Research – Results Article | J Vis Med. 2024;5:e0104

Category: Artificial Intelligence

Peer Reviewed | Open Access

10.30699/jovm.005.e0104

Journal of Visualized Medicine | ISSN:3041-9727

Machine Learning-Assisted Prediction of Biological Activities of Chemical Compounds on NMDA Receptors

Tahereh Mostashari-Rad[1*], Mohammadreza Momenzadeh[1]

1. *Department of Artificial Intelligence, Smart University of Medical Sciences, Tehran, Iran*

ABSTARCT

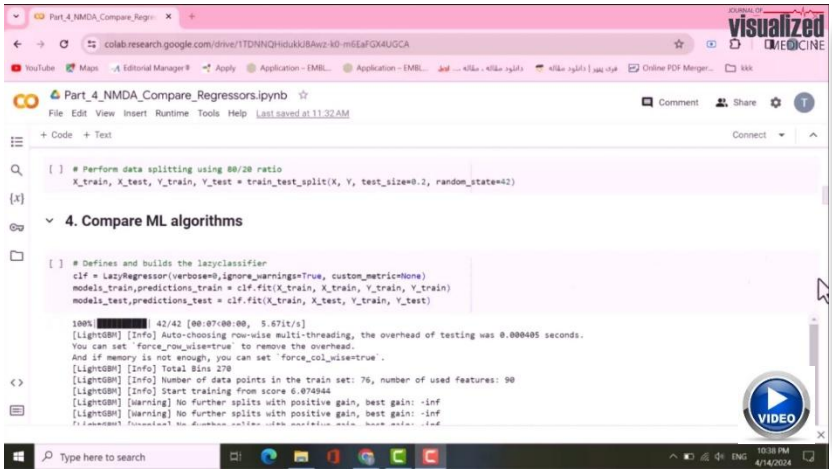
N-methyl-D-aspartate receptors (NMDARs) have strong effects on fast excitatory synaptic transmission in the brain and are essential for brain development, memory, and learning functions. NMDARs dysfunction cause not only complex neurological diseases and disorders (such as Alzheimer's disease, Autism Spectrum Disorder (ASD), depression, stroke, epilepsy, and schizophrenia) but also different cancers. Thus, introducing new potential compounds affecting NMDARs can facilitate the treatment of different diseases and disorders. In this study, different machine learning models were trained to predict the biological activities of chemical compounds on NMDARs. Application of these models provides the possibility of predicting the biological activities of new molecules in a short period of time and reducing the total costs of the drug discovery procedure.

Keywords: Artificial Intelligence, Machine Learning, Quantitative Structure-Activity Relationships, N-methyl-D-aspartate Receptors

Received: 02 Februray 2024; **Accepted**: 04 April 2024; **Published Online:** 05 May 2024;

Corresponding Information: **Tahereh Mostashari-Rad**, *Department of Artificial Intelligence, Smart University of Medical Sciences, Tehran, Iran* & Email: t.mostashari@gmail.com

Use DOI and watch the video article online



Highlights:

N-methyl-D-aspartate receptors (NMDARs) have strong effects on fast excitatory synaptic transmission in the brain and are essential for brain development, memory, and learning functions. NMDARs dysfunction cause not only complex neurological diseases and disorders (such as Alzheimer's disease, Autism Spectrum Disorder (ASD), depression, stroke, epilepsy, and schizophrenia) but also different cancers.

Use a device to scan and read the article online

Mostashari-Rad T, Momenzadeh M. Machine Learning-Assisted Prediction of Biological Activities of Chemical Compounds on NMDA Receptors. J Vis Med. 2024;5;e0104.

Download Citation: RIS | EndNote | Mendeley | BibTeX |

1. Introduction

Fast excitatory synaptic transmission in the brain has a complicated mechanism and can be facilitated by the function of ionotropic glutamate receptors (iGluRs) including a-amino-3-hydroxy-5-methyl-4- isoxazolepropionic acid (AMPA) receptors, kainate receptors, and N-methyl-D-aspartate receptors (NMDARs). The activity of these receptors is necessary for brain development, memory, and learning functions. NMDARs dysfunction cause complex neurological diseases and disorders including Alzheimer's disease, Autism Spectrum Disorder (ASD), depression, stroke, epilepsy, and schizophrenia **(1, 2)**. Many studies have shown that NMDARs are overexpressed in different kinds of cancers, such as neuroendocrine/ductal pancreatic cancers and breast cancers to mediate signaling for invasive tumor growth and brain metastasis, respectively **(3)**.

Therefore, there are pressing needs to introduce new potential therapeutic compounds affecting NMDARs to facilitate the treatment of not only neurological diseases and disorders, but also cancers.

The most important challenges and hurdles in drug design and development procedures are time consumption and high costs. Processing and analyzing the large amounts of data in genomics, proteomics, microarray data, and clinical trials are other obstacles in the process of drug design and development **(4)**. Artificial intelligence (AI) fields such as machine learning have paved the way to tackle these challenges and have been implemented in different drug discovery procedures. The greatest pharmaceutical companies around the world use AI and computer-aided methods in different stages of pharmaceutical industry, such as drug discovery, clinical trials, and manufacturing **(5-8)**. Using machine learning models to quantitatively predict the biological activity of chemical compounds on drug targets has created a great opportunity for rational drug design and discovery processes **(9)**.

In this study, different machine learning algorithms will be used to build predictive models for the biological activities of chemical compounds on Ionotropic Glutamate NMDA Receptors. The predictive quality of these models will be evaluated and compared with each other to distinguish the best model in this regard.

2. Protocol

2.1 Data Collection

Python programming language has been used to implement the entire process of this research. The ChEMBL web service package **(10)** was installed to retrieve bioactivity data from the ChEMBL Database. Then, the necessary libraries such as pandas were imported. After that, the target protein, which is the Ionotropic Glutamate NMDA Receptor, was searched and a "Target DataFrame" was created. The target in the form of a single protein belonging to Homo sapiens was selected for further investigations. A variable called "activity" was defined, and by applying two filters according to the target ChEMBL ID, the IC_{50}s were downloaded. The "Standard Type" column defines the IC_{50}s and the "Standard Value" column defines the potency of the compounds. Another DataFrame was written to get a CSV file including the bioactivity data. Finally, the data missing the IC_{50} values were eliminated.

2.2 Data Pre-processing

For data pre-processing of the bioactivity data, and for the benefit of creating machine learning models the compounds were categorized into three groups as "active", "inactive" and "intermediate" compounds. Compounds having values of less than 1000 nM were considered to be "active", while those greater than 10,000 nM were considered to be "inactive". The values between 1,000 and 10,000 nM were referred to as "intermediate". The condition codes were used to label the compounds as described. The rows containing the same molecule ChEMBL ID were removed to avoid redundancy **(11)**. Finally, a CSV file was created containing the molecule_chembl_id, canonical_smiles, standard_value, and bioactivity_class.

2.3 Exploratory Data Analysis (EDA)

In this step, RDkit and Mordred were installed. RDkit allows computing the molecular descriptors for the compounds in the dataset that have been compiled from previous steps. The Lipinski descriptors were calculated. The SMILES notations of the compounds were as inputs for this calculation. The SMILES notations contain the chemical information with exact atomic details of the molecule. The CSV file containing Lipinski descriptors was combined with the dataframe from the curated data in the previous part by using "pd.concat" function, to have the standard values and the bioactivity class columns together.

The standard values (IC_{50}) were converted to the pIC_{50} (negative logarithmic transformation) scale. Because the original IC_{50} values have an uneven distribution of the data points, thus to make a more even distribution, a negative logarithmic

transformation should be applied. After this step, the intermediate biological activity class was removed.

Exploratory data analysis (Chemical space Analysis) via Lipinski Descriptors allows us to investigate the chemical space. The seaborn and matplotlib libraries were imported to create the plots, such as the frequency plot of the two biological activity classes, a scatter plot of Molecular Weight (MW) versus LogP, and a pIC_{50} values distribution plot to see the distribution of the "active" and "inactive" compounds.

After that, the Mann-Whitney test was used to look at the difference between the two biological classes. It compares the "active" and the "inactive" classes to see whether there is a statistical significance for the pIC_{50} variable. All the files from the Mann-Whitney were saved as CSV and the box plots were saved as PDF files.

2.4 Calculating Additional Descriptors

The PADEL-descriptor software was downloaded to calculate the molecular descriptors. Pandas library was downloaded. In the next step, the calculation of molecular fingerprints (Pubchem fingerprints) was performed by running bash padel.sh. This program automatically cleans the chemical structures by removing salts and small organic acids from the chemical structures. After the calculation of the descriptors, the output file was prepared in the form of a CSV file. After that, the X and Y data matrices were created. The X data matrix is comprised of molecular descriptors, which are the Pubchem fingerprints. The Y data matrix contains the pIC_{50} values. Finally, the X and Y matrixes were combined.

2.5 Model Building

In order to make different machine learning models, the lazypredict library was installed. The other necessary libraries, such as pandas, seaborn, and also sklearn were imported. The dataset was downloaded from the previous steps. The data was split into X and Y variables. For data pre-processing, the low variance features were removed. The data were split in an 80/20 train and test ratio **(12)**. After that, about 30 machine learning models were built. The machine learning algorithm was assigned into a classifier variable. The results from the predictions were assigned to the train and test variables. In this way of model building the default parameters for all the machine learning algorithms were used. The model performances were visualized using bar plots of the coefficient of determination (R-squared or R^2), Mean Squared Error (RMSE) values, and the calculation time.

3. Results and Discussion

In this study, the Ionotropic Glutamate NMDA Receptor was selected as the biological target. NMDAR dysfunction can cause neurological diseases and disorders and also some kinds of cancers. The ChEMBL database was selected to retrieve the data related to the biological activities of the compounds on NMDA Receptor. ChEMBL is known as a reliable source of data in the fields of drug discovery and medicinal chemistry research. It provides different categories of data, such as standardized bioactivity, target, molecule, and drug data extracted from multiple sources. The biological activities related to the Ionotropic Glutamate NMDA receptors in the form of a single protein belonging to Homo Sapiens were downloaded. After pre-processing, the data was used to construct ML models, which are technically known as Quantitative Structure-Activity Relationship (QSAR). The development of such QSAR models holds great values for drug discovery projects. Particularly, it allows scientists to understand the origins of the biological activity. The interpretation of the model provides the opportunity to design a better drug. The biological activities were defined as IC_{50} values, which shows the potency of the compounds. The lower the number of IC_{50}s, the better the potency of the drug becomes. Thus, ideally, the compounds with lower IC_{50} values mean that the inhibitory concentration at fifty percent will have a low concentration. Labeling the compounds as "active", "inactive" and "intermediate" was performed according to IC_{50} values. Compounds having values of less than 1000 nM were considered to be "active", while those greater than 10,000 nM were considered to be "inactive". As for those values in between 1,000 and 10,000 nM were referred to as intermediate **(13)**.

The calculation of Lipinski's descriptors (MW, logP, No. H-bond donors and No. H-bond acceptors) provides the possibility to look at the chemical space. The plots, such as the frequency plot of the two biological activity classes, scatter plot of MW versus LogP and pIC_{50} values were created (Figure 1). The plots visualizing the biological activity classes show the distribution of the "active" and "inactive" classes. Because of the fact that, the threshold had been defined for the "active" and "inactive" classes (IC_{50} < 10000 nM = actives while IC_{50} > 10000 nM = inactives), it had been expected to see the difference in IC_{50} distribution of the two classes. The pIC_{50} distribution of the active compounds is vaster than inactive compounds. According to the distribution plots in Figure 2, the active and inactive compounds have completely different LogP values, but there is a small overlap between the values of molecular weights in active and inactive compounds.

The Mann-Whitney test was applied to look at the difference between the two biological classes. It compares the active class and the inactive class to see whether there is a statistical significance for the pIC_{50} variable. Taking a look at pIC_{50} values, the actives and inactives displayed statistically significant differences. Lipinski's descriptors: (MW, LogP, NumHDonors, and NumHAcceptors) show statistically significant differences between active and inactive compounds.

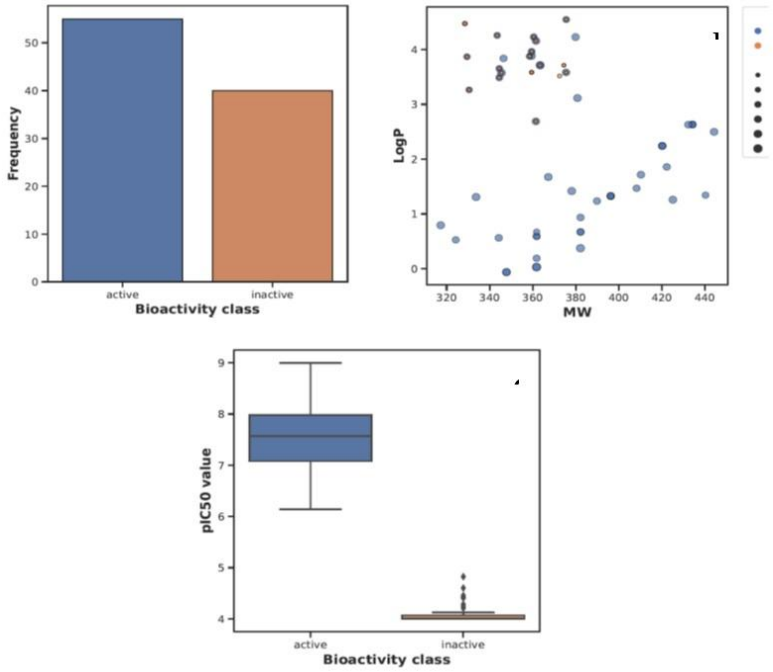


Figure 1. (A) The frequency plot of the two biological activity classes, (B) scatter plot of MW versus LogP and (C) pIC_{50} values distribution for the active and inactive compounds **(Design by Authors, 2024)**.

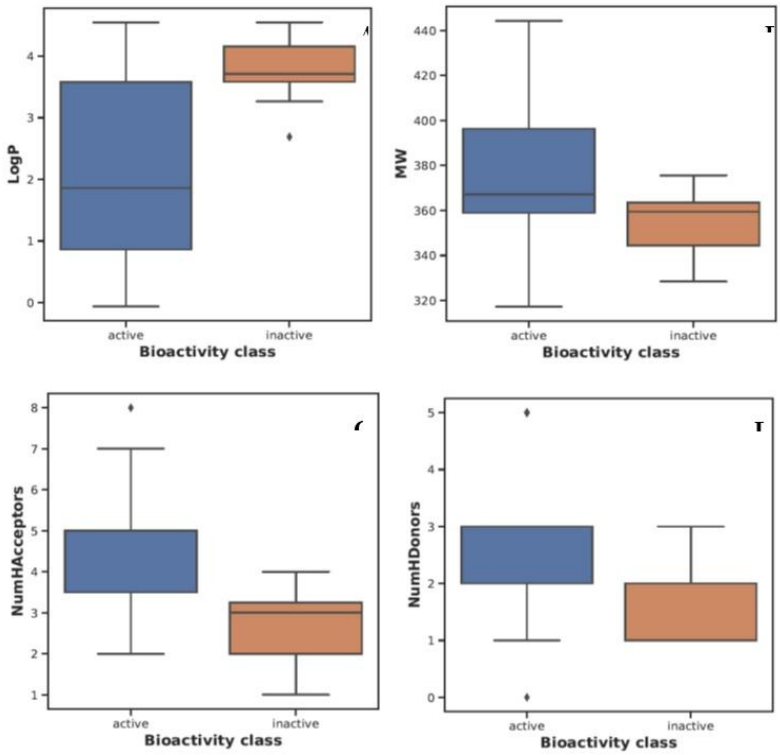


Figure 2. (A) LogP values distribution (B) MW values distribution (C) Number of hydrogen bond acceptors distribution (D) Number of hydrogen bond donors distribution for the active and inactive compounds **(Design by Authors, 2024)**.

The PADEL-descriptor software was applied to calculate PubChem fingerprints as molecular descriptors. The difference between the Lipinski's descriptors and the molecular fingerprints is that the Lipinski's descriptors will provide us with a set of simple molecular descriptors that essentially will give us a quick overview of the drug-like properties of the molecule. The PubChem fingerprints describe the local features of the molecule, but the Lipinski rule of five describes the global features of the molecule particularly the molecular size, the solubility, and the number of hydrogen bond donors and acceptors. The local features for the PubChem fingerprints will describe each molecule by the unique building blocks of the molecule. The way in which the building blocks are connected will create unique properties for the drug and that is the essence of drug discovery and drug design. Therefore, we have to find a way to rearrange the building blocks in such a way that the molecule provides the most potency toward the target protein, considering safety and toxicity to avoid side effects.

For model building, the lazypredict library was used. The data was comprised of 95 compounds with 881 descriptors. For data pre-processing the low variance features were removed. The number of features reduced from 881 to 121 descriptors. The data were split in an 80/20 train and test ratio. After that, about 30 machine learning models were used to predict the biological activities. Finally, R^2 and RMSE parameters were calculated for the evaluation of the models and selection of the best method that best fits with the present data.

R^2 is generally calculated for the evaluation of the regression machine learning models. For non-linear regression models, greater R^2 values mean that the model is fitted the data better and provides more accurate predictions. According to the R^2 values that are presented in Figure 3, Decision Tree Regressor, Extra Tree Regressor, Gaussian Process Regressor, are the best models for this dataset. RMSE is another way to evaluate the model performance. It is a measure of the average difference between the predicted values and the actual values. The lower the RMSE value, the better the performance of the model is for the prediction. Therefore, the best models in this study based on RMSE values are treebased regression models. As it is presented in Figure 3 and Figure 4 the results from the RMSE and the R^2 values completely match with each other.

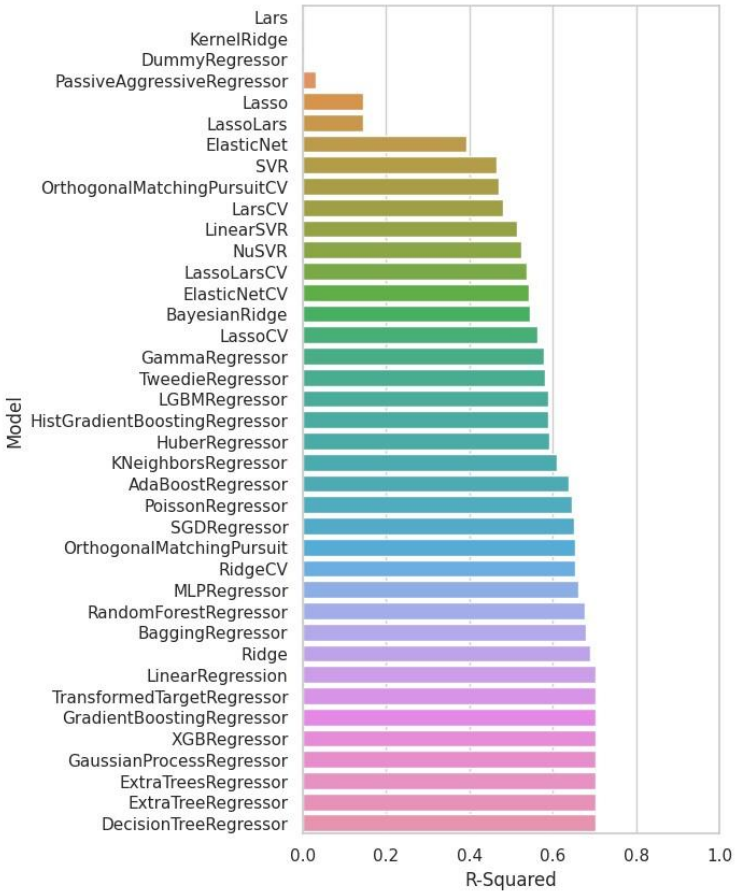


Figure 3. Bar plot of R-squared (coefficient of determination) values for the evaluation of the machine learning models **(Design by Authors, 2024)**.

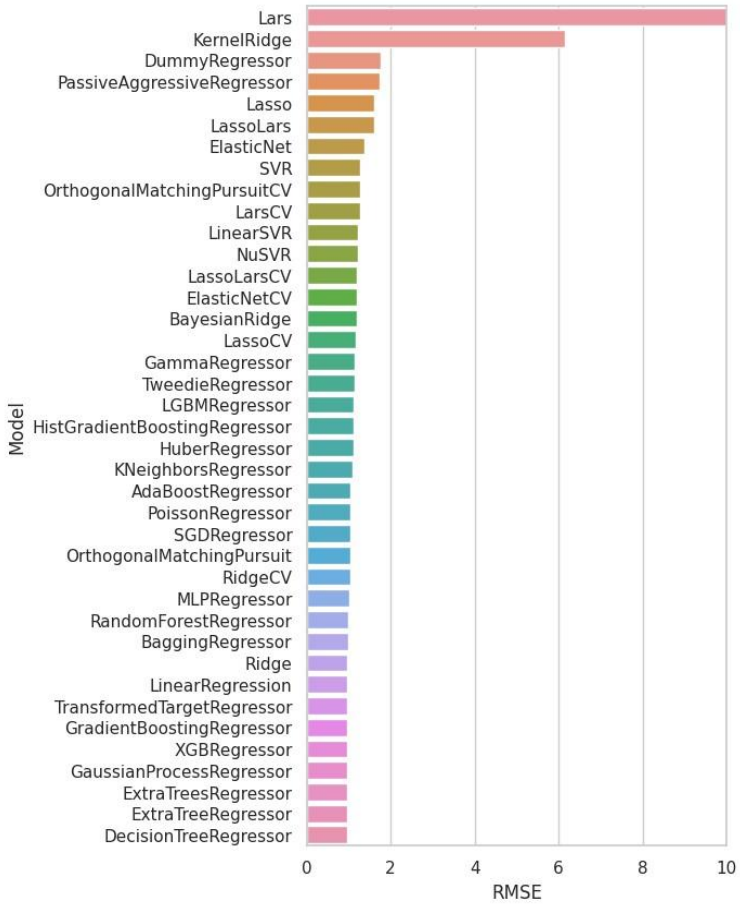


Figure 4. Bar plot of RMSE (Root Mean Squared Error) values for the evaluation of the machine learning **(Design by Authors, 2024)**.

4. Conclusion

In the present study, different machine learning models were trained for the prediction of biological activities of chemical compounds on NMDARs. Data was collected from the ChEMBL Database. The Lipinski descriptors and PubChem fingerprints descriptors were calculated as independent features for model training. The R^2 and RMSE values were calculated for the evaluation of the model performance. Consequently, the experimental results demonstrated that tree-based regression models have better performance than other methods. Moreover, application of these models provides the possibility of predicting the biological activities of new molecules in a short time and reducing the total costs of the drug discovery procedure.

5. Declarations

Acknowledgment

The authors of the present study would like to thank Smart University of Medical Sciences and the technical team of JoVm for their support.

Ethical Considerations

Not applicable.

Conflict of Interest

The authors declare no conflict of interest, financial or otherwise.

Financial Support and Sponsorship

This research received no specific grant from any funding agency in the public.

References

1. Chou TH, Tajima N, Romero-Hernandez A, Furukawa H. Structural Basis of Functional Transitions in Mammalian NMDA Receptors. Cell. 2020;182(2):357-71. [PMID] [PMCID] [DOI:10.1016/j.cell.2020.05.052]
2. Lee EJ, Choi SY, Kim E. NMDA receptor dysfunction in autism spectrum disorders. Curr Opin Pharmacol. 2015;20:8-13. [DOI:10.1016/j.coph.2014.10.007] [PMID]
3. Li L, Hanahan D. Hijacking the neuronal NMDAR signaling circuit to promote tumor growth and invasion. Cell. 2013;153(1):86-100. [DOI:10.1016/j.cell.2013.02.051] [PMID]
4. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers. 2021;25:1315-60. [PMID] [DOI:10.1007/s11030-021-10217-3] [PMCID]
5. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev. 2022;55(3):1947-99. [DOI:10.1007/s10462-021-10058-4] [PMID] [PMCID]
6. Tamiz N, Mostashari-Rad T, Najafipour A, Claes S, Schols D, Fassihi A. Synthesis, molecular docking and molecular dynamics simulation of 2-thioxothiazolidin-4-one derivatives against Gp41. Curr HIV Res. 2021;19(1):47-60. [PMID] [DOI:10.2174/1570162X18666200903172127]
7. Ghorayshian A, Danesh M, Mostashari-Rad T, Fassihi A. Discovery of novel RARα agonists using pharmacophore-based virtual screening, molecular docking, and molecular dynamics simulation studies. Plos One. 2023;18(8): e0289046. [DOI:10.1371/journal.pone.0289046] [PMID] [PMCID]
8. Mostashari-Rad T, Claes S, Schols D, Shirvani P, Fassihi A. Novel 2-alkylthio-1-benzylimidazole-5-carboxylic Acid Derivatives Targeting Gp41: Design, Synthesis, and In Vitro Anti-HIV Activity Evaluation. Curr HIV Res. 2022;20(5):380-96. [DOI:10.2174/1570162X20666220628154901] [PMID]
9. Staszak M, Staszak K, Wieszczycka K, Bajek A, Roszkowski K, Tylkowski B. Machine learning in drug design: Use of artificial intelligence to explore the chemical structure-biological activity relationship. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(2):e1568. [DOI:10.1002/wcms.1568]
10. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, et al. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res. 2015; 43(W1):W612-20. [DOI:10.1093/nar/gkv352] [PMID] [PMCID]
11. Momenzadeh M, Vard A, Talebi A, Mehri Dehnavi A, Rabbani H. Computer-aided diagnosis software for vulvovaginal candidiasis detection from Pap smear images. Microsc Res Tech. 2018; 81(1):13-21. [DOI:10.1002/jemt.22951] [PMID]
12. Vatankhah M, Momenzadeh M. Self-regularized Lasso for selection of most informative features in microarray cancer classification. Multimed Tools Appl. 2024;83(2):5955-70. [DOI:10.1007/s11042-023-15207-1]
13. Yu T, Nantasenamat C, Kachenton S, Anuwongcharoen N, Piacham T. Cheminformatic analysis and machine learning modeling to investigate androgen receptor antagonists to combat prostate cancer. ACS Omega. 2023;8(7): 6729-42. [DOI:10.1021/acsomega.2c07346] [PMID] [PMCID]