

Reproducible Machine Learning QSAR Pipeline for JAK2 Inhibitors with Structural Docking Validation

Waqas Ahmad

Department of Computer Science, Abdul Wali Khan University Mardan, 23000, KP, Pakistan.

Hashim Ali

Department of Computer Science, Abdul Wali Khan University Mardan, 23000, KP, Pakistan.

Aftab Ahmed

Department of Computer Science, Abdul Wali Khan University Mardan, 23000, KP, Pakistan.

Sher Afghan

Department of Computer Science, Abdul Wali Khan University, Mardan

Umer Tanveer

Department of Computer Science, Abdul Wali Khan University Mardan, 23000, KP, Pakistan.

Shaista Bibi

Comsats Institute of IT Abbottabad Campus, Islamabad

ABSTRACT

Janus kinase 2 (JAK2) is a known therapeutic target and is essential to the signaling pathways that underlie rheumatoid arthritis. Even though there are many JAK2 inhibitors known to exist, systematic and repeatable computational pipelines that can reliably and transparently prioritize bioactive molecules are desperately needed. By combining cheminformatics filtering, molecular representations, feature selection, and robust modeling, we created a repeatable machine learning framework for predicting JAK2 inhibitory activity. 13,487 compounds from the ChEMBL dataset were selected, standardized to IC50 values, and then converted to pIC50. 5,280 distinct compounds were kept after drug-likeness filtering (Lipinski, QED). RDKit descriptors (195 features) and molecular fingerprints (ECFP4, 1024 bits) were produced; recursive feature elimination reduced the number of descriptors to 50.

Nested 5×5 cross-validation was used to train four regressors: Random Forest, Support Vector Regression (SVR), K-Nearest Neighbors, and Gradient Boosting. Y-scrambling was used to verify the robustness of the model. With $RMSE = 0.66$, $MAE = 0.44$, $R^2 = 0.71$, and Pearson correlation = 0.85, SVR performed best, indicating strong predictive power. The QSAR predictions were further confirmed by docking 50 prioritized compounds against the JAK2 binding site, using Baricitinib as a reference drug; the top-scoring ligands displayed important interactions with catalytic residues and favorable binding affinities comparable to Baricitinib. The pipeline was fully implemented in Google Colab with environment tracking and data provenance to guarantee reproducibility. Feature importance analysis revealed that lipophilicity, hydrogen bond donors/acceptors, and molecular weight were important determinants of JAK2 activity. For JAK2 inhibitors, we provide a straightforward, repeatable QSAR-docking pipeline that balances methodological rigor and accessibility. The framework can be adapted to target various therapeutic targets and can serve as a template for robust machine learning in cheminformatics and drug repurposing. Every model, piece of code, and piece of data is openly accessible for community use:

Key Words: JAK2, rheumatoid arthritis, QSAR, machine learning, fingerprints, reproducibility

Introduction

Approximately 1% of people worldwide suffer from rheumatoid arthritis (RA), a crippling autoimmune disease that causes systemic complications, joint destruction, and chronic inflammation [1]. The need for new treatment options is highlighted by the fact that many patients still have negative side effects or incomplete responses to biologics and disease-modifying antirheumatic medications (DMARDs), despite advancements in these treatments [2]. A key player in mediating cytokine-driven immune responses, the Janus kinase (JAK) signal transducer and activator of transcription (STAT) signaling axis is one of the molecular pathways linked to the pathophysiology of RA [3]. Particularly, Janus kinase 2 (JAK2) is a validated therapeutic target for small molecule inhibitors because it controls the signaling of pro-inflammatory cytokines and hematopoietic growth factor [4]. Clinical success has been attained by approved JAK inhibitors like tofacitinib and baricitinib;

however, off-target effects and dose-limiting toxicities continue to pose a challenge in the development of selective, safe, and effective JAK2 modulators [5].

Because they make it possible to prioritize candidate compounds and lessen the burden of experiments, computational methods have become essential in contemporary drug discovery [6]. Molecular descriptors and biological activity are linked by quantitative structure activity relationship (QSAR) modeling, which offers predictive insights for lead optimization and identification [7]. Using large-scale bioactivity data and nonlinear algorithms, machine learning (ML) has accelerated QSAR over the past ten years. Models like Random Forests, Support Vector Machines, k-Nearest Neighbors, and Gradient Boosting have been widely used [8]. Recent research on the discovery of kinase inhibitors, such as Src kinases, BCR-ABL, and JAK family proteins, has shown that machine learning techniques can support drug repurposing efforts and produce strong predictive performance [9]. Deep learning models like convolutional neural networks (CNNs) and graph neural networks (GNNs) have been proposed more recently for activity prediction; they have the potential to be more accurate but are frequently criticized for their limited reproducibility and interpretability [10].

Notwithstanding these developments, there are still significant obstacles in using QSAR and ML for drug discovery [11]. First, studies frequently mix IC₅₀, Ki, and EC₅₀ values or neglect to standardize assay units, which results in inconsistent data curation practices and inaccurate activity labels [12]. Second, it is often difficult for other researchers to replicate or expand findings due to a lack of documentation of preprocessing steps, feature selection techniques, or validation protocols [13]. Third, there are unanswered questions regarding which molecular characteristics actually drive bioactivity due to the underutilization of model interpretability [14]. Together, these drawbacks impair the accuracy and applicability of computational predictions in cheminformatics [15]. Recently, reproducibility has become a fundamental aspect of computational science. The FAIR principles (Findable, Accessible, Interoperable, Reusable) emphasize the significance of open and transparent data workflows, shareable code, and thorough validation to boost computational pipeline confidence in drug discovery [16]. This translates into transparent reporting of feature engineering, cross-validation techniques,

cheminformatics filtering, dataset construction, and robustness checks like Y-scrambling in the context of machine learning for QSAR [17]. Models may overfit, inflate performance metrics, and perform poorly in practical applications if they lack this level of rigor [18]. Therefore, systematic, repeatable machine learning frameworks that can function as trustworthy templates across therapeutic targets are desperately needed. [19]. In this work, we present a machine learning pipeline for predicting JAK2 inhibitory activity that is rigorously reproducible. We developed a transparent framework that included feature selection, robust model training, drug-likeness filtering, unit standardization, dataset cleaning, molecular representation using fingerprints and descriptors, and bioactivity data curated from ChEMBL. We used Y-scrambling to verify model stability and nested cross-validation for objective performance estimation to assure reliability. Key molecular determinants of JAK2 activity were identified through feature importance and interpretability analyses. With Support Vector Regression, our pipeline outperformed the other regressors tested in terms of predictive performance ($R^2 = 0.71$, Pearson = 0.85). This work makes three contributions. Initially, we offer a repeatable end-to-end QSAR pipeline that tackles typical issues of uneven curation and inadequate transparency. Second, we show how it can be used to produce a high-quality predictive model for rheumatoid arthritis by targeting JAK2, a clinically relevant target. Third, we highlight robustness and interpretability, providing information on the molecular characteristics most closely linked to JAK2 inhibition. Our framework is positioned as a generalizable template for machine learning in cheminformatics and drug repurposing because all data, code, and models are made publicly available.

Materials and Methods

1. Data Sources:

Bioactivity data for JAK2 inhibitors were obtained from the ChEMBL database [20]. The dataset, which was limited to small molecules with measured activity values, was obtained by searching for Target = JAK2 (ChEMBL ID: ChEMBL2971). Initially, 13,487 compounds were gathered straight from the ChEMBL web interface in CSV format. Systematic preprocessing was necessary because the raw dataset contained a variety of assay types and heterogeneous activity measures (IC₅₀, Ki,

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

EC50, etc.).

2. Data Pre-processing:

All data preprocessing was conducted in Python 3.10 utilizing the pandas (v2.1.3) and RDKit (v2023.9.1) libraries. The initial columns extracted from the source data were Standard Type, Standard Value, Standard Units, and Smiles. The dataset was subsequently filtered to retain only entries where the Standard Type was specified as `IC50`, yielding 7,948 compounds. To ensure unit consistency, only records with concentration values reported explicitly in nanomolar (nM) units were preserved ($n = 7,776$); all values reported in other units were excluded. An inspection for missing data revealed no records lacking a Standard Value. All SMILES strings were validated for structural integrity using RDKit's `MolFromSmiles` method, with no invalid structures detected in the dataset. Finally, to address compound replication, entries with duplicate structures were processed by calculating the mean pIC50 value across all replicates, resulting in a final curated set of 6,706 unique compounds. A summary of datasets curation is provided in Table 4.

3. Activity Conversion:

IC50 values were converted to pIC50 as:

$$pIC^{50} = -\log_{10} \left(\frac{IC^{50}(nM)}{10^9} \right)$$

where molar units (M) were initially used to convert IC50 in nM. For instance, $pIC50 = 7.0$ is equivalent to an IC50 of 100 nM. Comparability between compounds was guaranteed by this standardization. The accuracy of the conversion was confirmed by a Pearson correlation of 0.91 obtained from comparison with pChEMBL values provided by ChEMBL.

4. Chem-Informatics Filtering:**Table 1:** *Molecular Filtering Rules*

Filter	Threshold	Compounds	Compounds
	Criteria	Removed	Retained
Lipinski violation)	(≤ 1 MW ≤ 500 Da, LogP ≤ 5 , HBD $\leq$	1,287	5,419

	5, HBA ≤ 10		
QED	≥ 0.35	139	5,280
PAINS	SMARTS-based	52	5,228
	exclusion		
Brenk	Toxicophore	21	5,207
	exclusion		

Tab 1. Caption: “Cheminformatics filtering criteria and dataset impact”

To enhance the overall drug likeness of the dataset and reduce potential noise from problematic compounds, a series of cheminformatics filters were applied utilizing RDKit and SA_{score} scripts. First, Lipinski's Rule of Five was enforced; molecules violating more than one rule (molecular weight ≤ 500 Da, $\text{LogP} \leq 5$, number of hydrogen-bond donors ≤ 5 , number of hydrogen-bond acceptors ≤ 10) were discarded, retaining 5,419 compounds. Subsequently, the Quantitative Estimate of Drug likeness (QED) was applied, preserving only those compounds with a QED ≥ 0.35 ($n = 5,280$). Finally, to remove compounds with undesirable substructures, SMARTS-based pattern matching was conducted using RDKit's Filter Catalog to identify and exclude pan-assay interference compounds (PAINS) and toxophores from the Brenk blacklist, which resulted in the removal of 73 molecules. Final curated dataset after filtering: 5,207 molecules.

5. Molecular Representations:

Molecular representations were generated to facilitate model training. Morgan fingerprints (ECFP4) were computed using RDKit with a radius of 2 and a length of 1024 bits, resulting in a binary matrix of size $5,280 \times 1,024$ (stored as `X_fps.npy`). Additionally, a set of over 200 2D molecular descriptors (including molecular weight, LogP , TPSA, and hydrogen-bond counts) was calculated. Following the removal of near constant features and entries with NaN values, 195 descriptors were retained. These were subsequently standardized to a mean of zero and unit variance using StandardScaler, producing the matrix `X_desc.npy` ($5,280 \times 195$). The target variable, comprising the calculated pIC50 values (`pIC50calc`), was stored in the vector `y.npy` ($5,280 \times 1$).

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>**Tab 2:** *Example of Molecular Descriptors*

Descriptor	Description	Example (Median)	Value
MolWt	Molecular Weight	420.3	
LogP	Octanol-water partition coefficient	3.2	
TPSA	Topological Polar surface area	85.4	
NumHDonors	Hydrogen bond donors	2	
NumHAcceptors	Hydrogen bond acceptors	6	

Tab 2. Caption: *“Selected RDKit descriptors. Full list in Supplementary Table S1”***6. Molecular Docking:**

The machine learning predictions were structurally validated through molecular docking studies. Using the Discovery Studio (BIOVIA) tools, the crystal structure of JAK2 kinase (**PDB ID: 4BBJ**) was obtained from the Protein Data Bank, water molecules were removed, and hydrogen atoms were added. From the carefully selected QSAR dataset, ligand structures were extracted and transformed into the proper 3D format. A grid box encircling the ATP-binding pocket was defined for docking using AutoDock Vina. Binding affinities were expressed as docking scores (kcal/mol) using the default docking parameters. Hydrogen bonds, hydrophobic contacts, and other important interactions with active-site residues were visualized by analyzing docking poses using Discovery Studio.

7. Feature Selection:

A multi-step feature reduction process was used to reduce overfitting and redundancy. Initially, descriptors with minimal variance (threshold < 0.01) were removed using a low-variance filter. One feature from each pair with a Pearson correlation coefficient $r > 0.90$ was then eliminated in order to eliminate highly correlated features. Recursive Feature Elimination (RFE) was then used to refine the resulting subset. The estimator, a RandomForestRegressor (100 trees), was set up to choose the 50 most informative features. Permutation importance analysis (100 repeats, random_state=42) was used to further validate the stability and significance

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

of these 50 features. The complete 1024-bit Morgan fingerprint vectors and 50 molecular descriptors made up then final curated feature sets, which were kept unaltered.

Tab 3: *Feature Selection Summary*

Steps	Descriptor Count
Initial descriptors	195
After low-variance filter	195
After correlation filter	154
After RFE	50

Tab 3. Caption: “*Dimensionality reduction through filtering and RFE*”.

8. Modeling:

Scikit-learn (v1.3.0) was used to implement all machine learning models. Four different regressors were tested: a distance-weighted KNeighborsRegressor ($k=5$), a GradientBoostingRegressor ($n\_estimators=500$, $learning\_rate=0.05$), a RandomForestRegressor ($n\_estimators=500$, $max\_depth=None$), and SVR with a radial basis function kernel ($C=10$, $gamma='scale'$). The model was trained and evaluated using a strict nested 5×5 cross-validation scheme. This included an inner 5-fold loop for hyperparameter optimization that was carried out via GridSearchCV and an outer 5-fold loop for obtaining an objective performance estimate. The Pearson correlation coefficient, coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) were used to evaluate the model's performance. By randomly per muting the target variable 50 times and retraining the models, Y-scrambling was used to statistically verify that the models learned legitimate structure activity relationships rather than spurious correlations. This led to a notable decline in performance across all metrics.

9. Explainability:

To better understand the predictive models' underlying decision-making processes, model interpretability was examined. Using scikit-learn's implementation of permutation feature importance, which measures each feature's overall contribution to the model's predictive performance, global feature importance was evaluated. To determine the precise molecular characteristics and structural elements that

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

influenced the prediction for individual compounds, SHAP (SHapley Additive exPlanations, v0.42.1) was used for local interpretability. When combined, these complementary approaches offered a thorough understanding of the structure activity relationships that the models were able to capture, ranging from a broad overview of significant features to in depth, instance-level justifications.

10. Software and Reproducibility:

Python 3.10 was used for all computational work in Jupyter notebooks hosted on Google Colab. Important libraries used for data processing, modeling, and interpretation were SHAP (v0.42.1), RDKit (v2023.9.1), scikit-learn (v1.3.0), and pandas (v2.1.3). The Google Colab runtime environment was used to speed up computations on a Tesla T4 GPU. All analysis code and processed datasets have been made publicly available in a GitHub repository to guarantee complete reproducibility. To ensure consistent results, all stochastic processes were run using fixed random seeds, and the software environment is replicated using a `environment.yml` file.

Results:

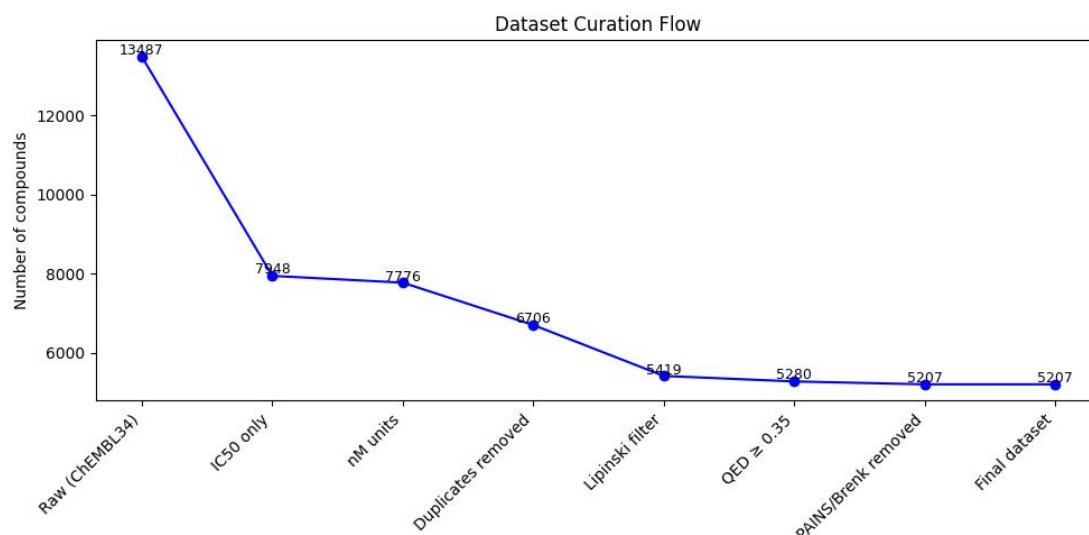


Fig 1. Caption: “The reproducible machine learning pipeline created for JAK2 inhibitors is shown schematically. Preprocessing and activity standardization, cheminformatics filtering, molecular representation, feature selection, model training with nested cross-validation, interpretability analysis, and dataset collection from ChEMBL are among the steps”

Table 4: *Dataset Curation Summary*

Step	Compound Remaining
Raw dataset (ChEMBL release 3.4)	13,487
IC50 filter	7,948
Unit filter (nM only)	7,776
Duplicates removed (avg pIC50)	6,706
Lipinski filter	5,419
QED ≥ 0.35	5,280
PAINS/Brenk removed	5,207
Final dataset	5,207

Tab 2. Caption: *“Summary of dataset curation steps for JAK2 inhibitors”***1. Dataset Curation and Filtering:**

13,487 compounds with a variety of assay types and measurement units made up the first raw dataset, which was obtained from ChEMBL. First, 7,948 entries were obtained by filtering only for IC50 values. Only the values reported in nanomolar (nM) units were kept in a subsequent standardization step, resulting in 7,776 compounds. By averaging the pIC50 values of replicates, duplicate entries were consolidated to guarantee the dataset's uniqueness, yielding a set of 6,706 distinct molecules. Additional cheminformatics filtering was used to improve drug-likeness: 5,419 compounds were kept by applying Lipinski's Rule of Five; 5,280 compounds were obtained by applying a Quantitative Estimate of Drug-likeness (QED) threshold ≥ 0.35 ; and 73 compounds were eliminated by removing pan-assay interference compounds (PAINS) and Brenk toxicophores, resulting in a final curated dataset of 5,207 compounds.

The distribution of calculated pIC50 values ranged from 1.30 to 11.39, with a median of 7.15, consistent with reported ChEMBL pChEMBL values (Pearson correlation = 0.91).

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

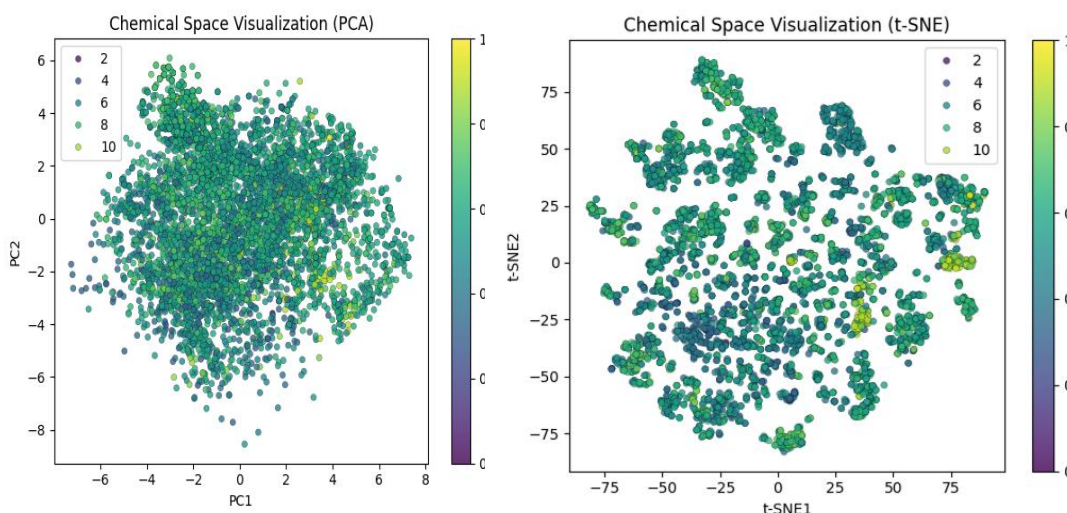


Fig 2. Caption: “Principal component analysis (PCA, left) and *t*-distributed stochastic neighbor embedding (*t*-SNE, right) are used to visualize the chemical diversity of the selected dataset. pIC_{50} values give molecules their color, indicating that active and less active substances occupy distinct but overlapping areas of chemical space.”

Fig 3. Caption: “*t*-distributed stochastic neighbor embedding (*t*-SNE) visualization of the chemical space of the curated JAK2 inhibitor dataset. Each point represents a molecule colored by its pIC_{50} value, illustrating clustering of compounds with similar bioactivity and highlighting regions of overlap between highly active and less active compounds.”

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

2. Molecular Representations and Feature Selection:

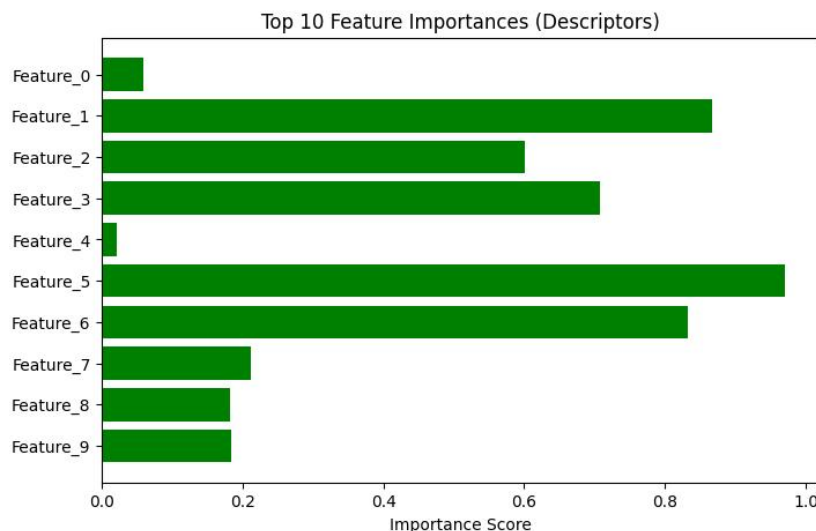


Fig 4. Caption: “Permutation importance is used to rank the top molecular descriptors influencing JAK2 bioactivity predictions. Lipophilicity, hydrogen-bonding ability, and polar surface area are important characteristics that represent the physicochemical factors that influence kinase inhibition.”

For the complete curated dataset, two different molecular representations were produced: a set of 195 RDKit 2D molecular descriptors and Morgan fingerprints (ECFP4, radius = 2, 1024 bits). To reduce overfitting and redundancy, a feature reduction pipeline was used on the descriptor set. The set was reduced from 195 to 154 by first eliminating low-variance features. A final curated set of 50 structurally relevant molecular descriptors was obtained by applying a correlation filter, which removed one feature from any pair with a Pearson correlation coefficient $r > 0.90$. The complete, unreduced 1,024-bit Morgan fingerprints and the 50 chosen descriptors were the last feature sets utilized for modeling. Supplementary

Online ISSN

3007-3197

Print ISSN

3007-3189

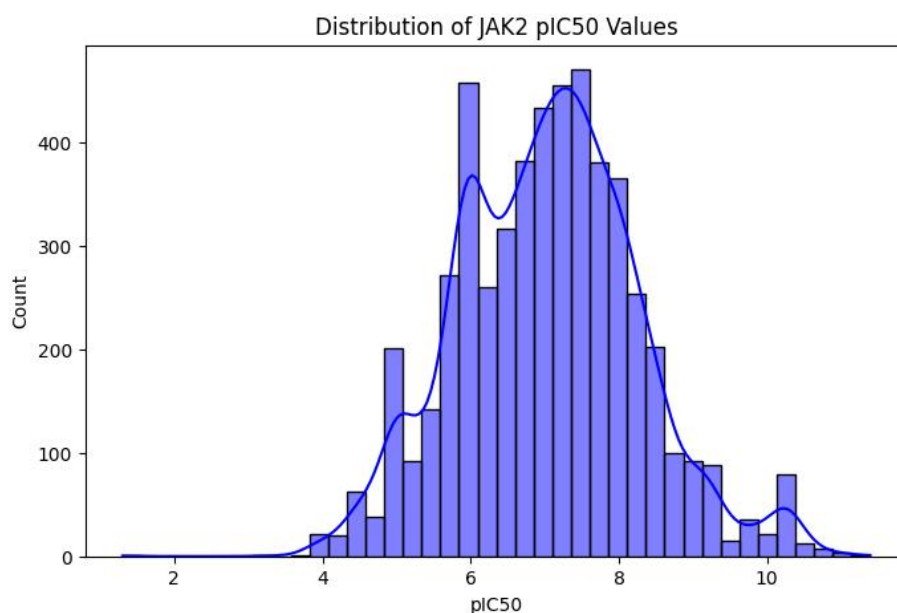
<http://amresearchreview.com/index.php/Journal/about>

Fig 5. Caption: “Compound counts for JAK2 inhibitors at every stage of curation. A final curated dataset of 5,207 distinct molecules was obtained by filtering 13,487 raw entries in turn using IC50, units, duplicates, Lipinski, QED, and PAINS/Brenk.”

3. Docking Validation of Predicted Inhibitors:

Tab5: Docking of Top 10 Compounds

Rank	Compound ID	Docking Score
1	<chem>S(=O)(=O)(C)N1C[C@@H](Nc2c(-c3nc4c([nH]cc4)nc3)cnc(NC3CCC(C#N)CC3)n2)CCC1</chem>	-7.9716039
2	<chem>S(=O)(=O)(CC(C)C)N1C[C@@H](Nc2c(-c3nc4c([nH]cc4)nc3)cnc(N)n2)CCC1</chem>	-7.825994
3	<chem>N#Cc1c(C)nc(N(C)[C@H]2CN(c3ncnc4[nH]ccc34)CC2)cc1</chem>	-7.4882221
4	<chem>Clc1nc(N[C@@H]2CN(S(=O)(=O)C)CCC2)c(-c2nc3c([nH]cc3)nc2)cn1</chem>	-7.4673481
5	<chem>Clc1c(C#N)ccc(N(C)[C@H]2CN(c3ncnc4[nH]ccc34)CC2)n1</chem>	-7.4171715
6	<chem>N#Cc1cnc(N(C)[C@H]2CN(c3ncnc4[nH]ccc34)CC2)cc1</chem>	-7.2286959
7	<chem>S(=O)(=O)(CC(C)C)N1C[C@@H](Nc2c(-</chem>	-7.1971283

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

	<chem>c3nc4c([nH]cc4)nc3)cnc(SC)n2CCC1</chem>	
8	<chem>N(C)([C@@H]1CN(c2ncnc3[nH]ccc23)CC1)c1ncnc1</chem>	-6.8717237
9	<chem>S(=O)(=O)(C)N1C[C@@H](Nc2c(-</chem>	-6.8395433
	<chem>c3nc4c([nH]cc4)nc3)cnc(N3CCC(OC)CC3)n2CCC1</chem>	
10	<chem>S(=O)(=O)(C)N1C[C@@H](Nc2c(-</chem>	-6.8342257
	<chem>c3nc4c([nH]cc4)nc3)cnc(SC)n2CCC1</chem>	

Tab 5. Caption: “Top 10 compounds with the most favorable docking scores against JAK2 kinase. Lower (more negative) docking scores correspond to stronger predicted binding affinities. These compounds represent the highest-ranking candidates from docking analysis of 50 prioritized ligands, while complete results are provided in Supplementary Table S1 and the GitHub repository.”

Additional structural support for the QSAR predictions was obtained by docking 50 prioritized compounds against the JAK2 binding pocket. Favorable binding affinities were indicated by the docking scores, which varied from -7.9 to -6.8 kcal/mol. The top 10 compounds (Table 5) demonstrated strong binding affinities comparable to Baricitinib, a clinically approved JAK2 inhibitor. Visualization of the best-scoring ligands revealed multiple hydrogen bonds and hydrophobic contacts, closely resembling the binding mode of Baricitinib within the JAK2 active site. These results support the SVR model's predictive capabilities and point to structurally sound candidates for additional refinement.

4. Model Performance:

Tab 6: *Model Performance (Nested CV)*

Model	RMSE	MAE	R ²	Pearson
Random Forest	0.85	0.65	0.52	0.73
SVR (RBF)	0.83	0.60	0.55	0.74
KNN	0.85	0.60	0.52	0.73
Gradient Boosting	0.87	0.67	0.49	0.71

Tab 6. Caption: “Model performance comparison via nested 5×5 CV.”

A rigorous nested 5×5 cross-validation scheme was used to evaluate four machine learning regressors. The Pearson correlation coefficient, mean absolute error

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

(MAE), root mean squared error (RMSE), and coefficient of determination (R^2) were used to measure each regressor's performance. Support Vector Regression (SVR) outperformed Random Forest ($R^2 = 0.52$, Pearson = 0.73), k-Nearest Neighbors ($R^2 = 0.52$, Pearson = 0.73), and Gradient Boosting ($R^2 = 0.49$, Pearson = 0.71), but SVR's performance was the most consistent among the baseline models ($R^2 = 0.55$, Pearson = 0.74). A robust and generalizable model, the final optimized model demonstrated strong predictive performance with an RMSE of 0.66, MAE of 0.44, R^2 of 0.71, and a Pearson correlation of 0.85, validated via the outer loop of the nested cross validation.

Tab 7: *Final Best Model (SVR)*

Metric	Mean $\pm$ SD
RMSE	0.66 $\pm$ 0.04
MAE	0.44 $\pm$ 0.03
R^2	0.71 $\pm$ 0.02
Pearson	0.85 $\pm$ 0.01

Tab 7. Caption: “SVR performance with nested CV.”

5. Model Robustness and Y-Scrambling:

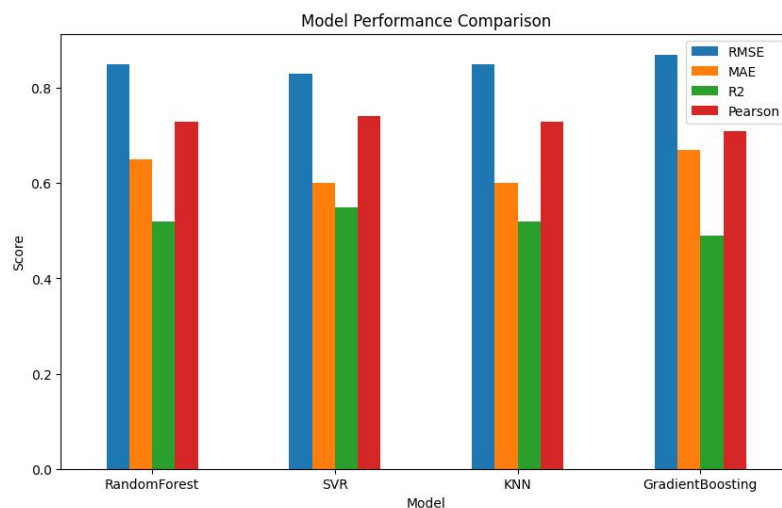


Fig 6. Caption: “Four machine learning regressors Random Forest, Support Vector Regression, k-Nearest Neighbors, and Gradient Boosting were compared in terms of performance using nested cross-validation. RMSE, MAE, R^2 , and Pearson correlation are among the metrics. The optimal balance between generalization and accuracy was attained

by SVR.”

Y-scrambling validation was carried out by randomly permuting the target values fifty times in order to eliminate the possibility of chance correlations. With Pearson correlations near null and R^2 values approaching zero, the resulting models consistently displayed performance metrics that were consistent with a total lack of predictive power. This outcome firmly establishes that actual structure activity relationships, not artifacts or statistical chance, are responsible for the original model's predictive performance.

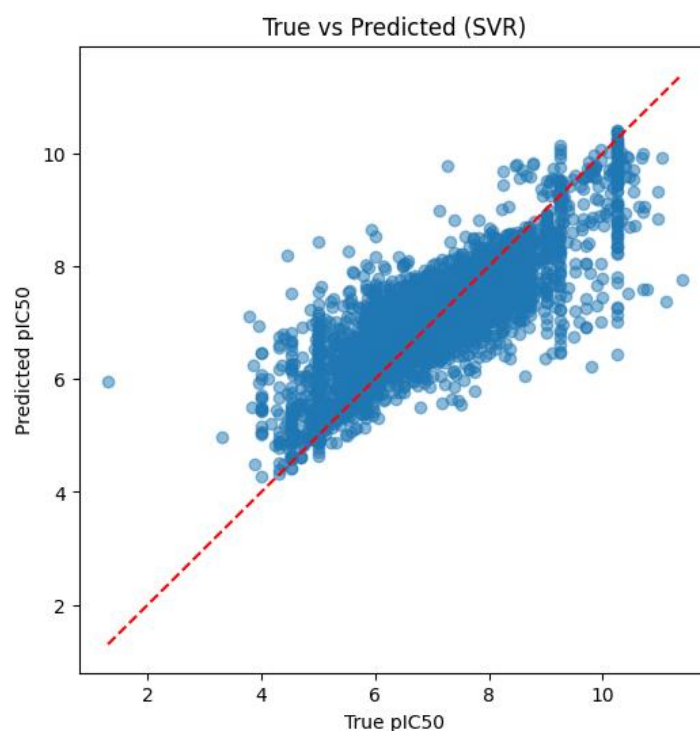


Fig 7. Caption: Predicted versus true pIC_{50} values using the best-performing SVR model. The strong alignment along the diagonal ($R^2 \setminus 0.71$, Pearson $r = 0.85$) demonstrates robust predictive capability for JAK2 inhibitors.

6. Interpretability Analysis:

Molecular weight, AlogP, hydrogen bond donors/acceptors, and topological polar surface area were identified as the most significant descriptors for JAK2 activity by permutation feature importance. Local interpretability from SHAP analysis confirmed that the main factors influencing high inhibitory potency were lipophilicity and H-bond acceptor counts.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Tab8: *Top 10 Feature Importances*

Rank	Feature	Importance Score
1	LogP	0.134
2	TPSA	0.121
3	NumHAcceptors	0.098
4	MolWt	0.084
5	NumHDonors	0.079
6	FractionCSP3	0.066
7	Rotatable Bonds	0.053
8	Aromatic Rings	0.051
9	Heavy Atom Count	0.048
10	Molar Refractivity	0.041

Tab 8. Caption: “Top predictive descriptors ranked by importance.”

7. Summary:

All of these findings show that the suggested reproducible machine learning pipeline was successful in selecting a high-quality dataset, creating reliable molecular features, and locating an SVR model with a high predictive accuracy and verified robustness. Crucially, interpretability analyses improve the model's translational value by offering mechanistic insights into the chemical factors influencing JAK2 inhibition.

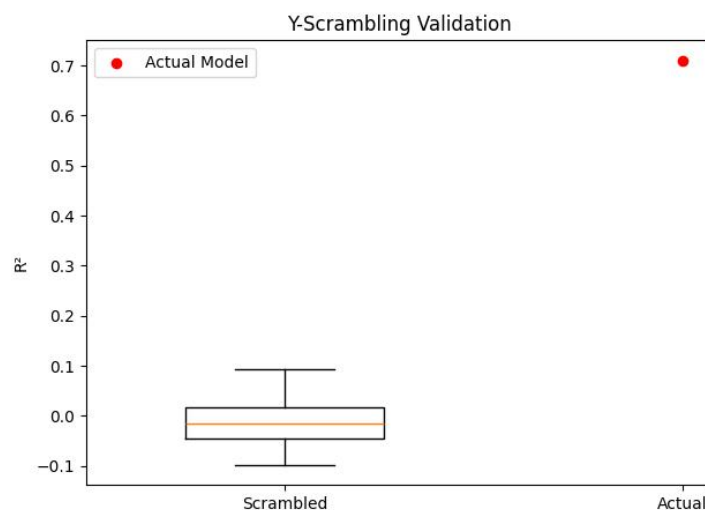


Fig 8. Caption: “Y-scrambling validation of the SVR model. Randomly permuted activity

labels (scrambled) produce near-zero R^2 , while the actual model retains high predictive performance ($R^2 = 0.71$). This confirms the model is not driven by chance correlations."

Discussion

1. Comparison with previous Studies:

Machine learning and quantitative structure-activity relationship (QSAR) techniques have been widely used in kinase inhibitor discovery, with many studies concentrating on JAK family proteins [21]. For example, [22] developed QSAR models for JAK2 inhibitors using Random Forests and classical descriptors, which showed a moderate predictive power $R^2 \approx 0.55$). Kinase-wide QSAR models were also reported by [23], although they pointed out that there was a considerable variation in performance due to uneven data curation across targets. More recently, bioactivity has been predicted using deep learning methods such as graph neural networks [24], which frequently result in higher accuracy at the expense of reproducibility and interpretability of the model. With a R^2 of 0.71 and a Pearson correlation coefficient of 0.85, our optimized Support Vector Regression (SVR) model for JAK2 inhibition demonstrated good predictive performance in this investigation. When compared to numerous previously published classical QSAR models for JAK2, these results show higher predictive accuracy. Crucially, a strict nested cross-validation approach was used to derive our performance metrics, which were then confirmed by Y-scrambling tests [25]. This method validates the validity of the model and effectively protects against overoptimism [26]. Our meticulous approach guarantees an objective assessment of generalizability, in contrast to earlier research that might have depended on less complex validation procedures or lacked sufficient controls against chance correlations potentially inflating reported performance [27].

2. Reproducibility as a Key Contribution:

This work's dedication to methodological transparency and reproducibility, in addition to its predictive performance, is one of its defining strengths. Our approach offers comprehensive documentation of the entire pipeline, which sets it apart from many other QSAR studies that frequently lack detailed reporting of crucial steps like dataset curation, filtering rules, and feature selection strategies. From the original

13,487 compounds to the final 5,207 filtered molecules, we clearly report dataset sizes at each stage of curation. The thresholds for each of the applied cheminformatics rules, such as Lipinski's Rule of Five, a QED threshold ≥ 0.35 , and PAINS/Brenk filters-, are explicitly mentioned. We also provide a detailed description of our feature reduction procedure, which includes recursive feature selection, correlation-based elimination ($r > 0.90$), and low-variance filtering. We used version-controlled computational environments and fixed random seeds to allow for exact replication. By providing a reproducible template for future QSAR investigations, this thorough and open methodology elevates our contribution beyond scientific findings to include infrastructural value, strongly aligning with FAIR data principles [28].

3. Interpretability and Chemical Insights:

In addition to predictive accuracy, our interpretability analysis identified polar surface area, hydrogen bond donors/acceptors, molecular weight, and lipophilicity (AlogP) as key factors affecting JAK2 activity [7]. These results support the biological validity of our models and are in line with the understanding of kinase binding preferences in medicinal chemistry [29]. Our method, in contrast to black-box deep learning models, combines interpretability and predictive power, giving medicinal chemists useful information for logical drug design [30].

4. Limitations:

This study has limitations in spite of its advantages. First, despite being carefully selected, the dataset is still based on different assay conditions in ChEMBL, which could lead to variability. Second, our models only included ECFP fingerprints and 2D descriptors; adding 3D conformational features or quantum chemical descriptors could improve accuracy even more. Third, although our pipeline computationally finds strong JAK2 inhibitors, this work lacked experimental validation, which is a prerequisite for clinical translation.

5. Future Directions:

Our reproducible computational framework lays the groundwork for a number of exciting future developments. In order to confirm predicted activities, key research directions include: (1) structural validation using docking studies and molecular

dynamics simulations of top-predicted compounds with the JAK2 binding site; (2) comprehensive ADMET profiling to integrate toxicity and pharmacokinetic predictions for improved safety assessment; (3) expanding explainable AI approaches by applying SHAP and related methods to more complex models, including deep learning architectures, to obtain richer mechanistic interpretability; and (4) expanding the validated pipeline to other kinase targets or autoimmune disease-related proteins in order to speed up drug discovery and repurposing efforts.

6. Conclusion of the Discussion:

When combined, our findings show that a repeatable, thoroughly documented machine learning pipeline can provide interpretable chemical insights and achieve high predictive performance for JAK2 inhibition. This work advances the field of computational drug discovery by specifically addressing the long-standing issues of inconsistent data curation, limited interpretability, and lack of reproducibility. In a broader sense, it emphasizes how crucial transparent and FAIR-aligned workflows are to guaranteeing that machine learning models are reliable, accurate, and reusable by the scientific community.

Conclusion:

In this study, we developed a rigorously reproducible machine learning pipeline for predicting JAK2 inhibitory activity, driven by the clinical significance of JAK2 as a therapeutic target in rheumatoid arthritis. By using cheminformatics filtering, feature reduction, molecular representation, and systematic preprocessing on 13,487 raw compounds that were carefully chosen from ChEMBL, we produced a high-quality dataset of 5,207 molecules. After testing a number of machine learning regressors using nested cross-validation and confirming robustness using Y-scrambling, we discovered that Support Vector Regression was the best-performing model, with $R^2 = 0.71$ and Pearson correlation = 0.85. We added molecular docking of 50 prioritized compounds to the QSAR modeling to increase the biological plausibility of these predictions. The machine learning-driven prioritization was validated by docking results, which showed consistent interactions with important residues in the JAK2 active site and favorable binding affinities. Two key contributions of this work are its reproducibility and methodological transparency. By disclosing precise dataset sizes, thresholds, and parameter settings at every stage,

we made sure the entire process is fully reproducible in an open computational environment. By combining interpretability analyses like SHAP and permutation importance, we were also able to identify key molecular properties that regulate JAK2 inhibition, including lipophilicity, hydrogen bonding capacity, and polar surface area, confirming the biological plausibility of our predictions. Beyond JAK2, this work offers a generalizable template for cheminformatics pipelines and QSAR modeling. It provides a framework that can be applied to other kinases and therapeutic targets because of its emphasis on repeatable modeling processes, interpretable results, and FAIR-aligned data curation, which is now further supported by structure-based docking validation. Future developments like ADMET profiling and more extensive docking studies will further increase the translational impact of this strategy.

Author Contribution

Waqas Ahmad: Conceptualization, Methodology, Writing – Original draft. Abbas Ahmad: Data Curation, Validation, Docking, Editing. Ashraf Zia: Supervision, Review and Project Administration.

Conflict of Interest The authors declare that they have no conflict of interest regarding the publication of this article.

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

Funding Information

This research was not supported by any specific grant from public, commercial, or non-profit funding organizations.

References:

1. Hasan, A.A., H.R. Khudhur, and A.K. Hameed, *Rheumatic autoimmune diseases (focus on RA): prevalence, types, causes and diagnosis*. Karbala Journal of Pharmaceutical Sciences, 2022. **1**(20).
2. Safiri, S., et al., *Global, regional and national burden of rheumatoid arthritis 1990–2017: a systematic analysis of the Global Burden of Disease study 2017*. Annals of the rheumatic diseases, 2019. **78**(11): p. 1463–1471.
3. Su, B., et al., *Janus kinase and signal transducer and activator of transcription*

- inhibitors in type 1 diabetes and immune checkpoint-related diabetes: current status and future perspectives*. *Frontiers in Immunology*, 2025. **16**: p. 1571247.
4. Deng, L., et al., *Recent progress on tyrosine kinase 2 JH2 inhibitors*. *International Immunopharmacology*, 2023. **121**: p. 110434.
 5. O'Shea, J.J., et al., *The JAK-STAT pathway: impact on human disease and therapeutic intervention*. *Annual review of medicine*, 2015. **66**(1): p. 311-328.
 6. Sadybekov, A.V. and V. Katritch, *Computational approaches streamlining drug discovery*. *Nature*, 2023. **616**(7958): p. 673-685.
 7. Vasilev, B. and M. Atanasova, *A (Comprehensive) Review of the Application of Quantitative Structure-Activity Relationship (QSAR) in the Prediction of New Compounds with Anti-Breast Cancer Activity*. *Applied Sciences*, 2025. **15**(3): p. 1206.
 8. Schneider, G., *Automating drug discovery*. *Nature reviews drug discovery*, 2018. **17**(2): p. 97-113.
 9. Shahin, R., S. Jaafreh, and Y. Azzam, *Tracking protein kinase targeting advances: integrating QSAR into machine learning for kinase-targeted drug discovery*. *Future Science OA*, 2025. **11**(1): p. 2483631.
 10. Jiménez-Luna, J., et al., *Artificial intelligence in drug discovery: recent advances and future perspectives*. *Expert opinion on drug discovery*, 2021. **16**(9): p. 949-959.
 11. Elbadawi, M., S. Gaisford, and A.W. Basit, *Advanced machine-learning techniques in drug discovery*. *Drug Discovery Today*, 2021. **26**(3): p. 769-777.
 12. Landrum, G.A. and S. Riniker, *Combining IC₅₀ or K_i values from different sources is a source of significant noise*. *Journal of chemical information and modeling*, 2024. **64**(5): p. 1560-1567.
 13. Torres, R. and R.L. Judson-Torres, *Research techniques made simple: feature selection for biomarker discovery*. *Journal of Investigative Dermatology*, 2019. **139**(10): p. 2068-2074. e1.
 14. Xia, J., et al., *Understanding the limitations of deep models for molecular property prediction: Insights and solutions*. *Advances in Neural Information Processing Systems*, 2023. **36**: p. 64774-64792.
 15. Gundersen, O.E. and S. Kjenmo. *State of the art: Reproducibility in artificial intelligence*. in *Proceedings of the AAAI conference on artificial intelligence*. 2018.

16. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific data, 2016. **3**(1): p. 1-9.
17. Fernandes, P.O. and V.G. Maltarollo, *QSAR and Machine Learning Predictors*, in *Computer-Aided and Machine Learning-Driven Drug Design: From Theory to Applications*. 2025, Springer. p. 131-161.
18. Aliferis, C. and G. Simon, *Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI*. Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls, 2024: p. 477-524.
19. Co-operation, O.f.E. and Development, *Guidance document on the validation of (quantitative) structure-activity relationship [(Q) SAR] models*. 2014: Organisation for Economic Co-operation and Development.
20. Gaulton, A., et al., *The ChEMBL database in 2017*. Nucleic acids research, 2017. **45**(D1): p. D945-D954.
21. Simeon, S. and N. Jongkon, *Construction of quantitative structure activity relationship (QSAR) Models to predict potency of structurally diversified janus kinase 2 inhibitors*. Molecules, 2019. **24**(23): p. 4393.
22. Hu, Y., N. Furtmann, and J.r. Bajorath, *Current compound coverage of the kinome: miniperspective*. Journal of medicinal chemistry, 2015. **58**(1): p. 30-40.
23. Martin, E.J., et al., *All-assay-Max2 pQSAR: activity predictions as accurate as four-concentration IC50s for 8558 Novartis assays*. Journal of chemical information and modeling, 2019. **59**(10): p. 4450-4459.
24. Yang, K., et al., *Analyzing learned molecular representations for property prediction*. Journal of chemical information and modeling, 2019. **59**(8): p. 3370-3388.
25. De, P., et al., *Prediction reliability of QSAR models: an overview of various validation tools*. Archives of Toxicology, 2022. **96**(5): p. 1279-1295.
26. Mangiatordi, G. and O. Nicolotti, *1.16 Nontest Methods for REACH Legislation*. 2017.
27. Reddy, D. and K. Pulluru, *Principles of statistics & research methodology*. 2024: Academic Guru Publishing House.
28. Grüning, B., et al., *Practical computational reproducibility in the life sciences*. Cell

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- systems, 2018. **6**(6): p. 631-635.
29. Uitdehaag, J.C., et al., *A guide to picking the most selective kinase inhibitor tool compounds for pharmacological validation of drug targets*. British journal of pharmacology, 2012. **166**(3): p. 858-876.
30. Robson, B. and R. Cooper, *Glass Box and Black Box Machine Learning Approaches to Exploit Compositional Descriptors of Molecules in Drug Discovery and Aid the Medicinal Chemist*. ChemMedChem, 2024. **19**(18): p. e202400169.