Received: 28 October 2022 | Revised: 18 February 2023 | Accepted: 24 February 2023

DOI: 10.1002/jcc.27103

RESEARCH ARTICLE

Journal of COMPUTATIONAL CHEMISTRY WILEY

Development of machine learning models based on molecular fingerprints for selection of small molecule inhibitors against JAK2 protein

Sharath Belenahalli Shekarappa[1] | Shivananda Kandagalla[2] | Julian Lee[1]

[1]School of Systems Biomedical Science and Department of Bioinformatics and Life Science, Soongsil University, Seoul, South Korea

[2]Laboratory of Computational Modeling of Drugs, Higher Medical & Biological School, South Ural State University, Chelyabinsk, Russia

Correspondence
Julian Lee, School of Systems Biomedical Science and Department of Bioinformatics and Life Science, Soongsil University, Seoul, South Korea.
Email: jul@ssu.ac.kr

Abstract

Janus kinase 2 (JAK2) is emerging as a potential therapeutic target for many inflammatory diseases such as myeloproliferative disorders (MPD), cancer and rheumatoid arthritis (RA). In this study, we have collected experimental data of JAK2 protein containing 6021 unique inhibitors. We then characterized them based on Morgan (ECFP6) fingerprints followed by clustering into training and test set based on their molecular scaffolds. These data were used to build the classification models with various supervised machine learning (ML) algorithms that could prioritize novel inhibitors for future drug development against JAK2 protein. The best model built by Random Forest (RF) and Morgan fingerprints achieved the G-mean value of 0.84 on the external test set. As an application of our classification model, virtual screening was performed against Drugbank molecules in order to identify the potential inhibitors based on the confidence score by RF model. Nine potential molecules were identified, which were further subject to molecular docking studies to evaluate the virtual screening results of the best RF model. This proposed method can prove useful for developing novel target-specific JAK2 inhibitors.

KEYWORDS

JAK2, machine learning, Morgan fingerprints, scaffolds, virtual screening

1 | INTRODUCTION

The Janus kinases (JAKs) are a class of cytoplasmic non-receptor tyrosine kinase molecule consisting of four subtypes, namely JAK1, JAK2, JAK3, and TYK2, which play predominant roles in the cytokine-mediated JAK-STAT signaling.[1,2] All the four subtypes contain a common structure region called Janus Kinase homology (JH2) domain, which mainly regulates the adjacent protein kinase domain (JH1).[3] Most of the known small molecules targeting JAKs bind to the adenosine triphosphate (ATP) site of the JH1 domain.[4] However, due to the high structural similarity of ATP binding site across JAK family, it is hard to discover potential molecule for a specific JAK family member by conventional methods. Therefore, selective inhibitors against each subtype of JAK protein are foremost important, as Bajusz and colleagues discovered subtype selective JAK inhibitors by structure-based drug design (SBDD) approach.[5] Pharmacophore- and docking-based VS approach for the design of JAK2/JAK3 dual inhibitors by Jasuja and colleagues.[6] Pharmacophore filtering and 3D-QSAR in the discovery of JAK2 inhibitors by Dhanachandra Singh and colleagues.[7]

Among the four JAK subtypes, JAK2 is crucial for cytokine receptor signaling in blood formation and immune response. This protein is a part of JAK-STAT signaling pathway, which transmits chemical signals from extracellular region to the nucleus resulting in DNA transcription.[3] The JAK2 V617F gene mutation results in the overproduction of JAK2 protein, which is crucial for controlling the production of blood cells from hematopoietic stem cells.[8] These mutations are associated with myelofibrosis, a condition where bone marrow is replaced by fibroblast. The V617F gene mutation is occasionally found in people with cancer or other bone marrow disorders. Hence, JAK2 has emerged as a potential therapeutic target for myeloproliferative disorders (MPD).[9] Currently, various drugs targeting JAK2 are in clinical and preclinical trials, among which some has been approved by the US Food and Drug

Administration (FDA), the European Medical Agency and other regulatory agencies in recent years. In 2013, ruxolitinib, a selective JAK2 inhibitor, was approved by FDA for the treatment of patients with intermediate and high-risk myelofibrosis.[10] Similarly, momelotinib, which is in Phase III clinical trial, was shown to be effective.[11] However, drugs such as AZD1480[12] and XL019[13] had to be canceled in early stage, owing to their severe side effects and poor bioavailability. Thus, novel JAK2 inhibitors with more drug-like properties are highly desirable in order to overcome these problems.

In recent years, with the development of computing power and the accumulation of experimental data, artificial intelligence (AI) has made great progress in the field of drug discovery.[14] These methods are based on Structure-Activity-Relationship/Quantitative Structure Relationship (SAR/QSAR) model, where pharmacological activities of the molecules are inferred their structural properties. The underlying theory is that molecules with same physical and chemical properties tend to have similar bioactivity.[15] Here, in order to design and develop still better and more effective JAK2 inhibitors, we develop and propose various classification methods based on machine learning (ML) using Extended-Connectivity Fingerprints (ECFP6) as input. These models would serve as potential tools to virtually screen JAK2 inhibitors from Drugbank molecules. It is expected that the identified virtual hits would provide some useful insights for the development of novel JAK2 inhibitors.

2 | MATERIALS AND METHODS

2.1 | Data curation for training and test set

The known inhibitors of JAK2 targets were retrieved from the ChEMBL (version 30) (https://www.ebi.ac.uk/chembl/). Six criteria were taken for data preparation: (1) Compounds without experimental activities were removed; (2) Removal of stereoisomers; (3) Removal of large and Invalid compounds; (4) Compounds with IC50, EC50, Ki, or Kd were taken, and converted into negative logarithmic scale (pIC50) for the uniform distribution of values; (5) Compounds with pIC50 value >7 were considered as highly active, while the compound ≤6 were considered as weakly active/inactive.[16] (6) Molecular normalization and tautomerization was performed using MolVS to standardize chemical structures for improving the data quality.[17]

2.2 | Preparation of training and test dataset

In order to evaluate the performance of a machine-leaning algorithm without bias, we have to check whether it correctly classifies new, unseen data. For this purpose, we have to construct a test set consisting of data that are dissimilar from the training set as much as possible. In order to evaluate the pairwise similarity between the molecules, a Murcko-scaffold analysis[18] was performed to reduce the chemical structure of a compound to its core components by removing the side chains and keeping only ring system and parts which link ring system together. The fingerprint score needed for the similarity score calculations were obtained using ECFP6 of length 2048 bits and Tanimoto coefficient (T_c), given by

$$T_c(a,b) = \frac{\mathrm{NC}}{N_a + N_b - N_c} \tag{1}$$

where N represents the number of attributes in each a and b molecules, and c is the common attribute in a and b. The range of T_c varies from 0 to 1, where 0 represents minimal and 1 maximal similarity. Using the Tanimoto similarity score, we performed the Butina clustering[19] of the scaffolds with a threshold of 0.7. We picked the first cluster as the test data, while the rest were used for training. This procedure ensures that the scaffolds of the molecules in the test set are dissimilar from those in the training set in terms of the Tanimoto score. All the analyses were performed using open-source cheminformatics library RDKit (2022.03.01) (http://www.rdkit.org). We also performed the principal component analysis (PCA) to assess the chemical dissimilarity of the training set and the test set, where Extended-Connectivity Fingerprints (ECFP6) of length 2048 were reduced to two dimensions by PCA algorithm to view the distributions of active and inactive molecules.

2.3 | Machine learning

Classification models were developed by eight machine learning (ML) algorithms including Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LogReg), Decision Tree (DT), Naïve Bayes (NB), Neural Network (MLP_NN) and Extreme Gradient Boosting (XGBoost). These methods exhibit high robustness, and therefore have been widely used in drug discovery. Scikit-learn ML python module was used for model building, tuning and validation. The developed ML models comprise of scaffolds as input data, which were grouped into training and test data based on similarity threshold ($T_c = 0.7$). Initially, Morgan fingerprints (ECFP) were computed for both training data and test data. Morgan fingerprints encode the chemical information of a compound as a sequence of binary bits. The ECFP represent chemical structure by means of circular atomic type connectivity, and their features represent the presence of substructures that can be found on different compound sets. It can theoretically characterize molecule of any size and any number of features. During the calculation, the connectivity of each atom is analyzed to a given radius and bits. Each element of the ECFP vector indicates the presence or absence of a specific feature in a compound. For this work, radius and bits were set to 3 and 2048 respectively, which is so called ECFP6, 2048 bits. The hyperparameter optimization was adjusted for top performing base models with grid search and 5-fold stratified cross validation on training data for better model generalization.

In order to explore the impact of balanced learning on model's performance, SMOTE (Synthetic Minority Oversampling Technique) was applied on the training set to balance the active/inactive class, and the test set was kept unbalanced. This technique aims to balance class distribution by randomly increasing minority class by replicating them between the existing minority classes. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each in the minority class.[20] Later, several classification models were applied for the processed data.

Random Forest is an ensemble method that combines randomly built multiple decision tree in the form of forest structure to improve the robustness over a single estimator. In order to classify the data, each decision tree is given a classification, the score of each tree is calculated to make the final decision.[21,22] SVM is one of the most popular and widely used algorithm, which uses sum function to transfer data into high-dimensional space and establish the optimal separation hyperplane.[23] KNN is an instance-based learning which uses k closet samples from the training dataset to assess the predicted value, which is also a non-parametric ML algorithm that works for both classifier and regressor methods.[24] LogReg is a linear model for classification rather than regression. The logical function was used to model the probabilities, which describe the possible outcomes of one single trail.[25] DT is a non-parametric supervised learning to build models that can learn decision rules from the input data and make predictions on the value of a target variable.[26] NB algorithms are supervised learning methods based on Bayes' theorem, that have an assumption of conditional independence between every pair of features. In this work, Bernoulli NB algorithm was applied to the datasets with fingerprints as features, represented as binary-valued vectors. The prior probability of the classes were set to none.[27] MLP_NN has the capacity to learn the nonlinear models in real time. It can have one or more nonlinear hidden layers between the input and output layers. For each layer's different numbers of hidden neurons can be assigned, which gives a weighted linear summation for the values from the previous layers followed by the activation of nonlinear function.[28] XGBoost are a kind of ensemble technique where predictors are ensembled sequentially one after the other. XGBoost is an optimized gradient boosting library designed to be highly efficient and flexible.[29]

2.4 | Model evaluation

In order to evaluate the model's performance, various metrics such as accuracy (ACC), Matthew's correlation coefficient (MCC), precision, recall, F1-score, and ROC-AUC score were calculate using Scikit-learn. The ration of true positive (TP)/true negative (TN) and false positive (FP)/false negative (FN) were used to describe these matrices. Briefly, TP and TN represent the correctly predicted active and inactive compounds, respectively. Whereas, FP indicates that inactive compounds have been incorrectly predicted as active compounds, and FN indicates that active compounds have been incorrectly predicted as inactive compounds.

Accuracy represents the proportion of correctly predicted classes to the total number of classes. MCC is generally used to measure the quality of classification models, it is a balanced measure that both TP/FP and TN/FN are considered. Precision indicates how many of the positive classes are the real positives. Recall measures the ability of a model to find out all of the positive classes. F1-score is the harmonic mean of recall and precision. The geometric mean (G-mean) is the squared root of the products of specificity and sensitivity, where specificity and sensitivity combined to give single score that balances both concerns. ROC-AUC score is the area under the ROC curve, AUC value of 1 specifies a perfect model while 0.5 specifies a random classifier. The formulas for all the metric calculation are as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{3}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

$$F1-\mathrm{score} = \frac{2\times(\mathrm{precision}\times\mathrm{recall})}{\mathrm{precision}+\mathrm{recall}} \tag{6}$$

$$G-\mathrm{mean} = \sqrt{\mathrm{Sensitivity}\times\mathrm{Specificity}} \tag{7}$$

2.5 | Virtual screening

Virtual screening of unknown dataset using the predictive model is a crucial step in identifying novel potential JAK2 inhibitors. The best performing model, which is the one obtained by the Random Forest (RF) method, was utilized to virtually screen the Drugbank database (https://go.drugbank.com/) (see Results). The Drugbank database contains 11,912 purchasable compounds. Before the application of the RF model for the activity prediction, the compounds were prioritized by Lipinski's rule of five (RO5) followed by removing PAINS (Pan-Assay Interference Compounds), invalid compounds (whose parameters could not be calculated by RDKit) and salts. The final set of 9136 compounds was left after this procedure, for which the descriptors and ECFP6 fingerprints were calculated using RDKit library (March 01, 2022). Then, the descriptors were used to predict the confidence of highly efficient compounds with the classification model.

2.6 | Molecular docking

The crystallographic structure of the human JAK2 with inhibitors were retrieved through Biotite[30] python module with search query "O60674" (UniProt ID). From RCSB Protein Data Bank (https://www.rcsb.org/), a total of 129 different JAK2 structures were identified and among them 127 JAK2 structures with non-covalently bound inhibitors were selected for the analysis. The crystallographic JAK2 structures were processed further by removing heteroatoms (cofactors, water

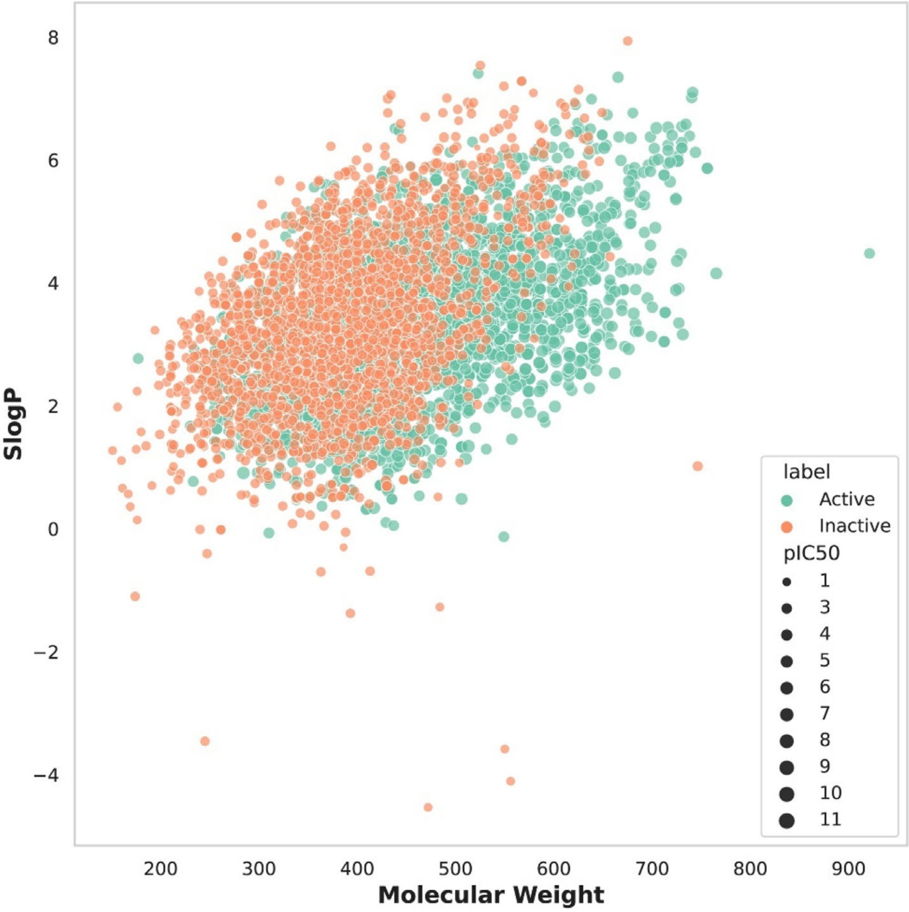


FIGURE 1 Visualization of logP (hydrophobicity) and molecular weight of all the molecules by pIC50 size distribution.

molecules, and metal ions) by using UCSF Chimera 1.14 (University of California, USA).[31] Only chain A of JAK2 structures with bound inhibitors were retained and superimposed using UCSF Chimera 1.14 to map the inhibitors' location. As is customary in the molecular docking procedure, we constructed a box shaped search region divided into a grid of evenly spaced points for evaluation of energy, called the grid box, which was centered on the bound inhibitors based on superimposed JAK2 structures (Figure S1). The JAK2 crystal structure (PDB id: 7LL4) was considered for docking due to its high resolution (1.31 Å). The JAK2 structure was prepared by adding polar hydrogens and parameterize it based on pKa of each aminoacid at pH 7.4 with AMBER99ff force field using pdb2pqr package, followed by deletion of non-polar hydrogens and conversion into PDBQT format using the prepare-receptor4.py Python scripts from Auto Dock Tools (ADT),[32] as described in the AutoDock Vina manual. Similarly, ligands were prepared using the prepare-ligand4.py Python scripts from ADT. Initially, the SMILES of the ligands were collected and processed in RDKit library as described in our earlier report.[33]

The docking was performed using GNINA 1.0[34] that is a fork of smina[35] and AutoDock (ADT) Vina,[36] a Convolution Neural Network (CNN)-based molecular docking algorithm with integrated support for conformational sampling, ligand optimization and scoring function. Here we used custom scoring function called vinardo,[37] a scoring function based on ADT vina that focus mainly on improving accuracy of docking. Validation of docking approach was assessed by cross-docking the crystallized ligand with other JAK2 protein (PDB id: 5AEP; 1.95 Å resolution). For crossdocking, we utilized various docking algorithms such as ADT-Vina,[36] QVina2,[38] QVina-W,[39] Smina,[35] Smina-vinardo, GNINA,[34] GNINA-vinardo to facilitate the validation of our docking studies. The post-docking analysis were performed using PyMOL and Chimera.

2.7 | Computing

All computing was performed using 8 core DDR4 64GiB (16 × 4), Intel(R) Core (TM) i7-6700 CPU @ 3.40GHz. The following packages were used in the study: Python 3.7.12, scikit-learn 1.0.2, RDKit 2022.03.2, Seaborn 0.11.2, pandas 1.3.4, numpy 1.21.6, Jupyter notebook 6.4.11, matplotlib 3.5.2, Autodock Vina, PyMOL, Chimera 1.14.

3 | RESULTS

We developed a ML-based drug discovery platform to predict the activity of compounds for the inhibition of JAK2 protein. Initially we built eight ML base models, then we finalized to top three based on the performance of the base models. Prior to modeling, we collected

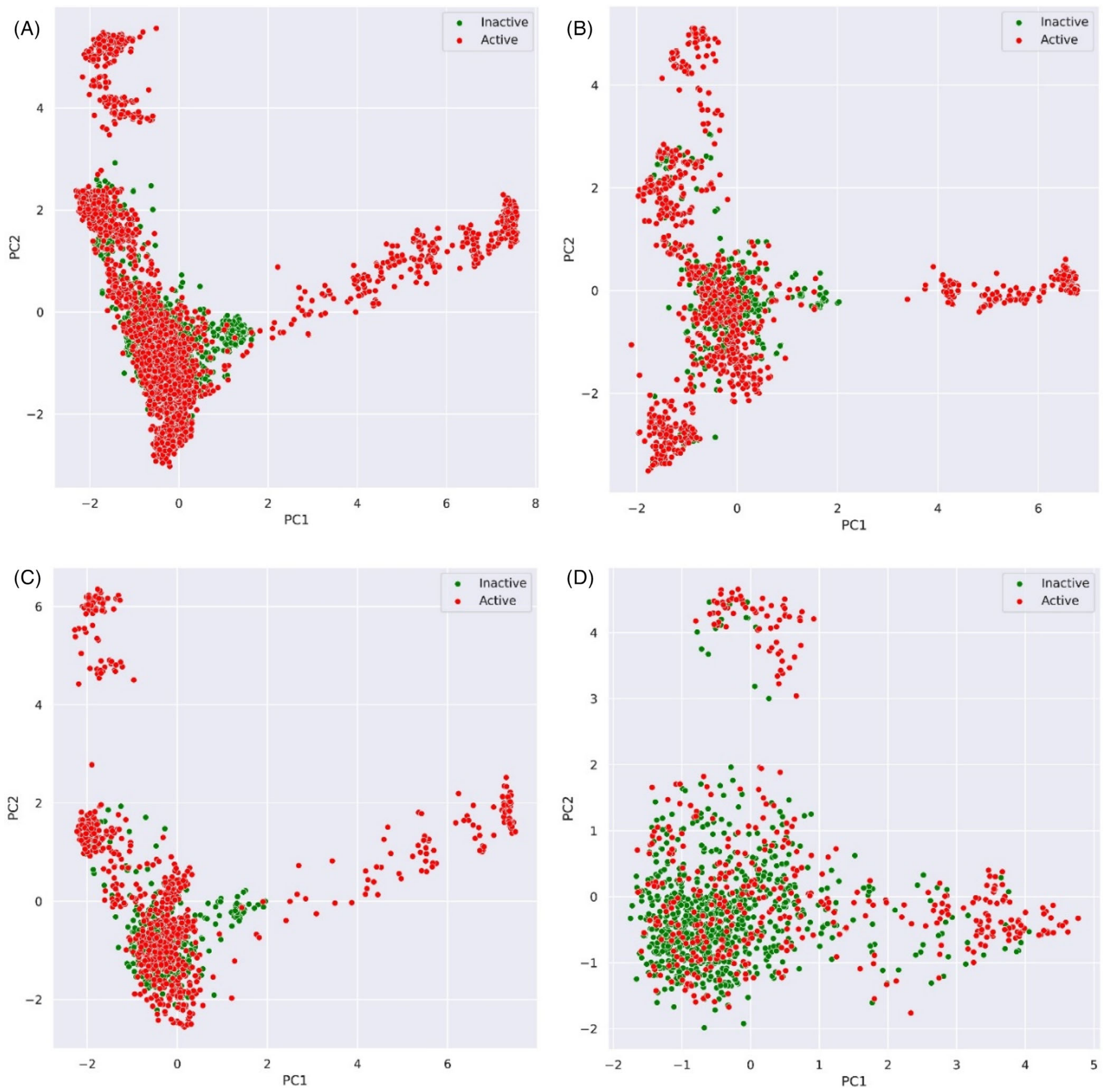


FIGURE 2 Visualization of the train and test data with PCA and ECFP6 for all the molecules: (A & B) Train data distributed by random split and scaffold split methods, respectively. (C & D) Test data distributed by random split and scaffold split, respectively.

bioactivity data of human tyrosine-protein kinase JAK2 from the ChEMBL database (target id: CHEMBL2971), which consists of 12,349 compounds. We curated the dataset by removing duplicates, stereoisomers and compounds without experimental activities (IC50, EC50, Ki, and Kd), deletion of large compounds (which behaves like outliers and make the training process longer) followed by molecular standardization and validation using MolVS for improving the data quality by identifying relationship between the molecules, which mainly includes metal ion disconnection, functional group normalization, reionization, removal of salts/solvents, tautomerization and charge neutralization. A total of 6021 unique compounds, including 3699 highly active and 2322 weakly active (inactive) inhibitors were selected for model building.

3.1 | Chemical diversity and scaffold-clustering analysis

To evaluate the chemical diversity of the final dataset, we calculated the logP and molecular weight of all the compounds. logP (the log of

TABLE 1 Performance metrics of base models on test dataset with default parameters.

Algorithm	ACC	Precision	Recall	*F*1-score	ROC AUC	MCC	*G*-mean
RF	**0.847**	**0.778**	**0.806**	**0.847**	**0.838**	**0.671**	**0.84**
DT	0.705	0.576	0.684	0.709	0.7	0.389	0.69
NB	0.801	0.849	0.547	0.789	0.746	0.56	0.72
KNN	**0.848**	**0.862**	**0.689**	**0.843**	**0.813**	**0.664**	**0.79**
LogReg	**0.816**	**0.728**	**0.783**	**0.817**	**0.809**	**0.609**	**0.81**
SVM	**0.839**	**0.761**	**0.806**	**0.84**	**0.832**	**0.656**	**0.83**
XGBoost	**0.798**	**0.696**	**0.783**	**0.8**	**0.795**	**0.577**	**0.79**
NN_MLP	**0.79**	**0.696**	**0.743**	**0.791**	**0.78**	**0.552**	**0.78**

Note: Bolded rows indicate the top three best base models taken for hyperparameter optimization.

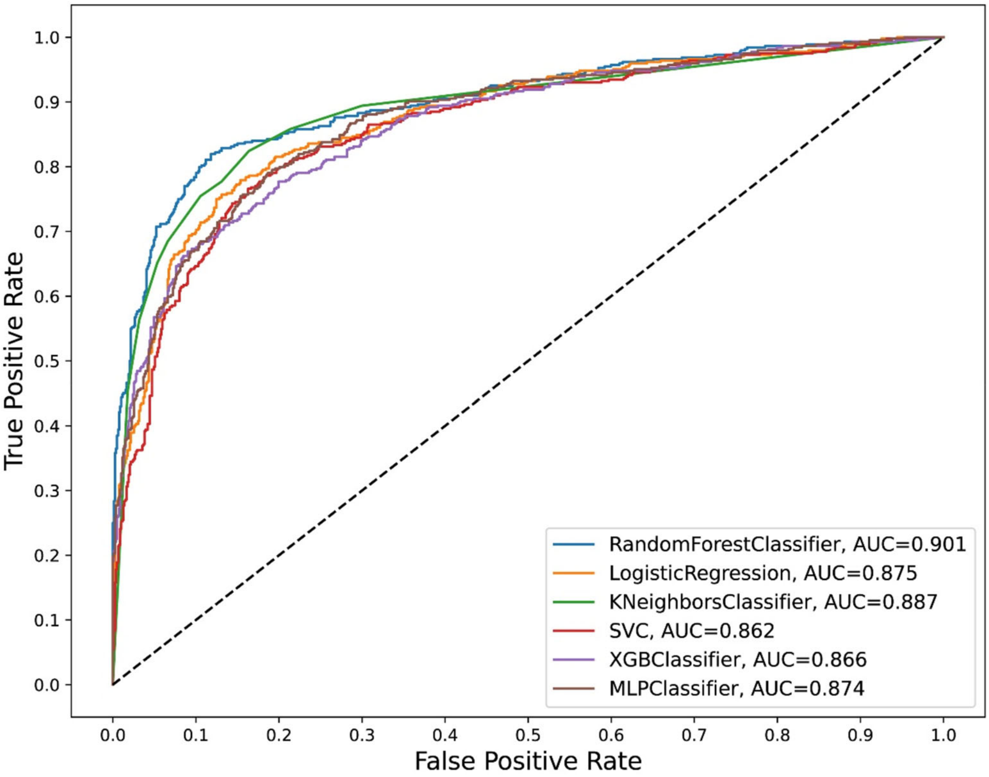


FIGURE 3 ROC curve comparison of top performing models with AUC value.

the partition coefficient), is the measure of hydrophobicity, defined as the logarithm of ratio of the concentrations of the compound in octanol and in water at equilibrium. The logP distribution of all the compounds was between −4.53 and 9.87, and the distribution of molecular weight was between 156.06 and 926.29. The scatter plot of logP against Molecular weight is shown in Figure 1. Here, each compound in the input data was clustered together by Murcko-scaffold such that the scaffolds in each cluster are as distant from the other. Clustering the scaffolds at 0.7 similarity index leads to reduction of the data from 6021 molecules to only 1906 cluster representatives. It can be seen from Supplementary material (Figure S2), that the distribution of the training and test set in both scaffold-split and random split are distant, indicating that the distribution by scaffold-split in chemical space has less overlap between the training and test set.

In order to assess the bioactivity distribution, we have constructed PCA plot for the distributions of active and inactive classes in the training and test set obtained from the scaffold-split, which are shown to be distinct (Figure 2B,D). For comparison, we also plotted the result for the random split, where we randomly selected 20% of the data as the test set (Figure 2A,C). We indeed see that the two distributions obtained from the scaffold split are more dissimilar compared to the result obtained from the random split. The significance of this result is that every single molecule in the test set obtained from the scaffold-split method is molecularly distant form the training set.

3.2 | Dataset building and machine learning

In order to increase the chemical diversity, training and test data were built using scaffold-split method (explained in Method section). Subsequently, the training set consisted of 3255/1535 compounds in the active and inactive class, respectively, while the test set has 444/787

TABLE 2 Evaluation metrices of top performing models on test set.

Algorithm	TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)	G-mean
RF	**358**	**700**	**87**	**86**	**80.6**	**88.9**	**84.7**	**0.84**
KNN	322	720	67	122	72.5	91.4	82	0.81
SVM	360	612	175	84	81	77.7	79.4	0.79
LogReg	352	657	130	92	79.2	83.4	81.3	0.81
XGBoost	357	603	184	87	80.4	76.6	78.5	0.77
MLP_NN	300	630	114	92	76.5	84.6	80.6	0.79

Note: Bolded rows indicate the best performing models based on G-mean value.

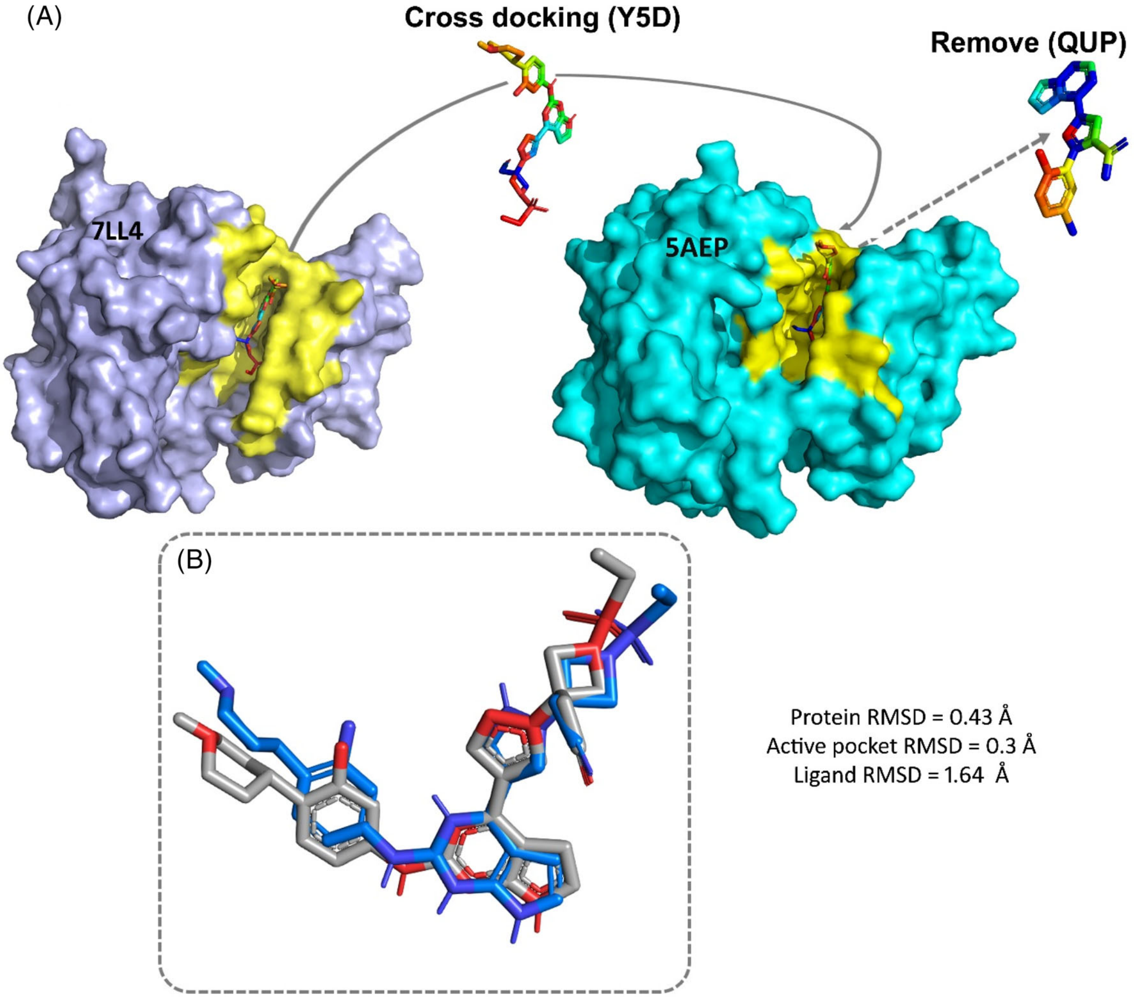


FIGURE 4 (A)) Crossdocking procedure. (B) Crossdocking analysis of co-crystallized structure (gray) and docked pose (blue) in the JH1 domain of JAK2 protein (PDB id: 5AEP).

active and inactive compounds, respectively. All the models constructed by scaffold-clustering method exhibited reasonably good performance, which authenticates the rationality of the training/test dataset constructed by the scaffold-clustering method.

In this work, we first developed all the models with default parameters (base models). Later, we choose best models from cross validation and hyperparameter tuning based on the training and test accuracy. The performance of the all the models except NB on the training set were above 0.9 (>90%) accuracy. However, the prediction results on the test dataset were rather varied (Table 1), where all the models show relatively good mean accuracy score on test set except DT and NB. Since there is imbalance in active and inactive compounds of the test set, the G-mean value was considered as reference metrics to determine the top models. The RF, kNN, LogReg, SVM, XGBoost, and NN_MLP models achieved a G-value value of 0.84, 0.79, 0.81, 0.83, 0.79, and 0.78, respectively. Then, the performance of top

TABLE 3 Docking score of top compounds against the JH1 domain of JAK2.

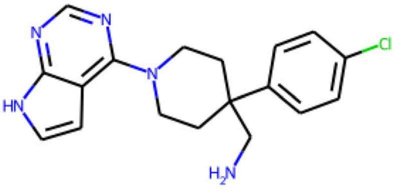
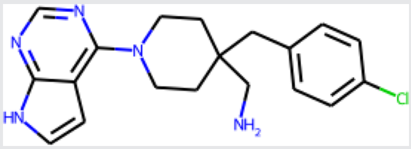
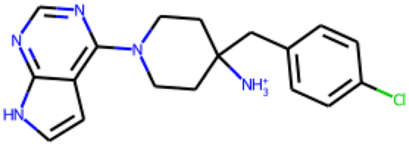
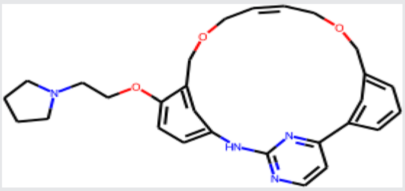
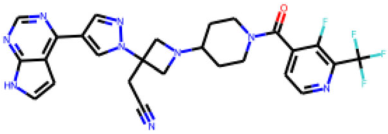
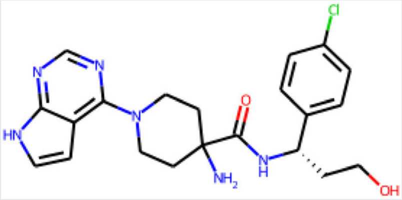
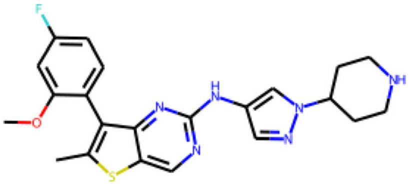
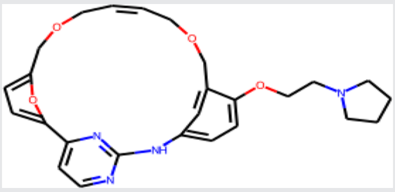
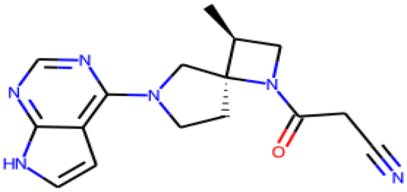
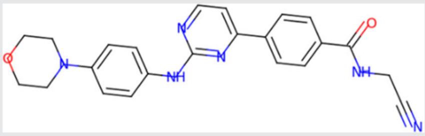
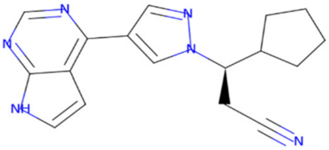
Drug bank ID	Structure of the molecule	Binding energy (kcal/mol)	ML confidence score	CNN pose score	CNN affinity
DB08148		−7.91	0.85	0.67	7.023
DB08149		−7.47	0.87	0.806	7.161
DB08150		−7.01	0.86	0.653	6.829
DB11697		−8.4	0.99	0.33	7.266
DB12154		−8.3	0.70	0.42	7.26
DB12218		−7.67	0.85	0.696	7.139
DB15191		−8.58	0.75	0.785	7.613
DB15294		−6.61	0.75	0.492	5.58
DB16133		−6.14	0.77	0.33	6.32

TABLE 3 (Continued)

Drug bank ID	Structure of the molecule	Binding energy (kcal/mol)	ML confidence score	CNN pose score	CNN affinity
Momelotinib		−7.79	0.72	0.78	6.707
Ruxolitinib		−9.03	0.92	0.986	7.608

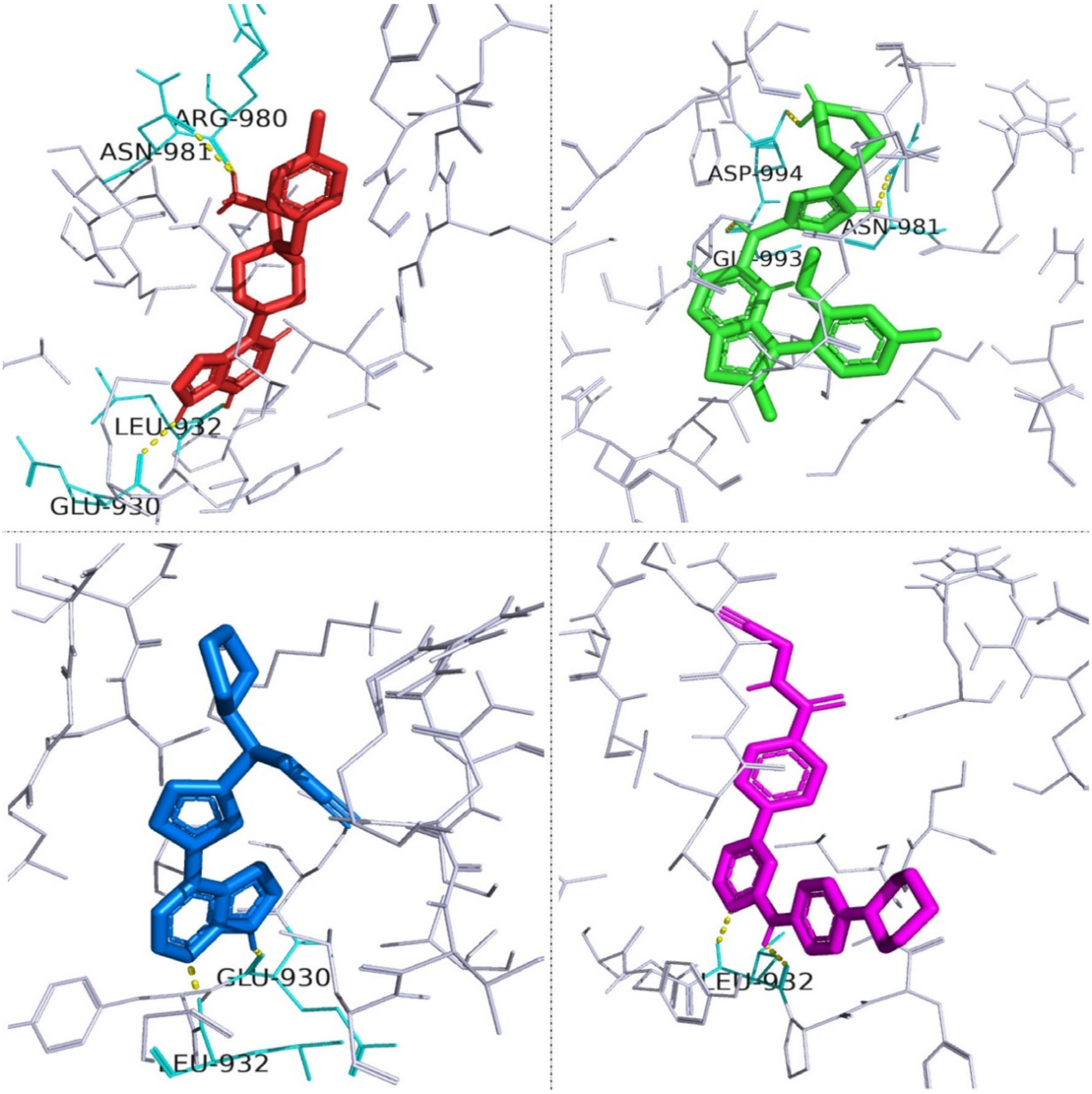


FIGURE 5 Docking pose and interaction comparison of potential JAK2 inhibitors DB08149 (red) and DB15191 (green) with known inhibitors momelotinib (blue) and ruxolitinib (magenta).

trained models was evaluated using 5-fold stratified cross validation (CV). The best estimators and CV score for the top models are displayed in Table S1 in the supplementary material.

All the top performing base models were improved significantly on the tests set after cross-validation and hyperparameter tuning, and have a remarkable identification ability of JAK2 inhibitors with low false positive rate. This suggest that, the RF, KNN, LogReg, SVM, XGBoost, and NN_MLP models performed relatively well in predicting positive and negative classes after the tuning.

The scikit-learn confusion metrics of the top models shows the percentage of TP/TN and FP/FN in Figure S3 in the Supplementary material, while their ROC curves with AUC value above 0.85 are displayed in Figure 3. Although~~,~~ the performance of RF, kNN, SVC, LogReg, XGBoost and NN_MLP models were comparable~~.~~, KNN was found to be least sensitive (72.5%), while RF and SVM showed a sensitivity of 80.6% and 81%, respectively. Similarly, KNN has showed high specificity of 91.4%, while specificity of RF and SVM were found to be 84.7% and 79.4%, respectively. However, since sensitivity and specificity do not provide accurate classification of positive and negative classes, the geometric-mean (*G*-mean) was calculated for all the top performing models. RF algorithm performed the best and had a *G*-mean value of 0.84 (Table 2). Such a model can be potentially used for drug repurposing. Altogether, the results of scaffold-clustering method can be used for drug discovery by reduced bias in chemical space of training/test set. Considering the best performance of these models by scaffold-clustering method in predicting compounds outside the training data, we used them to virtually screen external data such as Drugbank library, in order to prioritize compounds for experimental testing in future.

3.3 | Virtual screening by deploying the best model

Since RF model exhibited the best performance, we used this method to select the potential JAK2 protein inhibitor from the Drugbank compounds. By performing a pre-screening on 11,912 initial compounds to remove invalid compounds (see Section 2.5), we obtained 9136 molecules, whose activities were then predicted by the RF model. The top nine compounds predicted as actives by RF model are DB08148, DB08149, DB08150, DB11697, DB12154, DB12218, DB15191, DB15294, and DB16133, which mostly belongs to the class of piperidines and phenols. The potential activities of these nine molecules were further reassessed by molecular docking.

3.4 | Molecular docking analysis

Molecular docking is highly beneficial technique for understanding the protein-ligand interaction. All the nine molecules were docked sequentially on JAK2 protein (PDB id: 7LL4) to calculate their best binding affinity. Prior to docking, crossdocking of co-crystallized ligand Y5D was performed against other JAK2 protein (PDB id: 5AEP) by removing its original ligand QUP, followed by docking of Y5D in the active pocket of 5AEP (Figure 4A). The docking accuracy was assessed by calculating difference RMSD, which is shown in Table S2 in the supplementary material. Among all the docking algorithms, GNINA using vinardo scoring function has shown least RMSD difference of 1.64 Å, and position and orientation were in excellent agreement with the original docked structure (Figure 4B). This confirms the validation of the docking procedure. Henceforth, all the lead molecules identified based on ML confidence score were docked using GNINA-vinardo. Based on the binding affinity, the ligands were compared with known inhibitors such as Momelotinib and Ruxolitinib.

The detailed results of their binding affinity, pose scoring and CNN-affinity with ML confidence score are shown in Table 3. All the molecules except DB15294 and DB16133 have shown the binding affinity and CNN-affinity better than −7 kcal/mol and 7 pK_a for JAK2 subtype. Wherein, DB08149 and DB15191 have better hydrogen bond interactions, binding affinity (kcal/mol), CNN-pose score and CNN-affinity (pK_a) when compared to the known inhibitor Momelotinib and Ruxolitinib (Figure 5). The compounds we found could simultaneously bind to JH1 domain of JAK2, and inhibit the JAK2 activity by forming various intramolecular interactions (Figure S4). These analyses were consistent with the results of ECFP6 fingerprint analysis, indicating that our ML using ECFP6 fingerprint descriptors was accurate and reliable.

We also performed docking of all nine lead molecules against other subtypes of JAK family, such as JAK1 (PDB id: 6N7D, 1.78 Å), JAK3 (PDB id: 5LWM, 1.55 Å) and TYK2 (PDB id: 6AAM, 1.98 Å), where we see that they have reasonably good binding affinities and CNN-affinities for all the JAK subtypes (Table S3).

4 | DISCUSSION

Over the past decades, there has been dramatic increase in research expenditure on drug discovery and development for the treatment of various diseases. With the progress of computer-aided drug design (CADD), the application of machine learning (ML) in discovering novel molecules has increased rapidly. ML integrates vast amount of data sources to solve biological problems with the combination of computer science and statistics. Now a days, ML has become an essential tool for mining chemical information from large compound databases, to design novel drugs with important biological features. For instance, Sean Ekins et al. used ML to discover the novel antiviral compounds against yellow fever virus.[40] Akansha Rajput et al. used ML for the prediction of repurposed drugs for coronavirus (Covid-19).[41] Similarly, Zhao et al. used CADD to identify JAK2 inhibitors, which mainly utilizes structure-based drug designing approach.[42] Thus, the virtual screening combining both the ligand-based machine learning and structure based molecular docking is promising for the discovery of novel molecules.

In the present study, we have used molecular fingerprints (ECFP6) as descriptors to construct a binary classification model of JAK2 inhibitors. ECFPs are the most popular and efficient methods among other fingerprint descriptors. However, in some cases, the limited reach of the ECFP fingerprint radius does not resolve to different atom identities, which subsequently fails to distinguish the chirality of the molecules. The chirality of a molecule is an important feature that often determines the activity of known drugs or drug candidates. The respective stereoisomers may show different interactions with desired molecular targets, so this has become significance for the final bioactivity of chiral molecules. Therefore, in order to maintain the consistency in our data, we removed stereoisomers followed by molecular standardization using MolVS for improving the data quality. Here, various supervised models (RF, KNN, SVM, LogReg, XGBoost, MLP_NN, NB, and DT) were chosen to build the ML models. While the performance of RF, KNN, and LogReg models were comparable, overall, the RF algorithm performed better in terms of all the metrices (explained in Result section). A study by Minjian Yang et al.[43] on discovery of JAK2 inhibitors by ML lacks interpretation on the chemical diversity of training and test set, which is certainly important nowadays to use bias-free datasets for the accurate prediction.[44,45] Hence, in this study, we performed scaffold-clustering to distinguish training and test sets-based on molecular similarity cut-off ($T_c > 0.7$), which provides better insight to design novel JAK2 inhibitors, and generalize better than a random split method (explained in Result section). Later, by combining ML based on scaffold-clustering and structure-based molecular docking approach, nine potential JAK2 inhibitors were found, which mostly belongs to the class of piperidines and phenols. Among them, Pacritinib (DB11697; ML confidence score 0.9, binding affinity = −8.4 Kcal/mol and CNN-affinity of 7.26 pK_a) is mainly used to target JAK2 signaling in myelofibrosis (MF),[46] while the others are under investigational and experimental category. The compounds we found could simultaneously bind to JH1 domain of JAK2, and inhibit the JAK2 activity by forming various intramolecular interactions (Supplementary Figure S4). We also performed docking of all nine lead molecules against other subtypes of JAK family, such as JAK1 (PDB id: 6N7D, 1.78 Å), JAK3 (PDB id: 5LWM, 1.55 Å) and TYK2 (PDB id: 6AAM, 1.98 Å). Except DB16133 and DB15294, all other lead molecule has shown decent inhibition based on binding affinity and CNN-affinity for all the JAK subtypes. In addition, the molecules DB08148, DB08149, DB0150, DB116971, DB12154, DB12218, and DB15191 has shown the binding affinity and CNN-affinity above −7 kcal/mol and 7 pK_a for JAK2 subtype. Wherein, DB08149 and DB15191 are having better hydrogen bond interactions, binding affinity (kcal/mol), CNN-pose score and CNN-affinity (pK_a) when compared to known inhibitor momelotinib and ruxolitinib (Figure 6). The docking analysis of all the JAK family subtypes were furnished in Table S3 in the supplementary material.

There are several limitations to our study: it is worth nothing that the classification methods based on ML followed by molecular docking did not ensure compound activity, and even compounds with high docking score may have the possibility of false positive. However, this systematic study identified several potential novel JAK2 inhibitors through scaffold-clustering method, which can improve the efficiency of computational drug discovery. This approach not only augment prediction efficiency, but also generalization ability to predict the unknown data. Further experimental studies are required to determine the therapeutic potential and side effect of these compounds on JAK2 target.

5 | CONCLUSION

In this work, we developed a prediction protocol for potential JAK2 inhibitors by combining machine learning and molecular docking. Using this protocol, we could select nine novel inhibitor candidates from the DrugBank database. Our findings suggest that the developed classification model has a potential for distinguishing between active and inactive compounds. Overall, the current method has the benefits of being accurate, interpretable, and bias-free for quantitative prediction of selective JAK inhibitors.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea, funded by Ministry of Science and ICT, grant number NRF-2020R1A2C1005956. Dedicated to the memory of my colleague and friend, Sihyun Ham (J.L.).

DATA AVAILABILITY STATEMENT

Data available upon request

ORCID

Sharath Belenahalli Shekarappa https://orcid.org/0000-0003-2695-6484

Julian Lee https://orcid.org/0000-0002-6088-8406

REFERENCES

[1] W. J. Leonard, J. J. O'Shea, *Annu. Rev. Immunol.* **1998**, *16*, 293. https://doi.org/10.1146/ANNUREV.IMMUNOL.16.1.293

[2] S. Banerjee, A. Biehl, M. Gadina, S. Hasni, D. M. Schwartz, *Drugs* **2017**, *77*, 521. https://doi.org/10.1007/S40265-017-0701-9

[3] D. S. Aaronson, C. M. Horvath, *Science* **2002**, *296*, 1653. https://doi.org/10.1126/SCIENCE.1071545

[4] M. C. Bryan, N. S. Rajapaksa, *J. Med. Chem.* **2018**, *61*, 9030. https://doi.org/10.1021/ACS.JMEDCHEM.8B00667

[5] D. Bajusz, G. G. Ferenczy, G. M. Keseru, *J. Chem. Inf. Model.* **2016**, *56*, 234. https://doi.org/10.1021/ACS.JCIM.5B00634

[6] H. Jasuja, N. Chadha, M. Kaur, O. Silakari, *Mol. Divers.* **2014**,*18*, 253.

[7] K. D. Singh, M. Karthikeyan, P. Kirubakaran, S. Nagamani, *J. Mol. Graphics Model.* **2011**, *30*, 186. https://doi.org/10.1016/J.JMGM.2011.07.004

[8] C. James, *Hematology* **2008**, *2008*, 69. https://doi.org/10.1182/ASHEDUCATION-2008.1.69

[9] E. J. Baxter, L. M. Scott, P. J. Campbell, C. East, N. Fourouclas, S. Swanton, G. S. Vassiliou, A. J. Bench, E. M. Boyd, N. Curtin, M. A. Scott, W. N. Erber, A. R. Green, *Lancet* **2005**, *365*, 1054. https://doi.org/10.1016/S0140-6736(05)71142-9

[10] J. S. Fridman, P. A. Scherle, R. Collins, T. C. Burn, Y. Li, J. Li, M. B. Covington, B. Thomas, P. Collier, M. F. Favata, X. Wen, J. Shi, R. McGee, P. J. Haley, S. Shepard, J. D. Rodgers, S. Yeleswaram, G. Hollis, R. C. Newton, B. Metcalf, S. M. Friedman, K. Vaddi, *J. Immunol.* **2010**, *184*, 5298. https://doi.org/10.4049/JIMMUNOL.0902819

[11] A. Pardanani, J. Gotlib, A. W. Roberts, M. Wadleigh, S. Sirhan, J. Kawashima, J. A. Maltzman, L. Shao, V. Gupta, A. Tefferi, *Leukemia* **2018**, *32*, 1034.

[12] A. Pardanani, C. Harrison, J. E. Cortes, F. Cervantes, R. A. Mesa, D. Milligan, T. Masszi, E. Mishchenko, E. Jourdan, A. M. Vannucchi, M. W. Drummond, M. Jurgutis, K. Kuliczkowski, E. Gheorghita, F. Passamonti, F. Neumann, A. Patki, G. Gao, A. Tefferi, *JAMA Oncol.* **2015**, *1*, 643. https://doi.org/10.1001/JAMAONCOL.2015.1590

[13] S. Verstovsek, R. Hoffman, J. Mascarenhas, J. C. Soria, R. Bahleda, P. McCoon, W. Tang, J. Cortes, H. Kantarjian, V. Ribrag, *Leuk Res* **2015**, *39*, 157. https://doi.org/10.1016/J.LEUKRES.2014.11.018

[14] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, P. Kumar, *Mol. Divers.* **2021**, *25*(3), 2021.

[15] L. David, A. Thakkar, R. Mercado, O. Engkvist, *J. Cheminformatics* **2020**, *12*(1), 2020.

[16] B. Merget, S. Turk, S. Eid, F. Rippmann, S. Fulle, *J. Med. Chem.* **2017**, *60*, 474. https://doi.org/10.1021/ACS.JMEDCHEM.6B01611

[17] M. Swain, **2017**. https://doi.org/10.5281/ZENODO.260235

[18] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887. https://doi.org/10.1021/JM9602928

[19] D. Butina, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747. https://doi.org/10.1021/CI9803381

[20] N. v. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, *J. Artif. Intell. Res.* **2002**, *16*, 321. https://doi.org/10.1613/JAIR.953

[21] T. K. Ho. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. **1995**. https://doi.org/10.1109/ICDAR.1995.598994

[22] L. Breiman, *Mach. Learn.* **2001**, *45*(1), 2001.

[23] N. Cristianini, E. Ricci, in *Encyclopedia of Algorithms*, Springer, Boston, MA **2008**, pp. 928–932. https://doi.org/10.1007/978-0-387-30162-4_415

[24] A. Mucherino, P. J. Papajorgji, P. M. Pardalos, **2009**. https://doi.org/10.1007/978-0-387-88615-2_4

[25] S. Menard, in *Logistic Regression: From Introductory to Advanced Concepts and Applications*, SAGE Publications, Inc., London, UK **2014**. https://doi.org/10.4135/9781483348964

[26] L. Rokach, O. Maimon, in *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA **2005**, pp. 165–192. https://doi.org/10.1007/0-387-25465-X_9

[27] G. I. Webb, E. Keogh, R. Miikkulainen, R. Miikkulainen, M. Sebag, in *Encyclopedia of Machine Learning*, Springer, Boston, MA **2011**, pp. 713–714. https://doi.org/10.1007/978-0-387-30164-8_576

[28] P. Gupta, N. K. Sinha, in *Soft Computing and Intelligent Systems*, Academic Press, Cambridge, MA **2000**, pp. 337–356. https://doi.org/10.1016/B978-012646490-0/50017-2

[29] T. Chen, C. Guestrin. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672

[30] P. Kunzmann, K. Hamacher, *BMC Bioinform* **2018**, *19*, 346. https://doi.org/10.1186/S12859-018-2367-Z

[31] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, *25*, 1605. https://doi.org/10.1002/JCC.20084

[32] G. M. Morris, H. Ruth, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, A. J. Olson, *J. Comput. Chem.* **2009**, *30*, 2785. https://doi.org/10.1002/JCC.21256

[33] S. B. Shekarappa, H. Rimac, J. Lee, *Molecules* **2022**, *27*, 4887. https://doi.org/10.3390/MOLECULES27154887

[34] A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, D. R. Koes, *Aust. J. Chem.* **2021**, *13*, 43. https://doi.org/10.1186/S13321-021-00522-2

[35] D. R. Koes, M. P. Baumgartner, C. J. Camacho, *J. Chem. Inf. Model.* **2013**, *53*, 1893. https://doi.org/10.1021/CI300604Z

[36] O. Trott, A. J. Olson, *J. Comput. Chem.* **2010**, *31*, 455. https://doi.org/10.1002/JCC.21334

[37] R. Quiroga, M. A. Villarreal, *PLoS One* **2016**, *11*, e0155283. https://doi.org/10.1371/JOURNAL.PONE.0155183

[38] A. Alhossary, S. D. Handoko, Y. Mu, C. K. Kwoh, *Bioinformatics* **2015**, *31*, 2214. https://doi.org/10.1093/BIOINFORMATICS/BTV082

[39] N. M. Hassan, A. A. Alhossary, Y. Mu, C. K. Kwoh, *Sci. Rep.* **2017**, *7*(1), 2017.

[40] V. O. Gawriljuk, D. H. Foil, A. C. Puhl, K. M. Zorn, T. R. Lane, O. Riabova, V. Makarov, A. S. Godoy, G. Oliva, S. Ekins, *J. Chem. Inf. Model.* **2021**, *61*, 3804. https://doi.org/10.1021/ACS.JCIM.1C00460

[41] A. Rajput, A. Thakur, A. Mukhopadhyay, S. Kamboj, A. Rastogi, S. Gautam, H. Jassal, M. Kumar, *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3133. https://doi.org/10.1016/J.CSBJ.2021.05.037

[42] C. Zhao, S. H. Yang, D. B. Khadka, Y. Jin, K. T. Lee, W. J. Cho, *Bioorg. Med. Chem.* **2015**, *23*, 985. https://doi.org/10.1016/J.BMC.2015.01.016

[43] M. Yang, B. Tao, C. Chen, W. Jia, S. Sun, T. Zhang, X. Wang, *J. Chem. Inf. Model.* **2019**, *59*, 5002. https://doi.org/10.1021/ACS.JCIM.9B00798

[44] E. T. Swann, M. Fernandez, M. L. Coote, A. S. Barnard, *ACS Comb. Sci.* **2017**, *19*, 544. https://doi.org/10.1021/ACSCOMBSCI.7B00087

[45] M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy, B. da Mota, *Aust. J. Chem.* **2019**, *11*, 69. https://doi.org/10.1186/S13321-019-0391-2

[46] S. Hart, K. C. Goh, V. Novotny-Diermayr, Y. C. Tan, B. Madan, C. Amalini, L. C. Ong, B. Kheng, A. Cheong, J. Zhou, W. J. Chng, J. M. Wood, *Blood Cancer J.* **2011**, *1*(11), 2011.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: S. Belenahalli Shekarappa, S. Kandagalla, J. Lee, *J. Comput. Chem.* **2023**, 1. https://doi.org/10.1002/jcc.27103