Predicting novel functional roles of designed small biomolecules: an ML Approach utilizing PubChem Compound and Substance Identifiers (CID-SID ML model)

Ivanova, Mariya, Russo, Nicola, Mihaylov, Gueorgui and Konstantin, Nikolic ORCID logoORCID: https://orcid.org/0000-0002-6551-2977 (2025) Predicting novel functional roles of designed small biomolecules: an ML Approach utilizing PubChem Compound and Substance Identifiers (CID-SID ML model). In Silico Pharmacology. (Submitted)

Predicting Novel Functional Roles of Designed Small Biomolecules: an ML Approach Utilizing PubChem Compound and Substance Identifiers (CID-SID ML model)

Mariya L. Ivanova[1,*], Nicola Russo[1], Gueorgui Mihaylov[2] and Konstantin Nikolic[1]

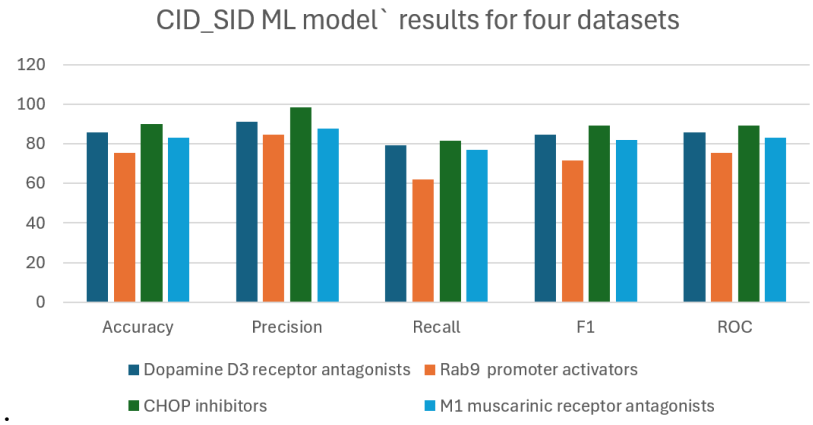
Author affiliations
[1]School of Computing and Engineering, University of West London, London, UK
[2]Haleon, London, UK
*Corresponding author mariya.ivanova@uwl.ac.uk

Abstract

Significance and Object: The proposed methodology aims to provide time- and cost-effective approach for the early stage in drug discovery. The machine learning models developed in this study used only the identification numbers provided by PubChem. Thus, a drug development researcher who has obtained a PubChem CID and SID can easily identify new functionality of their compound. The approach was demonstrated, using four bioassay which were on (i) the antagonists of human D3 dopamine receptors; (ii) the promoter Rab9 activators; (iii) small molecule inhibitors of CHOP to regulate the unfolded protein response to ER stress; (iv) antagonists of the human M1 muscarinic receptor.

Solution: The four bioassays used for demonstration of the approach were provided by PubChem. For each bioassay, the generated by PubChem CIDs, SIDs were extracted together with the corresponding activity. The resulting dataset was sifted with the dataset on a water solubility bioassay, remaining only the compounds common for both bioassays. In this way, the inactive compounds were reduced. Then, all active compounds were added, and the resulted dataset was later used for machine learning based on scikit learn algorithms.

Results: The average values of the ML models` metrics for the four bioassays were: 83.82% Accuracy with 5.35 standard deviation; 87.9% Precision with 5.04 standard deviation; 77.1% Recall with 7.65 standard deviation; 82.1% F1 with 6.44 standard deviation; 83.4% ROC with 5.09 standard deviation. Since the methodology was publicly available as a preprint, four more machine ML models have been developed. Their results are discussed in the "Results and Discussion" section



Key words: Machine Learning, D3 dopamine receptor antagonist, Rab9, CHOP, M1 muscarinic receptor, CID, SID, L-DOPA-induced dyskinesias, schizophrenia, Alzheimer`s, Parkinson`s, neurodegenerative disorders.

Introduction

One of the rules in ML is not to use identifications (IDs) of the samples during training, testing and evaluation of the machine learning (ML) models because by their nature, IDs are not suited for this task (Zhou, 2021). However, PubChem generates their IDs by applying an algorithm that considers the structure of the compounds/substances, their functionality and the similarity between them (Kim et al., 2016). On the other hand, nine out of ten drug candidates that have participated in clinical trials never apply for FDA approval, one of the reasons may be an unexpected side effect. (Jain, Subramanian and Rathore, 2022). In this way, drug candidates that would otherwise fail for such reasons would be discovered early in their development, thus preventing investment in unvalued formulations. In turn, this would prevent wasted investment and significantly reduce the overall cost of drug development that has been reported to require over 10 years and over USD 1 billion (Niazi and Mariam, 2025) per drug. Or it could detect functionalities of the designed small biomolecule that would enhance the knowledge regarding the capabilities of the new compound. Unlike ML models based on molecular fingerprints, the proposed approach does not require data conversion to numeric, as they are numeric.

To demonstrate the methodology, four bioassays provided by PubChem (NIH, 2025) were utilized. All of them have been obtained by quantitative high-throughput screening (qHTS) (Soon, Hariharan and Snyder, 2013) and thus contain a significant number of records, which made the development of the ML models in the presented study possible (Saha, Chauhan and Rastogi Verma, 2024). In summary, the considered bioassays, provided by PubChem were:

(i) AID 652054 regarding the dopamine receptor D3
(ii) AID 485297 regarding the promoter of the protein Rab9
(iii) AID 2732 regarding CHOP
(iv) AID 588852 regarding M1 muscarinic receptor

The first bioassay, PubChem AID 652054 (NIH, 2013) was primarily designed to discover novel antagonists of the dopamine receptor D3, whose drug to antibody ration (DAR) is a target for treatment of neuropsychiatric disorders, such as addictions, schizophrenia, psychosis (Grunze, 2023) and L-DOPA-induced dyskinesias (Chagraoui, Di Giovanni, Deurwaerdère, 2022). The bioassay`s dataset contained 364,367 rows with samples and 26 columns with their features. The results have been obtained by a luminometer reader using a 20 sec exposure time. The compounds whose activity have been <=-50 have been considered as active and these with activity >=-30 as inactive. Thus, 9,117 samples were defined as active, and 339,862 as inactive compounds. The inconclusive compounds with activity between these two values were not used in the study. For a comprehensive description of the bioassay, please refer to the bioassay`s documentation (NIH, 2013).

The second bioassay, PubChem AID 485297 (NIH, 2010a), was on the identification/discovery small chemical compounds that can modulate the expression of the endogenous protein Rab9 and provide new treatment modality for the neurodegenerative lipidosis such as Nieman Pick

Type C and Alzheimer`s disorder (Jordan, 2024). The original dataset of this bioassay contained 321,272 rows with samples and 11 columns with their features. The tests have been performed at concentration of the compound at 2.3μM, 11.40μM and 57.5 μM. The results were obtained by fitting the dose-response curve to the Hill equation. The compounds with activity <=-50 have been considered as active and these with activity >=-29 as inactive. For a comprehensive description of the bioassay, please refer to the bioassay`s documentation (NIH, 2010a).

The third bioassay, PubChem AID 2732 (NIH, 2010b) has been conducted with the intention small molecule inhibitors of DNA damage-inducible transcript 3, also known as C/EBP homologous protein **(**CHOP) to be discovered. The inhibition of CHOP is hypothesised to prevent programmed the unfolded protein response (UPR) cell death and thus having a therapeutic application to Alzheimer`s disorder, Parkinson`s disorder, haemophilia, lysosomal storage diseases, alpha-1 antitrypsin deficiency and diabetes. The original dataset of this bioassay contained 219,165 rows and 10 columns. The tests have been performed at 10μM concentration of the compounds. Using luciferase in the cell-line and following the protocol explained in bioassay`s documentation (NIH, 2010b), the luminescence signal was measured on an Envision Multilabel plate reader and analysed by an algorithm. From the entre dataset 8,243 samples had activity >70% and were considered as active. The rest of the compounds, i.e. 210 921 were ladled like inactive. For a comprehensive description of the bioassay, please refer to the bioassay`s documentation (NIH, 2010b).

The fourth bioassay, PubChem AID 588852 (NIH, 2012), identified antagonists of the human M1 muscarinic receptor which mediates the actions of Acetylcholine in the CNS and represent attractive therapeutic targets for cognition (Zhao et al,2018), Alzheimer's disease (Monaco, Trebesova and Grilli, 2024) and schizophrenia (Kingwell, 2024; Metz, Brines and Pavlov, 2024). The original dataset contained 359,484 rows of compounds and 9 columns with their features. The tests have been performed at the compound concentration of 3μM. A cutoff parameter which was a sum of average percent inhibition of the test compound wells and three times their standard deviation was used. The compound exhibition has been compared to this cutoff. The results were normalised to 100% and the compounds with a score over 80 was defined as active. Thus, 4,590 compounds were selected as active. For a comprehensive description of the bioassay, please refer to the bioassay`s documentation (NIH, 2012).

Additionally, to above-mentioned four bioassays, PubChem Bioassay 1996 (NIH, 2010c) was implemented in the study to assist in the handling the enormous balancing between the active and inactive compounds of the four bioassays. The PubChem Bioassay 1996 bioassay was on water solubility of small molecules. It played a role of a sieve of the inactive compounds, reducing them to suitable quantity. For a comprehensive description of the bioassay, please refer to the bioassay`s documentation (NIH, 2010c)

Computational approaches, lower the price and time necessary for drug discovery. Examples for such studies are Structure-Based Drug Design (SBDD) and Ligand-Based Drug Design (LBDD) (Bhujade et al.,2024); engineering of new compound features to expand the opportunity to forecast their functionalities (Ivanova, Russo, Djaid and Nikolic, 2024); utilising of existing drugs, using their safety profiles to discover new uses for treatment of variety of conditions (El-Atawneh and Goldblum, 2024); performing drug design with molecular dynamics and machine learning targeting dopamine D3 receptor (Ferraro et al, 2020); predicting selectivity of dopamine receptor ligands using three-dimensional biologically spectrum **(**Kuang et al., 2016); suggesting

a framework for AI-driven molecular design for discovering drugs against complex disorders (Cerveira et al, 2024); applying ML based on structural similarity for target identification, ChemMapper, and SwissTargetPrediction to identify muscarinic acetylcholine receptor M1 (Abdalfattah et al., 2024); using of mismatch negativity responses to predict of muscarinic receptor function, revealing the potential of generative models based on electrophysiological data (Schöbi et al., 2021). However, the method presented here has not been reported in the literature up to date. The ML part was conducted using the Python ML library scikit learn (Pedregosa et al., 2011) in the Jupyter notebook environment (Jupyter, 2024) and followed the best practice recommended in the domain (Vinuesa, 2024).

Methodology

For each bioassay listed above, the CIDs, SIDs and activity results were extracted, and a new dataset was created. Then, this new dataset was merged with the water solubility dataset, keeping only the compounds common for both bioassays. Thus, the first step towards balancing the data by reduction of the inactive compounds was completed. Further reduction of the inactive compounds continued with extracting every *n*-th compound from the inactive compounds dataset. The resulted dataset with inactive compounds then was concatenated with the active compounds' dataset and thus, the final dataset obtained. The value of *n* that defined which compound to be extracted was determine individually for each bioassay. To prepare the dataset for ML, equal number of target 1 and 0, corresponding to 10% of the entire final dataset, were extracted and concatenated. Thus, 20% test dataset was obtain with equal number of targets in order a reliable evaluation of the models to be provided. The remaining compounds created the train sets. Initially the datapoints of this sets were scaled, and then together with the target part of the train sets were balanced (He and Garcia, 2009) with Synthetic Minority Oversampling Technique (SMOTE) (Kabir et al, 2024) or Random Over Sampler (ROS) (Imbalanced Learn, 2024)

ML was performed with algorithms, provided by scikit learn, which were: Decision Tree Classifier (DTC) (Lee, Sim and Hong, 2024) Random Forest Classifier (RFC) (He et al., 2024), Support Vector Classifier(SVC) (Shin and Shin, 2024); Gradient Boosting Classifier (GBC) (Ibragimov and Vakhrushev, 2024) and XGBoosting Classifier (XGB) (Hanif, 2020). Cross validation (Bates, Hastie and Tibshirani, 2024) was used to estimate how well a model generalizes to unseen data. After that, comparing the train and test accuracy, the best performed model was scrutinised for overfitting (Ying, 2019) to show how well the chosen ML model generalise. The deviation between the test and train accuracy bigger than 5% was accepted as an indicator when overfitting starts. Two approaches for hyperparameter tuning were explored. The first was the hyperparameter tuning using grid search (Arnold et al., 2024) and the second was hyperparameter tuning with the real time running API Optuna (Akiba et al., 2019). The hyperparameters used for hyperparameter tuning were: (i) 'colsample_bytree' which control the fraction of features randomly selected for each tree during training; (ii) 'learning_rate' which defines the step size at which the model learns from each subsequent tree; (iii) 'n_estimator ' gives the number of trees that are constructed in the ensemble; (iv) 'subsample' is the fraction of samples used for training each individual tree in the ensemble; (v) 'max_depth' which defines the maximum depth of each individual decision tree within the forest; (vi) 'gamma' which is the minimum loss reduction required to make a further partition on a leaf node; (vii) 'reg_lambda' is the L2 regularization parameter; (viii) 'min_child_weight' which controls the minimum sum of instance weight needed in a child node to be further partitioned; (ix) 'min_samples_leaf' minimum number of samples required to be at a leaf node of a decision tree; (x) 'min_samples_split' which defines the minimum number of

samples required to split an internal node. Further, each ML hyperparameter tuned model was scrutinised for overfitting. The results were compared with each other and the best one was chosen for final one and visualised.

The metrics for evaluation of the ML models used in the study were (Opitz, 2024):

(i) Accuracy, showing the percentage of number of correct predictions divided by the total number of predictions.
(ii) Precision, showing the accuracy of positive predictions made by a model.
(iii) Recall is the ability of the ML model to correctly identify all actual positive instances within a dataset.
(iv) F1-score which is the metric that combine the Precision and Recall.
(v) ROC which shows the diagnostic ability of a binary classifier system when its discrimination threshold varied.

Confusion matrix, classification report, and plotting the learning curve and ROC visualised the final models.

The methodology graph is provided on Figure 1.

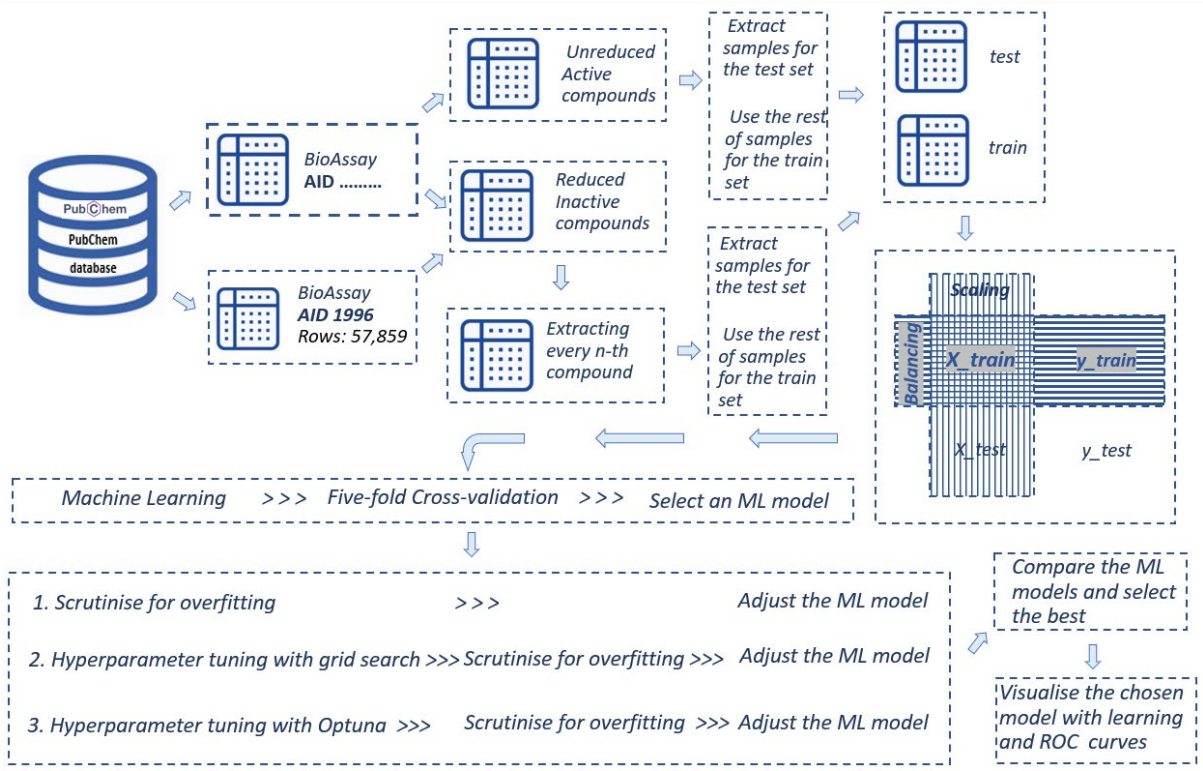


Figure 1. Methodology for development of the CID-SID ML model

Results and discussion

1.Predicting the human Dopamine D3 receptor antagonists.

1.1. Dataset After numerous simulations, it turned out that in order to achieve meaningful training of the models, the proportion between the target 1 and 0 in the train set should be

approximately 2:3. That is why, after the initial reduction of the inactive samples from 339,862 to 54,951 samples, which was achieved by crossing the original dataset with the solubility dataset, the reduction continued by extracting every 3rd sample by the inactive compounds. Thus, 18,317 inactive compounds remained and together with the active compounds, which were 9,117, created the final dataset of 27,434 compounds. To ensure an equal number of targets in the test sets to achieve a robust model, the inactive compounds were shuffled, and 10% of them were extracted, i.e. 2,750 samples. The same was done with the set of the active compounds. Thus, the test sets (X_test and y_test) became 5,500 compounds which was 20% from the final dataset and had an equal proportion of targets. The rest of the compounds, i.e. 12,817 samples, were used for the train sets. The compounds were scaled, and the sets were balanced with a Random Over Sampler. This increased the number of compounds from 12,817 to 31,134 samples.

1.2. ML. From the estimators listed in the Methodology section, XGBC presented best, obtaining 85.6% accuracy and 85.6% ROC, followed by RFC with 85.1 % accuracy and 85.1% ROC (Table SM1). Further, five-fold cross-validation nominated RFC with the best mean cross-validation score of 0.8851 with 0.0027 standard deviation which means that the mean accuracy obtained across the five folds by RFC was 88.5%. The next in the cross-validation order was XGBC, with a 0.885 cross-validation score and 0.0033 standard deviations (Table SM2). Scrutinising for overfitting of the XGBC showed that the deviation between the train and test accuracy started being bigger than 5% after the maximum depth of each individual decision tree within the forest, i.e. max_depth was equal to 9 (Figure SM1).

1.3. The hyperparameter tuning of the XGBC performed with grid search (Table SM3) and scrutinised for overfitting (Figure SM2) achieved accuracy of 85.9%, and ROC of 85.9%. The accuracy after the hyperparameter tuning with Optuna (Table SM4) and scrutinising for overfitting (Figure SM3) was 85.8% and ROC 85.8%, respectively.

1.4. The final ML model chosen based on the results presented above was the XGBC model with hyperparameters tuned by the greed search with max_depth equal to 4. The model obtained 85.8 % accuracy, 91% precision, 79.3% recall, 84.8% F1, 85.8% ROC. To visualise the results the learning curve, AUC, confusion matrix and classification report are provided in the supplementary materials (Figure SM4, SM5, SM6, Table SM5)

2. Predicting promoters of the Rab9 activator.

2.1. Dataset Simulations showed that to achieve meaningful training of the models, the proportion between the target 1 and 0 in the train set had to be approximately 1:3. That is why, after the initial reduction of the inactive samples from 301,951 to 47,918 samples, which was achieved by crossing the original dataset with the solubility dataset, the reduction continued by extracting every second sample by the inactive compounds. Thus, 22,701 inactive compounds remained and together with the active compounds, which were 9,138, created the final dataset of 31,939. To ensure an equal number of targets in the test set and achieve a robust model, the inactive compounds were shuffled, and 10% of them were extracted, i.e. 3,200 samples. The same was done with the set of the active compounds. Thus, the test sets (X_test and y_test) became 6,400 compounds, which was 20% from the final dataset and had an equal proportion

of targets. The rest of the compounds, i.e. 19,501 samples, were used for the training sets. The compounds were scaled, and the sets were balanced with SMOTE. This increased the number of compounds from 25,439 to 39,002 samples.

2.2. ML From the estimators listed in the Methodology section, RFC presented best, obtaining 75.7% accuracy and 75.8% ROC, followed by XGBC with 75.5 % accuracy and 75.5% ROC (Table SM6). Further, five-fold cross-validation ordered XGBC with the best mean cross-validation score of 0.8429 with 0.0032 standard deviation which means that the mean accuracy obtained across the five folds was 84.29%. The next in the cross-validation order was GBC, with a 0.8415 cross-validation score and 0.0038 standard deviations (Table SM7). Scrutinising for overfitting of the XGBC showed that the deviation between the train and test accuracy started being bigger than 5% after the maximum depth of each individual decision tree within the forest, i.e. max_depth was equal to 5 (Figure SM7).

2.3. The hyperparameter tuning of the XGBC performed with grid search (Table SM8) and scrutinised for overfitting (Figure SM8) achieved accuracy of 75.4%, ROC of 75.4% and mean cross-validation score of 0.843. The accuracy after the hyperparameter tuning with Optuna (Table SM9) and scrutinising for overfitting (Figure SM9) was 75.4%, ROC 75.4% and mean cross-validation score of 0.843., respectively.

2.4. The final ML model chosen based on the results presented above was the XGBC model with hyperparameter default values and with max_depth equal to 5. The model obtained 75.4 % accuracy, 84.7% precision, 62.1% recall, 71.7% F1, 75.4% ROC and mean cross-validation score of 0.843. To visualise the results the learning curve, AUC, confusion matrix and classification report are provided in the supplementary materials (Figure SM.10, SM11, SM12, Table SM10)

3. Predicting of small molecule inhibitors of CHOP to regulate the unfolded protein response to ER stress

3.1. Dataset The initial reduction of the inactive compounds decreased them from 210,921 to 24,188 samples, which was achieved by crossing the original dataset with the solubility dataset. Unlike the previous two cases, CHOP dataset did not need an additional reduction. The inactive 24,188 compounds together with the 8,243 active compounds created the final dataset of 32, 431. To ensure an equal number of targets in the test set and achieve a robust model, the inactive compounds were shuffled, and 10% of them were extracted, i.e. 3,243 samples. The same was done with the set of the active compounds. Thus, the test sets (X_test and y_test) became 6,486 compounds, which was 20% from the final dataset and had an equal proportion of targets. The rest of the compounds, i.e. 25,945 samples, were used for the training sets. The compounds were scaled, and the sets were balanced with a Random Over Sampler. This increased the number of compounds from 25,945 to 41,890 samples.

3.2. ML From the estimators listed in the Methodology section, GBC presented best, obtaining 89.9% accuracy and 89.9% ROC, followed by XGBC with 89.6 % accuracy and 89.6% ROC (Table SM11). Further, five-fold cross-validation ordered GBC with the best mean cross-validation score of 0.9414 with 0.0005 standard deviation which means that the mean accuracy obtained across

the five folds was 94.14%. The next in the cross-validation order was XGBC, with a 0.9407 cross-validation score and 0.0012 standard deviations (Table SM12). Scrutinising for overfitting of the GBC showed that the deviation between the train and test accuracy started being bigger than 5% after the maximum depth of each individual decision tree within the forest, i.e. max_depth was equal to 8 (Figure SM13).

3.3. The hyperparameter tuning of the GBC performed with grid search (Table SM13) and scrutinised for overfitting (Figure SM14) achieved accuracy of 89.9%, ROC of 89.9% and mean cross-validation score: 0.94. The accuracy after the hyperparameter tuning with Optuna (Table SM14) and scrutinising for overfitting (Figure SM15) was 89.3%, ROC 89.3% and mean cross-validation score: 0.932, respectively.

3.4. The final ML model chosen based on the results presented above was the GBC model with hyperparameter default values and with max_depth equal to 5. The model obtained 90.1 % accuracy, 98.3% precision, 81.7% recall, 89.2% F1, 89.2% ROC and mean cross-validation score: 0.943, respectively. To visualise the results the learning curve, AUC, confusion matrix and classification report are provided in the supplementary materials (Figure SM16, SM17, SM18, Table SM15)

4. Predicting antagonists of the human M1 muscarinic receptor (CHRM1).

4.1. Dataset. After numerous simulations, it turned out that in order to achieve meaningful training of the models, the proportion between the target 1 and 0 in the train set should be approximately 1:5. That is why, after the initial reduction of the inactive samples from 354,923 to 56,688 samples, which was achieved by crossing the original dataset with the solubility dataset, the reduction continued by extracting every fourth sample by the inactive compounds. Thus, 14,172 inactive compounds remained and together with the active compounds, which were 4,560, created the final dataset of 18,732. To ensure an equal number of targets in the test set and achieve a robust model, the inactive compounds were shuffled, and 10% of them were extracted, i.e. 1,880 samples. The same was done with the set of the active compounds. Thus, the test set (X_test and y_test) became 3 760 compounds, which was 20% from the final datasets and had an equal proportion of targets. The rest of the compounds, i.e. 14,972 samples, were used for the training sets. The compounds were scaled, and the sets were balanced with a Random Over Sampler. This increased the number of compounds from 14,972 to 24,584 samples.

4.2. ML From the estimators listed in the Methodology section, GBC presented best, obtaining 82.9% accuracy and 82.9% ROC, followed by XGBC with 82.4 % accuracy and 82.4% ROC (Table SM16). Further, five-fold cross-validation ordered XGBC with the best mean cross-validation score of 0.8834 with 0.0026 standard deviation which means that the mean accuracy obtained across the five folds was 88.34%. The next in the cross-validation order was XGBC, with a 0.8784 cross-validation score and 0.0028 standard deviations (Table SM17). Scrutinising for overfitting of the XGBC showed that the deviation between the train and test accuracy started being bigger than 5% after the maximum depth of each individual decision tree within the forest, i.e. max_depth was equal to 4 (Figure SM19).

4.3. The hyperparameter tuning of the XGBC performed with grid search (Table SM18) and scrutinised for overfitting (Figure SM20) achieved accuracy of 83%, ROC of 83% and mean cross-validation score: 0.883, respectively. The accuracy after the hyperparameter tuning with Optuna (Table SM19) and scrutinising for overfitting (Figure SM21) was 82.8%, ROC 82.8% and mean cross-validation score: 0.878, respectively.

4.4. The final ML chosen based on the results presented above was the XGBC model with hyperparameter default values and with max_depth equal to 5. The model obtained 83.2 % accuracy, 87.9% precision, 77.1% recall, 82.1% F1, 83.2% ROC. To visualise the results the learning curve, AUC, confusion matrix and classification report are provided in the supplementary materials (Figure SM22, SM23, SM24, Table SM20)

5. CID_SID ML models developed subsequently in other studies:

(i) 78.7% accuracy for predicting whether the compound is a histone-lysine N-methyltransferase (G9a) inhibitor (Ivanova, Russo and Nikolic, 2025a)

(ii) 80.2% accuracy for predicting whether the compound is a Human Dopamine D1 Receptor Antagonist (Ivanova, Russo and Nikolic, 2025b)

(iii) 85.2% accuracy for predicting whether the compound is a human tyrosyl-DNA phosphodiesterase 1 (TDP1) Inhibitors (Ivanova, Russo, Mihaylov and Nikolic, 2025c)

(iv) 81.5% accuracy for predicting whether the compound is a Transthyretin (TTR) transcription activator (Ivanova, Russo, Mihaylov and Nikolic, 2025d)

For more details about these four case studies, please refer to the relevant article.

Conclusion

The methodology presented in the study revealed that the information encoded in the PubChem SIDs and CIDs can be beneficial beyond their identification task. The results obtained by the ML models showed that the methodology can be a time- and cost- effective approach in the early stage of drug discovery. Once, the researcher has obtained the PubChem SID and CID for their new compound, these identifiers will be enough to predict new functionalities of the compound. For a demonstration of the idea and the approach in this study, four use cases were explored which ML models can be used by the researchers in drug discovery directly. Furthermore, the methodology is expected to be applicable to any PubChem bioassay which has significant number of records and well-defined targets that can be useful for ML training and testing.

Scientific contribution

1. A time- and cost-effective ML application, called CID_SID ML model, for predicting a side effect of already designed small biomolecules, using PubChem identifiers.
2. Eight CID_SID ML models can be used directly for predicting whether a small biomolecule has the potential to be:

- A DNA Damage-Inducible Transcript 3 (CHOP) inhibitor.
- A human Dopamine D1 receptor antagonist.
- A human Dopamine D3 receptor antagonist.

- A G9a inhibitor.
- A human M1 muscarinic receptor antagonist.
- A Raab promoter activator.
- A human TDP1 inhibitor.
- A TTR transcription activator.

Acknowledge

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

Author Contributions

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

Funding

MLI thanks the UWL Vice-Chancellor's Scholarship Scheme for their generous support. We sincerely thank PubChem for providing access to their database.

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/CID_SID_ML_model.git

Conflicts of Interest

The authors declare no conflict of interest.

References

Abdalfattah S, Knorz C, Ayoobi A, Omer EA, Rosellini M, Riedl M, Meesters C and Efferth T (2024) Identification of Antagonistic Action of Pyrrolizidine Alkaloids in Muscarinic Acetylcholine Receptor M1 by Computational Target Prediction Analysis. *Pharmaceuticals* 17(1):80. DOI: 10.3390/ph17010080.

Akiba T, Sano S, Yanase T, Ohta T and Koyama M (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. In: *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623-2631. DOI: 10.48550/arXiv.1907.10902

Arnold C, Biedebach L, Küpfer A and Neunhoeffer M (2024). The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*12(4): 841-848. DOI: 10.1017/psrm.2023.61.

Bates S, Hastie T and Tibshirani R (2024). Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association* 119(546), 1434-1445.

Bhujade PR, Shedame KG, Hatwar PR, Bakal RL, Nehar KN and Gawai AY (2024) A Review on Computer Aided Drug Design–In Silico. *Asian Journal of Pharmaceutical Research and Development* 12(6): 80-85. DOI: 10.22270/ajprd. v12i6.1467.

Cerveira A, Kremer F, Lourenço D and Corrêa UB (2024) Evaluation Framework for AI-driven Molecular Design of Multi-target Drugs: Brain Diseases as a Case Study. In: *2024 IEEE Congress on Evolutionary Computation (CEC),* Yokohama, Japan, 30 June – 5 July 2024, pp. 1-8. DOI: 10.1109/CEC60901.2024.10611839.

Chagraoui A, Di Giovanni G, De Deurwaerdère P (2022) Neurobiological and Pharmacological Perspectives of D3 Receptors in Parkinson's Disease. *Biomolecules* 12(2): 243. DOI:10.3390/biom12020243

El-Atawneh S and Goldblum A (2024) A Machine Learning Algorithm Suggests Repurposing Opportunities for Targeting Selected GPCRs. *International Journal of Molecular Sciences* 25(18):10230. DOI: 10.3390/ijms251810230.

Ferraro M, Decherchi S, De Simone A, Recanatini M, Cavalli A and Bottegoni G (2020) Multi-target dopamine D3 receptor modulators: Actionable knowledge for drug design from molecular dynamics and machine learning. *European Journal of Medicinal Chemistry* 188: 111975. DOI: 10.1016/j.ejmech.2019.111975.

Grunze H (2023) The role of the D3 dopamine receptor and its partial agonist cariprazine in patients with schizophrenia and substance use disorder. *Expert Opin Pharmacother* 24(18): 1985-1992. DOI: 10.1080/14656566.2023.2266359.

Hanif I (2020) Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction. In: *Proceedings of the1st International Conference on Statistics and Analytics (ICSA),* 2-3 August 2019, Bogor, Indonesia, p.10110. EAI. DOI: org/10.4108/eai.2-8-2019.2290338.

He H and Garcia EA (2009) Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21(9):1263-1284. DOI: 10.1109/TKDE.2008.239

He Z, Wang J, Jiang M, Hu L and Zou Q (2024) Random subsequence forests. *Information Sciences* 667: 120478. DOI: 10.1016/j.ins.2024.120478.

Ibragimov B and Vakhrushev A (2024) Uplift Modelling via Gradient Boosting. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Association for Computing Machinery, New York, NY, USA, pp.1177–1187. DOI: 10.1145/3637528.3672019.

Imbalanced Learn (2024) *RandomOverSapmpler.* Available at https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.RandomOverSampler.html (accessed: 04 Jan 2025)

Ivanova ML, Russo N, Djaid N and Nikolic K (2024) Application of machine learning for predicting G9a inhibitors. *Digital Discovery* 3(10):2010-2018. DOI: 10.1039/d4dd00101j.

Ivanova ML, Russo N and Nikolic K (2025a) Targeting Neurodegeneration: Three Machine Learning Methods for Discovering G9a Inhibitors Using PubChem and Scikit-Learn, *ArXiv*. DOI: 10.48550/arXiv.2503.16214

Ivanova ML, Russo N and Nikolic K (2025b) Leveraging 13C NMR Spectrum Data Derived from SMILES for Machine Learning-based Prediction of a Small Molecule Functionality: A Case Study on Human Dopamine D1 Receptor Antagonists. *ArXiv*. DOI: 10.48550/arXiv.2501.14044

Ivanova ML, Russo N, Mihaylov G and Nikolic K (2025c) Hierarchical Functional Group Ranking via IUPAC Name Analysis for Drug Discovery: A Case Study on TDP1 Inhibitors*, ArXiv,* DOI: 10.48550/arXiv.2503.05591

Ivanova ML, Russo N, Mihaylov G and Nikolic K (2025d) Comparative Analysis of Computational Approaches for Predicting Transthyretin Transcription Activators and Human Dopamine D1 Receptor Antagonists. *ArXiv.* DOI: 10.48550/arXiv.2506.01137

Jain R, Subramanian J and Rathore AS (2022) A review of therapeutic failures in late-stage clinical trials. *Expert Opin. Pharmacother*. 24(3):389–399. DOI:10.1080/14656566.2022.2161366

Jordan KL (2024) *Investigating the pathogenic roles of Rab GTPases in models of tauopathy*. PhD Thesis, University of Leicester.

Jupyter (2024) *Jupyter*. Available at: https://jupyter.org/ (accessed 4 Jan 2025)

Kabir MD, Ahmed MU, Begum S, Barua S, and Islam MR (2024) Balancing Fairness: Unveiling the Potential of SMOTE-Driven Oversampling in AI Model Enhancement. In *Proceedings of the 2024 9th International Conference on Machine Learning Technologies (ICMLT '24)*. Association for Computing Machinery, New York, NY, USA, 21–29. DOI: org/10.1145/3674029.3674034.

Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2016) PubChem Substance and Compound databases. *Nucleic Acids Research* 44(D1):D1202-13. DOI: 10.1093/nar/gkv951.

Kingwell K (2024) Muscarinic drugs breathe new life into schizophrenia pipeline. *Nature reviews. Drug discovery* 23(9):647-649. DOI: 10.1038/d41573-024-00129-w.

Kuang ZK, Feng SY, Hu B, Wang D, He SB and Kong DX (2016) Predicting subtype selectivity of dopamine receptor ligands with three-dimensional biologically relevant spectrum. *Chemical Biology & Drug Design* 88(6): 859-872. DOI:

Lee J, Sim MK and Hong JS (2024). Assessing Decision Tree Stability: A Comprehensive Method for Generating a Stable Decision Tree. *IEEE Access*. vol. 12, pp. 90061-90072, 2024, DOI: 10.1109/ACCESS.2024.3419228.

Metz CN, Brines M and Pavlov VA (2024) Bridging cholinergic signalling and inflammation in schizophrenia. *Schizophrenia*10(1):51. DOI: 10.1038/s41537-024-00472-2.

Monaco M, Trebesova H, Grilli M (2024) Muscarinic Receptors and Alzheimer's Disease: New Perspectives and Mechanisms. *Current Issues in Molecular Biology* 46(7): 6820-6835. DOI: 10.3390/cimb46070407.

Niazi SK and Mariam Z (2025) Artificial intelligence in drug development: reshaping the therapeutic landscape. *Ther. Adv. Drug Saf*. 16. DOI: 10.1177/204209862513217

NIH (2010a) *qHTS Assay for Rab9 Promoter Activators*. Available at: https://pubchem.ncbi.nlm.nih.gov/bioassay/485297 (accessed 4 Jan 2025)

NIH (2010b) *HTS for small molecule inhibitors of CHOP to regulate the unfolded protein response to ER stress.* Available at: https://pubchem.ncbi.nlm.nih.gov/bioassay/2732 (accessed 4 Jan 2025)

NIH (2010c) *Aqueous Solubility from MLSMR Stock Solutions*. Available at: https://pubchem.ncbi.nlm.nih.gov/bioassay/1996 (accessed 4 Jan 2025)

NIH (2012) *Fluorescence-based cell-based primary high throughput screening assay to identify antagonists of the human M1 muscarinic receptor (CHRM1).* Available at: https://pubchem.ncbi.nlm.nih.gov/bioassay/588852 (accessed 4 Jan 2025)

NIH (2013*) qHTS of D3 Dopamine Receptor Antagonist*: qHTS. Available at: https://pubchem.ncbi.nlm.nih.gov/bioassay/652054 (accessed 4 Jan 2025)

NIH (2025) *PubChem.* Available at: https://pubchem.ncbi.nlm.nih.gov/ (accessed 4 Jan 2025)

Opitz J (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics* 12: 820-836. DOI: 10.1162/tacl_a_00675

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12(10): 2825–2830.

Saha R, Chauhan A and Rastogi Verma S (2024) Machine learning: an advancement in biochemical engineering. *Biotechnology Letters* 46: 1-23. DOI: 10.1007/s10529-024-03499-8.

Schöbi D, Homberg F, Frässle S, Endepols H, Moran RJ, Friston KJ and Stephan KE (2021) Model-based prediction of muscarinic receptor function from auditory mismatch negativity responses. *NeuroImage* 237: 118096. DOI: 10.1016/j.neuroimage.2021.118096.

Shin J and Shin SJ (2024) A concise overview of principal support vector machines and its generalization. *Communications for Statistical Applications and Methods* 31(2): 235-246. DOI: 10.29220/CSAM.2024.31.2.235

Soon WW, Hariharan M and Snyder MP (2013) High-throughput sequencing for biology and medicine. *Molecular systems biology* 9(1): 640. DOI: 10.1038/msb.2012.61.

Vinuesa R, Rabault J, Azizpour H, Bauer S, Brunton BW, Elofsson A and Brunton SL (2024) Opportunities for machine learning in scientific discovery. *ArXiv*, abs/2405.04161.

Ying X (2019) An overview of overfitting and its solutions. *Journal of physics: Conference series* 1168: 022022. DOI: 10.1088/1742-6596/1168/2/022022.

Zhou, Z. H. (2021) *Machine learning*. Springer nature.

Zhao LX, Ge YH, Xiong CH, Tang L, Yan YH, Law PY, Qiu Y, Chen HZ. (2018) M1 muscarinic receptor facilitates cognitive function by interplay with AMPA receptor GluA1 subunit. *FASEB J* 32(8): 4247-4257. DOI: 10.1096/fj.201800029R

Supplementary material

1. Tables

Table SM1. Metrics evaluating the ML model predicting dopamine D3 receptor`s antagonists,

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
XGBoost	0.856	0.910	0.789	0.845	0.856
RandomForest	0.851	0.920	0.768	0.837	0.851
GradientBoost	0.851	0.908	0.782	0.840	0.851
Decision	0.831	0.883	0.764	0.819	0.831
K-nearest	0.828	0.843	0.807	0.825	0.828
SVM	0.808	0.904	0.690	0.783	0.808

Table SM2. Five-fold Cross-validation of the ML model predicting dopamine D3 receptor`s antagonists.

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8851	0.0027	[0.8815, 0.8879, 0.8883, 0.8824, 0.8852]
XGBoost	0.8850	0.0033	[0.885, 0.8815, 0.8907, 0.8817, 0.8861]
K-nearest	0.8758	0.0036	[0.8704, 0.8779, 0.8814, 0.8744, 0.8751]
GradientBoost	0.8731	0.0047	[0.8794, 0.8666, 0.8768, 0.869, 0.8739]
Decision	0.8511	0.0063	[0.8424, 0.8568, 0.8591, 0.8462, 0.8513]
SVM	0.8435	0.0033	[0.8455, 0.8373, 0.8456, 0.8427, 0.8463]

Table SM3. Grid search parameters used for hyperparameter tuning of the ML model predicting dopamine D3 receptor`s antagonists.

```
'max_depth': [4, 5, 7],
'learning_rate': [0.01, 0.1, 0.3],
'n_estimators': [100, 200, 300],
'subsample': [0.8, 1.0],
'colsample_bytree': [0.8, 1.0]
```

Table SM4. Optuna parameters used for hyperparameter tuning of the ML model predicting dopamine D3 receptor`s antagonists.

```
(max_depth=8,
 colsample_bytree=0.8076115597124519,
 learning_rate=0.05133112708737439,
 min_child_weight=3,
 n_estimators=356,
 subsample=0.8537278365573938,
 reg_lambda=0.018370514156021697,
 gamma=0.47079566916356813)
```

Table SM5. Classification report of the ML model predicting dopamine D3 receptor`s antagonists.

	precision	recall	f1-score	support
Active (target 1)	0.82	0.92	0.87	2750
Inactive (target 0)	0.91	0.80	0.85	2750
accuracy			0.86	5500
macro avg	0.86	0.86	0.86	5500
weighted avg	0.86	0.86	0.86	5500

Table SM6. Metrics evaluating the ML model predicting Rab9 promoter`s activators

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
RandomForest	0.757	0.837	0.640	0.725	0.758
XGBoost	0.755	0.842	0.628	0.719	0.755
GradientBoost	0.749	0.871	0.584	0.699	0.749
K-nearest	0.733	0.774	0.658	0.712	0.733
Decision	0.726	0.769	0.646	0.702	0.726
SVM	0.722	0.976	0.455	0.621	0.722

Table SM7. Five-fold Cross-validation of the ML model predicting Rab9 promoter`s activators.

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
XGBoost	0.8429	0.0032	[0.8439, 0.8368, 0.8452, 0.8456, 0.8431]
GradientBoost	0.8415	0.0038	[0.8386, 0.8364, 0.847, 0.8439, 0.8414]
RandomForest	0.8332	0.0048	[0.831, 0.8252, 0.8342, 0.8389, 0.8367]
SVM	0.8192	0.0035	[0.8175, 0.8139, 0.824, 0.8221, 0.8188]
K-nearest	0.8177	0.0054	[0.8166, 0.8108, 0.8226, 0.8252, 0.8136]
Decision	0.7849	0.0058	[0.7856, 0.7743, 0.7842, 0.7893, 0.7908]

Table SM8. Grid search parameter used for hyperparameter tuning of the ML model predicting Rab9 promoter`s activators

```
'max_depth': [4, 5, 6],
'learning_rate': [0.01, 0.1, 0.3],
'n_estimators': [100, 200, 300],
'subsample': [0.8, 1.0],
'colsample_bytree': [0.8, 1.0]
```

Table SM9. Optuna parameters used for hyperparameter tuning of the ML model predicting Rab9 promoter`s activators.

```
(max_depth=3,
 colsample_bytree=0.918671452698546,
 learning_rate=0.06934031885720114,
 min_child_weight=2,
 n_estimators=489,
 subsample=0.7923504306755987,
 reg_lambda=0.0017504557811672502,
 gamma=0.7884126281466084)
```

Table SM10. Classification report of the ML model predicting Rab9 promoter`s activators.

	precision	recall	f1-score	support
Active (target 1)	0.70	0.89	0.78	3200
Inactive (target 0)	0.85	0.62	0.72	3200
accuracy			0.75	6400
macro avg	0.77	0.75	0.75	6400
weighted avg	0.77	0.75	0.75	6400

Table SM11. Metrics, evaluating the ML model predicting CHOP`s inhibitors.

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.899	0.984	0.812	0.890	0.899
XGBoost	0.896	0.960	0.827	0.888	0.896
RandomForest	0.893	0.966	0.814	0.884	0.893
SVM	0.884	0.981	0.784	0.871	0.884
Decision	0.878	0.950	0.798	0.867	0.878
K-nearest	0.877	0.912	0.836	0.872	0.877

Table SM12. Five-fold Cross-validation of the ML model predicting CHOP`s inhibitors

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
GradientBoost	0.9414	0.0005	[0.9414, 0.9419, 0.9419, 0.9414, 0.9406]
XGBoost	0.9407	0.0012	[0.9402, 0.943, 0.9394, 0.9402, 0.9406]
RandomForest	0.9361	0.0018	[0.9373, 0.9386, 0.9332, 0.9354, 0.9359]
K-nearest	0.9315	0.0013	[0.9319, 0.9305, 0.9295, 0.9332, 0.9322]
SVM	0.9095	0.0015	[0.9112, 0.9076, 0.9104, 0.9106, 0.9078]
Decision	0.9068	0.0031	[0.9072, 0.9107, 0.9041, 0.9024, 0.9093]

Table SM13. Grid search parameters used for hyperparameter tuning of the ML model predicting CHOP`s inhibitors.

```
'n_estimators': [50, 100, 200],
'learning_rate': [0.01, 0.1, 0.2],
'max_depth': [3, 5, 7],
'min_samples_split': [2, 5, 10],
'min_samples_leaf': [1, 2, 4]
```

Table SM14. Optuna parameters used for hyperparameter tuning of the ML model predicting CHOP`s inhibitors.

```
max_depth=i,
min_samples_leaf=1,
learning_rate=0.2463083429915664,
min_samples_split=11,
n_estimators=179)
```

Table SM15. Classification report of the ML model predicting CHOP`s inhibitors

	precision	recall	f1-score	support
Active (target 1)	0.84	0.99	0.91	3243
Inactive (target 0)	0.98	0.82	0.89	3243
accuracy			0.90	6486
macro avg	0.91	0.90	0.90	6486
weighted avg	0.91	0.90	0.90	6486

Table SM16. Metrics evaluating the ML model predicting M1 muscarinic receptor`s antagonists.

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.829	0.872	0.771	0.819	0.829
XGBoost	0.824	0.868	0.765	0.813	0.824
RandomForest	0.799	0.913	0.662	0.767	0.799
SVM	0.798	0.859	0.713	0.779	0.798
K-nearest	0.797	0.839	0.735	0.784	0.797
Decision	0.785	0.899	0.641	0.749	0.785

Table SM17. Five-fold Cross-validation of the ML model predicting M1 muscarinic receptor`s antagonists

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
XGBoost	0.8834	0.0026	[0.8823, 0.8844, 0.8791, 0.8868, 0.8841]
GradientBoost	0.8784	0.0028	[0.8738, 0.8815, 0.8777, 0.878, 0.8812]
RandomForest	0.8758	0.0017	[0.877, 0.8754, 0.8767, 0.8727, 0.8775]
K-nearest	0.8712	0.0032	[0.8754, 0.8674, 0.8676, 0.8732, 0.8727]
SVM	0.8463	0.0022	[0.8433, 0.8468, 0.847, 0.8444, 0.8497]
Decision	0.8393	0.0037	[0.8409, 0.834, 0.8406, 0.8364, 0.8444]

Table SM18. Grid search parameters used for hyperparameter tuning of the ML model predicting M1 muscarinic receptor`s antagonists.

```
'max_depth': [3, 4, 5],
'learning_rate': [0.01, 0.1, 0.3],
'n_estimators': [100, 200, 300],
'subsample': [0.8, 1.0],
'colsample_bytree': [0.8, 1.0]
```

Table SM19. Optuna parameters used for hyperparameter tuning of the ML model predicting M1 muscarinic receptor`s antagonists.

```
max_depth=10,
colsample_bytree=0.8642446766992792,
learning_rate=0.016930731564735818,
min_child_weight=4,
n_estimators=260,
subsample=0.9558468050087313,
reg_lambda=0.16515384724293172,
gamma=0.4439768538612002)
```

Table SM20. Classification report of the ML model predicting M1 muscarinic receptor`s antagonists.

	precision	recall	f1-score	support
Active (target 1)	0.80	0.89	0.84	1880
Inactive (target 0)	0.88	0.77	0.82	1880
accuracy			0.83	3760
macro avg	0.84	0.83	0.83	3760
weighted avg	0.84	0.83	0.83	3760

2. Figures

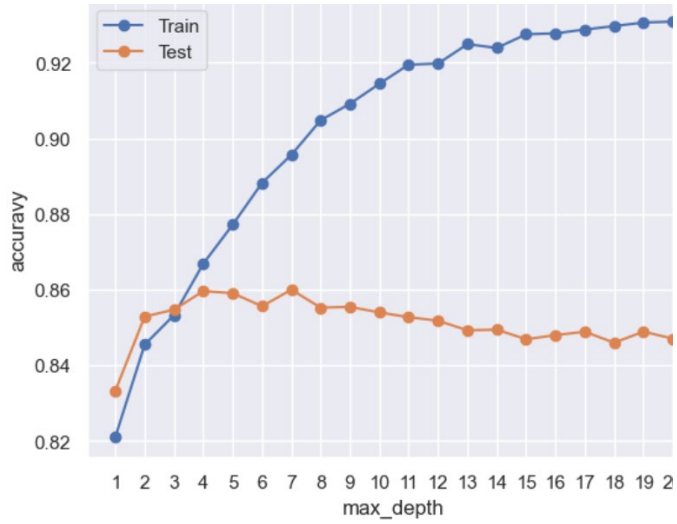


Figure SM1.Scrutinising for overfitting of the ML model predicting dopamine D3 receptor`s antagonists, which is with default values of its hyperparameters.

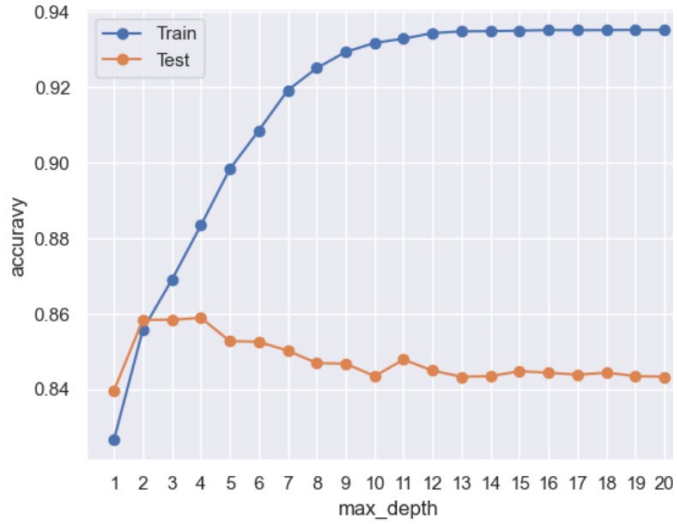


Figure SM2. Scrutinising for overfitting of the ML model predicting dopamine D3 receptor`s antagonists, which is a hyperparameter tuned by grid search.

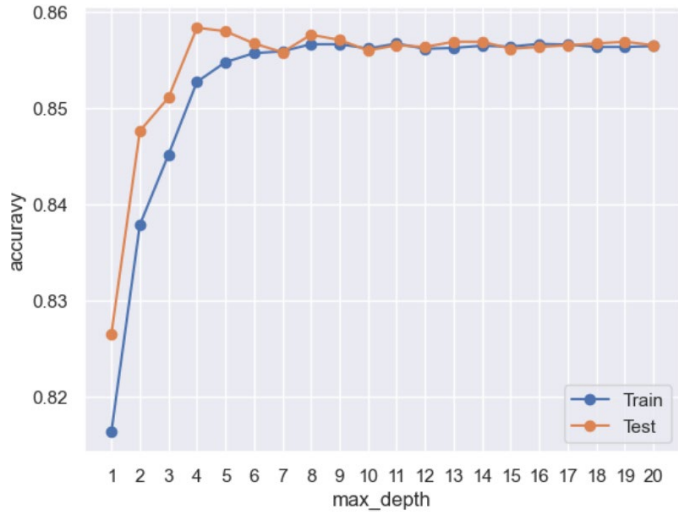


Figure SM3. Scrutinising for overfitting of the ML model predicting dopamine D3 receptor`s antagonists, which is a hyperparameter tuned by Optuna.

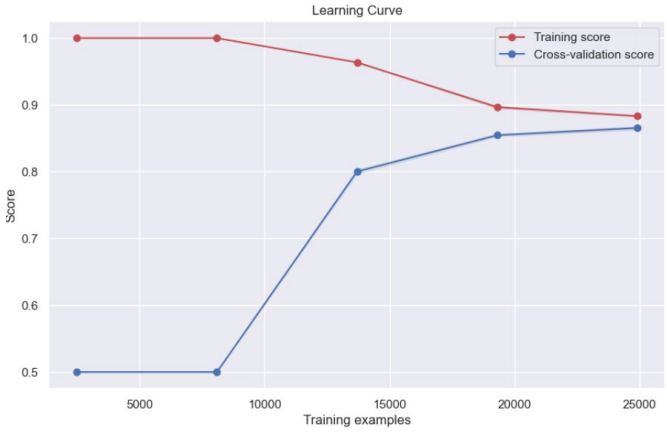


Figure SM4. Learning curve of the CID-SID ML model predicting dopamine D3 receptor`s antagonists.

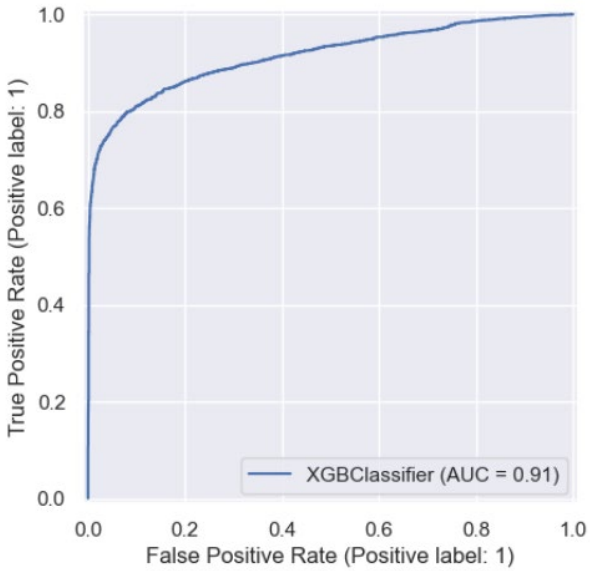


Figure SM5. ROC of the CID-SID ML model predicting dopamine D3 receptor`s antagonists.

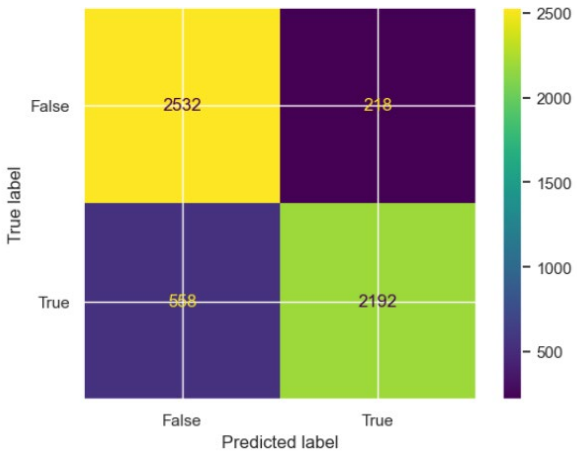


Figure SM6. Confusion matrix of the CID-SID ML model predicting dopamine D3 receptor `s antagonists.

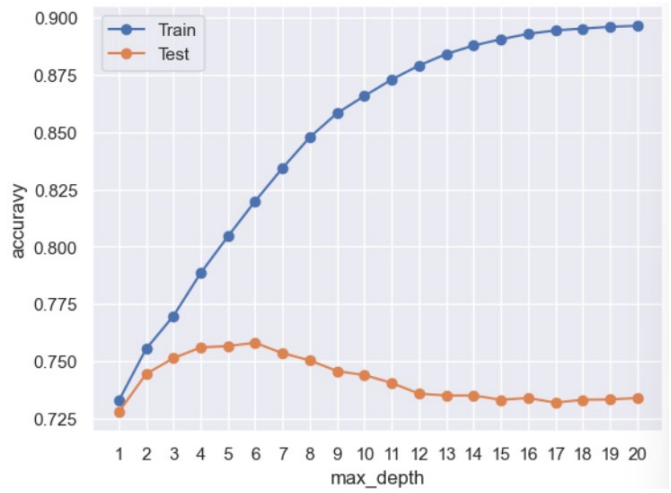


Figure SM7. Scrutinising for overfitting of the ML model predicting Rab9 promoter`s activators which is with default values of its hyperparameters.

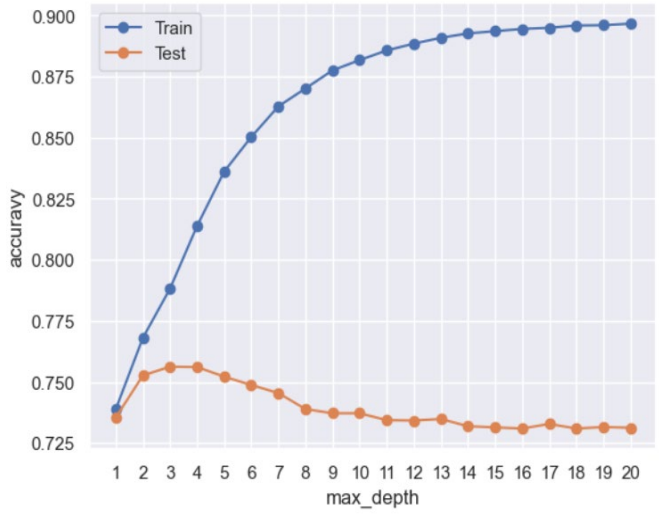


Figure SM8. Scrutinising for overfitting of the ML model predicting Rab9 promoter`s activators, which is a hyperparameter tuned by grid search.

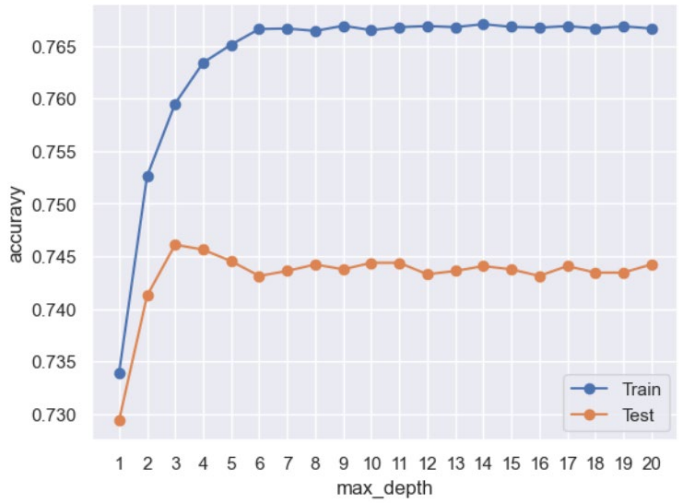


Figure SM9. Scrutinising for overfitting of the ML model predicting Rab9 promoter`s activators, which is a hyperparameter tuned by Optuna.

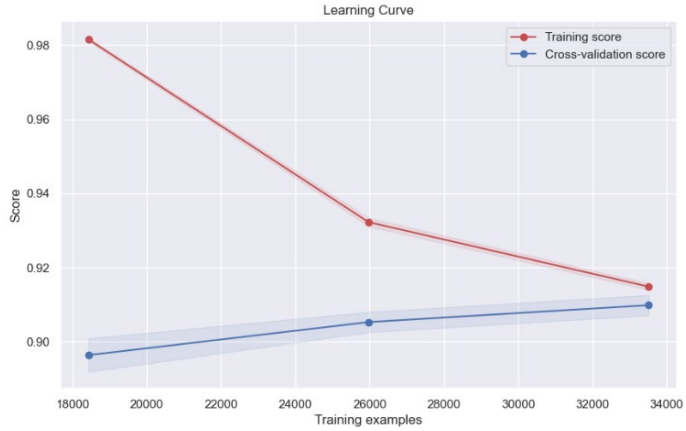


Figure SM 10. Learning curve of the CID-SID ML model predicting Rab9 promoter`s activators

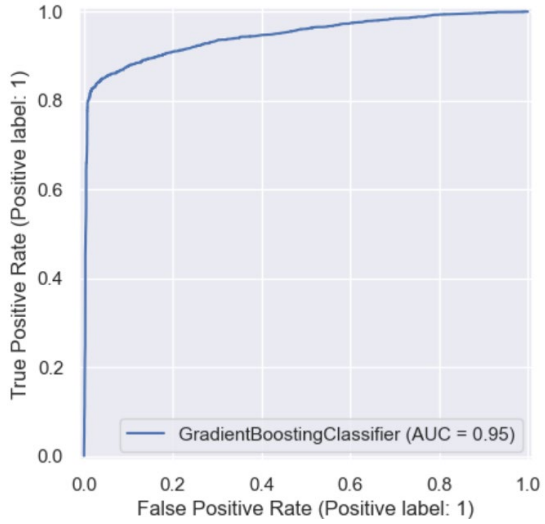


Figure SM11. ROC of the CID-SID ML model predicting Rab9 promoter`s activators

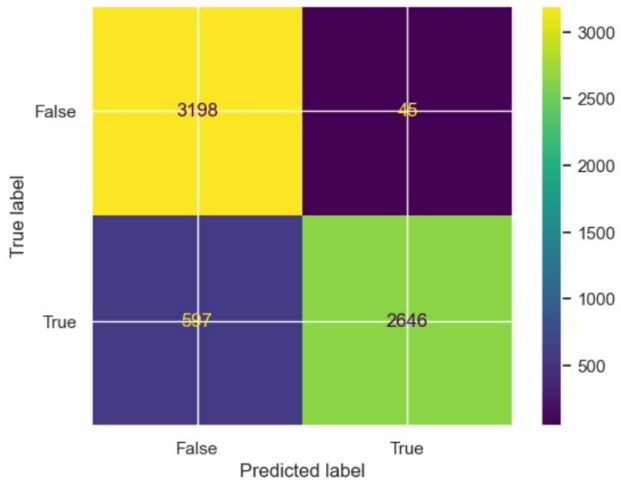


Figure SM12. Confusion matrix of the CID-SID ML model predicting Rab9 promoter`s activators

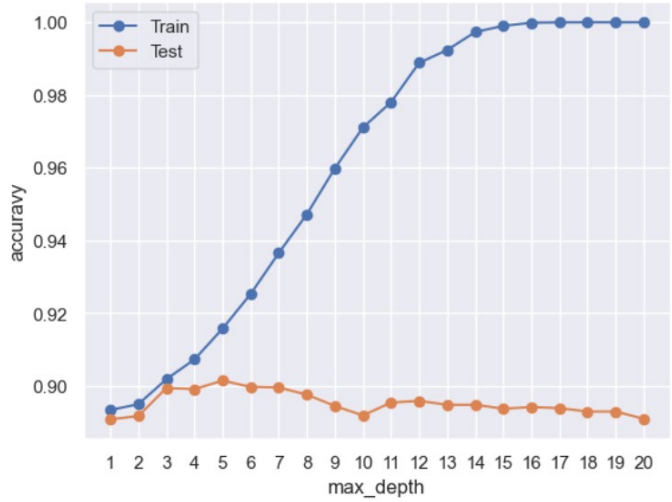


Figure SM 13. Scrutinising for overfitting of the ML model predicting CHOP`s inhibitors, which is with default values of its hyperparameters.

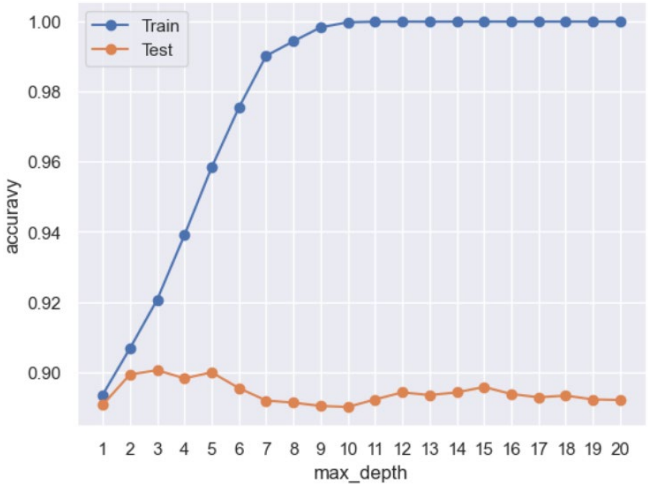


Figure SM14. Scrutinising for overfitting of the ML model predicting CHOP`s inhibitors., which is a hyperparameter tuned by grid search.

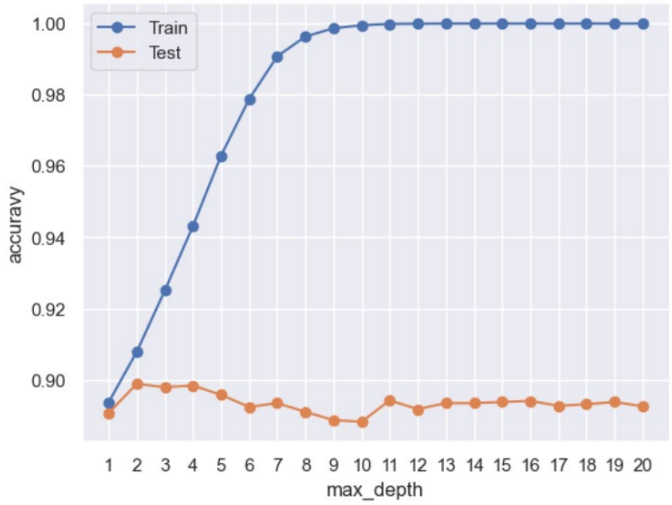


Figure SM15. Scrutinising for overfitting of the ML model predicting CHOP`s inhibitors, which is a hyperparameter tuned by Optuna.

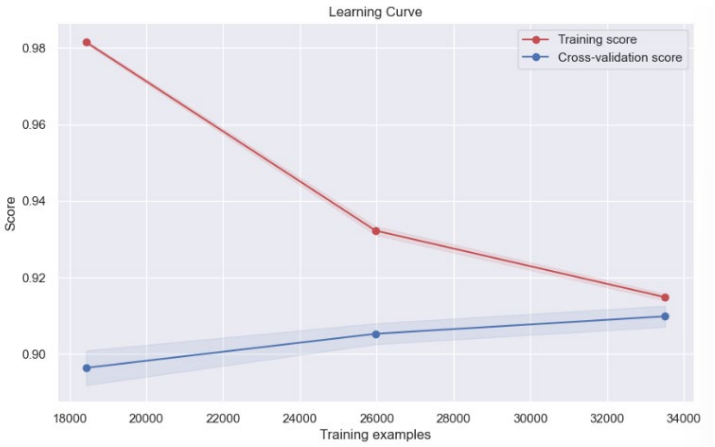


Figure SM 16. Learning curve of the CID-SID ML model predicting CHOP`s inhibitors.

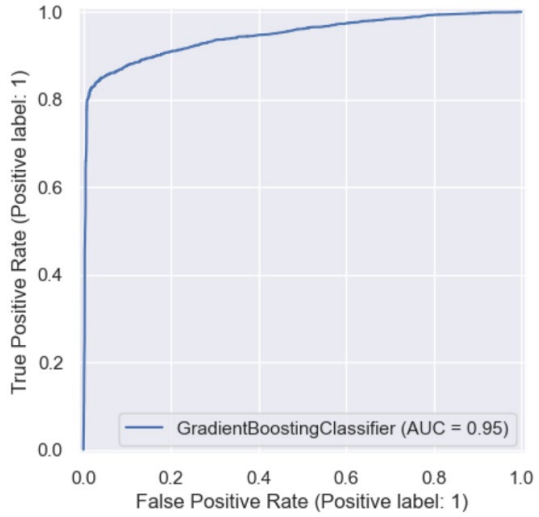


Figure SM17. ROC of the CID-SID ML model predicting CHOP`s inhibitors.

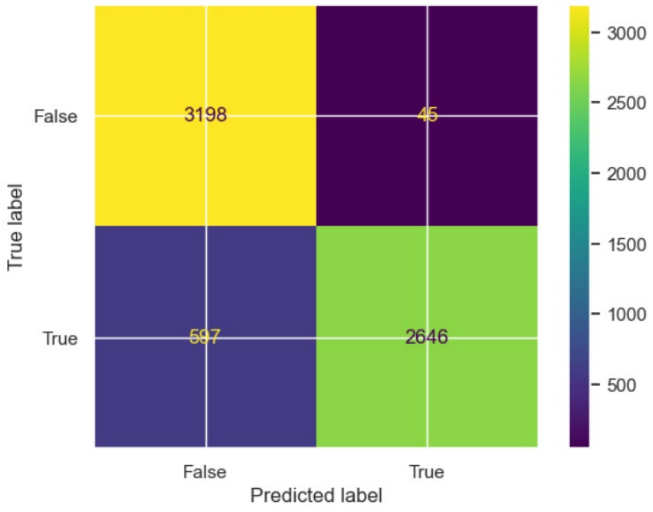


Figure SM18. Confusion matrix of the CID-SID ML model predicting CHOP`s inhibitors.

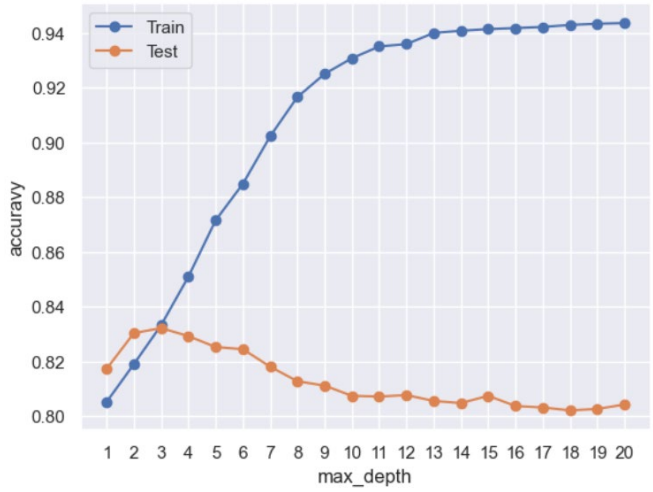


Figure SM19. Scrutinising for overfitting of the ML model predicting M1 muscarinic receptor`s antagonists, which is with default values of its hyperparameters.

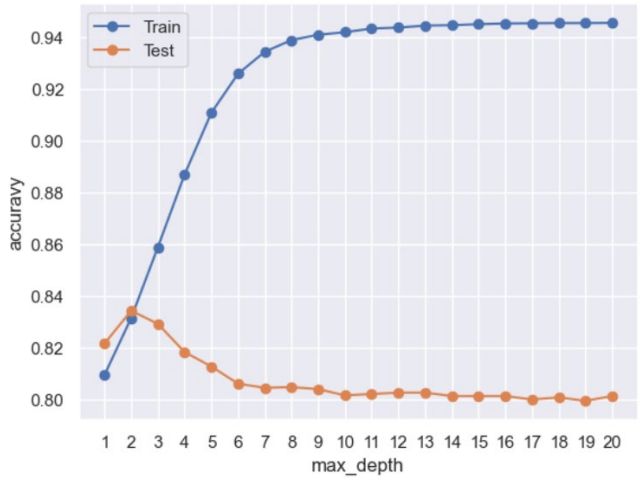


Figure SM20. Scrutinising for overfitting of the ML model predicting M1 muscarinic receptor`s antagonists, which is a hyperparameter tuned by grid search.

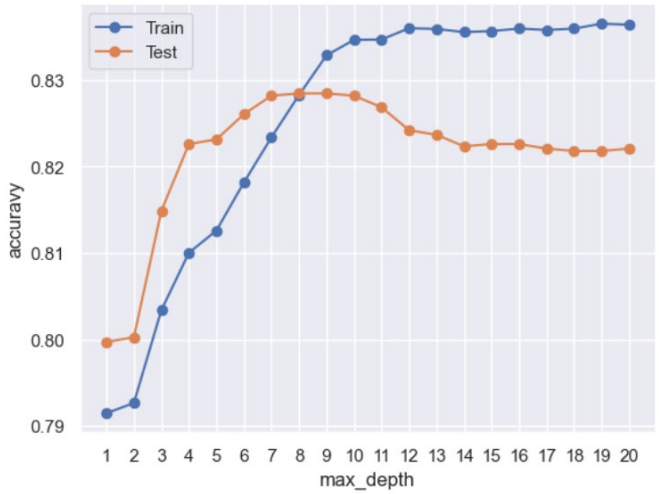


Figure SM21. Scrutinising for overfitting of the ML model predicting M1 muscarinic receptor`s antagonists, which is a hyperparameter tuned by Optuna.

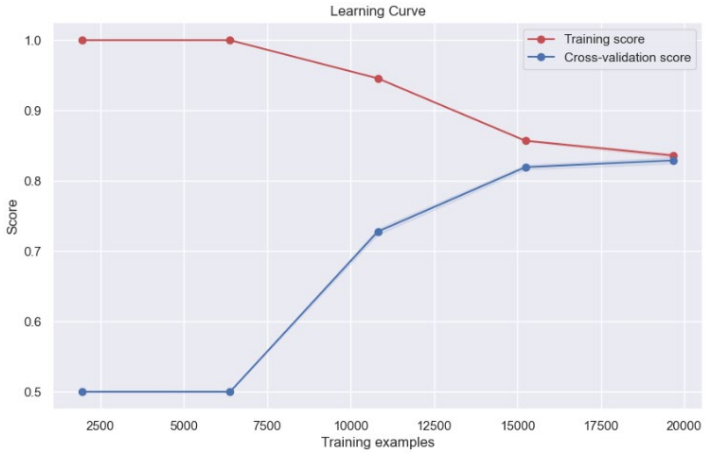


Figure SM 22. Learning curve of the CID-SID ML model predicting M1 muscarinic receptor`s antagonists.

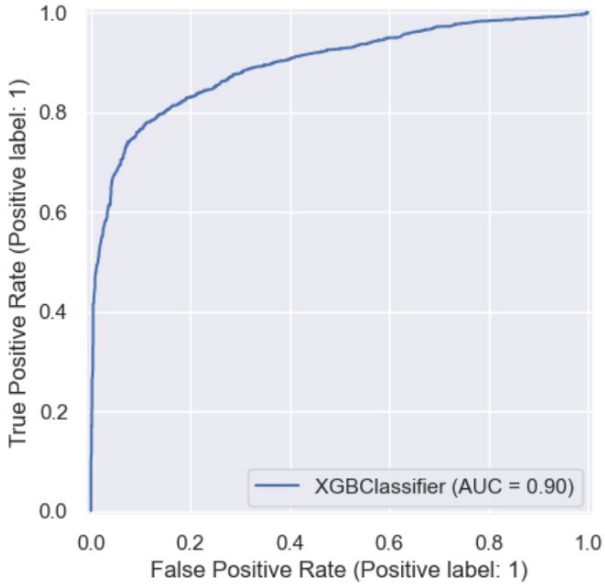


Figure SM23. ROC of the CID-SID ML model predicting M1 muscarinicr eceptor`s antagonists.

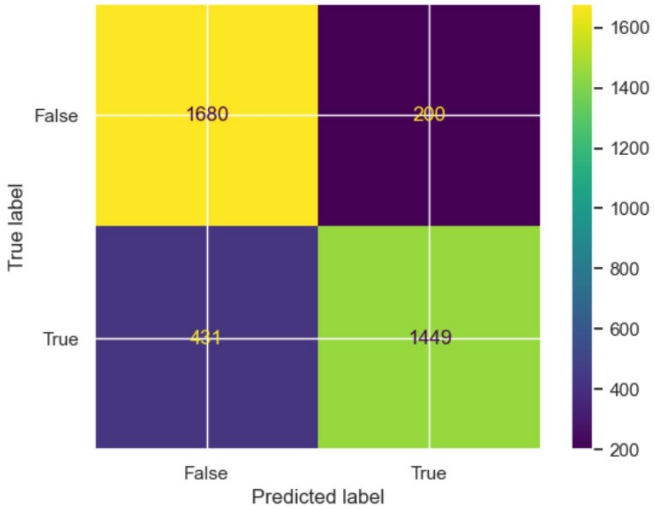


Figure SM24. Confusion matrix of the CID-SID ML model predicting M1 muscarinic receptor`s antagonists.