Comparative analysis of computational approaches for predicting human neuronal Transthyretin (TTR) transcription activators and human dopamine D1 receptor antagonists

Supplementary materials

Comparative analysis of computational approaches for predicting human neuronal Transthyretin (TTR) transcription activators and human dopamine D1 receptor antagonists

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

Tables

Table ESM1. Results of the ML metrics regarding the TTR case performed without dimensionality reduction (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR/4.1.ML_notPCA.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.568	0.592	0.435	0.502	0.568
SVM	0.560	0.596	0.374	0.459	0.560
K-nearest	0.550	0.596	0.312	0.409	0.550
RandomForest	0.538	0.691	0.138	0.230	0.538
Decision	0.521	0.527	0.403	0.457	0.521

Table ESM2. A five-fold cross-validation for the TTR case performed without dimensionality reduction (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR/4.1.ML_notPCA.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8637	0.0612	[0.7923, 0.8039, 0.8559, 0.9465, 0.9198]
Decision	0.7573	0.0352	[0.7285, 0.7221, 0.737, 0.8083, 0.7905]
SVM	0.7288	0.0307	[0.6899, 0.7043, 0.7311, 0.7786, 0.74]
K-nearest	0.6875	0.0431	[0.6543, 0.6256, 0.6895, 0.7355, 0.7325]
GradientBoost	0.6509	0.0195	[0.6365, 0.6627, 0.6196, 0.6686, 0.6672]

Table ESM3. Results of the ML metrics regarding the TTR case performed with dimensionality reduction with PCA (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR/4.2..ML_withPCA.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.584	0.654	0.356	0.461	0.584
SVM	0.572	0.626	0.359	0.456	0.572
Decision	0.563	0.597	0.388	0.471	0.563
K-nearest	0.534	0.551	0.368	0.441	0.534
RandomForest	0.531	0.769	0.088	0.158	0.531

Table ESM4. A five-fold cross-validation for the TTR case performed with dimensionality reduction achieved with PCA (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR/4.2..ML_withPCA.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8833	0.0568	[0.8145, 0.8262, 0.8826, 0.9554, 0.9376]
Decision	0.7834	0.0361	[0.7329, 0.7533, 0.786, 0.8187, 0.8262]
GradientBoost	0.7448	0.0337	[0.6988, 0.7147, 0.7489, 0.786, 0.7756]
SVM	0.7374	0.0391	[0.7122, 0.6895, 0.7207, 0.7964, 0.7682]
K-nearest	0.6661	0.0458	[0.6128, 0.6196, 0.6627, 0.7266, 0.7088]

Table ESM5. Results of the ML metrics regarding the dopamine D1 receptor antagonist predictions using a dataset with reduced number of samples and without dimensionality reduction (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/4.1.ML_noPCA_3177.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.657	0.703	0.544	0.614	0.657
GradientBoost	0.650	0.692	0.541	0.607	0.650
RandomForest	0.601	0.822	0.259	0.394	0.601
Decision	0.571	0.588	0.471	0.523	0.571
K-nearest	0.512	0.520	0.309	0.387	0.512

Table ESM6. A five-fold cross-validation of ML model predicting the dopamine D1 receptor antagonists based on a reduce dataset samples and without dimensionality reduction (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/4.1.ML_noPCA_3177.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8817	0.0495	[0.847, 0.8083, 0.8856, 0.9242, 0.9435]
Decision	0.7809	0.0334	[0.7533, 0.7519, 0.7652, 0.7935, 0.8408]
SVM	0.7753	0.0291	[0.7504, 0.7385, 0.7816, 0.7845, 0.8214]
GradientBoost	0.7045	0.0217	[0.6805, 0.6805, 0.7058, 0.7207, 0.7351]
K-nearest	0.6938	0.0287	[0.6746, 0.6612, 0.6776, 0.7207, 0.7351]

Table ESM7. Results of the ML metrics regarding the dopamine D1 receptor antagonist predictions using a dataset with reduced number of samples and PCA dimensionality reduction (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/4.2.ML_withPCA_3177.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.647	0.692	0.529	0.600	0.647
GradientBoost	0.600	0.705	0.344	0.462	0.600
Decision	0.537	0.559	0.350	0.430	0.537
RandomForest	0.534	0.677	0.129	0.217	0.534
K-nearest	0.519	0.532	0.321	0.400	0.519

Table ESM8. A five-fold cross-validation of the ML models predicting the dopamine D1 receptor antagonists based on data with reduced samples and PCA dimensionality reduction of the features (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/4.2.ML_withPCA_3177.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8764	0.0565	[0.8306, 0.7994, 0.8767, 0.9227, 0.9524]
Decision	0.7854	0.0537	[0.7801, 0.7266, 0.7296, 0.8276, 0.8631]
GradientBoost	0.7747	0.0522	[0.7281, 0.7073, 0.7831, 0.8024, 0.8527]
SVM	0.7485	0.0362	[0.7221, 0.6969, 0.7637, 0.7578, 0.8021]
K-nearest	0.6917	0.0226	[0.6672, 0.6657, 0.6969, 0.7043, 0.7247]

Table ESM9 Metrics results of ML models based on a full dataset of 13CNMR spectroscopy data, predicting dopamine D1 receptor antagonists(GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/6.1.ML_noPCA_fullSet.ipynb)

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

Table ESM10. Five-fold cross-validation of ML based on a full dataset of 13CNMR spectroscopy data, predicting dopamine D1 receptor antagonists

https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/6.1.ML_noPCA_fullSet.ipynb)

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

Table ESM11. Metric results of ML with a full dataset of 13CNMR data, dimensionality reduced by PCA for predicting dopamine D1 receptor antagonists (GitHub file:

https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/6.2..ML_withPCA_fullSet.ipynb

)

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 ESM12. Five-fold cross-validation of ML based on a full dataset with 13CNMR, reduced by PCA for predicting dopamine D1 receptor antagonists (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/6.2..ML_withPCA_fullSet.ipynb)

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 ESM13. Metric results of ML models, predicting TTR transcription antagonists based on a dataset with added molecular features (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR/5.1.ML_noPCA_withPubChem_data.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.671	0.740	0.526	0.615	0.671
RandomForest	0.656	0.794	0.421	0.550	0.656
SVM	0.604	0.671	0.409	0.508	0.604
Decision	0.596	0.638	0.441	0.522	0.596
K-nearest	0.547	0.562	0.429	0.487	0.547

Table ESM14. A five-fold cross-validation for predicting TTR transcription antagonists based on a dataset with added molecular features (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR/5.1.ML_noPCA_withPubChem_data.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8899	0.0369	[0.8556, 0.8507, 0.8756, 0.9331, 0.9347]
Decision	0.8032	0.0429	[0.75, 0.7792, 0.7792, 0.86, 0.8476]
SVM	0.7245	0.0272	[0.6879, 0.7045, 0.7201, 0.7496, 0.7605]
GradientBoost	0.6931	0.0186	[0.6599, 0.6874, 0.6983, 0.7123, 0.7076]
K-nearest	0.6645	0.0499	[0.6335, 0.6112, 0.6345, 0.7465, 0.6967]

Table ESM15. ML metric results of ML models predicting TTR transcription activators based on a dataset with 13C NMR and molecular feature and dimensionality reduced by PCA (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR/5.2..ML_withPCA_withPubChem_data.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.650	0.707	0.512	0.594	0.650
GradientBoost	0.604	0.625	0.521	0.568	0.604
K-nearest	0.534	0.536	0.503	0.519	0.534
Decision	0.519	0.528	0.359	0.427	0.519
RandomForest	0.519	0.528	0.359	0.427	0.519

Table ESM16. A five-fold cross-validation for predicting TTR transcription activators based on a dataset with 13C NMR and molecular feature and dimensionality reduced by PCA (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR/5.2..ML_withPCA_withPubChem_data.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
GradientBoost	0.7172	0.0127	[0.7061, 0.7023, 0.7138, 0.7352, 0.7286]
RandomForest	0.7146	0.0099	[0.7126, 0.7072, 0.7023, 0.7303, 0.7204]
SVM	0.6952	0.0093	[0.6847, 0.6891, 0.6908, 0.7007, 0.7105]
K-nearest	0.6449	0.0203	[0.6338, 0.6382, 0.6266, 0.6414, 0.6842]
Decision	0.6090	0.0092	[0.5944, 0.6102, 0.6168, 0.6036, 0.6201]

Table ESM17. ML metric results based on a reduced dataset of 13CNMR and molecular features for predicting dopamine D1 receptor antagonists (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/5.1.ML_noPCA_3177_withPubChem_data.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.751	0.835	0.626	0.716	0.751
RandomForest	0.746	0.871	0.576	0.694	0.746
SVM	0.690	0.774	0.535	0.633	0.690
Decision	0.659	0.694	0.568	0.625	0.659
K-nearest	0.594	0.654	0.400	0.496	0.594

Table ESM18. Five-fold cross-validation of ML based on a reduced dataset with 13CNMR and molecular features for predicting dopamine D1 receptor antagonists (GitHub file:

https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/5.1.ML_noPCA_3177_withPubChem_data.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.9138	0.0287	[0.9094, 0.8692, 0.9004, 0.945, 0.9449]
Decision	0.8404	0.0303	[0.8009, 0.8083, 0.8499, 0.8707, 0.872]
SVM	0.8166	0.0190	[0.7964, 0.792, 0.8262, 0.8276, 0.8408]
GradientBoost	0.7878	0.0128	[0.7875, 0.7637, 0.792, 0.8009, 0.7946]
K-nearest	0.7331	0.0507	[0.6835, 0.6627, 0.7533, 0.7771, 0.7887]

Table ESM19.ML metric results based on ML based on a reduced dataset with 13CNMR data and molecular features, dimensionality reduced by PCA for predicting dopamine D1 receptor antagonists (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/7.2..ML_withPCA_fullSet_withPubChem_data.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.704	0.833	0.512	0.634	0.704
SVM	0.687	0.759	0.547	0.636	0.687
RandomForest	0.672	0.868	0.406	0.553	0.672
Decision	0.585	0.626	0.424	0.505	0.585
K-nearest	0.576	0.621	0.391	0.480	0.576

Table ESM20. Five-fold cross-validation of ML based on a reduced dataset with 13CNMR data and molecular features, dimensionality reduced by PCA for predicting dopamine D1 receptor antagonists (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/7.2..ML_withPCA_fullSet_withPubChem_data.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.9031	0.0378	[0.8618, 0.8692, 0.8886, 0.9421, 0.9539]
GradientBoost	0.8347	0.0264	[0.7994, 0.8187, 0.8247, 0.8648, 0.8661]
Decision	0.8220	0.0352	[0.7771, 0.789, 0.8217, 0.8559, 0.8661]
SVM	0.8121	0.0240	[0.7845, 0.7979, 0.7994, 0.8276, 0.8512]
K-nearest	0.7307	0.0522	[0.6835, 0.6761, 0.7073, 0.7979, 0.7887]

Table ESM21. ML metric results based on a full dataset of 13CNMR and molecular features for predicting dopamine D1 receptor antagonists (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/7.1.ML_noPCA_fullSet_withPubChem_data.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.758	0.842	0.636	0.724	0.758
RandomForest	0.756	0.872	0.600	0.711	0.756
SVM	0.747	0.822	0.631	0.714	0.747
Decision	0.687	0.715	0.621	0.665	0.687
K-nearest	0.615	0.657	0.483	0.556	0.615

Table ESM22. Fi-fold cross-validation of ML based on a full dataset with 13CNMR and molecular features for predicting dopamine D1 receptor antagonists (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/7.1.ML_noPCA_fullSet_withPubChem_data.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8878	0.0469	[0.8475, 0.8473, 0.8547, 0.9362, 0.9532]
Decision	0.8085	0.0400	[0.7715, 0.7781, 0.7783, 0.8521, 0.8623]
SVM	0.8007	0.0210	[0.7785, 0.7799, 0.7946, 0.8218, 0.8288]
GradientBoost	0.7607	0.0039	[0.7599, 0.7548, 0.7636, 0.7589, 0.7662]
K-nearest	0.7104	0.0391	[0.6742, 0.6812, 0.6804, 0.7599, 0.7564]

Table ESM23. Metrics results of ML with a full dataset of 13CNMR and molecular features and dimensionality reduced by PCA for predicting dopamine D1 receptor antagonists (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/7.2..ML_withPCA_fullSet_with PubChem_data.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.743	0.820	0.623	0.708	0.743
GradientBoost	0.739	0.809	0.624	0.705	0.739
RandomForest	0.719	0.847	0.534	0.655	0.719
Decision	0.657	0.689	0.572	0.625	0.657
K-nearest	0.620	0.662	0.491	0.564	0.620

Table ESM24. Five-fold cross-validation of ML based on a full dataset with 13CNMR and molecular feature and dimensionality reduced by PCA for predicting dopamine D1 receptor antagonists (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/D1/7.2..ML_withPCA_fullSet_with PubChem_data.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8701	0.0509	[0.8277, 0.8249, 0.8337, 0.925, 0.9393]
SVM	0.7931	0.0173	[0.7748, 0.7758, 0.7904, 0.804, 0.8204]
Decision	0.7840	0.0447	[0.746, 0.7482, 0.7489, 0.8306, 0.8463]
GradientBoost	0.7512	0.0052	[0.7474, 0.7439, 0.7538, 0.7521, 0.7589]
K-nearest	0.7056	0.0382	[0.6714, 0.6734, 0.679, 0.7472, 0.7571]

Table ESM25. CID_SID ML model`s metrics predicting TTR transcription activation capability of compounds designed initially for another purpose (GitHub file:

https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR_CID_SID_ML_model/TTR_CID_SID_ML_model.ipynb)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.815	0.946	0.668	0.783	0.815
K-nearest	0.791	0.844	0.715	0.774	0.791
RandomForest	0.785	0.844	0.700	0.765	0.785
SVM	0.774	0.970	0.565	0.714	0.774
Decision	0.740	0.774	0.676	0.722	0.740

Table ESM26. Five-fold cross-validation of CID_SID ML models predicting TTR transcription activation capability of compounds that have been initially designed for another purpose (GitHub file: https://github.com/articlesmli/NMR_ML_TTR_D1/blob/main/TTR_CID_SID_ML_model/TTR_CID_SID_ML_model.ipynb)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
GradientBoost	0.8458	0.0085	[0.8472, 0.852, 0.8297, 0.847, 0.8533]
RandomForest	0.8301	0.0055	[0.8205, 0.8362, 0.8312, 0.8281, 0.8344]
K-nearest	0.8175	0.0067	[0.8205, 0.8157, 0.8076, 0.8155, 0.8281]
SVM	0.8108	0.0139	[0.8079, 0.8205, 0.8155, 0.7855, 0.8249]
Decision	0.7799	0.0078	[0.7811, 0.7874, 0.7681, 0.7744, 0.7886]

Figures

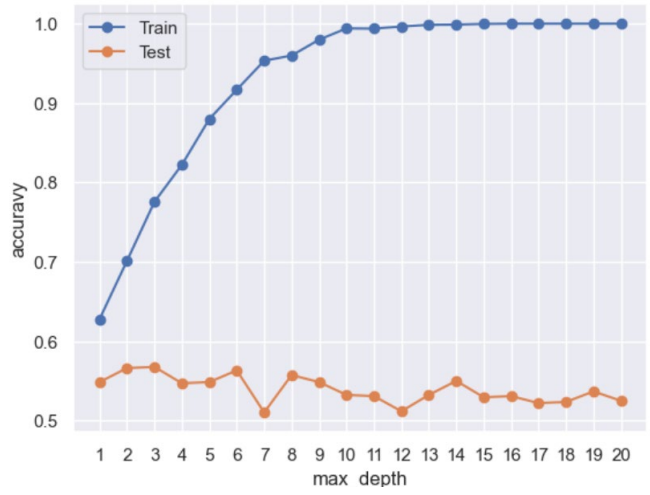

Figure ESM1. Tracing the deviation between train and test accuracies for the TTR case without dimensionality reduction

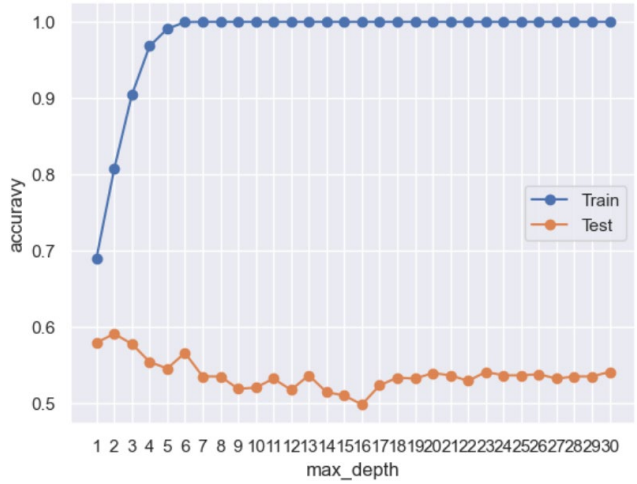

Figure ESM2. Tracing the deviation between train and test accuracies for the TTR case with dimensionality reduction performed by PCA

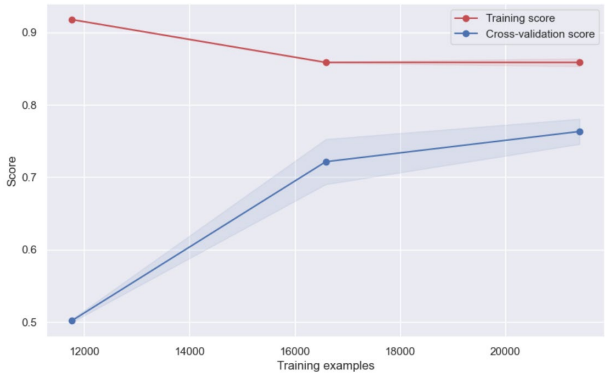

Figure ESM3. Learning curve dopamine D1 receptor case without dimensionality reduction

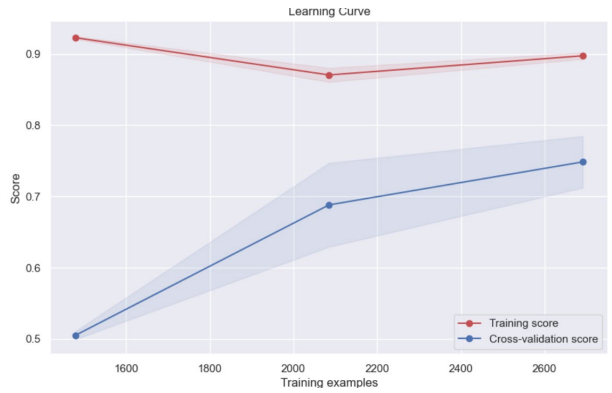

Figure ESM4. Learning curve of ML model predicting dopamine D1 receptor antagonists with dimensionality reduction performed by PCA

Figure ESM5. Tracing the deviation between train and test accuracies for the TTR case with molecular features added without dimensionality reduction

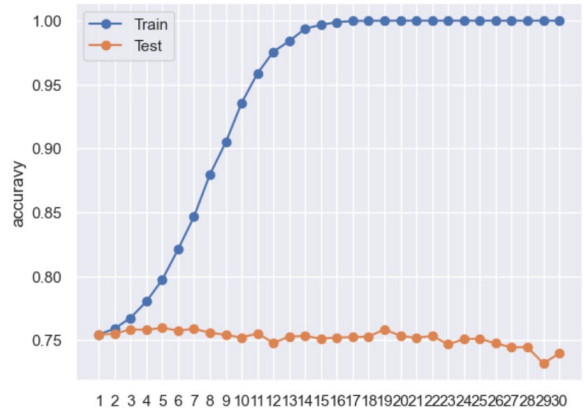

Figure ESM6. Tracing the deviation between train and test accuracies for the dopamine D1 receptor antagonist and molecular features case without dimensionality reduction

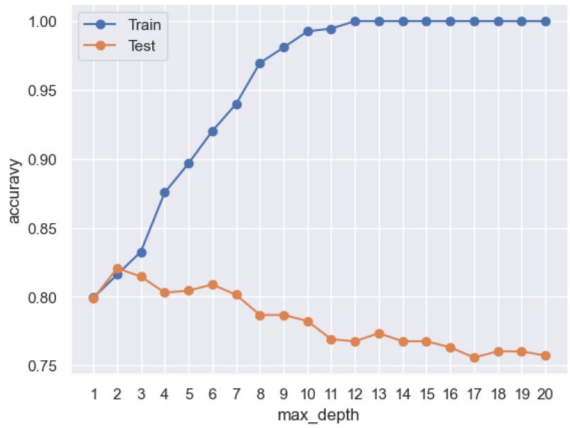

Figure ESM7. Tracing the deviation between train and test accuracies of the CID_SID ML model based on a dataset focused on predicting TTR transcription activators.