT3 Inhibitor VII (STAT3-IN-8) drug is used for head and neck cancer treatment. Some indirect STAT3 inhibitors like AZD-1480 and Ruxolitinib (FDA-approved) also show similar trends. **Figure 3** shows the presence of these functional groups in the chemical 2-D-structures of known STAT3 inhibitors, i.e., STAT3 Inhibitor VII, Ruxolitinib, AZD-1480, and BBI608. These findings suggest that the researcher can further utilize this analysis to design novel drug candidates to be used as an inhibitor for the STAT3 signaling pathway.

Prediction models

One of the major challenge in this type of study is to classify chemicals based on the 2-D, 3-D, and FP descriptors is that all descriptors are not relevant. Thus, several feature selection techniques have been

used to get the best set of features for the classification. After selecting best feature, we developed several prediction models with the help of machine learning based classifiers such as RF, DT, LR, XGB, SVM, and GBM. The complete architecture of the study is shown in **Figure 4.**

Performance of classification models

2-D descriptors

We compute 1444 2-D descriptors for the positive and negative dataset. After removing low variance and highly correlated features, we get 74 features. With this feature set, we have developed the classification models. RF attains maximum performance with balanced sensitivity and specificity on the training dataset (AUC = 0.84) and validation (AUC = 0.84) dataset, and the complete information is available in Supplementary Table S1. Further, we obtained 41 2-D descriptors with the help of the SVC-L1 method. After reducing the features, there is a slight change in the AUC 0.83 and 0.84; accuracy 76.35% and 75.46% on training and validation dataset with the RF classifier (Table 1).

Table1: Performance of machine-learning models on training and validation dataset with best 41 2-D descriptors.

Classifier	Training Dataset					Validation Dataset				
	Sensitivity	Specificity	Accuracy	AUC	MCC	Sensitivity	Specificity	Accuracy	AUC	MCC
DT	64.15	64.29	64.22	0.69	0.28	72.24	59.74	66.20	0.73	0.32
RF	76.10	76.58	76.35	0.83	0.53	74.63	76.36	75.46	0.84	0.51
LR	69.68	69.00	69.32	0.75	0.39	71.64	69.01	70.37	0.77	0.41
XGB	71.55	71.80	71.68	0.78	0.43	72.54	70.93	71.76	0.80	0.44
KNN	70.33	70.40	70.36	0.77	0.41	70.75	70.93	70.83	0.79	0.42
GNB	65.20	66.13	65.69	0.70	0.31	69.55	68.05	68.83	0.73	0.38
SVM	74.80	73.79	74.27	0.81	0.49	71.34	74.76	72.99	0.81	0.46

3-D descriptors

Based on nine selected 3-D descriptors, RF performs best among all the classifiers with balanced sensitivity, specificity on training (AUC = 0.75), and validation dataset (AUC = 0.74), as shown in Supplementary Table S2. After removing four features with the help of SVC,-L1, the performance is computed on the best five 3-D descriptors. In this case, RF outperforms all other classifiers with highest AUC (0.741 and 0.729) on training and testing data. Whereas, XGB and performers quite well AUC~0.73 on training data and AUC~0.71 on validation data, as shown in Table 2.

Table 2: Performance of ML-based models 5 selected 3-D descriptors on training and validation dataset

Classifier	Training Dataset					Validation Dataset				
	Sensitivity	Specificity	Accuracy	AUC	MCC	Sensitivity	Specificity	Accuracy	AUC	MCC
DT	64.80	62.00	63.33	0.68	0.27	67.16	51.76	59.72	0.66	0.19
RF	67.15	66.35	66.73	0.74	0.34	66.27	65.18	65.74	0.73	0.31
LR	65.77	65.54	65.65	0.71	0.31	65.67	64.54	65.12	0.70	0.30
XGB	65.29	66.94	66.15	0.73	0.32	65.67	66.13	65.90	0.72	0.32
KNN	68.21	67.01	67.58	0.74	0.35	69.85	62.62	66.36	0.73	0.33
GNB	65.85	65.69	65.77	0.71	0.32	67.46	61.98	64.82	0.70	0.30
SVM	66.91	66.50	66.69	0.73	0.33	66.87	65.18	66.05	0.71	0.32

FP descriptors

Further, we developed classification models based on FP descriptors. Firstly, we used 1622 features after removing low variance and highly correlated descriptors. RF algorithm achieves maximum performance with AUC (0.86) on balanced sensitivity and specificity on both training and validation dataset. In this case, SVM also achieve comparable performance i.e., AUC (training data = 0.84 and testing data = 0.85), results of XGB, GBM, LR, DT, KNN is provided in Supplementary Table S3. We also developed models 116 features selected using SVC-L1 method and achieved nearly same performance (Table 3). These results shows that FP based models outperform all the classification models based on 2-D and 3-D chemical features.

Table 3: The performance of machine learning models on 116 FP based features on training and validation dataset

Classifier	Training Dataset					Validation Dataset				
	Sensitivity	Specificity	Accuracy	AUC	MCC	Sensitivity	Specificity	Accuracy	AUC	MCC
DT	64.96	65.24	65.11	0.71	0.30	67.46	61.66	64.66	0.70	0.29
RF	78.46	77.61	78.01	0.86	0.56	79.40	77.96	78.70	0.86	0.57
LR	75.85	76.66	76.28	0.83	0.53	72.84	76.68	74.69	0.81	0.50
XGB	77.32	77.54	77.43	0.84	0.55	77.91	80.83	79.32	0.86	0.59
KNN	76.18	75.04	75.58	0.83	0.51	77.02	73.80	75.46	0.83	0.51
GNB	73.98	74.08	74.03	0.81	0.48	69.55	73.80	71.61	0.79	0.43
SVM	78.62	78.35	78.48	0.86	0.57	77.31	80.19	78.70	0.86	0.58

Performance of Hybrid model

In order to improve the performance, we combine 2-D (41 features), 3-D (5 features), and FP(116 features) descriptors and developed the models using 162 features. The performance of RF based models using combined features was 0.87 and 0.88 AUC on training and validation dataset respectively (See Supplementary Table S4). We further perform feature ranking on the combined 162 features with the

e obtained a minimum set of features which have almost

similar performance as the above mentioned models. First we rank the features and then check the performance of top-10, 20, 30,.....162 features. Finally, we select top-49 descriptors (i.e., 14 2-D, 1 3-D and 34 FP) out of 162 feature-set as represented in Supplementary Table S4. Top-49 features perform almost similar as 162 features. RF obtained maximum AUC of 0.87, and accuracy >78.5 on both training and testing dataset with minimum sensitivity and specificity difference. The results of all other classifiers i.e., SVM, DT, KNN, LR, XGB and GBM is shown in the Table 4.

Table 4: The performance of machine learning based on hybrid model (2-D+3-D+FP) descriptors on training and validation dataset

Classifier	Training Dataset					Validation Dataset				
	Sensitivity	Specificity	Accuracy	AUC	MCC	Sensitivity	Specificity	Accuracy	AUC	MCC
DT	68.22	68.03	68.12	0.74	0.36	66.67	72.70	69.91	0.74	0.39
RF	78.42	78.61	78.52	0.87	0.57	79.00	78.16	78.55	0.87	0.57
LR	77.00	76.34	76.66	0.84	0.53	75.67	77.87	76.85	0.83	0.54
XGB	77.31	77.10	77.20	0.85	0.54	80.00	75.29	77.47	0.85	0.55
KNN	74.94	75.89	75.43	0.83	0.51	78.00	75.58	76.70	0.83	0.53
GNB	74.23	74.00	74.11	0.81	0.48	75.33	72.99	74.07	0.80	0.48
SVM	77.71	77.55	77.63	0.86	0.55	78.33	76.72	77.47	0.85	0.55

Repurposing of FDA-approved drugs to target STAT3

In order to identify the potential drug candidates for the inhibition of the IL6/STAT3 pathway, we downloaded 1102 FDA-approved drug molecules from the Drug Bank database [30]. Then we trace the PubChem CID (compound ID) of the FDA-approved drugs. Out of 1102 drugs the 2-D structures were available for only 842 drugs. Further, we use SDF files of 842 molecules, for the identification of potential drug candidates. We have utilized “Predict” module of our web server “STAT3In” (with default parameters, i.e., Random Forest Threshold =0.48). Our model is able to predict 19 potential drug candidates that can be used as STAT3 inhibitors. Several past studies also support our findings that these drugs act as potential inhibitors in various diseases which are linked with IL6/STAT3 activation [31-35]. We identify eight potential drugs (such as warfarin, dexpanthenol, perindopril, tamoxifen, pentagastrin, duloxetine, ledipasvir, and, olopatadine) which are utilized in the treatment of severe disease conditions like tumor progression, angiogenesis, COVID-19 progression, by inhibiting IL6/STAT3 pathway, as depicted in **Table 5.**

Table 5: Potential FDA-approved drug candidates predicted by our web server (STAT3In) for STAT3 inhibition

Drug Bank ID	FDA-Approved Drugs	STAT3In Prediction	Functions
DB00682	Warfarin	Inhibitor	Inhibition of IL6/STAT3-dependent fibrin production in severe listeriosis [32]
DB09357	Dexpanthenol	Inhibitor	Inhibition of LPS-induced neutrophils influx, protein leakage, and release of TNF-α and IL6 in bronchoalveolar lavage fluid (BALF) in acute lung injury [33]
DB00790	Perindopril	Inhibitor	It regulates the inflammatory mediators, NF-κB/TNF-α/IL6, and apoptosis in renal diseases [34] and inhibit the activation of STAT3 [35]. ACE inhibitor perindopril-inhibited tumor growth was associated with the suppression of angiogenesis [36].
DB00675	Tamoxifen	Inhibitor	Treatment of ER-positive breast cancer with tamoxifen by inhibiting the IL6/STAT3 signal pathway, Inhibition of tumor growth and angiogenesis [37, 38]. Anticancer drugs that have shown potential activity in both MERS and SARS-CoV [31].
DB00183	Pentagastrin	Inhibitor	Anti-malarial, anti-fungal, anti-bacterial, and anti-inflammatory [39].
DB00476	Duloxetine	Inhibitor	Inhibit overexpression of IL6 mRNA in anxiety- and major depressive disorder (MDD), anti-inflammatory action against IL6 [40-42].
DB09027	Ledipasvir	Inhibitor	Anti-viral activity against COVID-19 [43], (sofosbuvir, and ledipasvir) inhibited STAT3 protein levels to cure HCV infections [44].
DB00768	Olopatadine	Inhibitor	Inhibit CHMCs activation and release of IL6, tryptase, and histamine and use as anti-allergy drug [45].

Webserver Implementation

In order to assist the scientific society, we have developed a webserver, “STAT3In,” with the capability to classify STAT3 inhibitors. We have used HTML5, JAVA, CSS3, and PHP scripts to build the web server’s front- and back end. The STAT3In web server is compatible with various devices such as mobile, iPad, tablet, and desktop, and different browsers. We have implemented the random forest model developed using hybrid chemical descriptors as the input features, in the back-end of the server. There are three major modules in the web server, defined as “Predict,” “Draw,” and “Analog design”. The comprehensive description of each module is proffered below.

Predict

The predict module allows users to classify the uncharacterized chemical compound as STAT3 inhibitor or non-inhibitor. This module can accept chemical compounds in various formats, such as SDF, SMILES, and MOL, from the users and also allow them to select the desired threshold. The users can enter either a

upload a file consisting of multiple chemical compounds.

The output page provides the class(es) of the submitted compound(s) as STAT3 inhibitor or non-inhibitor, along with their machine learning score. The result is downloadable in comma-separated value (CSV) format and also allows users to search or sort the output table.

Draw

This module allows users to draw or alter the chemical molecule structure and provide it to the prediction model to classify the molecule as a STAT3 inhibitor or non-inhibitor. In order to make the process interactive, we have implemented Ketcher [46], which is an open-source web-based chemical structure editor. The user is allowed to select the threshold as per their suitability. The output page shows the predicted class of the molecule in the tabular form, which is downloadable in CSV format.

Analog Design

The analog design module allows users to generate the analogs using combination of submitted scaffolds, building blocks, and linkers. We have implemented SmiLib [47] software to generate the analogs. Subsequently, the generated analogs are classified into STAT3 inhibitors or non-inhibitors based on the selected threshold. The result page exhibits the class of the generated analogs as inhibitors and non-inhibitors along with their machine learning score in the tabular form, which is downloadable in the CSV format.

Discussion And Conclusion

STAT3 is one of the most crucial transcription factors and oncogene, which plays a significant role in the onset and progression of the tumor. Hence, it could be an excellent therapeutic target for various cancer therapies due to its versatile regulatory pathways and crucial biological roles in cancer [48]. Moreover, it has been shown in the literature that the level of IL6 is quite elevated in coronavirus infected patients, who are increasing exponentially worldwide. Cytokine IL6 mediates its effect via JAK/STAT3 pathway; therefore, it is the need of the hour to develop the computation methods that can predict a chemical molecule's potency to be a STAT3 inhibitor. There are numerous method developed in the past which exploits the structure-activity relationship of the chemical molecules to predict the potential of a chemical molecule to be an inhibitor such as EGFRPred [20] which predicts the EGFR inhibitor potential of a molecule, DrugMint [21] predicts if a molecule could be a potential drug candidate, using machine learning methods, etc. In this study, we consider inhibitors and non-inhibitors IL6-mediated STAT3 inhibitors. From the functional group analysis, we identify that aromatic and rings found in most of the STAT3 inhibitors and R2NH, R3N, ROR groups are significantly elevated in STAT3 non-inhibitors compounds. Literature evidences also prove that the presence of such type of functional groups in the STAT3 inhibitor drugs such as AZD-1480, Ruxolitinib, Napabucasin, and STAT3-In-8. So, with the knowledge of such type of information biologist can design novel inhibitors against STAT3 activity. In this study, we take STAT3 inhibitors and non-inhibitors as positive and negative dataset for the generation of prediction models. Random forest based models achieve maximum performance (i.e., AUC sensitivity and specificity on validation dataset using hybrid

descriptors. Further, we have taken 842 FDA-approved drugs, for the identification of potential drug candidate against STAT3 activation. We identify few drugs as mentioned in Table 5, which can inhibit IL6/STAT3 activation and used as drug candidate against cytokine storm [49, 50] associated with COVID-19. Using machine learning techniques with minimal features derived from chemical molecules a webserver named, “STAT3In”, is developed to predict and design the potential STAT3 inhibitors. Considering the importance of STAT3 in the biological system, we hope this method will aid researchers working in the field of cancer therapy and infectious diseases, including COVID-19.

Declarations

Acknowledgements

Authors are thankful to the Department of Science and Technology (DST-INSPIRE) and DBT for fellowships and the financial support and Department of computational biology, IIITD New Delhi for infrastructure and facilities.

Funding

Not applicable

Conflict of Interest

The authors declare no competing financial and non-financial interests.

Availability of data and material

All the datasets generated for this study are either included in this article are available at the “STAT3In” webserver, https://webs.iiitd.edu.in/raghava/stat3in/dataset.php.

Code availability

Not applicable

Authors’ contributions

AD, and SP collected and processed the datasets. AD implemented the algorithms and developed the prediction models. AD, SP, NS and GPSR analysed the results. SP created the back-end of the web server and AD created the front-end user interface. AD, NS, NLD, SP and GPSR penned the manuscript. GPSR conceived and coordinated the project, and gave overall supervision to the project. All authors have read and approved the final manuscript.

Ethics approval

Not applicable

Consent for publication

Not applicable

References

1. Jafarzadeh A, Nemati M, Jafarzadeh S (2021) Contribution of STAT3 to the pathogenesis of COVID-19. Microb Pathog 154:104836. DOI 10.1016/j.micpath.2021.104836
2. Calo V, Migliavacca M, Bazan V, Macaluso M, Buscemi M, Gebbia N, Russo A (2003) STAT proteins: from normal control of cellular events to tumorigenesis. J Cell Physiol 197:157–168. DOI 10.1002/jcp.10364
3. Levy DE, Lee CK (2002) What does Stat3 do? J Clin Invest 109:1143–1148. DOI 10.1172/JCI15650
4. Ma JH, Qin L, Li X (2020) Role of STAT3 signaling pathway in breast cancer. Cell Commun Signal 18:33. DOI 10.1186/s12964-020-0527-z
5. Lee H, Jeong AJ, Ye SK (2019) Highlighted STAT3 as a potential drug target for cancer therapy. BMB Rep 52:415–423
6. Corvinus FM, Orth C, Moriggl R, Tsareva SA, Wagner S, Pfitzner EB, Baus D, Kaufmann R, Huber LA, Zatloukal K et al (2005) Persistent STAT3 activation in colon cancer is associated with enhanced cell proliferation and tumor growth. Neoplasia 7:545–555. DOI 10.1593/neo.04571
7. Wong ALA, Hirpara JL, Pervaiz S, Eu JQ, Sethi G, Goh BC (2017) Do STAT3 inhibitors have potential in the future for cancer therapy? Expert Opin Investig Drugs 26:883–887. DOI 10.1080/13543784.2017.1351941
8. Furqan M, Mukhi N, Lee B, Liu D (2013) Dysregulation of JAK-STAT pathway in hematological malignancies and JAK inhibitors for clinical application. Biomark Res 1:5. DOI 10.1186/2050-7771-1-5
9. Banerjee K, Resat H (2016) Constitutive activation of STAT3 in breast cancer cells: A review. Int J Cancer 138:2570–2578. DOI 10.1002/ijc.29923
10. Buettner R, Mora LB, Jove R (2002) Activated STAT signaling in human tumors provides novel molecular targets for therapeutic intervention. Clin Cancer Res 8:945–954
11. Gao H, Guo RF, Speyer CL, Reuben J, Neff TA, Hoesel LM, Riedemann NC, McClintock SD, Sarma JV, Van Rooijen N et al (2004) Stat3 activation in acute lung injury. J Immunol 172:7703–7712. DOI 10.4049/jimmunol.172.12.7703
12. Yang XO, Panopoulos AD, Nurieva R, Chang SH, Wang D, Watowich SS, Dong C (2007) STAT3 regulates cytokine-mediated generation of inflammatory helper T cells. J Biol Chem 282:9358–9363. DOI 10.1074/jbc.C600321200
13. Shao S, He F, Yang Y, Yuan G, Zhang M, Yu X (2012) Th17 cells in type 1 diabetes. Cell Immunol 280:16–21. DOI 10.1016/j.cellimm.2012.11.001

14. Gubernatorova EO, Gorshkova EA, Polinova AI, Drutskaya MS (2020) IL-6: Relevance for immunopathology of SARS-CoV-2. Cytokine Growth Factor Rev 53:13–24. DOI 10.1016/j.cytogfr.2020.05.009
15. Jafarzadeh A, Jafarzadeh S, Nozari P, Mokhtari P, Nemati M (2021) Lymphopenia an important immunological abnormality in patients with COVID-19: Possible mechanisms. Scand J Immunol 93:e12967. DOI 10.1111/sji.12967
16. Chen X, Tang J, Shuai W, Meng J, Feng J, Han Z (2020) Macrophage polarization and its role in the pathogenesis of acute lung injury/acute respiratory distress syndrome. Inflamm Res 69:883–895. DOI 10.1007/s00011-020-01378-2
17. Zinzalla G, Haque MR, Basu BP, Anderson J, Kaye SL, Haider S, Hasan F, Antonow D, Essex S, Rahman KM et al (2010) A novel small-molecule inhibitor of IL-6 signalling. Bioorg Med Chem Lett 20:7029–7032. DOI 10.1016/j.bmcl.2010.09.117
18. Zou S, Tong Q, Liu B, Huang W, Tian Y, Fu X (2020) Targeting STAT3 in Cancer Immunotherapy. Mol Cancer 19:145. DOI 10.1186/s12943-020-01258-7
19. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. DOI 10.1002/jcc.21707
20. Singh H, Singh S, Singla D, Agarwal SM, Raghava GP (2015) QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. Biol Direct 10:10. DOI 10.1186/s13062-015-0046-9
21. Dhanda SK, Singla D, Mondal AK, Raghava GP (2013) DrugMint: a webserver for predicting and designing of drug-like molecules. Biol Direct 8:28. DOI 10.1186/1745-6150-8-28
22. Ke G et al (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30:3146–3154
23. Dhall A, Patiyal S, Kaur H, Bhalla S, Arora C, Raghava GPS (2020) Computing Skin Cutaneous Melanoma Outcome From the HLA-Alleles and Clinical Characteristics. Front Genet 11:221. DOI 10.3389/fgene.2020.00221
24. Sharma N, Patiyal S, Dhall A, Pande A, Arora C, Raghava GPS (2020) AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. Brief Bioinform. DOI 10.1093/bib/bbaa294
25. Patiyal S, Agrawal P, Kumar V, Dhall A, Kumar R, Mishra G, Raghava GPS (2020) NAGbinder: An approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. Protein Sci 29:201–210. DOI 10.1002/pro.3761
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12:2825–2830
27. Kaur H, Dhall A, Kumar R, Raghava GPS (2019) Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data. Front Genet 06

28. Bhalla S, Kaur H, Dhall A, Raghava GPS (2019) Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. Sci Rep 9:15790. DOI 10.1038/s41598-019-52134-4
29. Cao Y, Charisi A, Cheng LC, Jiang T, Girke T (2008) ChemmineR: a compound mining framework for R. Bioinformatics 24:1733–1734. DOI 10.1093/bioinformatics/btn307
30. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074–D1082. DOI 10.1093/nar/gkx1037
31. Huang J, Rohatgi A, Schneider J, Braunstein M (2020) Considerations for the Management of Oncology Patients During the COVID-19 Pandemic. Oncology (Williston Park) 34:432–441. DOI 10.46883/ONC.2020.3410.0432
32. Nishanth G, Deckert M, Wex K, Massoumi R, Schweitzer K, Naumann M, Schluter D (2013) CYLD enhances severe listeriosis by impairing IL-6/STAT3-dependent fibrin production. PLoS Pathog 9:e1003455. DOI 10.1371/journal.ppat.1003455
33. Li-Mei W, Jie T, Shan-He W, Dong-Mei M, Peng-Jiu Y (2016) Anti-inflammatory and Anti-oxidative Effects of Dexpanthenol on Lipopolysaccharide Induced Acute Lung Injury in Mice. Inflammation 39:1757–1763. DOI 10.1007/s10753-016-0410-7
34. Shalkami AS, Hassan MIA, Abd El-Ghany AA (2018) Perindopril regulates the inflammatory mediators, NF-kappaB/TNF-alpha/IL-6, and apoptosis in cisplatin-induced renal dysfunction. Naunyn Schmiedebergs Arch Pharmacol 391:1247–1255. DOI 10.1007/s00210-018-1550-0
35. Bhat SA, Goel R, Shukla R, Hanif K (2016) Angiotensin Receptor Blockade Modulates NFkappaB and STAT3 Signaling and Inhibits Glial Activation and Neuroinflammation Better than Angiotensin-Converting Enzyme Inhibition. Mol Neurobiol 53:6950–6967. DOI 10.1007/s12035-015-9584-5
36. Yang Y, Ma L, Xu Y, Liu Y, Li W, Cai J, Zhang Y (2020) Enalapril overcomes chemoresistance and potentiates antitumor efficacy of 5-FU in colorectal cancer by suppressing proliferation, angiogenesis, and NF-kappaB/STAT3-regulated proteins. Cell Death Dis 11:477. DOI 10.1038/s41419-020-2675-x
37. Xing J, Li J, Fu L, Gai J, Guan J, Li Q (2019) SIRT4 enhances the sensitivity of ER-positive breast cancer to tamoxifen by inhibiting the IL-6/STAT3 signal pathway. Cancer Med 8:7086–7097. DOI 10.1002/cam4.2557
38. Kim JW, Gautam J, Kim JE, Kim JA, Kang KW (2019) Inhibition of tumor growth and angiogenesis of tamoxifen-resistant breast cancer cells by ruxolitinib, a selective JAK2 inhibitor. Oncol Lett 17:3981–3989. DOI 10.3892/ol.2019.10059
39. Balakrishnan V, Lakshminarayanan K (2020) Screening of FDA Approved Drugs Against SARS-CoV-2 Main Protease: Coronavirus Disease. Int J Pept Res Ther: 1–8. DOI 10.1007/s10989-020-10115-6
40. Zhang X, Wang Q, Wang Y, Hu J, Jiang H, Cheng W, Ma Y, Liu M, Sun A, Zhang X et al (2016) Duloxetine prevents the effects of prenatal stress on depressive-like and anxiety-like behavior and hippocampal expression of pro-inflammatory cytokines in adult male offspring rats. Int J Dev [illegible]yneu.2016.09.005

41. Dionisie V, Filip GA, Manea MC, Manea M, Riga S (2021) The anti-inflammatory role of SSRI and SNRI in the treatment of depression: a review of human and rodent research studies. Inflammopharmacology 29:75–90. DOI 10.1007/s10787-020-00777-5
42. Jansen van Vuren E, Steyn SF, Brink CB, Moller M, Viljoen FP, Harvey BH (2021) The neuropsychiatric manifestations of COVID-19: Interactions with psychiatric illness and pharmacological treatment. Biomed Pharmacother 135:111200. DOI 10.1016/j.biopha.2020.111200
43. Chen YW, Yiu CB, Wong KY (2020) Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL (pro)) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. F1000Res 9:129. DOI 10.12688/f1000research.22457.2
44. Aydin Y, Kurt R, Song K, Lin D, Osman H, Youngquist B, Scott JW, Shores NJ, Thevenot P, Cohen A et al (2019) Hepatic Stress Response in HCV Infection Promotes STAT3-Mediated Inhibition of HNF4A-miR-122 Feedback Loop in Liver Fibrosis and Cancer Progression. Cancers (Basel) 11. DOI 10.3390/cancers11101407
45. Kempuraj D, Huang M, Kandere K, Boucher W, Letourneau R, Jeudy S, Fitzgerald K, Spear K, Athanasiou A, Theoharides TC (2002) Azelastine is more potent than olopatadine n inhibiting interleukin-6 and tryptase release from human umbilical cord blood-derived cultured mast cells. Ann Allergy Asthma Immunol 88:501–506. DOI 10.1016/s1081-1206(10)62389-7
46. Karulin B, Kozhevnikov M (2011) Ketcher: web-based chemical structure editor. J Cheminformatics 3:3
47. Schüller A, Hähnke V, Schneider G (2007) SmiLib v2. 0: a Java-based tool for rapid combinatorial library enumeration. QSAR Comb Sci 26:407–410
48. Yuan J, Zhang F, Niu R (2015) Multiple regulation pathways and pivotal biological functions of STAT3 in cancer. Sci Rep 5:17663. DOI 10.1038/srep17663
49. Patiyal S, Kaur D, Kaur H, Sharma N, Dhall A, Sahai S, Agrawal P, Maryam L, Arora C, Raghava GPS (2020) A Web-Based Platform on Coronavirus Disease-19 to Maintain Predicted Diagnostic, Drug, and Vaccine Candidates. Monoclon Antib Immunodiagn Immunother 39:204–216. DOI 10.1089/mab.2020.0035
50. Dhall A, Patiyal S, Sharma N, Usmani SS, Raghava GPS (2021) Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. Brief Bioinform 22:936–945. DOI 10.1093/bib/bbaa259

Figures

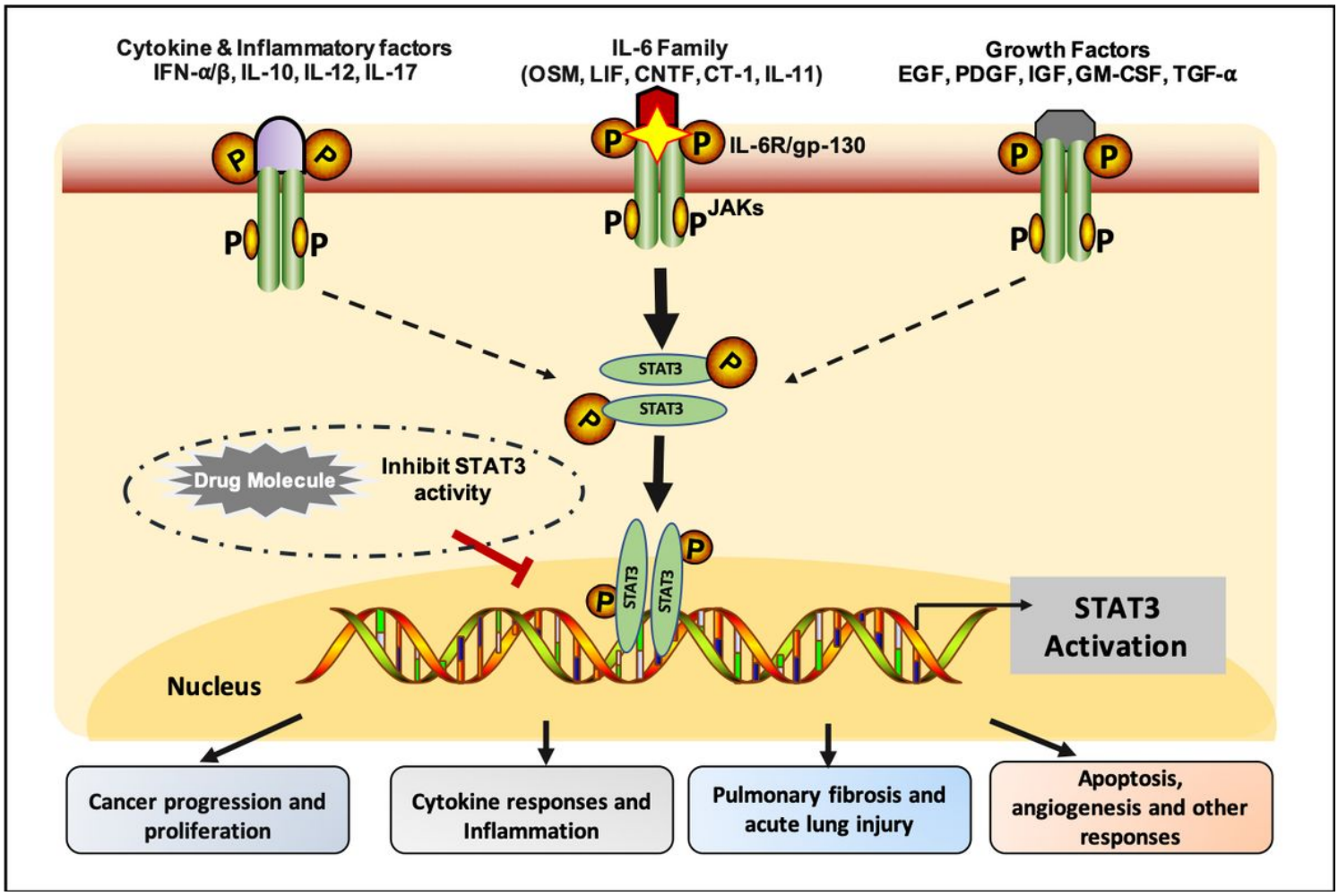

Figure 1

The binding of these factors to the cell surface receptor results in phosphorylation of JAKs by adding the phosphate group. Further, phosphorylated STAT3 monomers form a homodimer molecule which translocate into the nucleus and binds to the specific target gene promoters to regulate the gene transcription process [4], as shown in Figure 1

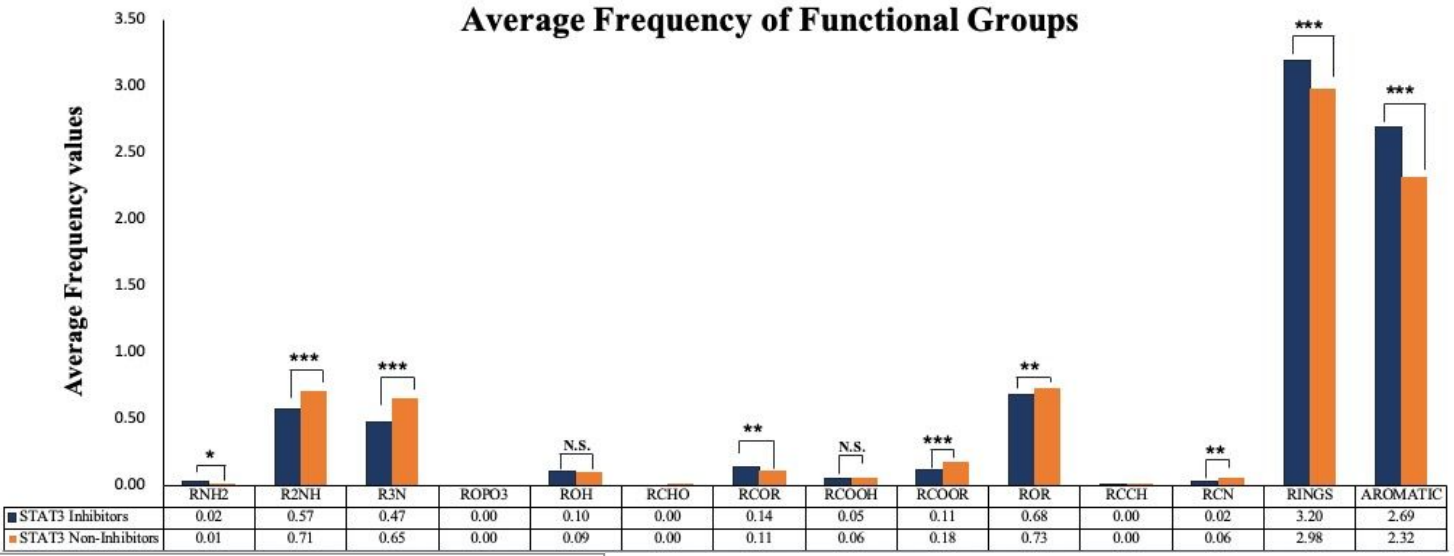

Figure 2

We analysed average frequency values and found that in the positive dataset, the abundance of rings, and aromatic groups is significantly higher as compared to non-inhibitors. On the other hand, the frequency of secondary amines (R2NH), tertiary amines (R3N), and ester (ROR) groups is significantly elevated in non-inhibitors compounds, i.e., STAT3 non-inhibitors, as represented in Figure 2.

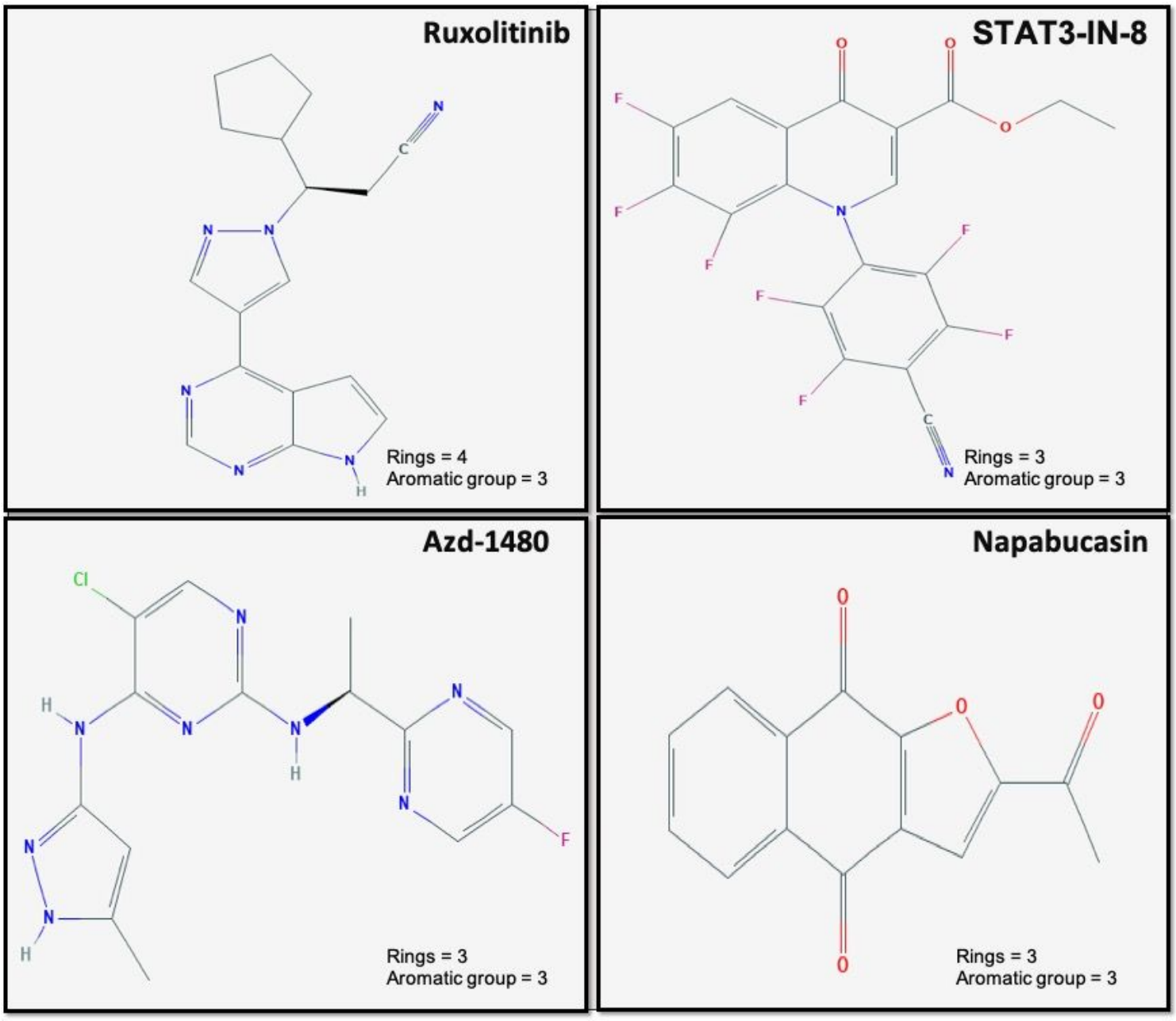

Figure 3

Shows functional groups in the chemical 2-D-structures of known STAT3 inhibitors, i.e., STAT3 Inhibitor VII, Ruxolitinib, AZD-1480, and BBI608.

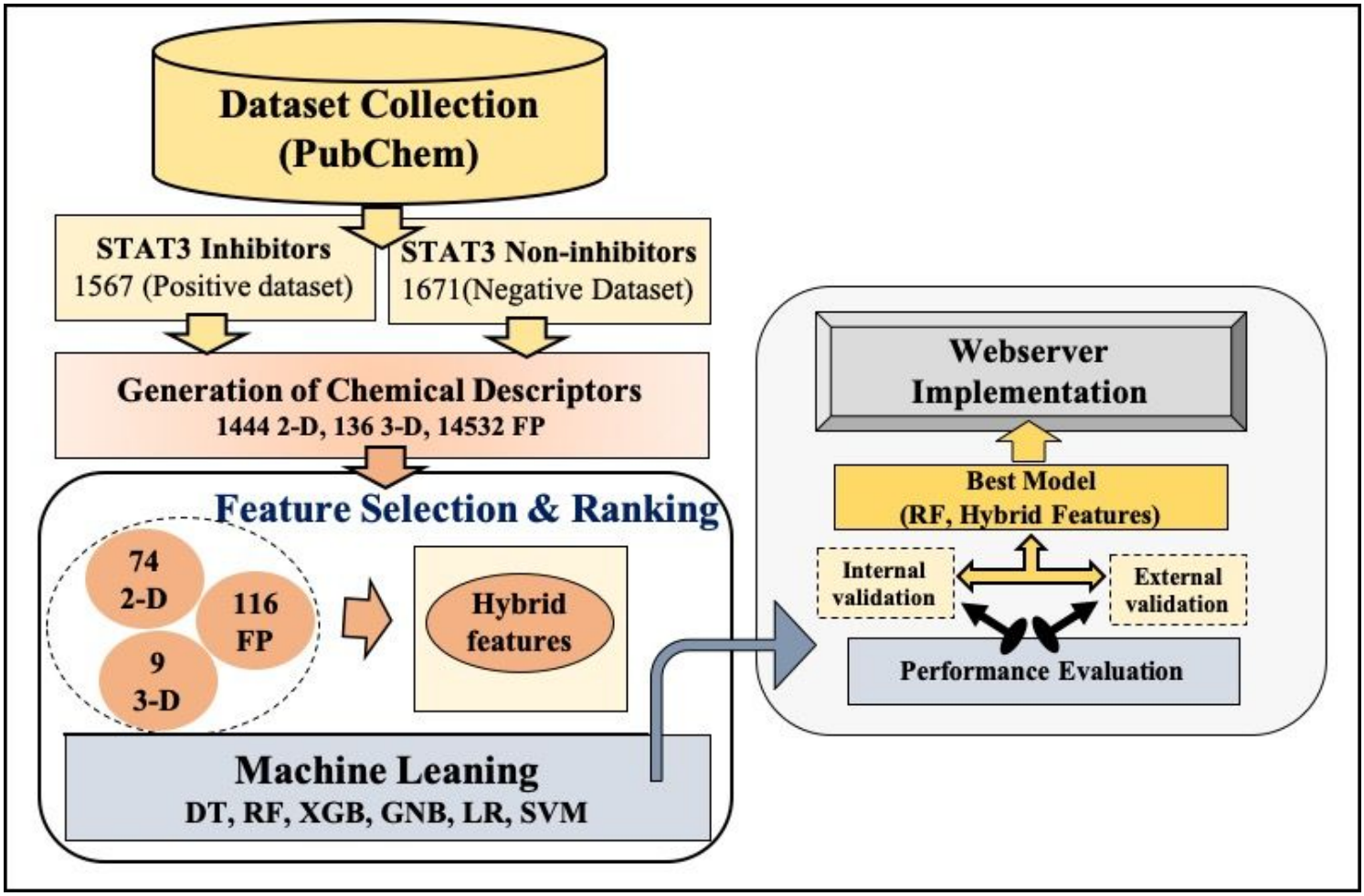

Figure 4

The complete architecture of the study is shown in Figure 4.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.docx