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

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

Mariya L. Ivanova, School of Computing and Engineering, University of West London, United Kingdom; mariya.ivanova@uwl.ac.uk

Nicola Russo, School of Computing and Engineering, University of West London, United Kingdom; 21485661@student.uwl.ac.uk

Gueorgui Mihaylov, Haleon, United Kingdom; gueorgui.m.mihaylov@haleon.com

Konstantin Nikolic, School of Computing and Engineering, University of West London, United Kingdom; konstantin.nikolic@uwl.ac.uk

Abstract

This study is part of a larger, ongoing research effort to develop a global machine learning (ML) model for drug discovery and development. The methodology utilised the scikit-learn ML library and combined NMR spectroscopy data (derived from SMILES notations) with molecular features from PubChem to predict the functionality of small biomolecules. The approach's effectiveness was demonstrated using a human dopamine D1 receptor antagonist case study and compared to a case study predicting neuronal Transthyretin (TTR) transcription activators. Additionally, a CID_SID

ML model was developed to predict TTR transcription activation capabilities of compounds, based solely on their PubChem CID and SID. The enhanced ML model predicting dopamine D1 receptor antagonists obtained 75.8% Accuracy, 84.2% Precision, 63.6% Recall, 72.5% F1-score and 75.8 % ROC, trained on 25,532 samples and tested on 5,466. The hypothetical ML model predicting neuronal TTR transcription activators achieved 67.4% Accuracy, 74.0% Precision, 53.5% Recall, 62.1% F1-score and 67.4 % ROC, if it could be trained with 25,532 samples and tested with 5,466. CID_SID ML model achieved 81.5% Accuracy, 94.6% Precision, 66.8% Recall, 78.3% F1-score and 81.5 % ROC. Overall, an improved machine learning (ML) approach was developed by incorporating additional molecular features. The comparison to another case study revealed that the effectiveness of the ML approach is dependent on the specific case study. Additionally, the CID_SID ML model yielded promising results, demonstrating its potential for predicting side effects related to TTR transcription activation.

Keywords: CID_SID ML model, neurodegenerative disorders, 13C NMR spectroscopy, drug discovery and development, machine learning.

Introduction

Nuclear Magnetic Resonance (NMR) Spectroscopy has been considered as one of the most powerful and informative techniques currently available for small biomolecule analysis [1]. It provides detailed information about the arrangement of atoms within a molecule; identification of functional groups; three-dimensional arrangement of atoms in space; dynamics and interaction with other molecules; quantitative analyses [2]. Enabling computers to learn from data through analysis and eliminating explicit programming, ML has been applied to data obtained via NMR spectroscopy such as employing message passing neural networks (MPNNs) to predict 13C NMR chemical shifts in small molecules [3]; decreasing the ML errors of the NMR spectroscopy chemical shift predictions by using Deuterated chloroform [4]; Computational structural elucidation aided by NMR spectroscopy and machine learning techniques [5]; combining AlphaFold and chemical shift prediction in a deep learning model for faster protein NMR assignment [6]; built from over 8,300 carbon atoms in diverse environments and without limitations on molecular complexity, a general-case neural network model was created for predicting (estimating) 13C NMR spectra [7].

Beyond ML, a range of computational methods listed below have been employed to predict NMR [8 - 9], such as:

(i) Computational "ab initio" techniques, meaning they operate "from first principles," determine a system's properties using only fundamental laws, in contrast to methods relying on empirical data. One such application is the development of a 13C NMR chemical shift prediction procedure for small molecule structure elucidation, which employed gauge invariant atomic orbitals with density functional theory and incorporated empirical systematic error correction terms [10].

(ii) Additive increment-based approaches represent the initial methods for predicting chemical shifts where the core atomic shift has been subsequently modified by positive or negative contributions from attached substituents. Such rules, for instance, have been implemented in the CASPER program [11] to predict 13C, 1H, and 15N shifts of polysaccharides.

(iii) Hierarchical Organisation of Spherical Environments (HOSE) code characterises an atom's molecular environment by iteratively describing concentric spheres of neighbouring atoms. Essentially, this approach operates as a nearest neighbour search, where the HOSE

code serves as the metric for defining structural similarity (9)

The online tool NMRDB [12] employed in the study to convert SMILES into 13C NMR spectroscopy data has been based on HOSE code and databases of existing NMR spectra and chemical structures reviewed by a board of reviewers [13]. Recognising that NMRDB-derived information represents a first approximation compared to experimental data, its employment was sufficient for the demonstration of the proposed methodology idea within the current and previous studies.

The presented research had to solve two tasks: first, to assess the applicability of a pre-existing methodology [14] to a different case study; and second, to determine how the inclusion of molecular features would affect the accuracy of the investigated machine learning model. The existing methodology was an ML approach based on NMR spectroscopy data derived from SMILES notations, demonstrated by the case study on human dopamine D1 receptor antagonists, PubChem AID 504652 [15], which data was provided by PubChem, the world`s largest database for freely available chemical information [16]. The core idea of the existing ML approach was to leverage the relationship between a small biomolecule's carbon skeleton and the information provided by 13C NMR spectroscopy.

The NMR spectrum, with its peaks, indicates atomic bonds and ligands and elucidates the molecular environment of the carbon atoms. This data was numerically encoded for machine learning, enabling its association with labels. So, it was anticipated and respectively demonstrated that 13C NMR spectroscopy could effectively contribute to the ML-based prediction of small biomolecule functionality.

The new case study, PubChem AID 1117267 bioassay [17] used for the comparison was also provided by PubChem. It was focused on TTR transcription activators containing 91,943 samples, 1,155 of which were defined as TTR activators. Data was obtained by High-throughput screening (HTS), which is an automated method in drug discovery and toxicology that rapidly tests vast compound libraries for biological activity using robotics and automated systems for parallel processing and analysis. The bioassay was conducted to identify small biomolecules activating neuronal TTR transcription. Compounds are tested in concentration of 16.7 uM. For more information about the screening protocol, please refer to the bioassay documentation [17].

The protein TTR was called “the servant of many masters”. The primary functions of TTR are the transport of thyroid hormones and retinol through the bloodstream and cerebrospinal fluid [18].

However, its stability is critical, and misfolding can lead to amyloid disease, causing amyloidosis of nerves, ligaments, heart and arterioles [19] that could lead to polyneuropathy where multiple peripheral nerves become damaged [20], and/or cardiomyopathy affecting the heart muscles [21]. Additionally, TTR has been reported to play an independent neuroprotective role in both the peripheral and central nervous systems, contributing to nerve function and repair [22 - 23 - 24]. Using TTR Knockout mouse [25], it has been found out that the absence of transthyretin leads to increased levels of neuropeptide Y (NPY) in the brain and peripheral nervous system. Despite the need for additional research to fully elucidate the specific effects in TTR knockout mice, the study confirmed the presence of alterations in feeding behaviour, anxiety levels, stress responses, metabolic parameters, and potential changes in nerve function and repair [26].

So, overall, the protein TTR is a subject of research for many scientists and clinicians due to its crucial roles in the body and its association with a significant and debilitating group of diseases known as transthyretin amyloidosis (ATTR).

Regarding the second objective of the presented research, which expanded the study beyond the comparison, molecular features obtained from PubChem were added to the NMR spectroscopy data

to explore their influence on the ML models. A similar approach, which has improved the ML model, has been performed in studies regarding G9a [27 - 28]. These molecular features were as follows:

(i) Rotatable Bond Count (RBC) is determined by counting the single bonds that are not part of a ring and connect two non-hydrogen atoms that are also not at the end of a chain.

(ii) Hydrogen Bond Acceptor Count (HDAC) refers to the number of hydrogen bond acceptors present in the given small biomolecule.

(iii) Hydrogen Bond Donor Count (HBDC) refers to the number of hydrogen bond donors present in the given small biomolecule.

(iv) Topological Polar Surface Area is computed as the surface sum over polar atoms in the molecule [29]

(v) XLogP3 (XL) represents a predicted octanol-water partition coefficient [30].

(vi) Molecular Weight (MW) is defined as the sum of the mass of all constituent atoms [31].

Furthermore, an additional ML application was developed for researchers interested in TTR transcription activators aimed to predict whether a small biomolecule, initially designed for another

purpose, could also function as a TTR transcription activator. This ML prediction was based exclusively on PubChem CID and SID data of the small biomolecule, using the methodology [32] that has previously demonstrated success in predicting the case studies related to CHOP inhibitors, dopamine D1 receptor antagonist, dopamine D3 receptor antagonist, TDP1 inhibitors, M1 muscarinic receptor antagonists, Rab 9 promoter activators and G9a enzyme inhibitors While identifiers are generally not used in ML, PubChem identifiers can be an exception. This is because their generation process employs an algorithm that considers the structure and similarity between substances and compounds [33]. For more details, please refer to the parent study [32].

Last, but not least, since the ML model of interest was classification, i.e. heavily dependent on the balance between the classes, the severe imbalance between the active and inactive biomolecules in the TTR bioassay dataset was handled through three steps. One of these steps incorporated PubChem AID 1996 bioassay focused on water solubility of small biomolecules [34], whose dataset was used as a sieve for the reduction of the small biomolecules without considering the level of aqua solubility of the samples.

Methodology

The methodology followed whose applicability exploration was the aim of the research. For this purpose, the CID, SMILES notations, and labels were extracted from the PubChem AID 1117267 bioassay dataset, which contains 91,909 rows of small biomolecules and nine columns of features describing these molecules. The samples detected as TTR transcription activators were 1,155, called Active and 90,755 detected as Inactive compounds. The severe imbalance between the active and inactive compounds was initially decreased by merging the dataset of bioassay AID 1117267 with the dataset of the PubChem AID 1996 bioassay. Since the latter contained 57,856 samples of small biomolecules, this margin of both datasets played the role of a sift, reducing the number of inactive compounds. Further, after shuffling of the inactive compounds only every eighth sample was retained. The reduced inactive compounds, 2,023 samples, were then concatenated with all active compounds, i.e. 1,155 samples from the PubChem AID 1117267 bioassay dataset and the final dataset with 3,177 samples in total was obtained. Further, the SMILES notations of the resulting dataset were converted into numerical spectroscopy data employing the online tool NMRDB that has been designed for this purpose. The process requested each SMILES to be uploaded on the website where the NMRDB tool will generate 13C NMR spectroscopy prediction, which was used to

obtain the necessary data for ML. The scale of the chemical shifts, i.e. the pics, started from 0 and reached beyond 200. The scale was deviated by natural numbers, which defined subranges. These subranges were used as a feature and generated the columns in the data frame. The presence of a pick or picks in each range was counted and placed as a value for the corresponded subrange of the relevant small biomolecule. Once all SMILES notations were converted to 13C NMR spectroscopy data, the dataset was merged with the labels based on the SMILES notations. In this way, since the canonical SMILES are unique, misconnection between isomers was avoided.

To ensure the reliability of the evaluation, an equal number of samples from each class was extracted and the testing set created. The remaining samples were balanced with oversampling. Although SMOTE is the technique that is generally recommended for balancing of data because it is expected to reduce overfitting and improve generalisation, simulations for this study showed that Random Over Sampler was more effective. This involved increasing the minority class size to match the majority class by duplicating minority class samples. The balanced dataset was used for ML training with the RFC, DTC, GBC, and SVC estimators provided by the scikit learn ML library. The results of prediction were evaluated by the classification

metrics, which are based on the four possible scenarios: true positive (TP) and true negative (TN), where the prediction corresponds to the actual values, and false positive (FP) and false negative (FN), where they do not. So, the evaluation was done based on the metrics:

(i) Accuracy – the total of correct predictions to the total of all predictions.

(ii) Precision – true positive out of the total of all positive predictions.

(iii) Recall – true positive out of all positive instances, giving the sensitivity of the ML model

(iv) F-1 score – indicates the balance between precision and recall

(v) ROC (Receiver Operating Characteristic) shows the true positive against the false positive when the threshold varies.

The comparison of the results ranked the most optimal estimator for the case. The chosen ML model was scrutinised for overfitting, tracing for the deviation between training and testing accuracy to be smaller than 5%.

To ensure fairness in comparison between two cases, data set used for the ML model based on human dopamine D1 receptor antagonist

data was reduced to the same number of active and inactive small biomolecules used in the TTR transcription activators prediction. Then, the difference in accuracy between the ML model based on the reduced D1 receptor antagonist dataset and the dataset used in the previous study was used to calculate how much the accuracy had grown due to the increase in dataset samples. It was hypothesised that the existence of the same number of samples for the TTR case would increase the accuracy of the ML model by the same percentage. ML with the enriched dataset was performed in the same manner explained above for the dopamine D1 receptor antagonist and TTR transcription activators.

Principal component analysis (PCA) was applied with the intention to reduce the noise of data by identifying the highest variance in the data, and due to effective filtering of the noise, data will become clean, and the ML model will be able to learn the actual underlying patterns more effectively [35]. Also, the reduction of dimensionality is expected to handle the data sparsity, helping the ML model to find a meaningful relationship. Last but not least, PCA removes multicollinearity, which can lead to more reliable metrics.

During the development of the prior study (14), which demonstrated the prediction of human dopamine D1 receptor antagonists, it was observed that the model depended on the number of samples. To

ensure a fair comparison in the current study, ML based on dopamine D1 receptor antagonists was conducted, using the same number of samples as the ML prediction of TTR transcription activators. Moreover, in the current study, the molecular features were added to both the reduced dataset and the entire dataset for dopamine D1 receptor antagonists. he former was used for the purpose of comparison, and the latter was used to explore the effect that molecular features have on accuracy in general. The methodology is illustrated in Figure1.

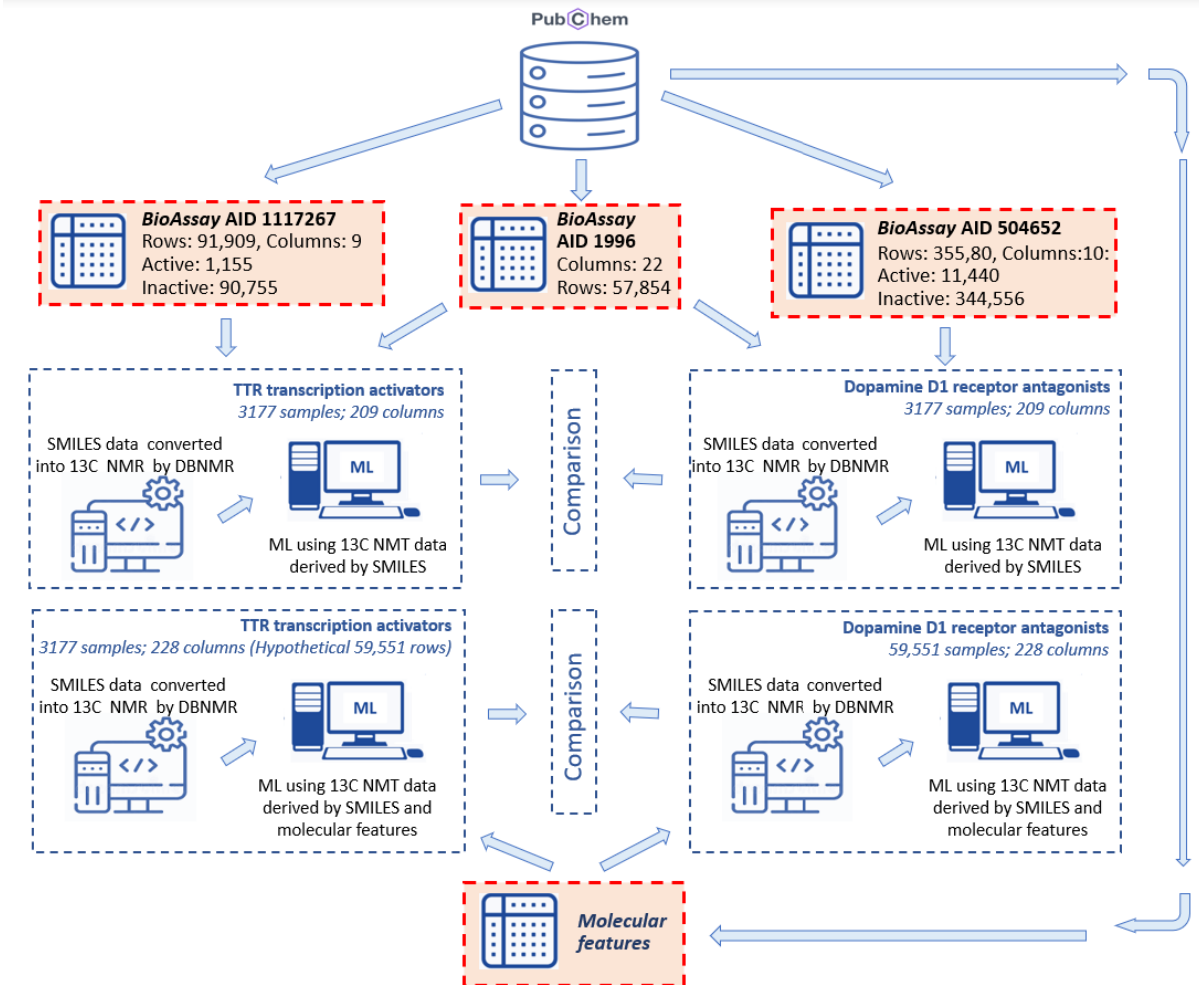

Figure 1. Methodology of comparison of dopamine D1 receptor antagonist and TTR activators based on 13C NMR spectroscopy data and molecular features.

Results and discussion

The results of comparison of the ML models are summarised in Table 1, providing connection with the relevant tables or figures in the electronic supplementary material (ESM).

Leveraging the methodology outlined above, the initial collection of inactive compounds from the TTR bioassay was narrowed down to 2,023 small biomolecules. Combining this reduced set of inactive samples with the full set of 1154 active compounds resulted in a data set with a total of 3177 rows and 209 columns. From this dataset, a set of 340 samples per class was extracted for testing. The remaining samples**,** after balancing their label proportion with SMOTE, were allocated for training. As a result of this, the ML models were trained with 2497 samples and tested with 680 samples. From the classifiers listed above, GBC was the optimal estimator, achieving 56.8% accuracy, 59.2% precision, 43.5% recall, 50.2% F1-score, 56.8 % ROC (Table ESM1) and a five-fold cross-validation score of 0.651 with 0,02 standard deviation (Table ESM2). The deviation between the ML model performance and the cross validation implied overfitting, which was confirmed by comparing the training and testing accuracies (Figure ESM1). Applying PCA slightly improved the accuracy to 58.4%, precision to 65.4% and ROC to 58.4.%. However, the low recall of 35.3% decreased the F1-score to 46.1% F1-score, which was an indicator of an increase of the bias of the ML model towards the majority class (the inactive compounds), and thus, the ML model can miss crucial instances of the minority class (the active compounds) (Table ESM3). The increased five-fold cross-validation

score of 0.745 with 0,04 standard deviation (Table ESM4) compared to the single ML model accuracy of 58.4% (Table ESM3) suggested that the ML model was overfitted, which was subsequently confirmed by tracing the deviation between train and test accuracy (Figure ESM2).

To perform a fair comparison between the two case studies, based on the random principle, equal samples of 1,948 inactive and 1,093 active compounds were extracted from both the dopamine D1 receptor antagonist and TTR transcription activator case studies. After the rows of the dopamine D1 receptor antagonist dataset were reduced to correspond to the TTR dataset, ML was performed in the same manner as it was done for the TTR transcription activators case. The optimal estimator was SVC, which obtained 65.7 % accuracy, 70.3% precision, 54.4 % recall, 61.4% F1-score, 65.7 % ROC (Table ESM5) and a five-fold cross-validation score of 0.775 with 0.02 standard deviation (Table ESM6). Expectedly, because the difference between the single and cross-validation score was significant, the ML model was overfitted (Figure ESM3). However, since the ML model based on full set of the available data for the dopamine D1 receptor antagonist was not overfitted [see the original research (14)] it was hypothesised that the increase of the samples of the TTR case equal to the same amount of samples used in the dopamine D1 receptor

case, i.e. 10,542 active and 46,496 inactive compounds, would improve the TTR ML model with the same percentage as the accuracy of the dopamine D1 ML model. So, since the metrics of the ML model predicting the dopamine D1 receptor antagonist were accuracy of 71.5 %, precision of 77.4%, recall of 60.6 %, F1-score of68.0 % and ROC71.5% (Table ESM9) and a five-fold cross-validation score of 0.748±0.003 standard deviation (Table ESM10) the percentage of improvement of the ML model based on the increase of the number of samples was calculated as follow: the accuracy of the ML model with reduced rows plus the accuracy of the ML model with reduced rows multiplied by the unknown percentage should be equal to the accuracy of the ML model with non-reduced rows (Figure 2). The increase of 5.8 in the accuracy of the ML model due to an increase in the number of samples was 8.85%. So, it was hypothesised that the increase the number of samples for TTR case to the level of samples for the D1 case, i.e. 57,038 compounds would increase the accuracy of the ML model that predicts TTR transcription activators to 61.82%.

In order to improve the performance of the ML models, their complexity was mitigated by the PCA dimensionality reduction. For the TTR transcription activators case (Table ESM1 and Table ESM2 without PCA and Table ESM3 and Table ESM4 with PCA) based on the difference between the single ML metrics and the cross-validation

score can be concluded that the ML models were overfitted. This was confirmed by scrutinising them for overfitting, where the difference between training and testing accuracy was higher than 5 % since the beginning, i.e. max_depth=1. For the ML model without PCA, the training accuracy was 62.8% while the testing accuracy was 54.9% (Figure ESM1). For the ML model with PCA, the training accuracy at max_depth=1 was 69% compared to the testing accuracy of 57.9% (Figure ESM2). Similarly to the ML models predicting TTP transcription activators, the presence of PCA did not improve the ML models with a reduced dataset.

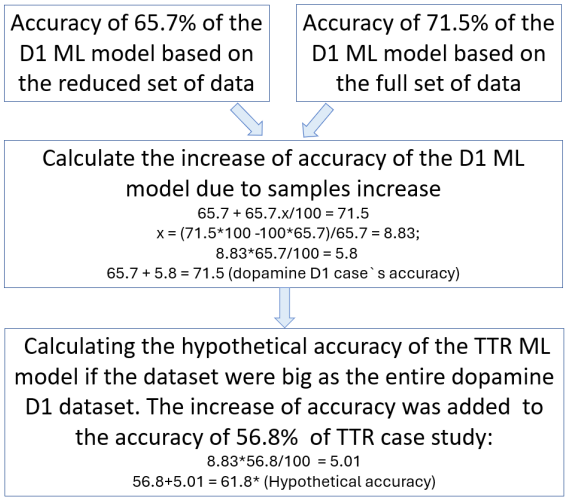

Figure 2. Calculating the hypothetical accuracy of ML model predicting TTR activators based on 13C NMR spectroscopy data.

The accuracy without PCA was 65.7% (Table ESM5) with 0.775 cross-validation score (Table ESM6) compared to the accuracy with PCA

that was 64.7% (Table ESM7) with a cross-validation score of 0.748 (Table ESM8). The addition of samples to the dataset used for ML without PCA dimensionality reduction improved the ML model performance to 71.5% of accuracy (Table ESM9) with cross validation of 0.749 (Table ESM10), and significant overfitting has not been detected (Figure ESM3). The increase in the number of samples in the PCA dimensionality reduced case, on the other hand, did not increase the accuracy. It was accuracy of 64.7% (Table ESM7) with a five-fold cross-validation score of 0.748 (Table ESM8) and become 64.2% accuracy (Table ESM11) with 0.683 five-fold cross-validation score (Table ESM12), however the decrease of the difference between the single ML model accuracy (Table ESM11) and the five-fold cross-validation score (Table ESM12) was an indicator that due to the application of PCA the ML model became more robust than the ML model based on reduced dataset (Figure ESM4).

The next major step in the study, the addition of above-mentioned molecular features to the 13C NMR spectroscopy data, improved the ML models` performance in both the TTR transcription activators and dopamine D1 receptor antagonists cases. Regarding the TTR transcription activators dataset, which had 3177 samples, merging it with the molecular data led to a decrease in the dataset to 3,041 samples (1,948 inactive and 1,093 active compounds). The optimal

ML model for this case was GBC, achieving 67.1% accuracy, 74.0% precision, 52.6% recall, 61.5% F1-score, 67.1 % ROC (Table ESM13) and a five-fold cross-validation score of 0.693 with 0,019 standard deviation (Table ESM14). The scrutiny for overfitting revealed the optimal performance of GBC was at max_depth=2. At this depth, the model achieved 62.4%accuracy, 66.7% precision, 49.4% recall, 56.8% F1-score and 62.4% ROC (Figure ESM5). The PCA dimensionality reduction in case of TTR dataset with 13CNMR spectroscopy data and molecular features worsened the performance of the GBC ML model to 60.4% accuracy, 62.5% precision, 52.1% recall, 56.8% F1-score, 60.4 % ROC (Table ESM15) and five-fold cross-validation score of 0.717 with ±0.013 standard deviation (Table ESM16). Regarding the ML models predicting whether a compound is a dopamine D1 receptor antagonist, the ML model accuracy increased from 65.7% (Table ESM 5) to 75.1% (Table ESM17) with a five-fold cross-validation score of 0.788≠0.013 standard deviation (Table ESM18). The dimensionality reduction with PCA in this case worsened the results to an accuracy of 70% (Table ESM19) and a five-fold cross-validation score of 0.835±0.038 standard deviation (Table ESM20). The increased of the gap between the accuracy of the single ML model and the cross-validation score indicated for increase in overfitting.

As mentioned above, since there was a significant quantity of real data about the dopamine D receptor antagonist, the addition of molecular features was explored for the whole dataset as well, with and without PCA. Expectedly, the increase in the number of samples improved the performance of the optimal ML model GBC, obtaining 75.8% accuracy, 84.2% precision, 63.6% recall, 72.4% F1-score, 75.8 % ROC (Table ESM21) and five-fold cross-validation score 0.761 ≠0.0039 (Table ESM22). However, the increase in the number of samples in case of dimensionality reduction did not lead to substantial changes in the ML model metrics (Table ESM23 for a single ML model and Table ESM24 for five-fold cross-validation) (Figure ESM6).

Hyperparameter tuning for each of the cases was performed by Optuna, but the suggested hyperparameters did not improve the ML models. For more details regarding the hyperparameter tuning, please refer to the code provided on GitHub. So,ooverall, it can be observed that, in both study cases, i.e. ML for predicting whether a compound is a human dopamine D1 receptor antagonist or a TTR transcription activator, the addition of the aforementioned molecular features to the 13CNMR spectroscopy dataset improved the performance of the ML models. The total number of features was 227. The followed investigations for overfitting revealed that the optimal

ML model, predicting whether a compound is a TTR transcription activator was GBC with default hyperparameter features and max_depth=3, achieving accuracy 75.8%, precision 84.2%, recall 63.6%, F1-score 72.5%, ROC 75.8%. The ML model was trained on 25.532 samples and tested on 5,466 samples. The optimal ML model, predicting whether a compound is a TTR transcription activator was GBC with default hyperparameter features and max_depth=2, achieving accuracy 67.4%, precision74%, recall 53.5%, F1-score 62.1%, ROC 67.4%. The ML model was trained on 3,216 samples and tested on 680 samples.

Table 1. Initial ML results obtained before estimating the generalisation through scrutinising for overfitting. The TTR ML values marketed with * hypothetical value calculated theoretically based on the percentage of increase of metrics for dopamine D1 receptor antagonists based on increase of the samples.

	Train with 27,756 samples		**Test with 5,466 samples**	
Case study & dataset	**Accuracy [%]**	**Cross-val.**	**Accuracy [%]**	**Cross-val.**
TTR & no PCA	GBC 56.8 (Table ESM1)	0.651 (Table ESM2)	61.82*	-
TTR & with PCA	GBC 58.4 (Table ESM3)	0.745 (Table ESM4)	-	-
TTR & no PCA &	GBC 67.1	0.693	67.1*	-

PubChem mol. data	(Table ESM13)	(Table ESM14)		
TTR & with PCA & PubChem mol. data	SVC 65.0 (Table ESM 15)	 0.695 (Table ESM 16)	-	-
D1 & no PCA	SVC 65.7 (Table ESM5)	 0.775 (Table ESM6)	SVC 71.5 (Table ESM9)	 0.749 (Table ESM10)
D1 & with PCA	SVC 64.7 (Table ESM7)	 0.7485 (Table ESM8)	SVC 64.2 (Table ESM11)	 0.683 (Table ESM12)
D1 & no PCA & PubChem mol. data	GBC 75.1 (Table ESM17)	 0.788 (Table ESM18)	GBC 75.8 (Table ESM21)	 0.761 (Table ESM22)
D1 & with PCA & PubChem mol. data	GBC 70.4 (Table ESM19)	 0.835 (Table ESM20)	SVC 74.3 (Table ESM23)	 0.793 (Table ESM24)

The additional computational approach, CID_SID ML model, developed to predict the TTR transcription activators achieved initially accuracy of 81.5%, precision of 94.6%, recall of 66.8%, F1-score of 78.3% and ROC of 81.5% with GBC (Table ESM 25) and cross-validation score 0.8458 with ≠0.0085 standard deviation (Table ESM 26). Optuna performed the hyper-parameter tuning of GBC. However, the suggested best parameters, which were results of one hundred study trials, did not improve the ML model. On the other hand, the scrutiny for overfitting revealed that the optimal max_depth=2(Figure

ESM7). The final model, with default hyperparameter values and max_depth=2, achieved an accuracy of 81.5%, precision of 94.6%, recall of 66.8%, F1-score of 78.3%, and an ROC of 81.5%. The ML model was trained with 3.358 and tested with 680 samples. Regarding the CID_SID ML model predicting dopamine D1 receptor antagonists, the ML model has already been developed (14), obtaining an accuracy of 80.2%, precision of 86.3%, recall of 70.4%, F1-score of 77.6%, and ROC of 79.9% when trained with 19,438 samples and tested with 4,723 samples.

Conclusion

The lower number of samples used for training and testing of the ML model to predict the human dopamine D1 receptor antagonists achieved lower results than the ML model trained and tested with the full sets of small biomolecules. This scalability dependence was used to calculate the hypothetical accuracy for predicting the TTR transcription activators. However, the big difference of the ML models` results for the small dataset of these two case studies implied that the ML approach is a case study dependent, so that the hypothetical accuracy is recommended to be tested with real data. Given that such data is not available, the investigation of the ML approach reminded open for exploration with other bioassays that

have significant number of samples. On the other hand, the addition of atomic features of the biomolecule for both bioassays improved the ML models except the variant with atomic features and a full dataset on human dopamine D1 receptor antagonists. It performed worse than the ML model with a reduced number of features. The hypothesis is that the model's complexity obstructs it from achieving better results.

The computational approach developed beyond the main course, i.e. the CID_SID ML model, offered a cost-effective and rapid way to gain early knowledge about potential side effects of drug candidates with respect to TTR transcription activators.

Although the utilized 13C NMR spectroscopy data is typically classical, the encoded into it information corresponds to the interaction of the quantum mechanical nature of matter and the light, providing deeper insights of the interaction considered in the case studies.

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.

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/NMR_ML_TTR_D1.git

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

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

References

[1] Jacobsen, N. E. NMR Spectroscopy Explained: Simplified Theory, Applications and Examples for Organic Chemistry and Structural Biology, **2007**, John Wiley & Sons, Inc., DOI: 10.1002/9780470173350

[2] Marion, D. An introduction to biological NMR spectroscopy. Mol¡. Cel¡.Proteomics **2013**, 78.(11), 3006-25, DOI: 10.1074/mcp.o113.030239

[3] Williamson, D.; Ponte, S.; Iglesias, I.; Tonge, N.; Cobas, C.; Kemsley, E. K. Chemical shift prediction in ^{13}C NMR spectroscopy using ensembles of message passing neural networks (MPNNs). J¡. Magn¡.Reson¡ **2024**, 9❷❹, 107795, DOI: 10.1016/j.jmr.2024.107795

[4] Rull, H.; Fischer, M.; Kuhn, S. NMR shift prediction from small data quantities. Jornal.of.Cheminform¡ **2023**, 7❶, 114, DOI: 10.1186/s13321-023-00785-x

[5] Cortés, I.; Cuadrado, C.; Hernández D. A.; Sarotti, A. M. Machine learning in computational NMR-aided structural elucidation. Front¡. Nat¡.Prod¡.**2023**, 8, 1122426, DOI: 10.3389/fntpr.2023.1122426

[6] Klukowski, P.; Riek, R.; Güntert, P. Time-optimized protein NMR assignment with an integrative deep learning approach using AlphaFold and chemical shift prediction.Sci¡.Adv¡.**2023**?5, eadi9323. DOI:10.1126/sciadv.adi9323

[7] Bret, C. L. (2000). A General ^{13}C NMR Spectrum Predictor Using Data Mining Techniques. SAR.QSAR.Environ¡.Res. **2000**, 77(3–4), 211–234, DOI: 10.1080/10629360008033232

[8] Jonas, E.; Kuhn, S.; Schlörer, N; Prediction of chemical shift in NMR: A review. Magn.Reson.Chem. **2022**, 26(11), 1021-1031. DOI: 10.1002/mrc.5234

[9] Jonas, E.; Kuhn. S.; Schlörer, N. Prediction of chemical shift in NMR: A review, Magn¡.Reson¡.Chem¡ **2022**, 26(11), 1021, DOI: 10.1002/mrc.5234

[10] Xin, D.; Sader, C., A.; Chaudhary, O.; Jones, P.; Wagner,K.; Tautermann, K.,S; et al Development of a ^{13}C NMR Chemical Shift Prediction Procedure Using B3LYP/cc-pVDZ and Empirically Derived Systematic Error Correction Terms: A Computational Small Molecule Structure Elucidation Method J¡.Org¡.Chem¡.**2017**, 48 (10), 5135-5145 DOI: 10.1021/acs.joc.7b00321

[11] CASPER Reilly, D., Wren, C., Giles, S., Cunningham, L., & Hargreaves, P. (2016). CASPER¿Computer.Assisted.Search. Prioritisation.and.Environmental.Response.Application.

[12] Emerenciano V., P.; Diego, D., G.; Ferreira M., J., P.; Scotti, M., T.; Comasseto, J., V,; Rodrigues, G., V. Computer-aided prediction of 125Te and13CNMRchemicalshiftsofdiorgano tellurides. J¡.Braz¡. Chem¡.Soc¡ **2007**,7❹(6),11831188. DOI: 10.1590/S0103-50532007000600012

[13] NMRDB https://nmrshiftdb.nmr.uni-koeln.de/ Accessed May 18, 2025.

[14] Ivanova, M., L.; Russo, N.; Nikolic, K. Leveraging 13C NMR spectroscopic data derived from SMILES to predict the functionality of small biomolecules by machine learning: a case study on human Dopamine D1 receptor antagonists, ArXiv, Preprint at DOI: 10.48550/arXiv.2501.14044

[15] National Center for Biotechnology Information. PubChem Bioassay Record for AID 504652, Antagonist of Human D 1 Dopamine Receptor: qHTS, Source: National Center for Advancing Translational Sciences (NCATS). https://pubchem.ncbi.nlm.nih.gov/bioassay/504652. Accessed May 18, 2025.

[16] National Institutes of Health, PubChem https://pubchem.ncbi.nlm.nih.gov/ Accessed May 18, 2025.

[17] National Center for Biotechnology Information. PubChem Bioassay Record for AID 1117267, Source: The Scripps Research Institute Molecular Screening Center. https://pubchem.ncbi.nlm.nih.gov/bioassay/1117267. Accessed Apr. 20, 2025.

[18] Buxbaum, J.N.; Reixach, N. Transthyretin: The Servant of Many Masters. Cell.Mol.Life.Sci. **2009**, 66(19), 3095-101, DOI: 10.1007/s00018-009-0109-0

[19] Ueda, M. Transthyretin: Its Function and Amyloid Formation, Neurochem¡.Int. **2022,** 7[illegible], 105313, DOI: 10.1016/j.neuint.2022.105313

[20] Liz, M., A.; Coelho, T.; Bellotti, V.; Fernandez-Arias, M., I; Mallaina, P.; Obici, L. A Narrative Review of the Role of Transthyretin in Health and Disease. Neurol.Ther. **2020**,.❺(2), 395-402, DOI: 10.1007/s40120-020-00217-0

[21] Nikitin, D.; Wasfy, J., H; Winn, A., N.; Raymond, F.; Shah, K., K.; Kim, S. et al. The effectiveness and value of disease-modifying therapies for transthyretin amyloid cardiomyopathy: A summary from the Institute for Clinical and Economic Review's Midwest Comparative Effectiveness Public Advisory Council, J¡.Manag¡.Care. Spec¡.Pharm¡ **2025**, 97(3), 323-8, DOI: https://www.jmcp.org/doi/abs/10.18553/jmcp.2025.31.3.323

[22] Fleming CE, Saraiva MJ, Sousa MM. Transthyretin enhances nerve regeneration. J Neurochem. **2007**,103, 831–839. doi: 10.1111/j.1471-4159.2007.04828.x

[23] Fleming, C.E.; Mar, F.M.; Franquinho, F.; Saraiva, M., J.; Sousa, M., M. Transthyretin internalization by sensory neurons is megalin mediated and necessary for its neuritogenic activity, J.Neurosci. **2009**, 8❺, 3220–3232. DOI: 10.1523/JNEUROSCI.6012-08.2009

[24] Rios, X.; Gómez-Vallejo, V.; Martín, A.; Cossío, U.; Morcillo, M. A.; Alemi, M. et.al¡ Radiochemical examination of transthyretin (TTR) brain penetration assisted by iododiflunisal, a TTR tetramer stabilizer and a new candidate drug for AD. Sci.Rep **2019**, ❺, 13672, DOI: 10.1038/s41598-019-50071-w

[25] Nunes, A., F.; Saraiva M. J.; Sousa M., M. Transthyretin knockouts are a new mouse model for increased neuropeptide Y?. FASEB.J **2007**, 20(1), 166-168, DOI: 10.1096/fj.05-4106fje

[26] Magalhães, J.; Eira, J; Liz, M., A. The role of transthyretin in cell biology: impact on human pathophysiology. Cell.Mol.Life.Sci¡ **2021**, 78(17-18), 6105-6117, DOI: 10.1007/s00018-021-03899-3

[27] Ivanova, M., L.; Russo, N.; Djaid, N.; Nikolic, K. Application of machine learning for predicting G9a inhibitors, Digital.Discovery, **2024**, 3(10), 2010-2018, DOI: 10.1039/D4DD00101J

[28] Ivanova, M., L.; Russo, N.; Djaid, N.; Nikolic, K. Targeting Neurodegeneration: Three Machine Learning Methods for Discovering G9a Inhibitors Using PubChem and Scikit-Learn, ArXiv, **2025**, Preprint at DOI: 10.48550/arXiv.2503.16214

[29] Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J.Med. Chem. **2000**, 43(20), 3714-7, DOI: 10.1021/jm000942e

[30] Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X. Computation of octanol-water partition coefficients by guiding an additive model with knowledge, J¡.Chem¡.Inf¡.Model¡ **2007**, 47(6), 2140-8, DOI: 10.1021/ci700257y

[31] van der Veen, A., M., H.; Meija, J.; Possolo, A.,A; Hibbert, D., B. Interpretation and use of standard atomic weights (IUPAC Technical Report), Pure.Appl¡.Chem¡.**2021**, **89**.(5), 629-646, DOI: 10.1515/pac-2017-1002

[32] 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**, Preprint at DOI: 10.48550/arXiv.2501.02154

[33] Kim, S.; Thiessen, P., A; , Bolton, E., E; Chen, J.; Fu, G. Gindulyte, A. et al. PubChem Substance and Compound databases, Nucleic.Acids.Res. **2016,**.00, D1202-13, DOI: 10.1093/nar/gkv951

[34] National Center for Biotechnology Information. PubChem Bioassay Record for AID 1996, Aqueous Solubility from MLSMR Stock Solutions, Source: Burnham Center for Chemical Genomics. https://pubchem.ncbi.nlm.nih.gov/bioassay/1996. Accessed May 18, 2025.

[35] scikit-learn. PCA. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html Accessed May 18, 2025.