Leveraging 13C NMR spectrum data derived from SMILES for machine learning-based prediction of a small biomolecule functionality: a case study on human Dopamine D1 receptor antagonists

Leveraging 13C NMR spectrum data derived from SMILES for machine learning-based prediction of a small biomolecule functionality: a case study on human Dopamine D1 receptor antagonists

Mariya L. Ivanova[1,*, ORCID], Nicola Russo[1, ORCID], Gueorgui Mihaylov[2, ORCID], Konstantin Nikolic[1, ORCID]
Author affiliations
[1]School of Computing and Engineering, University of West London, London, UK
*Corresponding author mariya.ivanova@uwl.ac.uk

Abstract

This study contributes to ongoing research which aims to predict small biomolecule functionality using Carbon-13 Nuclear Magnetic Resonance (^{13}C NMR) spectrum data and machine learning. A bioassay on human dopamine D1 receptor antagonists was used to demonstrate the approach. The Simplified Molecular Input Line Entry System (SMILES) notations of the compounds were extracted and converted into spectroscopic data using purpose-built software. This data was then used for machine learning with scikit-learn algorithms. The ML models were trained with 27,756 samples and tested with 5,466. Of the estimators tested (K-Nearest neighbor, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, XGBoost Classifier, and Support Vector Classifier), the Support Vector Classifier was found to be the most effective, achieving 71.5% accuracy and a cross-validation score of 0.749. The methodology can be applied to predict the functionality of any compound, provided relevant data are available. It was also hypothesized that an increase in sample numbers would lead to increased accuracy. Additionally, a time- and cost-efficient CID_SID ML model was also developed, allowing compounds to be checked for D1 receptor antagonist activity using only their PubChem identifiers. This model's metrics were 80.2% accuracy and a five-fold cross-validation score of 0.8071.

Keywords: scikit-learn, drug development, drug discovery, CID-SID ML model, neurotransmitter

Introduction

The study is based on data obtained through quantitative high throughput screening (qHTS) on human dopamine receptor D1 (D1 DAR) antagonists (1). The information regarding this screening was provided by PubChem (2), the world's biggest freely available database for chemical data. The mentioned screening, registered as PubChem AID 504652 bioassay, contained 359,035 rows of samples and 10 columns with the features of these samples, such as PubChem identifiers of compounds (CIDs) or substances (SIDs), SMILES notations and the outcomes of the screening. The bioassay aimed to select and develop allosteric modulators of D1 DAR for in vivo and in vitro applications. It was performed by measuring the compound D1 DAR antagonism with an EC80 addition of dopamine by tracking calcium flux in a force-coupled inducible Hek293 Trex D1 cell line. The compound concentration utilized in this bioassay was 10.0microM. For more information regarding the screening protocol, please refer to the bioassay`s documentation (1).

Since the ratio of active and inactive antagonists in the PubChem AID 504652 bioassay (1) was severely imbalanced, another bioassay, PubChem AID 1996 for small molecule water solubility (3), was used as a filter to reduce the number of inactive antagonists. This water solubility bioassay was also provided by PubChem and had 57,859 rows of samples and 30 columns with their characteristics.

For the purpose of the study, 13C NMR spectroscopy was considered. This spectroscopy utilises the magnetic property of the 13C isotope (4), measuring the atomic nuclei's absorption and emission of radiofrequency radiation. The 13C nucleus contains six protons and seven neutrons. The extra ½ spin makes the 13C isotope behave as a little magnet. As a result, NMR spectroscopy is sensitive to differences in the chemical structure of the organic molecule, giving information for the connectivity and relative position of the carbon atoms in the molecule. So, given the Carbon skeletons of the organic molecule (5) and the technology of 13C NMR spectroscopy, it was hypothesised that 13C NMR spectroscopic data would contribute to the development of robust, reliable, cost- and time-effective machine learning (ML) model that would predict the functionalities of small biomolecules.

To test the hypothesis, the SMILES notations from the PubChem AID 504652 bioassay were extracted and converted into ^{13}C NMR spectroscopic data by the NMRDB software (6), which data was subsequently used for ML.

Various studies have been conducted exploring the combination of ML and 13C NMR spectroscopy, such as the study of correlations between peak patterns and chemical structure using machine learning (7); a computational approach to establish correspondence between molecular graphs and NMR spectra (8); prediction of 13C NMR chemical shifts with a message passing neural network (MPNN) (9); detection of chemical shifts of benzene compounds (10); software predicting 13C spectroscopic data by uploading SMILES, international chemical identifier (inChl) or drawing the structure of the compound (11); prediction of NMR spectra with the COLMARppm software using NMR data obtained in aqueous solution (12); use of an artificial neural network (ANN) to generate new structures based on information obtained from 13C NMR (13).

ML has been increasingly integrating into the search for innovative, robust, and reliable approaches that can significantly reduce the time and cost of laboratory experiments and, consequently, the time and cost of the research itself. Some examples of such approaches are AlphaFold3 for modelling the structure of proteins and their interactions (14); deep learning for structure-based drug design (15); generation of ML attributes based on atomic characteristics (16); machine learning approaches for pharmaceutical decision-making processes (17). On the other hand, ML applications regarding the dopamine D1 receptor antagonists are classification conducted using topological descriptors (18); effective identification of novel agonists of dopamine receptors has been obtained using predictive or generative geometric ML (19); agonists and antagonists have been predicted, using topological fragment spectra (TFS)-based support vector machine (SVM) (20). However, the approach presented in the current article has not been reported in the available literature.

The methodology applied in the study is explained in the section Methodology below and illustrated in Figure 1.

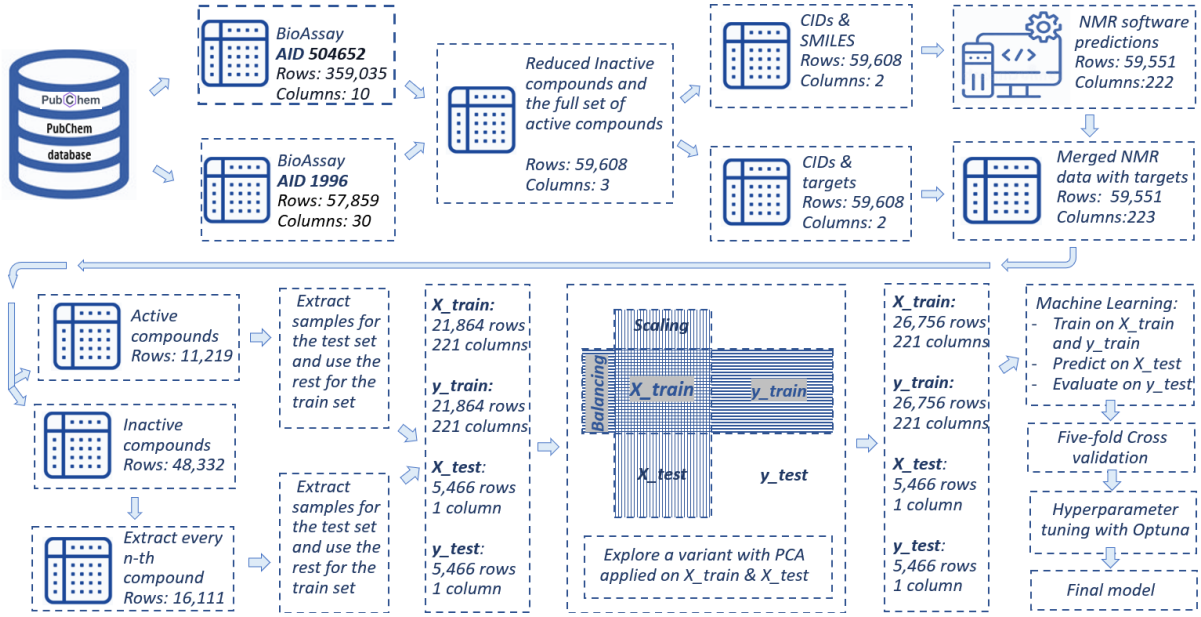

Figure 1.
^{13}C NMR spectroscopy data-based ML model
for predicting of human dopamine D1receptor antagonists' methodology

To complement the main study, the CID_SID ML model (21) was developed to assist researchers interested in D1 DAR antagonists. This ML model can predict the functionality of a compound based on the information encoded in its PubChem CID and SID, thus providing drug researchers with a cost- and time-effective tool for early-stage drug development. The generated code is available on GitHub for direct use by interested parties.

Methodology

From the PubChem bioassay AID 504652 (1), the columns with SMILES notations, outputs of the screening and CIDs of the compounds were extracted, and a new dataset was created. This dataset, in turn, was merged with the water solubility PubChem AID 1996 bioassay dataset (3), and only the unique identifiers (CIDs) of the substances for both datasets were kept. Thus, the inactive compounds were reduced to some extent. Then, the reduced inactive compounds were concatenated with all active compounds from the PubChem bioassay AID 504652 (1). The newly formed dataset was used to create two datasets. One of these two datasets, called 'smiles', contained the CIDs and SMILES notations of the compounds and was utilised to scrap the 13C NMR spectrum data. The other dataset, called 'targets', contained the CIDs and the relevant outputs, which, at the end of the dataset generation process, completed the dataset for ML by adding the labels to the 13C NMR spectrum data

The SMILES strings from the dataset ‘smiles’ were used to calculate the 13C NMR spectroscopic data by the NMRDB software (6). From the generated by the software lists of signals, i.e. the predicted 13C NMR spectroscopy picks, the picks in each range were counted. The length of the ranges, which, in turn, also established the number of ranges, had been determined by consecutive whole numbers from zero to three hundred. The resulting data were then merged with the dataset ‘targets’, and a dataset containing 13C NMR spectroscopy information and targets was obtained. The inactive compounds in the resulting dataset were then additionally reduced, keeping only every third inactive compound. Thus, the final dataset necessary for ML was completed. From it, an equal number of samples were extracted for each target and created the test sets. The rest of the samples formed the train sets. Further, the datasets were balanced in order to ensure that the ML model would learn from both classes effectively (22) and enhance the accuracy of the ML models. The ML binary classification estimators used in the study originated from scikit-learn (23) and were K-Nearest neighbour (KNN) (24), Random Forest Classifier (RFC) (25), XGBoosting Classifier (XGBC) (26), Decision Tree Classifier (DTC) (27), Gradient Boosting Classifier (GBC) (28) and Support Vector Classifier (SVC) (29). Principle component analysis (PCA) was used to reduce the dataset's high dimensionality that arose due to the conversion of SMILES notations into spectroscopy data. The number of components suitable for the given dataset was calculated using unsupervised ML, and subsequently, the suggested number of components was applied for PCA dimensionality reduction.

The metrics suitable for binary classification (30) used to evaluate the ML models were as follows:

(i) Accuracy was calculated by dividing the correct prediction by the total number of predictions, unveiling the overall correctness of the ML model.
(ii) Precision exposes how many of the positive predictions are true positive
(iii) Recall reveals how many of the positive cases in the data set have been identified correctly
(iv) F1-score is the harmonic mean of the precision and the recall
(v) ROC curves illustrate how a binary classifier's performance changes as its decision threshold for classifying instances varies.

Due to the size of the dataset, hyperparameter tuning (31) would take a long time. To reduce the time without making quality compromises, from the final dataset, a representative dataset of 1000 compounds was extracted and used for hyperparameter tuning performed by Optuna (32). The hyperparameters considered during the hyperparameter tuning were as follows:

(i) The ‘rbf’ kernel (radial basis function kernel) finds boundaries and achieves high accuracy by mapping data into a higher-dimensional space, separating non-linearly separable data; the kernel ‘linear’ - the choice of the SVM when the data can be effectively separated by a linear boundary.
(ii) The parameter C minimises the training error, controlling the balance between the closest data points and the separating hyperplane.
(iii) Gamma, which controls the influence of a single training example

The difference between the train and test accuracy was traced because it provided information about the presence of overfitting, i.e. when the model performs well on the train set but not on unseen data. This scrutinising for overfitting was the last step before the final model was selected and visualised by plotting the learning curve, matrix and area under the curve (AUC).

Results and discussion

Crossing both bioassay datasets, PubChem AID 504652 (1) and AID 1996 (2), and keeping only the samples common for both resulted in a new dataset of 49,744 samples. The additional reduction of the inactive compounds, keeping only every third inactive compound, and concatenating the resulting set of inactive compounds with the full set of active compounds from PubChem AID 504652 (1) led to a dataset of 27,330 samples. For the test sets, 2,733 were extracted for each target, making testing sets 20% of the entire dataset. The training sets created by the remaining samples became 26,756 samples after their balancing with the Random Over Sampler algorithm (33). The number of columns corresponding to the ranges in which the ^{13}C NMR spectroscopic picks were counted dropped from 300 to 243 after the column containing only zeroes was removed.

From the above-listed ML classifiers, SVC was the most suitable for the study. It obtained 71.5% accuracy, 77.4% precision, 60.6% recall, 68% F1 and 71.5% ROC, followed by XGBC with Accuracy 68.8% accuracy, 73.7% precision, 58.7 % recall, 65.3% F1 and ROC 68.8% ROC (Table 1). The mean cross-validation ordered the classifiers in the same order where the cross-validation score of SVC based on accuracy was 0.7487 with 0.003 standard deviations for SVC and 0.7207 with 0.0054 standard deviations for XGBC, i.e. the SVC model across the data set obtained mean 74.87% accuracy and for XGBC the accuracy was 72.07% (Table 2). The five-fold cross-validation results were higher than the ML model results and since the difference between them was less than 5%, this indicated that the model generalized well.

The hyperparameter tuning with Optuna suggested kernel='rbf', C=493.2744094205687, gamma = 0.00040591543979010816. However, the model performed with these values obtained 68.51% accuracy, which is lower than the accuracy performed with the default set of values, i.e. kernel ='rbf', C=1, gamma = 'scale', where the value 'scale' means that the algorithm automatically will define the best reasonable initial value for gamma. This hyperparameter is calculated by number of features in the dataset multiply by the variance of the features in the dataset.

When using PCA, unsupervised ML learning suggested the number of components to be reduced from 221 to 23. However, the dimensionality reduction with PCA with 23 components did not improve the performance of the ML model. The SVC achieved 64.1% accuracy, 68.6% precision, 51.8 % recall, 59.1% F1 and 64.1% ROC, followed by RFC with 61.9% accuracy, 70.4% precision, 41.1 % recall, 51.9% F1, 61.9% ROC (Table 3). The five-fold cross-validation score of SVC was 0.683 with 0.0052 standard deviations, followed by RFC with 0.6713 cross-validation score and 0.0072 standard deviations (Table 4). Visualisation of the SVC model is illustrated by the learning curve (Figure 2), AUC (Figure 3), Confusion matrix (Figure 4) and the Classification report (Table 5).

So, the final ML model chosen in the study was SVC with default hyperparameters that achieved 71.5% Accuracy, 77.4% Precision, 60.6% Recall, 68% F1, 71.5% ROC and cross-validation score 0.749 with 0.0036 standard deviations. The final model was visualised with the learning curve (Figure 5), AUC (Figure. 6), Confusion matrix (Figure 7) and the Classification report (Table 6).

During the hyperparameter tuning, which was performed with a smaller number of data set samples, it was observed that the metrics of the ML model were lower than the final results of the study. With training sets of 2366 samples and test sets of 440 samples, the accuracy was 65.7%, precision 63.6%, recall 73.2%, F1 68.1% and ROC 65.7%, which were lower than the performance of the ML model when trained and tested with the entire data set. This observation gave reason to conclude that there is a high probability that increasing the number of samples will increase the robustness and reliability of the ML model.

The CID_SID ML model (21), which was not a primary object of exploration in the presented study, was developed in favour of the researchers interested in human dopamine D1 receptor antagonists. The best-presented estimator was XGBC achieving Accuracy 80.2%, Precision 86.3%, Recall 70.4%, F1 77.6%, ROC 79.9% (Table 7) and cross-validation score 0.8071 with 0.0047 Standard deviation (Table 8), followed by GBC with Accuracy 80.1%, Precision 88.5%, Recall 67.7%, F1 76.7%, ROC79.7% (Table 7), and a five-fold cross-validation score 0.8031 with 0.0038 standard deviation (Table 8). Thus, a drug developer can check if their compound is also a human dopamine D1 receptor antagonist using only its PubChem CID and SID. Given the methodology requirements (21), the dataset contained only CIDs, SIDs and targets of the considered compounds. The CID_SID ML model was trained with 19,438 samples and tested with 4,723. Standardisation of data was processed by scikit-learn (23) standard scalar function, which transformed data so that each feature has a mean of 0 and a standard deviation of 1, preventing the occurrence of feature prioritising. Balancing of train data was performed with a random over sampler (33) approach randomly selecting instances from the target 1 and creating duplicates of them in order to handle the imbalanced dataset. The ML model was hyperparameter tuned by grid search, and the resulting hyperparameter values were as follows:

(i) colsample_bytree=1.0 (that randomly selects a subset of features for each tree, reducing the correlation between trees in the ensemble and thus preventing overfitting and improving the generalisation)
(ii) learning_rate = 0.1 (determine how much the model`s parameters are adjusted during each iteration of the training process)
(iii) max_depth=5 (limits the number of levels or nodes the model has from the root node to the leaf nodes)
(iv) min_child_weight=5 (control the minimum number of samples required to create a new node in a tree)
(v) n_estimators=200 (denotes the number of decision trees within SVC)
(vi) subsamples=0.9 (It controls the fraction of training data randomly sampled for each tree in the ensemble)

Tracing the deviation between test and train accuracy was performed, and it was found that the overfitting started for max_depth > 12 where the test accuracy was 80%, revealing that the ML

model was not overfitted (Figure 8). The final CID_SID ML model was visualised by plotting the learning curve (Figure 9), AUC (Figure 10), confusion matrix (Figure 11), and classification report (Table 9).

Conclusion

Although the spectroscopic data used in the study was obtained by software, the expectations are that the methodology presented in the paper would be applicable when the data is obtained through actual 13C NMR spectroscopy. Although the paper was considered D1 DAR antagonists, the methodology can be applied for predicting any functionality of any small biomolecule when the data necessary for ML meets the data generation requirements. It was observed that the increase in the samples positively influenced the ML model, which gave a hint that increasing the dataset would improve the ML model. The additionally developed CID_SID ML model can give an insight into whether a compound is a human D1 DAR antagonist solely based on its PubChem identifiers. Further investigations are in process to improve the presented ML models.

Author Contributions

MLI, NR, GM and KN conceptualized the project and designed the methodology. MLI and NR wrote the code. MLI, NR and GM processed the data. KN supervised the project. All authors were involved with the writing of the paper.

Acknowledge

MLI thanks the UWL Vice-Chancellor's Scholarship Scheme for their generous support. We thank NIH for providing access to their PubChem database. The article is dedicated to Luben Ivanov.

Data and Code Availability Statement

The raw data used in the study is available through the PubChem portal:
https://pubchem.ncbi.nlm.nih.gov/

The code generated during the research is available on GitHub:
https://github.com/articlesmli/13C_NMR_ML_model_D1.git

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Antagonist of Human D 1 Dopamine Receptor: qHTS. *PubChem*. National Institutes of Health. **2011**. https://pubchem.ncbi.nlm.nih.gov/bioassay/504652 (accessed 2025-01-14) (PubChem AID: 504652)

2. *PubChem.* National Institutes of Health. https://pubchem.ncbi.nlm.nih.gov/docs/statistics (accessed 2025-01-14)

3. Aqueous Solubility from MLSMR Stock Solutions. *PubChem*. National Institutes of Health. **2010**. https://pubchem.ncbi.nlm.nih.gov/bioassay/1996 (accessed 2025-01-14) (PubChem AID 1996)

4. Felli, I.C.; Pierattelli, R. 13C Direct Detected NMR for Challenging Systems. *Chem. Rev*. **2022**, *122* (10), 9468-9496. DOI: 10.1021/acs.chemrev.1c00871

5. Bai, H.; Han, L.; Ma, H. Unraveling the Influence of the Carbon Skeleton Structure and Substituent Electronic Effects on the Nontraditional Intrinsic Luminescence Properties of Nonconjugated Polyolefins. *Macromolecules.* **2024**, *57* (4), 1819-1828. DOI: 10.1021/acs.macromol.3c02131

6. *NMRDB Tools for NMR spectroscopists*; Predict 13C NMR. https://nmrdb.org/13c/index.shtml?v=v2.138.0 (accessed 2025-10-14)

7. Džeroski, S.; Schulze-Kremer, S.; Heidtke, K.R.; Siems, K.; Wettschereck, D. Diterpene Structure Elucidation from 13C NMR-Spectra with Machine Learning. In *Intelligent Data Analysis in Medicine and Pharmacology* Lavrač, N., Keravnou, E.T., Zupan, B. Eds.; The Springer International Series in Engineering and Computer Science Vol. 414; Springer, Boston, MA. US, **1997**; pp. 207–225. DOI :10.1007/978-1-4615-6059-3_12

8. Xu, H.; Zhou, Z.; Hong, P. Enhancing Peak Assignment in 13C NMR Spectroscopy: A Novel Approach Using Multimodal Alignment. ***ArXiv***. 2024. DOI:10.48550/arXiv.2311.13817

9. Williamson, D.; Ponte, S.; Iglesias, I.; Tonge, N.; Cobas, C.; Kemsley. E.K. (2024) Chemical Shift Prediction in 13C NMR Spectroscopy Using Ensembles of Message Passing Neural Networks (MPNNs). *J. Magn. Reason.* **2024**, *368*, 107795. DOI: 10.1016/j.jmr.2024.107795.

10. Duprat, F.; Ploix, J.L.; Dreyfus, G. Can Graph Machines Accurately Estimate 13C NMR Chemical Shifts of Benzenic Compounds? *Molecules*. **2024**, *29*(13), 3137. DOI: 10.3390/molecules29133137

11. *CASPRE*; CASPRE - 13C NMR Predictor. https://caspre.ca/ (accessed 2025-01-14)

12. Rigel, N.; Li, D.W.; Brüschweiler, R. COLMARppm: A Web Server Tool for the Accurate and Rapid Prediction of 1H and 13C NMR Chemical Shifts of Organic Molecules and Metabolites. *Anal. Chem*. **2024**, *96* (2), 701-70. DOI: 10.1021/acs.analchem.3c03677

13. Meiler, J.; Will, M. Genius: A Genetic Algorithm for Automated Structure Elucidation from 13C NMR Spectra. J. Am. Chem. Soc. **2002**, *124* (9), 1868-1870. DOI: 10.1021/ja0109388

14. Abramson, J.; Adler, J.; Dunger, J. et al. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature*. **2024**, *630*, 493–500. DOI: 10.1038/s41586-024-07487-w

15. Turzo, S.B.A.; Hantz, E.R.; Lindert, S. Applications of Machine Learning in Computer-aided Drug Discovery. *QRB Discov*. **2022**, *3*, e14. DOI: 10.1017/qrd.2022.12

16. Ivanova, M.L.; Russo, N.; Djaid, N.; Nikolic, K. Application of Machine Learning for Predicting G9a Inhibitors. *Digit. Discov*. **2024**, *3*(10), 2010-2018. DOI: 10.1039/d4dd00101j.

17. Dara, S.; Dhamercherla, S.; Jadav, S.S.; Babu, C.H.M.; Ahsan, M.J. Machine Learning in Drug Discovery: A Review. *Artif. Intell. Rev.* **2022**, *55*, 1947–1999. DOI: 10.1007/s10462-021-10058-4

18. Kim, H.J.; Cho, Y.S.; Koh, H.Y.; Kong, J.Y.; No, K.T.; Pae, A.N. Classification of Dopamine Antagonists Using Functional Feature Hypothesis and Topological Descriptors. *Bioorg. Med. Chem*. **2006**, *14*(5), 1454-1461. DOI: 10.1016/j.bmc.2005.09.072

19. Sobodu, T.; Yusuf, A.; Kiel, D.; Kong, D. Discrimination vs. Generation: The Machine Learning Dichotomy for Dopaminergic Hit Discovery. *ArXiv.* **2024**. DOI: arxiv.org/pdf/2409.12495

20. Fujishima, S.; Takahashi, Y.; Nishikori. K.; Katoh, T.; Okada, T. Extended Study of the Classification of Dopamine Receptor Agonists and Antagonists using a TFS-based Support Vector Machine. *New Gener. Comput.* **2007**, *25*, 203–212. DOI: 10.1007/s00354-007-0012-x

21. Ivanova, M. L; Russo, N.; Nikolic, K. Predicting Novel Pharmacological Activities of Compounds Using PubChem IDs and Machine Learning (CID-SID ML Model). *ArXiv*. **2025**. DOI: 10.48550/arXiv.2501.02154

22. Gnip P, Vokorokos L and Drotár P (2021) Selective Oversampling Approach for Strongly Imbalanced Data. *PeerJ Comput. Sci.* **2021**, *18*(7), e604. DOI: 10.7717/peerj-cs.604

23. Scikit-learn Home Page https://scikit-learn.org/stable/index.html (accessed 2025-01-14)

24. Cunningham, P.; Delany, S. J. K-Nearest Neighbour Classifiers - A Tutorial. *ACM Comput. Surv*. **2021**, *54*(6): Article 128 (July 2022), 25 pages. DOI: 10.1145/3459665

25. Manzali, Y.; Elfar, M. Random Forest Pruning Techniques: A Recent Review. *Oper. Res. Forum*. **2023**, *4* (2), 43. DOI: 10.1007/s43069-023-00223-6

26. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, D. A Comparative Analysis of Gradient Boosting Algorithms. *Artif. Intell. Rev*. **2021**, *54*, 1937–1967. DOI: 10.1007/s10462-020-09896-5

27. Klusowski. J. M.; Tian, P. M. Large Scale Prediction with Decision Trees. *J. Am. Stat. Assoc*. **2024**, *119*(545), 525-537. DOI: 10.1080/01621459.2022.2126782

28. Mushava, J.; Murray, M. (2024) Flexible Loss Functions for Binary Classification in Gradient-Boosted Decision Trees: An Application to Credit Scoring. *Expert. Syst. Appl*., **2024**, *238*, 121876. DOI: 10.1016/j.eswa.2023.121876

29. Guido, R.; Ferrisi, S.; Lofaro, D.; Conforti, D. An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information*., **2024**, *15*(4), 235. DOI: 10.3390/info15040235

30. Rainio, O.; Teuho, J.; Klén, R. Evaluation Metrics and Statistical Tests for Machine Learning. *Sci. Rep*. **2024**, *14*, 6086. DOI: 10.1038/s41598-024-56706-x

31. Panda, B. Hyperparameter Tuning. *Research Gate*. **2019**. DOI: 10.13140/RG.2.2.11820.21128

32. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. ArXiv, **2019**. DOI: 10.48550/arXiv.1907.10902

33. RandomOverSampler *Imbalanced Learn* https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.RandomOverSampler.html (accessed 2025-01-14)

Table1. Results of ML with ^{13}C NMR data of human dopamine D1 receptor antagonists.

	1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
0	SVM	0.715	0.774	0.606	0.680	0.715
4	XGBoost	0.688	0.737	0.587	0.653	0.688
2	RandomForest	0.653	0.792	0.416	0.546	0.653
3	GradientBoost	0.650	0.674	0.582	0.625	0.650
5	K-nearest	0.612	0.671	0.441	0.532	0.612
1	Decision	0.578	0.596	0.483	0.534	0.578

Table 2. Five-fold cross-validation results of ML with ^{13}CNMR data of human dopamine D1 receptor antagonists.

	1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
0	SVM	0.7487	0.0030	[0.7508, 0.7461, 0.7536, 0.7472, 0.7461]
4	XGBoost	0.7207	0.0054	[0.7252, 0.7144, 0.7236, 0.7265, 0.7141]
2	RandomForest	0.7073	0.0021	[0.7065, 0.7084, 0.7102, 0.7073, 0.704]
3	GradientBoost	0.6885	0.0048	[0.6925, 0.6872, 0.6861, 0.6954, 0.6817]
5	K-nearest	0.6429	0.0059	[0.6509, 0.6356, 0.6476, 0.6432, 0.637]
1	Decision	0.6012	0.0065	[0.6096, 0.6059, 0.5939, 0.6034, 0.5933]

Table 3. Results of ML with ^{13}C NMR data of human dopamine D1 receptor antagonists reduced by PCA.

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.642	0.696	0.506	0.586	0.642
RandomForest	0.619	0.704	0.411	0.519	0.619
GradientBoost	0.619	0.645	0.530	0.582	0.619
XGBoost	0.615	0.659	0.478	0.554	0.615
K-nearest	0.580	0.612	0.438	0.511	0.580
Decision	0.544	0.555	0.444	0.493	0.544

Table 4. Five-fold cross-validation results of ML with ^{13}CNMR data of human dopamine D1 receptor antagonists reduced by PCA.

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
SVM	0.6830	0.0052	[0.6881, 0.6808, 0.6782, 0.6903, 0.6778]
RandomForest	0.6713	0.0072	[0.6736, 0.6601, 0.6753, 0.6809, 0.6667]
GradientBoost	0.6663	0.0074	[0.6775, 0.659, 0.6654, 0.6718, 0.6581]
XGBoost	0.6616	0.0040	[0.6676, 0.6595, 0.6561, 0.6645, 0.6604]
K-nearest	0.6205	0.0043	[0.6239, 0.6147, 0.6158, 0.6239, 0.6242]
Decision	0.5780	0.0094	[0.5878, 0.5752, 0.5856, 0.5801, 0.5615]

Table 5. Classification report of the SVC ML model with ^{13}CNMR data of human dopamine D1 receptor antagonists reduced by PCA.

	precision	recall	f1-score	support
Active (target 1)	0.61	0.78	0.69	2733
Inactive (target 0)	0.70	0.51	0.59	2733
accuracy			0.64	5466
macro avg	0.65	0.64	0.64	5466
weighted avg	0.65	0.64	0.64	5466

Table 6. Classification report of the SVC ML model with ^{13}CNMR data of human dopamine D1 receptor antagonists.

	precision	recall	f1-score	support
Active (target 1)	0.68	0.82	0.74	2733
Inactive (target 0)	0.77	0.61	0.68	2733
accuracy			0.71	5466
macro avg	0.73	0.71	0.71	5466
weighted avg	0.73	0.71	0.71	5466

Table 7. Results of CID_SID ML models with ^{13}C NMR data of human dopamine D1 receptor antagonists.

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
XGBoost	0.802	0.863	0.704	0.776	0.799
GradientBoost	0.801	0.885	0.677	0.767	0.797
RandomForest	0.790	0.812	0.738	0.774	0.789
K-nearest	0.774	0.784	0.738	0.761	0.773
SVM	0.766	0.827	0.655	0.731	0.763
Decision	0.757	0.748	0.752	0.750	0.756

Table 8. Five-fold cross-validation results of the CID_SID ML models with 13CNMR data of human dopamine D1 receptor antagonists.

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
XGBoost	0.8071	0.0047	[0.8133, 0.7995, 0.8084, 0.8097, 0.8046]
GradientBoost	0.8031	0.0038	[0.8052, 0.7991, 0.8092, 0.7997, 0.8022]
RandomForest	0.7903	0.0059	[0.7957, 0.7813, 0.7957, 0.7938, 0.7853]
K-nearest	0.7810	0.0036	[0.7811, 0.7745, 0.783, 0.7853, 0.7811]
Decision	0.7524	0.0090	[0.7648, 0.7457, 0.7578, 0.7546, 0.7394]
SVM	0.7439	0.0032	[0.7434, 0.7425, 0.7502, 0.7427, 0.7408]

Table 9. Classification report of the XGBC CID_SID ML mode with $^{13}CNMR$ data of human dopamine D1 receptor antagonists.

	precision	recall	f1-score	support
Active (target 1)	0.76	0.90	0.82	2430
Inactive (target 0)	0.87	0.70	0.78	2293
accuracy			0.80	4723
macro avg	0.81	0.80	0.80	4723
weighted avg	0.81	0.80	0.80	4723

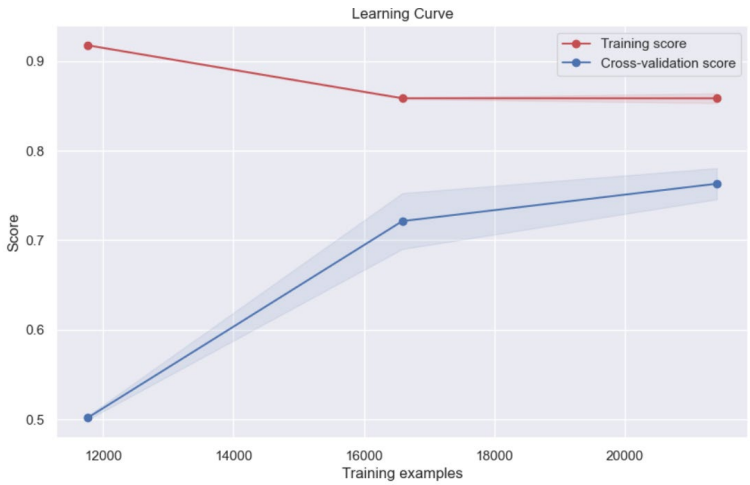

Figure 2. Learning curve off ML with 13CNMR data of human dopamine D1 receptor antagonists

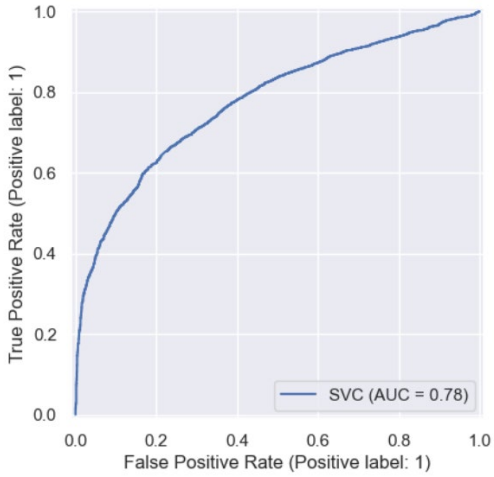

Figure 3. ROC curve of ML with 13CNMR data of human dopamine D1 receptor antagonists

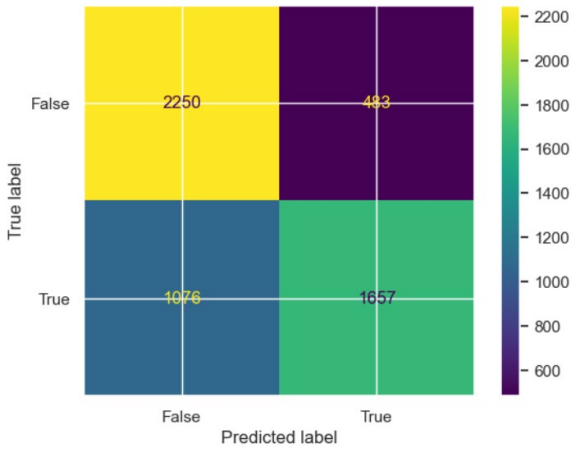

Figure 4. Matrix off ML with 13CNMR data of human dopamine D1 receptor antagonists

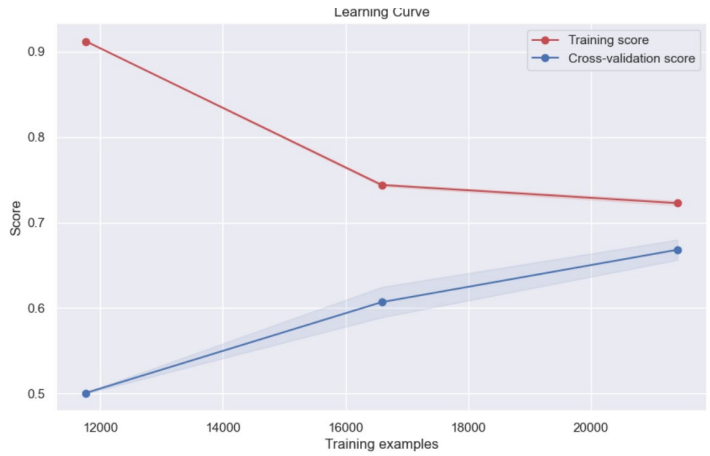

Figure 5. Learning curve off ML with 13CNMR data of human dopamine D1 receptor antagonists reduced by PCA

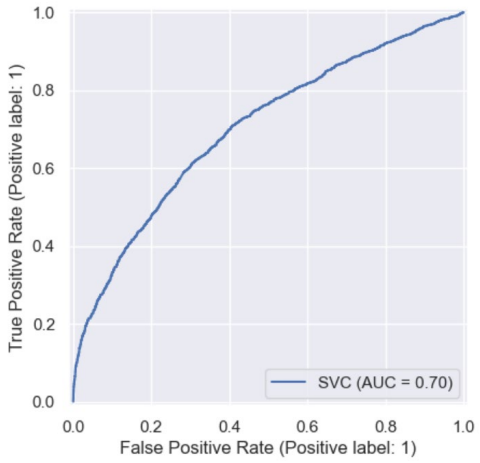

Figure 6. ROC curve of ML with 13CNMR data of human dopamine D1 receptor antagonists reduced by PCA

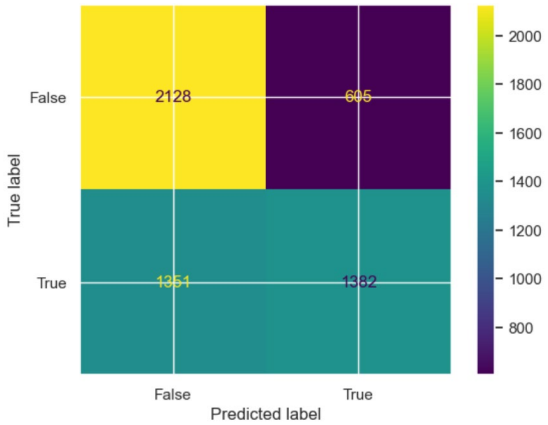

Figure 7. Matrix off ML with 13CNMR data of human dopamine D1 receptor antagonists reduced by PCA

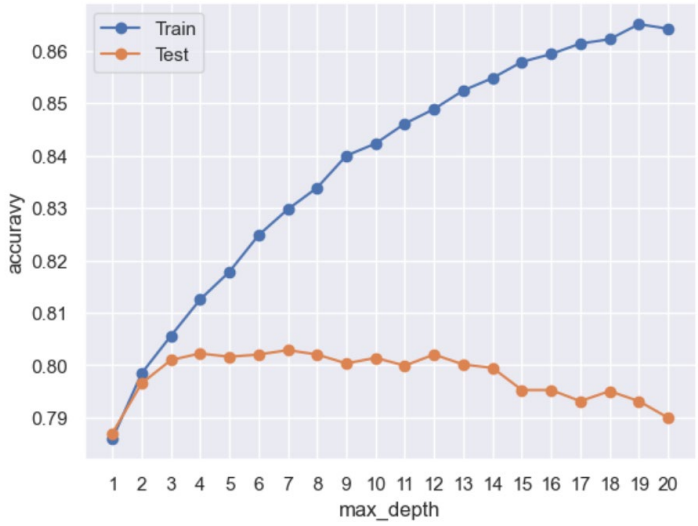

Figure 8. Tracing the deviation between train and test accuracy of the XGBC CID_SID ML model with 13CNMR data of human dopamine D1 receptor antagonists

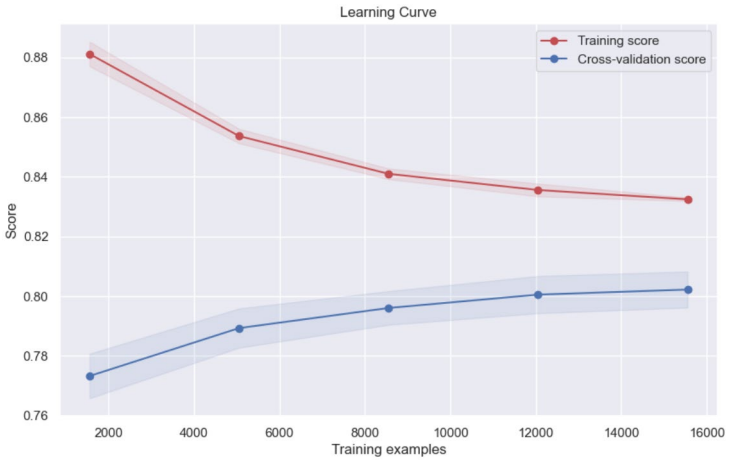

Figure 9. Learning curve of the XGBC CID_SID ML model with 13CNMR data of human dopamine D1 receptor antagonists

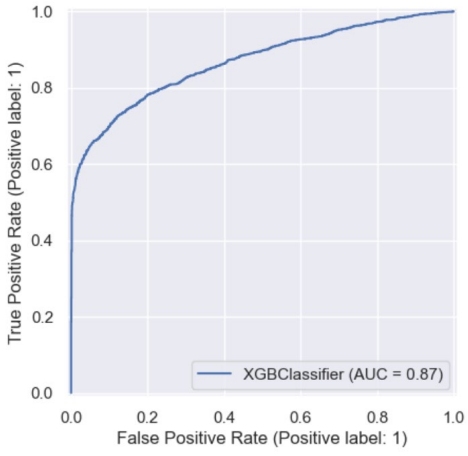

Figure 10. ROC curve of the XGBC CID_SID ML model with 13CNMR data of human dopamine D1 receptor antagonists

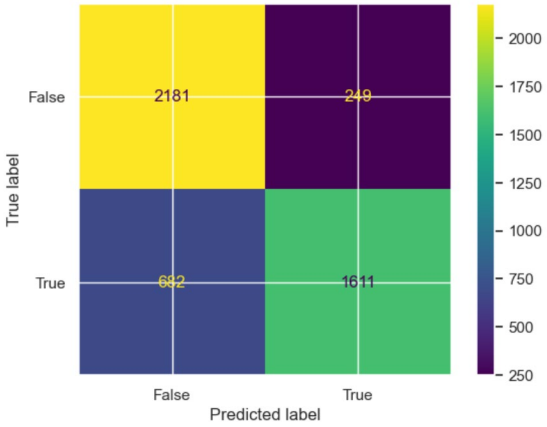

Figure 11. Matrix of the XGBC CID_SID ML model with 13CNMR data of human dopamine D1 receptor antagonists