MinKLIFSAI: simple machine learning approach toward selective kinase inhibitor

Mohamed Abdelalim[1]

[1]Independent Researcher: Ann Arbor, USA

Corresponding author: mohamed.abdelaleem97@gmail.com, moh.acad@protonmail.com

Abstract:

The aim of achieving selectivity in kinase inhibition is a big challenge within the realm of drug discovery, particularly due to the structural similarities between various kinases. Can machine learning be leveraged to overcome this hurdle? Utilizing different fingerprints may indeed lead to improved results. However, is there a single machine-learning approach that can effectively address selectivity across all kinases. In this study, the author collect kinase activity data from PubChem database (January 2023) using Uniprot IDs for each kinase. Each Uniprot ID is associated with its unique dataset, and duplicate points were removed to ensure accuracy. The data was then appended together, and any datasets containing fewer than 120 points were discarded. Each data point was categorized as either Active (1) or Inactive (0) based on the activity data. Two fingerprinting approaches were employed for predictions: MACCS fingerprints and Morgan2 (ECFP2) with a 2048-bit representation. The combined dataset was then divided into two subsets, one featuring imbalance data and another with balanced data. Random Forest and Artificial Neural Network models were applied to both datasets. To evaluate the performance of these models, various metrics were employed, including accuracy, sensitivity, specificity, and area under the curve (AUC). The results showed that Morgan fingerprinting performed slightly better than MACCS fingerprinting. A total of 480 target IDs was produced, with 452 unique IDs identified. On each dataset(balance and imbalance), two models were developed for both fingerprints, resulting in a combined total of 1920 predictions. Interestingly, the imbalance data yielded higher specificity compared to the balanced data. Each model has been deployed and made publicly available at (github.com/phalem/minKLIFSAI). However, the current data on all kinases is not yet sufficient to enable machine learning to reliably discover selective inhibitors.

Related work:

This work is a continuation of [1], with the ultimate goal of developing an app that can generate, predict, and explain selective inhibitors for any protein target of interest. The application of machine learning to kinases is not a novel concept; previous studies have explored this approach to create kinase inhibitor apps using different datasets [2-5]. Others have established valuable data sources as well [6]. For this work, the data was collected from PubChem in January 2023, utilizing Uniprot IDs provided in the supplementary section. The author provides both the app and the data as starting points for future research in the bioassay and docking studies, available at(https://github.com/phalem/KLIFSAI). The name KLIFSAI was inspired by KLIFS [7], a valuable resource for kinase-related research.

Introduction:

Kinases are among the most crucial proteins within a cell, involved in various processes such as cell survival, signaling, and proliferation. Alongside GPCRs, they represent one of the largest protein target families[8]. The development of selective kinase inhibitors is particularly challenging due to the structural similarities between different kinases. Manning et al. have categorized them into over 508 distinct groups [9]. One of the primary reasons for striving towards kinase selectivity is the risk of off-target effects, as exemplified by Imatinib's impact on opening up new avenues for discovering selective kinase inhibitors [9-10]. Molecular Fingerprints are a valuable source of information that captures the presence or absence of specific features within compounds. Examples include MACCS fingerprints, which involve the mathematical representation of functional groups [12], and Morgan fingerprints (also known as ECFP2), which involve creating an environment around atoms to produce either 1024-bit or 2048-bit representations [13]. To evaluate such models, several metrics have been developed. For instance, accuracy is a straightforward metric that represents the number of correct predictions made by a model. Other metrics commonly used in binary classification scenarios include specificity, sensitivity, and the Area Under the Curve (AUC) [14].

Method:

1.Data collection and preparation:

Data collected from PubChem[15] based on Uniprot ID of each Gen ID using PubChem GUI. Every ID has its own data separated. Data was appended together, with duplicate CID removed to make it easy to retrieve the SMILES string for each compound. SMILES was canonicalized using RDKit to remove invalid molecules as well as duplicated SMILES were removed. After that, each SMILES was aligned with each CID in the original data. According to the Activity label of PubChem data, active compounds were assigned a value of 1 and inactive compounds were assigned a value of 0 for further model classification tasks. Any target with less than 20 active points and 100 inactive points was removed. Data was separated into two groups:

(1) Imbalanced data contained different ratios of active and inactive data without applying any enrichment values, with some data reduced to 150K points for both.

(2) Balanced data, in which the number of inactive compounds was less than or equal to the number of active compounds.

For each model, the number of compounds used are stored in a dictionary in the model directory in the supplementary files. For each target ID, molecular descriptors were calculated using MACCS Fingerprint and Morgan2 or ECFP2 fingerprint based on 2048 bits using RDKit. Data was split into 80% for the training set and 20% for the test set, respectively, using scikit-learn[16] library's model_selection module and train_test_split method.

2. Model:

In order to reproduce the results, the author used code inspired by TeachopenCADD [17]. The code and notebook are provided in the github for reproducibility. For each Uniprot ID, Random Forest (RF)[18] and Artificial Neural Network(ANN) were applied to both imbalanced and balanced data using scikit-learn[16] library. The number of estimators was set equal to 10, and the criterion equal to entropy. For ANN, the hidden layer size was set equal to (30,3). In the case of balanced data, hyperparameter optimization (named as balanced_opt in the supplementary materials) was performed for each number of estimator from 1 to 20, and hidden layer sizes were considered to be one of the following: (5,3), (10,3), or (30,3).

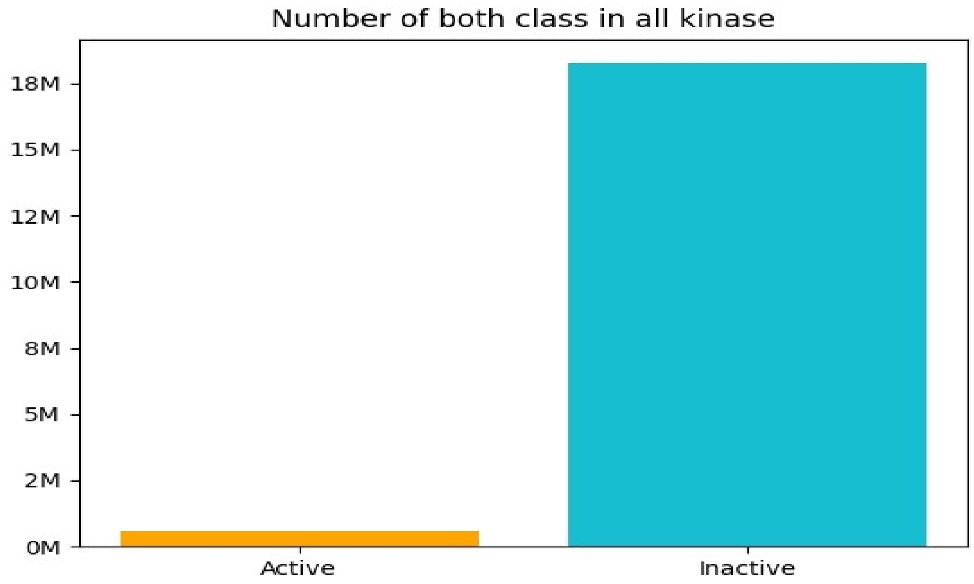
3.Evaluation:

Accuracy, sensitivity, specificity, and area under the curve (AUC) were calculated using the accuracy_score, recall_score, and roc_auc_score functions from scikit-learn. ROC curves were plotted using matplotlib [19] for each model based on the test set predictions.

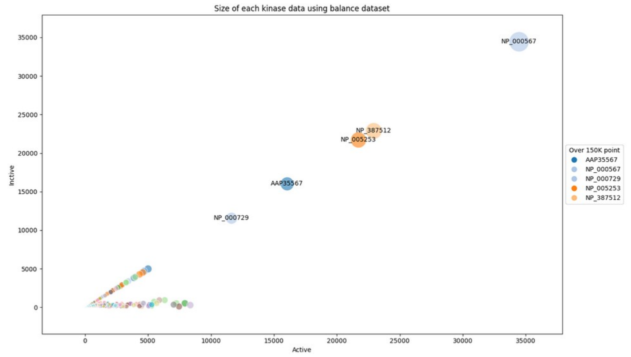
4.Deployed: Streamlit (https://streamlit.io) was used for real-time prediction, both offline and online, at (github.com/phalem/minKLIFSAI). An example of a single data with code provided to reproduce the results of each model.

Result:

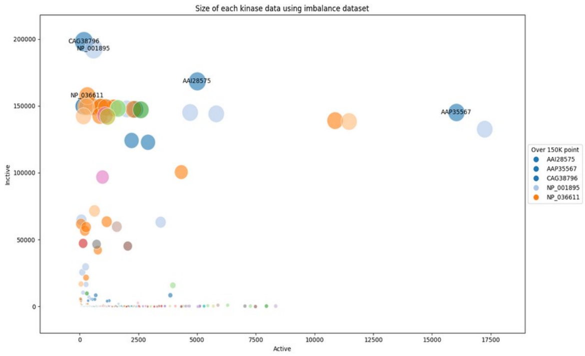
Data: With a total of 18.8 million data points, comprising 18 million inactive and 0.5 million active points, each has a target, after removing duplicated SMILES for all data the data become 1.2 million unique compounds, consisting of 300k active and 900K inactive compounds. Figure (1a) shows the ratio of both active and inactive data in the figure. Figures (b,c) represent each target ID, with the size of the representation corresponding to the amount of data available for that target ID. The obtained data are small compared to typical datasets, but were reduced to 480 target IDs, which correspond to 452 unique Uniprot IDs.



A)Number of Active and Inactive compounds for all targets.

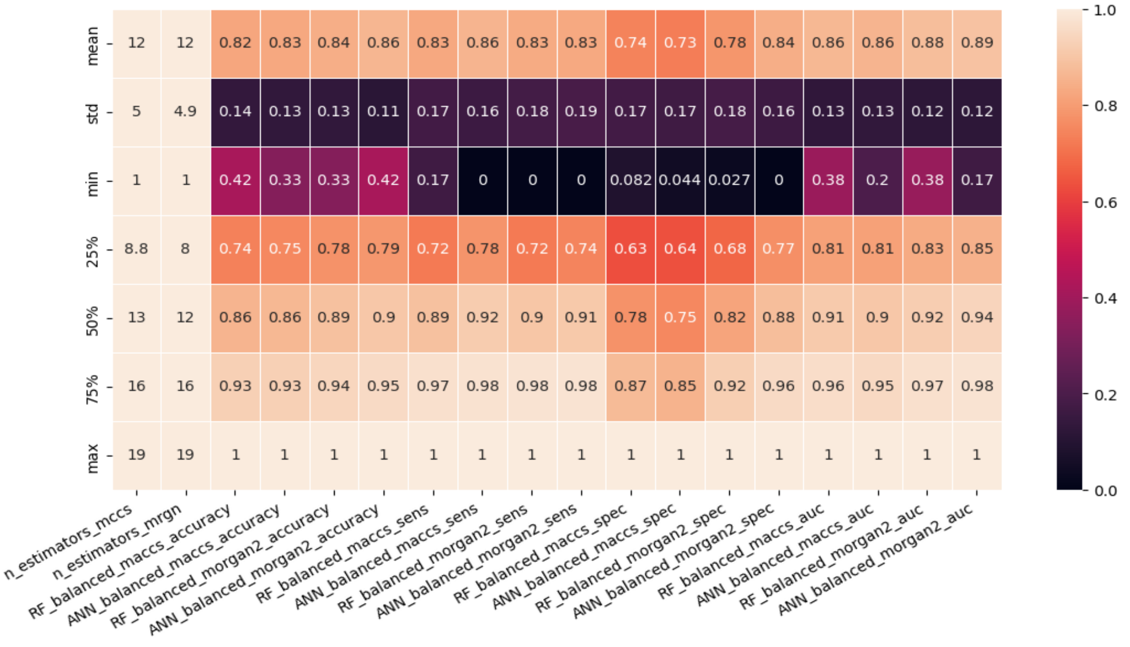


b)Number of data points for balance data

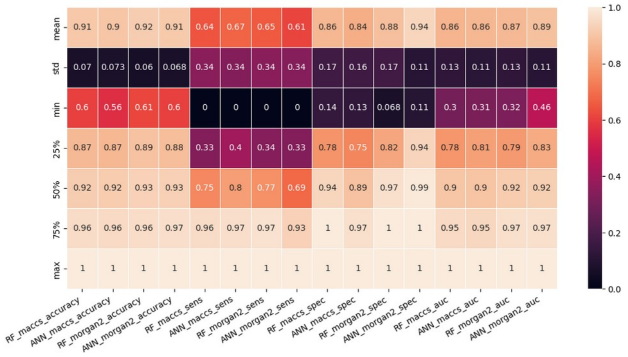


c)Number of data points for imbalance data

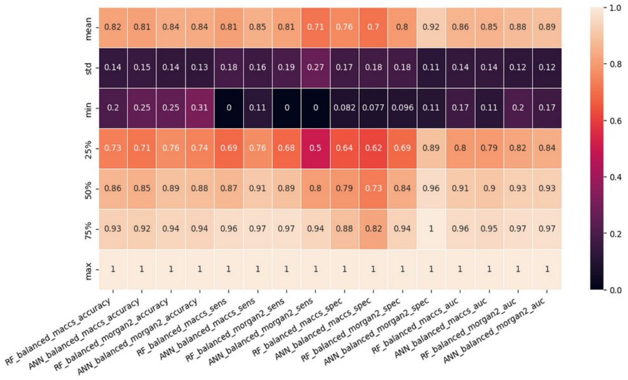
Models: For each of the 480 target IDs, there are four models trained: two Random Forest models and two Artificial Neural Network (ANN) models, based on both MACCS and Morgan Fingerprints, respectively. Each target ID has a target analysis dictionary containing information about the target data, including the number of data points, ROC curves, and other parameters and evaluation metrics for both MACCS and Morgan2 models. This information can be found in the supplementary materials. Figure(2) is a summary that describes the analysis of each model for all 480 targets. Further details for each target ID and data type can also be found in the supplementary materials.



a)Model evaluation on the balance dataset with hyperparameter optimization



b)Model evaluation on the imbalanced data



c)Model evaluation on the balanced data.

Models trained on imbalanced data tend to have high specificity, but lower accuracy and sensitivity compared to those trained on balanced data. Models based on Morgan2 perform slightly better than MACCS on both imbalanced and balanced datasets. Hyperparameter optimization of balanced data led to a slight increase in all evaluation metrics, particularly when dealing with limited data. Most Random Forest models performed well on the balanced dataset using 16 estimators as the optimal number. Even with high evaluation metrics, the small size of the dataset make it difficult to determine whether the model would perform well in real-time scenarios. Further examples can be found in the supplementary materials for additional research and use. The Streamlit app provides the capability to utilize the model in both online and offline modes, generating target predictions as well as scripts that aid in identifying active targets against certain molecules, which can be easily downloaded. The app and models are provided with the hope of reducing the cost of cancer treatments by discovering selective kinase inhibitors or making it available for free.

Limitation:

Although these models are not among the most accurate ones published, they can still be useful tools. The models have been evaluated, but not thoroughly tested using either structure-based or ligand-based approaches. Unfortunately, some targets have very limited data available, which makes it challenging to predict active compounds or selective inhibitors using this model alone. Since the model has not been experimentally validated, further experimental validation is necessary to confirm its effectiveness.

Conclusion:

In this paper, the author demonstrate the feasibility of developing a machine learning model for each kinase. By providing a user-friendly interface app that is ready to use, with the aim to make the process much easier for researchers to discover and utilize these models. The ability to predict active compounds for 480 targets represents a significant first step towards realizing KLIFSAI's dream of generating active compounds, predicting bioactivity, and performing kinase docking prediction. However, further activity prediction, optimization, and experimental validation will be necessary to advance this process. The way may be long, but small steps can make it farther possible, and the dreams of yesterday will become the realities of tomorrow.

Abbreviations:

GPCR: G protein-coupled receptors

MACCS key: Molecular ACCess System keys

ECFP: Extended connectivity fingerprints

SMILES: Simplified Molecular Input Line Entry System

RF: Random forest

ANN : Artificial Neural Network

AUC: Area Under the Curve

ROC: Receiver-operating characteristic curve

GUI: Graphical User Interface

Reference:

[1] "Src kinase app: valid inhibitor generation and prediction with explanation using predictive model and selfies | Biological and Medicinal Chemistry | ChemRxiv | Cambridge Open Engage.". Available: https://chemrxiv.org/engage/chemrxiv/article-details/62d9b695fe12e3a114adc997

[2] G. De Simone, D. S. Sardina, M. R. Gulotta, and U. Perricone, "KUALA: a machine learning-driven framework for kinase inhibitors repositioning," *Sci. Rep.*, vol. 12, no. 1, p. 17877, Oct. 2022, doi: 10.1038/s41598-022-22324-8.

[3] J. Wu *et al.*, "Large-scale comparison of machine learning methods for profiling prediction of kinase inhibitors," *J. Cheminformatics*, vol. 16, no. 1, p. 13, Jan. 2024, doi: 10.1186/s13321-023-00799-5.

[4] Z.-Y. Yang, Z.-F. Ye, Y.-J. Xiao, C.-Y. Hsieh, and S.-Y. Zhang, "SPLDExtraTrees: robust machine learning approach for predicting kinase inhibitor resistance," *Brief. Bioinform.*, vol. 23, no. 3, p. bbac050, May 2022, doi: 10.1093/bib/bbac050.

[5] "Prediction of kinase inhibitors binding modes with machine learning and reduced descriptor sets – PMC." Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7804204/

[6] O. J. M. Béquignon, B. J. Bongers, W. Jespers, A. P. IJzerman, B. van der Water, and G. J. P. van Westen, "Papyrus: a large-scale curated dataset aimed at bioactivity predictions," *J. Cheminformatics*, vol. 15, no. 1, p. 3, Jan. 2023, doi: 10.1186/s13321-022-00672-x.

[7] "KLIFS: an overhaul after the first 5 years of supporting kinase research | Nucleic Acids Research | Oxford Academic." Available: https://academic.oup.com/nar/article/49/D1/D562/5934416?login=false

[8] P. Cohen, "Protein kinases — the major drug targets of the twenty-first century?," *Nat. Rev. Drug Discov.*, vol. 1, no. 4, pp. 309–315, Apr. 2002, doi: 10.1038/nrd773.

[9] "The protein kinase complement of the human genome – PubMed." Available: https://pubmed.ncbi.nlm.nih.gov/12471243/

[10] P. Cohen, D. Cross, and P. A. Jänne, "Kinase drug discovery 20 years after imatinib: progress and future directions," *Nat. Rev. Drug Discov.*, vol. 20, no. 7, pp. 551–569, Jul. 2021, doi: 10.1038/s41573-021-00195-4.

[11] L. Xue and J. Bajorath, “Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening,” *Comb. Chem. High Throughput Screen.*, vol. 3, no. 5, pp. 363–372, Oct. 2000, doi: 10.2174/1386207003331454.

[12] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, “Reoptimization of MDL keys for use in drug discovery,” *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 6, pp. 1273–1280, Dec. 2002, doi: 10.1021/ci010132r.

[13] “Extended-Connectivity Fingerprints | Journal of Chemical Information and Modeling.”Available: https://pubs.acs.org/doi/10.1021/ci100050t

[14] “What’s under the ROC? An Introduction to Receiver Operating Characteristics Curves.” Available: https://journals.sagepub.com/doi/epdf/10.1177/070674370705200210

[15] S. Kim *et al.*, “PubChem Substance and Compound databases,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1202-1213, Jan. 2016, doi: 10.1093/nar/gkv951.

[16] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” Jun. 05, 2018, *arXiv*: arXiv:1201.0490. doi: 10.48550/arXiv.1201.0490.

[17] D. Sydow, A. Morger, M. Driller, and A. Volkamer, “TeachOpenCADD: a teaching platform for computer-aided drug design using open source packages and data,” *J. Cheminformatics*, vol. 11, no. 1, p. 29, Apr. 2019, doi: 10.1186/s13321-019-0351-x.

[18] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[19] “Matplotlib: A 2D Graphics Environment | IEEE Journals & Magazine | IEEE Xplore.” Available: https://ieeexplore.ieee.org/document/4160265