OXFORD

Briefings in Bioinformatics, 23(1), 2022, 1–13

https://doi.org/10.1093/bib/bbab377
Problem Solving Protocol

Ensemble modeling with machine learning and deep learning to provide interpretable generalized rules for classifying CNS drugs with high prediction power

Tzu-Hui Yu, Bo-Han Su, Leo Chander Battalora, Sin Liu and Yufeng Jane Tseng

Corresponding author. Department of Computer Science and Information Engineering, National Taiwan University, No.1 Sec.4, Roosevelt Road, Taipei, Taiwan 106; Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No.1 Sec.4, Roosevelt Road, Taipei, Taiwan 106. Tel.: +886.2.3366.4888#529; Fax: +886.2.23628167; E-mail: yjtseng@csie.ntu.edu.tw

Abstract

The trade-off between a machine learning (ML) and deep learning (DL) model's predictability and its interpretability has been a rising concern in central nervous system-related quantitative structure–activity relationship (CNS-QSAR) analysis. Many state-of-the-art predictive modeling failed to provide structural insights due to their black box-like nature. Lack of interpretability and further to provide easy simple rules would be challenging for CNS-QSAR models. To address these issues, we develop a protocol to combine the power of ML and DL to generate a set of simple rules that are easy to interpret with high prediction power. A data set of 940 market drugs (315 CNS-active, 625 CNS-inactive) with support vector machine and graph convolutional network algorithms were used. Individual ML/DL modeling methods were also constructed for comparison. The performance of these models was evaluated using an additional external dataset of 117 market drugs (42 CNS-active, 75 CNS-inactive). Fingerprint-split validation was adopted to ensure model stringency and generalizability. The resulting novel hybrid ensemble model outperformed other constituent traditional QSAR models with an accuracy of 0.96 and an F1 score of 0.95. With the power of the interpretability provided with this protocol, our model laid down a set of simple physicochemical rules to determine whether a compound can be a CNS drug using six sub-structural features. These rules displayed higher classification ability than classical guidelines, with higher specificity and more mechanistic insights than just for blood–brain barrier permeability. This hybrid protocol can potentially be used for other drug property predictions.

Key words: machine learning; deep learning; CNS drug; interpretability; graph convolutional network; BBB permeability

Tzu-Hui Yu is an undergraduate at National Taiwan University in Bio-Industry Communication and Development, with interests in cheminformatics and precision medicine.
Bo-Han Su holds a postdoctoral position at the Department of Computer Science and Information Engineering of National Taiwan University, with interests in cheminformatics and AI-assisted drug design.
Leo Chander Battalora is an undergraduate at Temple University in Philadelphia, USA, with interests in robotics, system administration and HPC.
Sin Liu is a PhD student at the Graduate Institute of Biomedical Electronics and Bioinformatics of National Taiwan University. His recent research interests are mainly focused on AI-assisted drug design.
Yufeng Jane Tseng is a Professor at the Graduate Institute of Biomedical Electronics and Bioinformatics, Department of Computer Science and Information Engineering and School of Pharmacy at National Taiwan University and is currently the Director of the Drug Research Center and Associate Director of the Neuroscience Center at National Taiwan University. She specializes in metabolomics and AI-assisted drug design and discovery.
Submitted: 17 June 2021; **Received (in revised form):** 30 July 2021

Introduction

Over the past several decades, the growing needs of neurodegenerative diseases and neuropsychiatric disorders have become increasingly evident. Due to a rapidly aging population, Alzheimer's, Parkinson's and other neurodegenerative diseases have played a substantial role in the growing healthcare expenditures [1–4]. Neuropsychiatric disorders, such as schizophrenia, depression and autism, among other mood disorders, are the leading risk factors for suicide and are responsible for a large portion of the global disease burden [4–6]. These factors have all resulted in a greater demand for CNS drug development [1].

CNS drug development has several significant challenges. The most significant is the presence of the blood–brain barrier (BBB), a selective semipermeable membrane that serves the purpose of preventing injury to the CNS from external insults and toxins [1, 3, 7, 8]. Consequently, the BBB could make drugs that had previously been proven effective in other body parts unusable for CNS targets [7]. Additionally, the limited clinically relevant animal models for CNS drug testing and an incomplete understanding of complex CNS pathogenesis have further hindered the progress of CNS drug development [3, 9–11].

Quantitative structure–activity relationship (QSAR) modeling has been widely used for CNS drug research over the past decade [4, 7, 8, 11, 12]. QSAR modeling is used to predict the effectiveness of drug candidates and to provide useful insight to scientists such as structural features for the BBB penetration [11, 13]. There are many descriptors calculated by previous QSAR models that play a significant role in determining absorption and permeability. Lipophilicity was one of the first significant properties discovered, specifically keeping a LogP under five [14–17]. Low molecular weight has also been shown to result in a higher likelihood of passive lipid-mediated transport across the BBB [14]. High hydrogen bonding was found to inhibit BBB penetration; hence, a low hydrogen bond donor (HBD) count correlated well [8, 17]. Many researchers have used TPSA as a predictor of BBB penetration, with CNS drugs having significantly lower TPSAs [1, 14]. Strong acids such as carboxylic acids have performed poorly in CNS applications, with pKa being another potent descriptor [1, 8, 14]. Pajouhesh and Lenz determined that keeping a low number of rotatable bonds ($\leq$10) improved permeability. Researchers had dedicated themselves to incorporate existing knowledge into simple practical rules for CNS drug design. For instance, a set of rules for good CNS penetration has been proposed by Lipinski; Molecular weight ≤ 400; Log $p \leq 5$; HBD $\leq$3; Hydrogen bond acceptor $\leq$7 [14].

With the advanced machine learning (ML) development, ML algorithms are used more often to construct CNS-QSAR models such as support vector machine (SVM) [18, 19], decision tree (DT) [20, 21] and random forest (RF) [22, 23]. SVM is a non-probabilistic binary linear classifier, meaning it separates samples into one category or the other [7, 11, 24, 25]. This separation is achieved by mapping the samples into a higher order feature space and then finding a linear hyperplane with as large of a gap as possible, separating the two categories. Once the model is trained, new inputs can be mapped into that same space and are measured to see which side of the hyperplane they fall into. A DT is a white-box model that constructs a binary tree of decision nodes that can either use regression or classification. DT's main advantage lies in the visual interpretability of the trained models, which makes it especially appealing to scientists [11, 12]. RF is an ensemble learning algorithm that utilizes the building of multiple DTs. This process helps circumvent DT's main drawback of overfitting to the training set, causing RF to be very popular in scientific works [11, 12]. With the recent improvement in computing capabilities, a different type of learning called deep learning (DL) has become more popular. Among all DL models, the graph convolutional network (GCN) has become a prevalent one in drug discovery. The GCN employs a DL technique of using weighted nodes to make decisions in an attempt to mimic the neural networks of biological brains [24, 26–28]. This method can extract meaningful features from simple descriptions of the graph structure to form molecular-level representations that can be used in place of fingerprint descriptors from conventional ML applications.

Other than ML, DL has been shown to produce incredible results in computer vision [29], speech recognition [30], reinforcement learning [31], text analysis [32] and even CNS drug research [33]. A recent study by Miao in BBB drug classification employed DL algorithms to achieve an accuracy of 0.97, an AUC ROC of 0.98 and an F1 score of 0.92, which is a benchmark result among CNS drug studies [33]. While this result is impressive, it succumbs to the same problems as most DL models, namely, the lack of interpretability. This black-box quality does not improve scientists' understanding of CNS drug design [7]. Furthermore, the size of the collected compound collection used for the training and validation of DL models in Miao's study was restricted to a relatively small dataset. This size results in a limitation of the scope and applicability domains of the models [33].

Most importantly, Miao's study adopted a 5-fold cross-validation based on random partitioning of datasets. This could lead to significant redundancies between training and validation data, resulting in undetected overfitting and overestimation of model performance. Studies suggested that the recent accuracy improvements in QSAR modelling were due to memorization of the similarity among scaffold-based substructures, which has been a long-existing issue in ML model training known as 'data leakage' [34]. Results based on random split methods were less reflective of real-world drug research settings and there existed concerns over model generalizability. A scaffold-split [35] is therefore recommended, which allocates different compound scaffolds to the training and validation sets. Similar splitting methods, such as a fingerprint splitter, could also provide challenging arrangements for evaluating model performance [36, 37].

In this study, we developed and demonstrated a new hybrid approach first using a DL GCN technique to generate an additional descriptor of probability showing whether a compound is considered 'a CNS compound' for each compound and then combine this additional descriptor along with other structural descriptors to generate an SVM model. This ensemble protocol is able to utilize the high prediction power from DL GCN and structural descriptors selected from the final SVM model. Eventually, a set of six sub-structural features were proposed to also provide a quick assessment for CNS drug classification. This set of rules displayed higher classification ability than classical guidelines, with higher specificity and more mechanistic insights.

Method

Compound collection

We initially collected a high-quality, diverse, heterogeneous compound set ($N = 943$ from the literature by Arup *et al.* as our main dataset [38] and an external dataset ($N = 125$) from Seelig's studies as our extended testing set used to evaluate our model's applicability [17]. The four redundant compounds

Table 1. The summary of our collected dataset

	CNS drug	Non-CNS drug
Main Set ($N = 940$) from Arup *et al.* [39]	$n = 315/940$	$n = 625/940$
External Testing Set ($N = 117$) from Seelig *et al.* [17]	$n = 42/117$	$n = 75/117$

(valproic acid, meprobamate, amobarbital and amantadine) between the two datasets were removed from the external dataset. We only included the approved drugs in our dataset since any lead compounds or drug candidates that fail during preclinical toxicity studies and investigational drugs that fail during clinical trials can distort our models. The classification of those compounds has been carefully investigated by previous studies [17, 38]. In the two datasets, the CNS drugs were known to have a strong tendency of penetration into the brain according to proven clinical or experimental evaluations, while the non-CNS drugs were validated with low or no tendency of BBB penetration or with no binding affinity or pharmacologic effect to any receptors in the brain. Arup *et al.* confirmed that the 943 compounds were all oral drugs. Drugs with CNS side effects at therapeutic doses were also removed from the non-CNS drugs. We further checked and curated the whole dataset again according to these conditions and literature surveys. A small fraction of compounds for which classification was ambiguous in both sets was also disregarded. A summary of the two finalized datasets is shown in Table 1. Our main set included 940 compounds (315 CNS and 625 non-CNS drugs), while there were 117 compounds (42 CNS and 75 non-CNS drugs) in the external testing set. The full datasets are provided in Supplementary Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/.

Descriptor calculation

In our study, molecular descriptors and several fingerprints were used, alone or in combination, to develop classification models for the identification of CNS drug candidates. All the descriptors were calculated by PaDEL-Descriptor [39], software developed by the National University of Singapore. PaDEL descriptors were calculated with the aid of the Chemistry Development Kit and several additional programs, including atom-typed electrotopological state descriptors, Crippen's LogP, molar refractivity, extended topochemical atom descriptors and molecular linear free energy relation descriptors. The PaDEL descriptors contain 1444 1D and 2D descriptors, 431 3D descriptors and 12 types of fingerprints. We chose only 1D and 2D descriptors, along with nine types of fingerprints in our studies. We also removed some descriptors that cannot be normally computed by PaDEL-Descriptor. The final list of 17 473 descriptors was composed of 1381 1D and 2D molecular properties and 16 092 fingerprints. The frequency distribution of these descriptors was analyzed in the results section to ensure the diversity of the chemical and physiochemical properties in our dataset.

QSAR ML model construction

The protocol of this study is illustrated in Figure 1. We developed and evaluated a series of QSAR models to identify and predict potential CNS drugs. We first systematically built three nonlinear categorical models based on traditional ML methods, including DT, RF and SVM. These models were implemented using the Python Scikit-learn package [40]. The main set of 940 drugs was divided into a training set and a validation set at an 80:20 ratio, based on the Tanimoto similarity of their ECFP4 fingerprints [41, 42], a 1024D extended connectivity fingerprint system which is not included in the PaDEL descriptors. This fingerprint partitioning method aims to make molecular properties in each dataset as different as possible from the ones in the other datasets. We further used a web-based visualization tool [43] to visualize how our datasets were split in different datasets based on the ECFP4 fingerprints. The rendering was provided in Supplementary Figure S1, see Supplementary Data available online at http://bib.oxfordjournals.org/. The training, validation and testing sets were, respectively, projected into a 3D chemical space obtained by principal component analysis. The figure showed that the training and validation set were generally distributed in a diverse and heterogeneous region. The training set was used to generate and fit the models, while the validation set was used to tune and optimize the models. We tuned model hyperparameters according to the performance of the validation dataset, avoiding our models generating predictions via memorization instead of generalization, a long-existing issue in ML model training known as 'data leakage' [34]. We further imported an external testing set ($N = 117$) to evaluate model performance.

The models were optimized mainly via descriptor selection and parameter adjustment. Model performances are evaluated based on the accuracy of the validation set. According to Lo *et al.* [28], a small and non-redundant descriptor set could not only provide a more cost-effective calculation of descriptors but also help to understand the underlying relationship between those physicochemical descriptors and the property of interest. As PaDEL descriptors comprise many inter-correlated molecular features, potentially causing multi-collinearity or overfitting. As a result, we screened out irrelevant descriptors according to the literature reviewed in the introduction section and used descriptor importance to select a minimum subset of descriptors that are most related to the prediction of CNS drugs. The descriptor importance is calculated in different ways across various ML models. In DT and RF models, the importance of a descriptor can then be evaluated by the mean decreased Gini index on that descriptor across the trees [44]. We utilized the feature importance information to understand underlying mechanisms behind our ML models and provide insights for feature selection. In SVM models, the classification result is a hyperplane that separates the classes as best as possible. There are weights representing this hyperplane, by giving the coordinates of a vector that is orthogonal to the hyperplane—these are the coefficients given by SVM models. We interpreted these coefficients as the importance of different features.

In terms of parameter adjustment, we chose hyperparameters that could provide general and accurate predictions. We tuned DT models by hyperparameters including the maximum

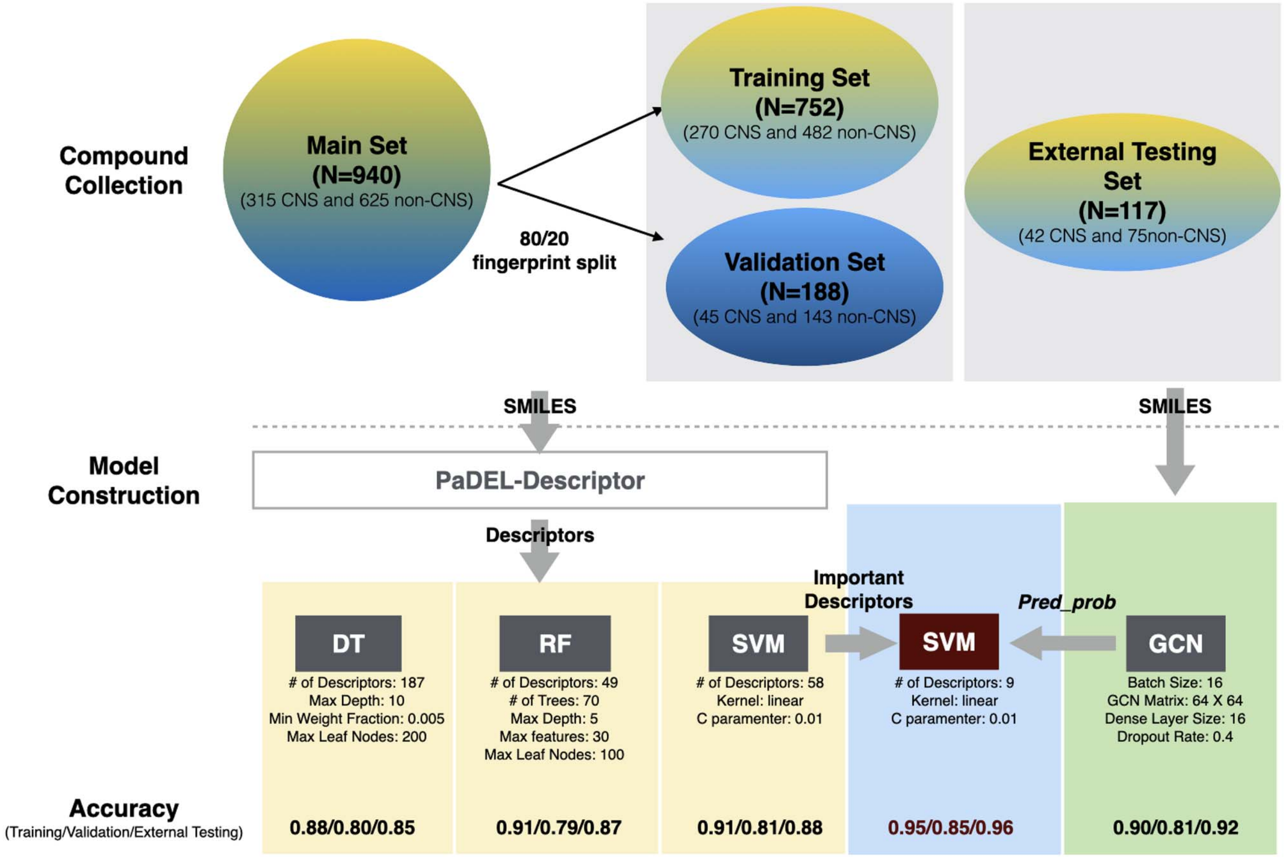


Figure 1. Overall protocol of the studies. (i) Compound collection: Our main data set contained 940 market drugs and was later split into training ($N = 752$) and validation ($N = 188$) sets using fingerprint-based partitioning. An additional external dataset of 117 market drugs was imported to evaluate model performance. (ii) Model construction: Three traditional ML models (DT, RF and SVM) were built using PaDEL descriptors. A Graph Convolution Network DL model was also built, which input SMILES notations and output the predicted probability of each compounds being a CNS active drug (Prob_pred). Lastly, a hybrid ensemble SVM model took in Prob_pred and important descriptors learned from traditional ML models to predict CNS-activeness. (iii) Accuracy: The hybrid ensemble SVM model exhibited the highest accuracy. Training Set ($N = 752$).

depth, the maximum number of leaf nodes and the minimum weighted fraction of the sum total of weights required to be at a leaf node. The RF model hyperparameters being tuned included the number of trees in the forest, the maximum depth and the max number of features considered for splitting a node. The SVM model was tuned mainly by the regularization parameter C that determines the trade-off between tolerating training errors and forcing rigid margins [40]. Detailed hyperparameter ranges are exhibited in Supplementary Table S3, see Supplementary Data available online at http://bib.oxfordjournals.org/.

GCN DL model construction

We further adopted a GCN architecture [27] to build a CNS drug classifier. The same datasets used in the ML models were applied in this study, with a main dataset (split by fingerprints) and an external testing dataset. The GCN model was implemented by DeepChem 2.0 [41], a popular open-source library for DL implementation in drug discovery. DL approaches perform automatic feature extractions during the learning processes. We directly imported the simplified molecular input line entry system (SMILES) notations of the compounds as input into the GCN model without prior calculations nor a manual selection of descriptors.

We used Adam algorithms [45] to optimize the GCN models with a learning rate of 0.001. The training procedures were tuned by different hyperparameters, including batch size, size of graph convolution matrix, dense layer size and dropout rate. The ranges of these hyperparameters are exhibited in Supplementary Table S2, see Supplementary Data available online at http://bib.oxfordjournals.org/. Each model was trained with an epoch size of 300 due to limitations in time and graphics processing unit efficiency. The GCN model also calculated the predicted probability of each compound being a CNS active drug (*Prob_pred*).

Comprehensive model construction combined with ML and DL approaches

Following the results from the former section, we developed a novel hybrid model, adding the predicted probability of CNS activity (*Prob_pred*) generated by the GCN to the best performing ML model (SVM) as an additional descriptor input. We used the same training, validation and external testing sets as those in previous ML and DL models. The process for tuning this final model was similar to the process used for our QSAR models, in which several descriptor sets and algorithm hyperparameters were tested to achieve the best model performance.

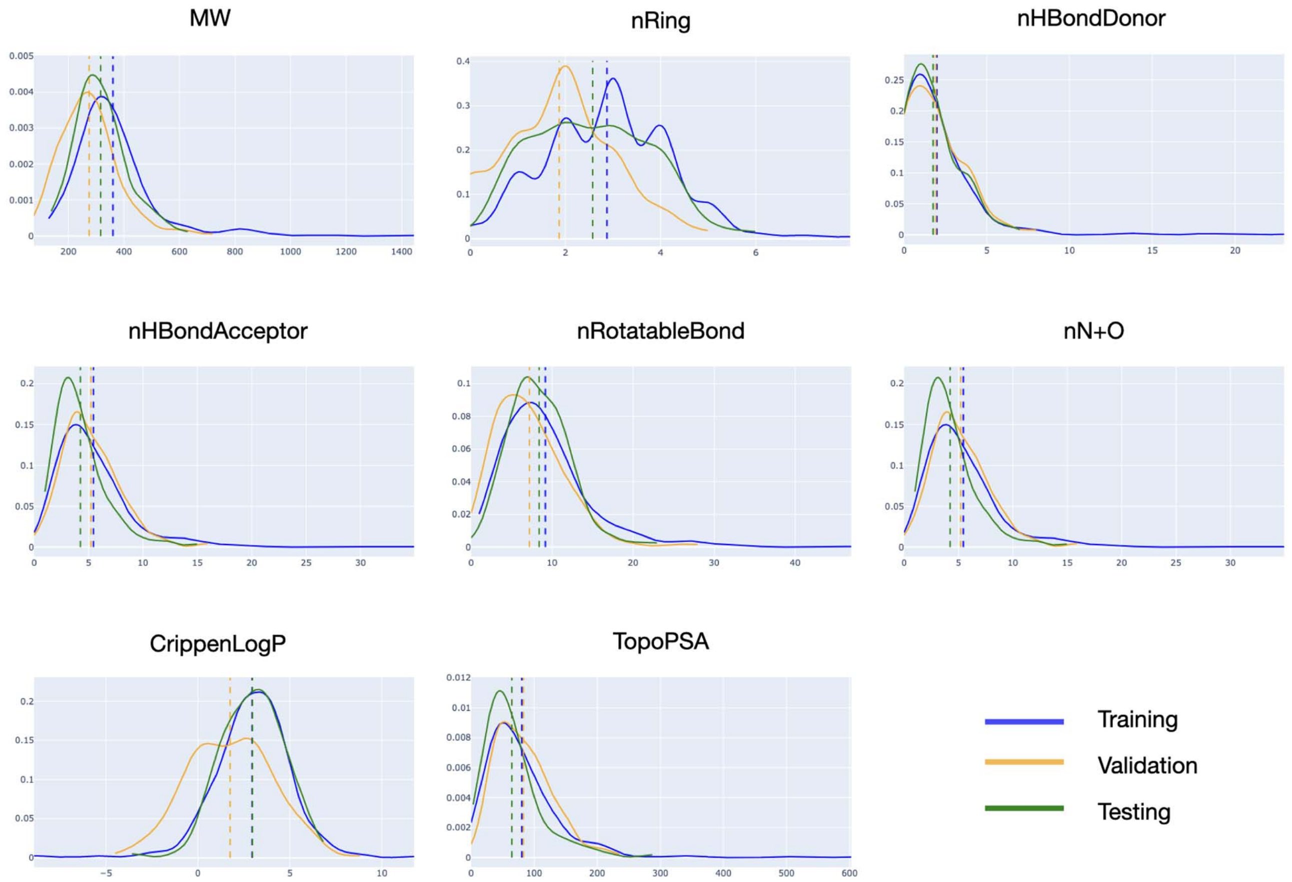


Figure 2. Distribution of important descriptors in the three datasets. The *x*-axis represented values of the selected nine descriptors including MW, nRing, nHBondDonor, nHBondAcceptor, nOrtatableBond, nN + O, CrippenLogP and TopoPSA. The *y*-axis represented a likelihood that a molecule obtains the corresponding value of the *x*-axis in the data set, and the values were evaluated by probability density function. The blue, orange and green curves illustrated the distribution in the training, validation and testing data sets, respectively. The plots demonstrated the training set and validation set covers diverse, heterogeneous distribution.

Finally, descriptor importance scores were calculated, which we reviewed in Results section.

Results

Chemical space

First, we analyzed the descriptors known to be determinants of CNS activity to compare the chemical space of our three datasets [11]. In Figure 2, we used probability density function to demonstrate the distribution of the eight descriptors in the three datasets. The molecules in the validation set have explicitly different distribution in molecular weight, number of rings, number of rotatable bonds and logP compared with the training set. The distribution of compounds in testing set covers obvious diversity compared with the training and validation set in most of the descriptors, except for the number of HBDs and logP. This result implied that the splitting method was effective in generating two datasets with different molecular properties, so predicting the validation sets would require extrapolating far outside the training data, which could potentially prevent data leakage caused by overfitting, ensuring the generalization of the ML and DL models. Furthermore, the selected external testing set was suitable to confirm the performance of the model.

Overall results from the ML models

Using three different algorithms and multiple subsets of the available molecular descriptors, hundreds of CNS activity prediction QSAR models were trained. The best performance in each ML model was listed in Table 2.1 and the optimized hyperparameters were described in Figure 1. Our benchmark ML model was the linear SVM classification model with regularization parameter C equal to 0.01, using 58 physicochemical descriptors. This hyperparameter and descriptor setting avoided overfitting and allowed this SVM model to achieve an accuracy of 0.81 on the validation set and 0.88 on the external testing set, performing the best among all other ML models (see Figure 3).

Results from the GCN models

In the configuration of the final optimum model, the batch size was 16, the size of the graph convolution matrix was 64 * 64, the dense layer size was 16 and the dropout rate was 0.4. Table 2.2 shows the scores of our benchmark GCN model, whose performance is comparable with the optimum SVM model. Compared with the former ML models, the GCN approach provided higher accuracy in the validation set.

Table 2. Comparison among the five models in the study

		Fingerprint-split validation			Random-split validation		
		Accuracy[a]	AUC	F1 score	Accuracy[a]	AUC	F1 score
Table2.1 ML models							
SVM	**Training Set**	0.912 (±0.000)	0.962	0.877	0.937 (±0.000)	0.974	0.893
	Validation Set	0.809 (±0.000)	0.820	0.667	0.872 (±0.000)	0.948	0.853
	External Testing Set	**0.882 (±0.000)**	**0.948**	**0.837**	**0.907 (±0.000)**	**0.957**	**0.801**
DT	**Training Set**	0.883 (±0.016)	0.974	0.893			
	Validation Set	0.798 (±0.023)	0.753	0.825			
	External Testing Set	0.851 (±0.013)	0.913	0.856			
RF	**Training Set**	0.978 (±0.000)	0.998	0.969			
	Validation Set	0.792 (±0.000)	0.818	0.636			
	External Testing Set	0.874 (±0.000)	0.957	0.835			
Table 2.2 DL models (GCN)							
GCN	**Training Set**	0.963 (±0.000)	0.976	0.948	0.949 (±0.000)	0.998	0.925
	Validation Set	0.803 (±0.002)	0.901	0.734	0.867 (±0.006)	0.871	0.733
	External Testing Set	**0.916 (±0.001)**	**0.878**	**0.909**	**0.924 (±0.000)**	**0.873**	**0.905**
Table2.3 Comprehensive models (SVM + GCN)							
SVM + GCN	**Training Set**	0.949 (±0.000)	0.987	0.925	1.000 (±0.000)	1.00	1.000
	Validation Set	0.846 (±0.000)	0.842	0.781	0.989 (±0.000)	1.000	0.984
	External Testing Set	**0.958 (±0.000)**	**0.983**	**0.955**	**0.974 (±0.000)**	**0.978**	**0.963**

[a] Each accuracy data field displays average ± std of 50 repeated trials' accuracy scores, while the AUC and the F1 score fields only showed average score for simplicity.

Results from the hybrid model

After testing many descriptor subsets to optimize the hybrid ensemble model, we settled on a set of eight descriptors. The set of molecular features with the highest importance scores from our optimum ensemble model are listed in Table 3 and Figure 4. The determinant descriptors included heterocycles (*MACCSFP121*), hydrogen bonding (*nHBDon_Lipinski, MLFER_BH*), acidity (*nAcid*), aromatic structure (*MACCSFP65, MACCSFP19*), charge (*MACCSFP49*) and electronegativities (*Mpe*). The frequency distribution of these descriptors (see Table 3) is significantly different among CNS and non-CNS data sets ($P < 0.01$), indicating the high importance of these descriptors to distinguish CNS drugs. These properties were also in line with previous studies which pointed out these properties' impact on drug permeability through the BBB [46, 47] but not fairly reflected in the GCN model.

For comparison, we also trained the model using random-split validation methods. The results revealed a potentially inflated predictivity under the random split method, with an accuracy 1.00 on the training set and 0.99 on the validation set (see Table 2.3).

Discussion

Advantages in hybrid model construction and evaluation

In this work, the new hybrid ensemble model outperformed any other separate conventional ML/DL models. The separate models encountered accuracy bottlenecks despite a series of tuning and optimization processes [48]. The proposed hybrid ensemble protocol, where both SVM and GCN classification algorithms were used, broke the glass ceiling of model performance. Compared with the benchmark ML model, the hybrid ensemble model increases the accuracy of the training, validation and external testing sets, by 4, 4 and 8%, respectively. The performance was achieved mainly through improvements in model specificity. Past studies suggested hybrid models performed better in a probabilistic approach, rather than deterministic. This modeling method could combine complementary properties exhibited by single existing models and further augment the generalization capability [49].

Another advantage of our study was that we strictly evaluated the model's performance to ensure generalization capability. We adopted a fingerprint-splitting method to generate an internal validation data set. This splitting method could ensure a better examination of ML/DL models' true learning abilities [35]. Based on our validation set, a variety of hyperparameters and descriptor sets were adopted to tune the model. The result revealed that even in the most extreme case, in which the distribution of the testing data was extremely different from the training data, we could still guarantee accuracy of 0.85 (see Table 2.3), showing that the hyperparameters were tuned effectively to help models adapt and react properly to unseen data. This could provide future researchers insights in tuning models when considering enhancing model generalization ability.

Key feature interpretation

The optimal hybrid ensemble suggested eight determinant descriptors for CNS drug classification. While literature had pointed out the relationship between CNS activeness and these eight descriptors, it did not adequately emphasize their significant impacts based on the results of our analysis. Herein, we would discuss these features one by one, in the order of their importance ranking suggested by our model. We performed

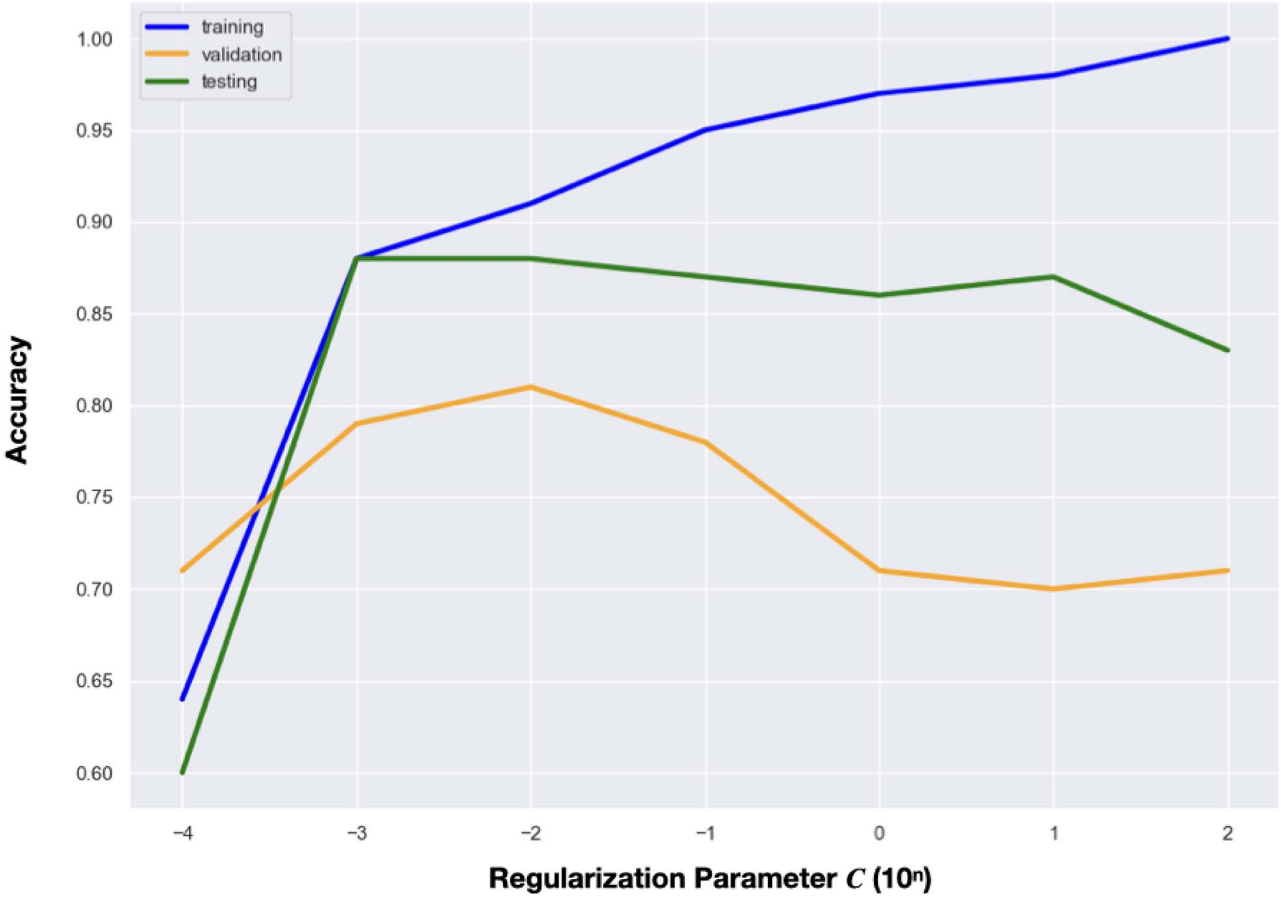


Figure 3. The correlation between regularization parameter and accuracy of the SVM model. With greater value of C, the SVM algorithm more correctly got all the training compounds classified while leading to a stifled predictive performance in the validation and test sets. When C is equal to 0.01, the model achieved the best performance for the validation set and still kept a higher accuracy for the external testing set.

Table 3. Predictor importance for the SVM classification model

	Predictors (abbreviation)	Predictor Interpretation	Coeff	Mean of CNS Drugs	Mean of Non-CNS Drugs	P-value
1	pred_prob	The probability of becoming CNS drugs, predicted by the benchmark GCN model	1.504	0.80	0.05	<0.01
2	MACCSFP121[a]	Nitrogen heterocycle	0.244	0.76	0.49	
3	nHBDon_Lipinski	Number of HBDs	−0.185	0.93	2.48	
4	nAcid	Number of acidic groups	−0.164	0.04	0.4	
5	MACCSFP65[a]	Aromatic bond between carbon and nitrogen, denoted by C%N	−0.149	0.18	0.27	
6	MACCSFP19[a]	Seven-membered ring	0.129	0.23	0.05	
7	MACCSFP49[a]	charge	−0.113	0.02	0.08	
8	Mpe	Mean atomic Pauling electronegativities	−0.098	0.96	0.98	
9	MLFER_BH	Overall or summation solute hydrogen bond basicity	−0.056	1.36	1.79	

[a] The MACCS Fingerprints are binary encoded descriptors, using the digits 0 and 1 to represent the presence or absence of particular functional groups, atoms or fragments.

frequency distribution analysis on these physicochemical descriptors using all compounds ($N = 1057$) in our study, including the main set ($N = 940$) and the external testing set ($N = 117$). By this, we could see how these physicochemical properties could separate CNS drugs from non-CNS drugs.

First, *MACCSFP121* (the existence of the nitrogen heterocycle) was the most important molecular feature. In our study, 76% of CNS drugs had at least one nitrogen heterocycle, while only 49% of non-CNS drugs had any nitrogen heterocycles. Nitrogen-based heterocycles could display a broad range of biological

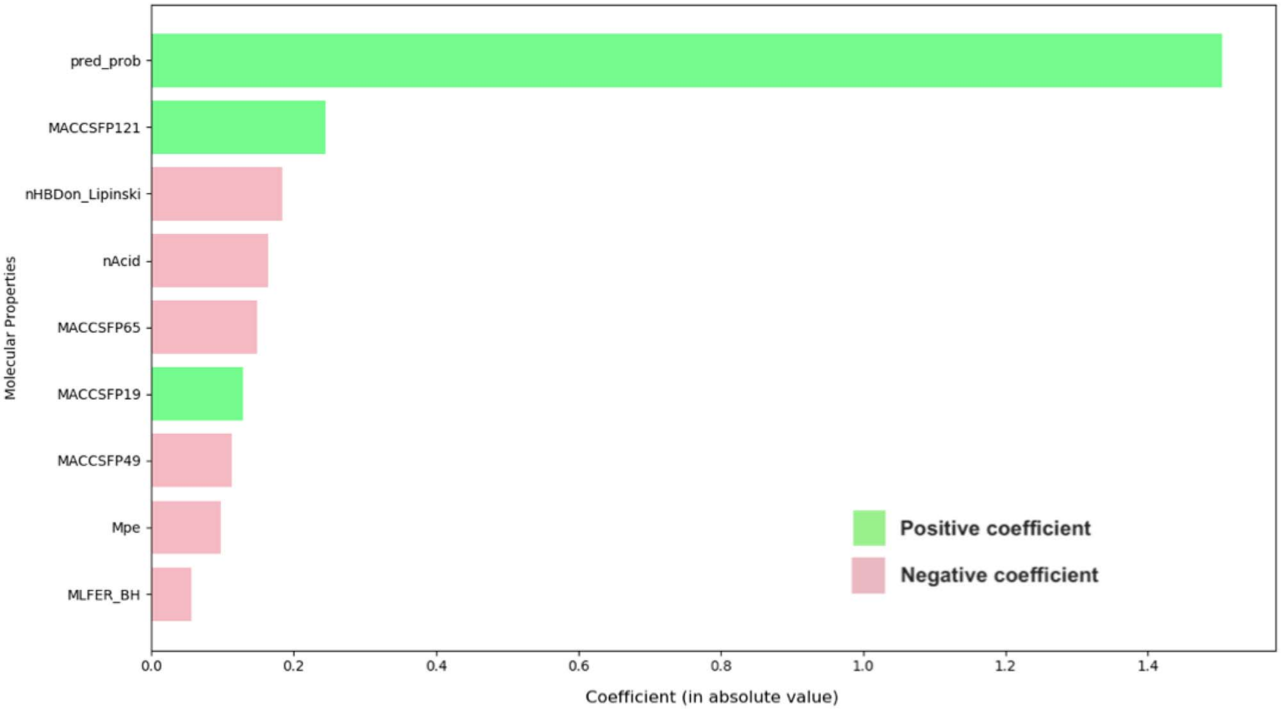
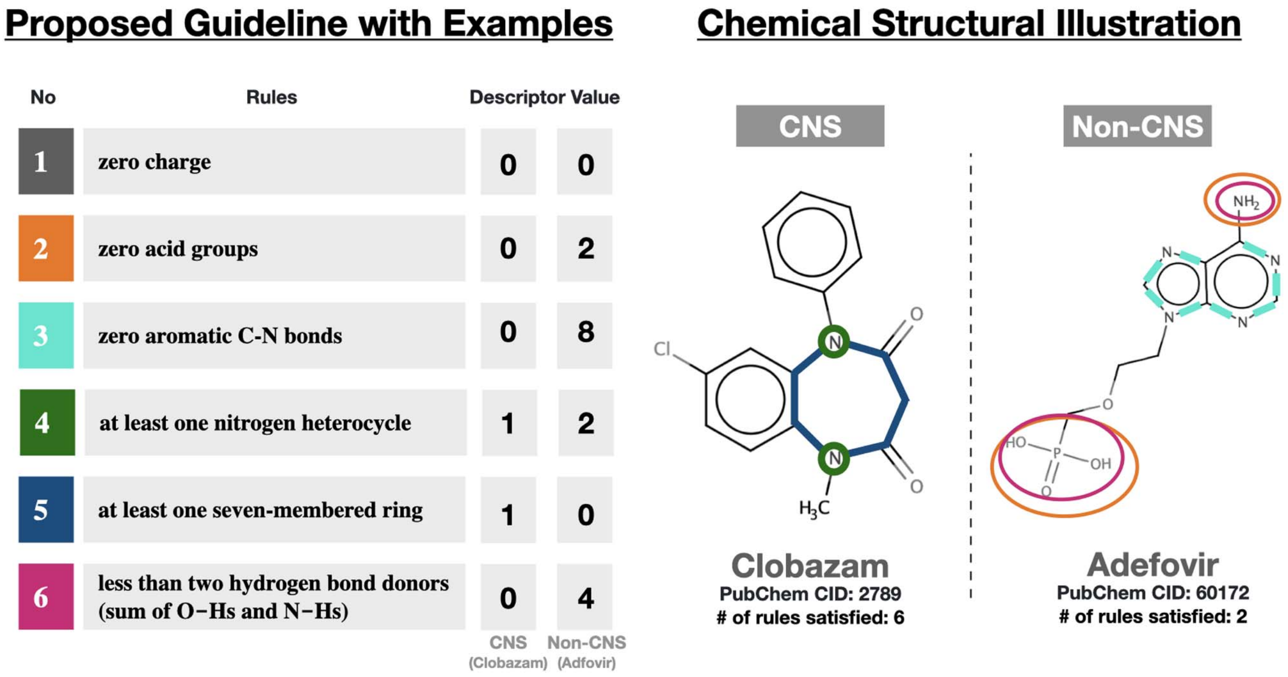


Figure 4. Predictor importance for the SVM classification model. The chart depicted model coefficients—a measure of feature importance in an SVM—in the order of importance with green bars representing positive coefficients and pink bars representing negative coefficients.



No	Rules	Descriptor Value	
		CNS (Clobazam)	Non-CNS (Adfovir)
1	zero charge	0	0
2	zero acid groups	0	2
3	zero aromatic C-N bonds	0	8
4	at least one nitrogen heterocycle	1	2
5	at least one seven-membered ring	1	0
6	less than two hydrogen bond donors (sum of O−Hs and N−Hs)	0	4

Figure 5. The illustration of proposed guideline to determine CNS activeness. A well-known CNS drug, Clobazam and a common non-CNS drug Adefovir were taken as examples to illustrate how our guideline classifies CNS drugs. Proposed guideline on the left lists rulesets proposed and the calculated descriptor values of the two molecules on the right. Each rule was marked in different colors. Corresponding features in the molecule which matched to the rule were also highlighted with the same color of the rule in the structure scheme.

activities and have therefore become a useful tool to manipulate lipophilicity, polarity and hydrogen bonding capacity, due to the broad range of biological activities. Among all nitrogen heterocyclic compounds, morpholine—with the presence of a nitrogen at the opposite position of an oxygen atom—have a peculiar pK value and a flexible structure, helping it reach a hydrophilic–lipophilic balance, enhancing blood solubility and brain permeability [50, 51].

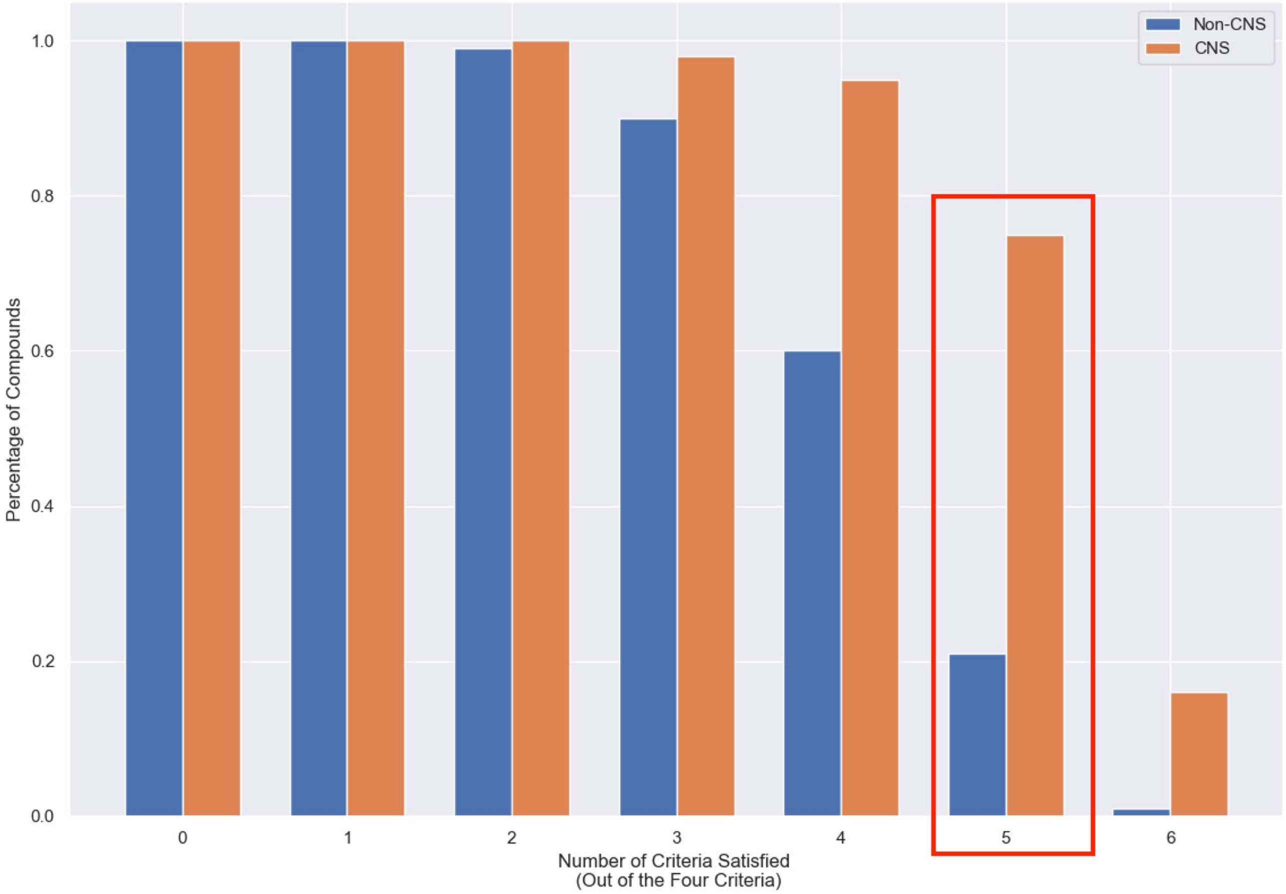


Figure 6. Frequency distribution of CNS drugs based on the proposed guideline. The *x*-axis represented the number of rules satisfied (out of the six rules specified above) and the *y*-axis represented the percentage of compounds satisfying the rules. The bars outlined in red were the suggested threshold of this classification guideline. There were two groups of compounds: CNS (orange) and Non-CNS (blue).

The following descriptors were *nHBDon_Lipinski* (the number of HBDs) and *nAcid* (the number of acid groups). The importance of HBDs had been emphasized in most CNS QSAR modeling, representing hydrogen bond-forming potential. Increasing hydrogen bonding decreases BBB penetration. Many CNS drug research measure hydrogen bonding ability by simply counting the number of hydrogen bond acceptors or donors [14]. In our study, 74% of the CNS drugs were with less than or equal to 2 HBDs, while the percentage was about half as many among the non-CNS drugs (39%), indicating a significant difference. In terms of *nAcid*, strong acids tend to be precluded from BBB penetration [52]. Most of the CNS drugs (97%) in our data set had no acid groups, and only a few of them (3%) possessed one acid group. Interestingly, nearly half of the compounds with one acid group possessed an amino acid, for example Gabapentin. Several amino acids transport ligands into the CNS via selective efflux transporter P-glycoprotein [38].

The rest of the determinant descriptors were less discussed in the literature, yet they were still proven to be related to CNS activeness. *MACCSFP 65* (the existence of an aromatic bond between carbon and nitrogen, denoted by C%N) was evaluated as a negative factor in our model that could decrease a compound's likelihood to overcome the BBB. The addition of aromatic rings could lead to an increase in log P, enhancing lipophilicity; however, it was also likely to increase the topological polar surface area of the molecule and therefore compromise lipophilicity [1, 14, 53]. The combination of the two descriptors, *MACCSFP 121* (N-Heterocycle) and *MACCSFP 65* (C%N), could help reach an optimal balance of log P and TPSA. With the existence of nitrogen heterocycles, a compound's basic level of log P was ensured; on the other hand, the absence of aromatic bonds between carbons and nitrogens could limit a TPSA increase. Empirically, among all nitrogen heterocyclic compounds, 76% of CNS drugs did not have any C%N substructures (44% among non-CNS drugs). This indicated that a compound with nitrogen heterocycles, but without C%N, displayed higher brain uptake.

MACCSFP 19 represents the existence of seven-membered rings. Common seven-membered rings, including azepines, benzodiazepines and diazepines, demonstrate a broad range of biological activities and have been important structural scaffolds of CNS drugs [54]. In our study, 23% of CNS drugs had seven-membered ring structures, while that of non-CNS drugs was only 5%. In terms of *MACCSFP 49* (charge), a significant electrostatic charge could prohibit passive diffusion into the BBB [55]. In our study, 97% of CNS drugs have zero net charges. *Mpe* (Mean atomic Pauling electronegativities) and *MLFER_BH* (Overall or summation solute hydrogen bond basicity) were computed values related to hydrogen bonding properties, which had been discussed above in the *nHBDon_Lipinski* subsection [56–58].

Seeing that structural features have higher importance scores in our model and can provide synthetically tractable insights; here, we thoroughly discussed the top six descriptors. A compound is more likely to be CNS-active if it has

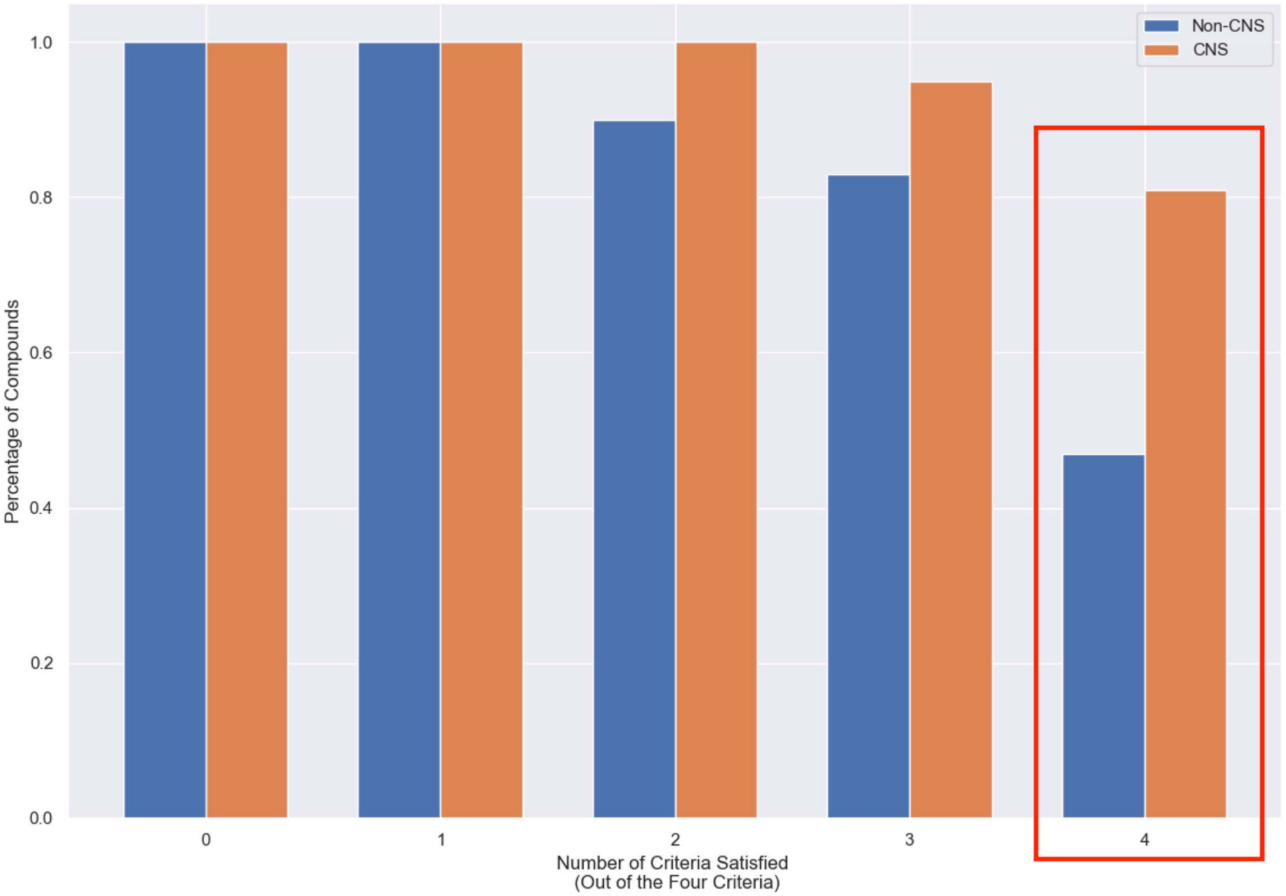


Figure 7. Frequency distribution of CNS drugs based on RoCNS. The x-axis represented the number of rules satisfied (out of the four rules in RoCNS) and the y-axis represented the percentage of compounds satisfying the rules. The bars outlined in red were the suggested threshold of this classification guideline. There were two groups of compounds: CNS (orange) and Non-CNS (blue).

(i) zero charge
(ii) zero acid groups
(iii) zero aromatic C-N bonds
(iv) at least one nitrogen heterocycle
(v) at least one seven-membered ring
(vi) less than two HBDs (sum of O-Hs and N-Hs).

To further show that the above rules are effective for the classification of CNS drugs versus non-CNS drugs, a *t*-Distributed Stochastic Neighbor Embedding method (*t*-SNE) [59] was employed, projecting the whole dataset from the nine descriptors listed in Table 3 into a 2D chemical space (see Supplementary Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/). Supplementary Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/, was used to represent the similarity of the features between the compounds on a 2D embedding space. We specifically selected five clusters of compounds that are highlighted in Supplementary Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/, for further the interpretation of our rules. Detailed physiochemical properties of each cluster were provided in Supplementary Table S2, see Supplementary Data available online at http://bib.oxfordjournals.org/, and discussed in the supplementary file. Each representative compound in the five cluster was also listed in Supplementary Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/, to enhance the interpretability of the features. From our analysis, the properties in the representative five clusters were explicitly matched to our proposed six rules for the prediction of CNS drugs.

Clobazam [60], a well-known benzodiazepines for treating anxiety that was correctly classified as a CNS drug in our dataset, is chosen as an example to demonstrate how the CNS rules were applied. Adefovir [61], a common non-CNS drug, is chosen as a negative example to show how our proposed guideline rules out a compound failing to penetrate BBB. The rulesets and the corresponding features are reported in Figure 5. The figure shows that the marketed CNS drug, Clobazam, does not have any electric charges, acid groups, aromatic C-N bonds and hydrogen bonds while having a seven-membered ring with two nitrogen atoms; this satisfies all six proposed rules to become a CNS drug. In fact, Clobasam is not the only member of Benzodiazepines that satisfies the rules. All benzodiazepines are characterized by their seven-membered nitrogen heterocycles, which could serve as physiological ligands of specific neuronal binding sites and ensure pharmacological heterogeneity. Most benzodiazepines have fewer HBDs and do not have charge and acid groups [62, 63]. These features all contribute to enhance the BBB permeability of benzodiazepines, making them common CNS drugs. On the other hand, Adefovir has two acid groups, eight aromatic

C-N bonds, four H-bond donors and zero seven-membered rings. These properties violate four of our proposed rules (out of six), making Adefovir less likely to become a CNS drug.

From our analysis, when a compound meets five or more criteria out of the six, it would exhibit better BBB permeability (accuracy: 0.77, sensitivity: 0.75, specificity: 0.78). The frequency distribution in Figure 6 revealed a strong relationship between the number of criteria met and the likelihood to become a CNS drug. Compared with most CNS studies, this paper provides a new angle to this topic, instead of focusing on conventional descriptors such as TPSA, MW and pKa. We leverage simple structural features to provide practical insights for novel lead molecule synthesis and appropriate structural modifications.

In comparison to prevalent theories such as Lipinski's Rule for CNS drugs (RoCNS) [14], our rule could better distinguish CNS and non-CNS-active compounds. In our data set, 75% of CNS drugs adhere to our guideline, and only 21% of non-CNS drugs follow this guideline. On the other hand, although 80% of CNS drugs comply with Lipinski's RoCNS, 45% of non-CNS drugs also satisfy the condition (Figure 7), indicating a poor performance in separate two types of drugs.

Conclusion

This study tackled the classical black-box learning models for CNS drug classification by (i) novel model ensemble approach, (ii) knowledge-based descriptor selection and (iii) rigid evaluation criteria. We used the predicted probabilities (*pred_prob*) as a way to access the probabilities of whether a compound can be a CNS compound generated by the GCN model. These probability values of each compound were used further as an additional descriptor in a linear SVM model with other eight determinant molecular features. The model stringency was ensured by adopting a fingerprint splitter to split the dataset into training and validation sets, two data subsets with different molecular properties. Using the validation set for hyperparameter tuning, the model could extrapolate far outside the training data, rewarding generalization. An external data set was used to evaluate model performance. Our model pursued interpretability and generalizability with due consideration of predictability. The optimum model reported the accuracy of 0.96, ROC AUC of 0.98 and F1 score of 0.94, comparable with past benchmark CNS QSAR studies.

Noticeably, the reduced feature dimensionality not only avoided overfitting but also provided structural insights. Lipinski's Rule for CNS drugs was supported by past studies although the challenges also lied on its over-simplified rules in sub-structural features. Compared with Lipinski's rules, the six sub-structural features proposed by this work have higher discriminative power for CNS drug classification and provide more details in sub-structures. This hybrid protocol can also be used for other endpoints not limited to CNS drug classification.

Key Points

- A new protocol to combine the power of ML and DL was developed to provide interpretable generalized rules with high prediction power for CNS drug classification.
- The resulting hybrid ensemble model outperformed past benchmark CNS QSAR studies with an accuracy of 0.96 and an F1 score of 0.95 while maintaining high interpretability.
- Our model laid down a set of simple physicochemical rules to determine CNS activeness using six substructural features with higher discriminative power and more mechanistic insights.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgements

We thank the Laboratory of Computational Molecular Design and Metabolomics and the Department of Computer Science and Information Engineering of National Taiwan University provided resources in performing these studies.

Associated Content

The code to implement ML and DL models is available at https://github.com/tzuhuiyuatntu/cnsstudy2019.

Funding

Ministry of Science and Technology, Taiwan (grant numbers MOST 109–2320-B-002-040 and MOST 109–2823–8-002-010); National Taiwan University (grant number NTU-CC-110 L890803 and NTU-110 L8809); Ministry of Health and Welfare (grant number MOHW110-FDA-D-114-000611).

References

1. Rankovic Z. CNS drug design: balancing physicochemical properties for optimal brain exposure. *J Med Chem* 2015;**58**:2584–608.
2. Prinz M, Priller J. The role of peripheral immune cells in the CNS in steady state and disease. *Nat Neurosci* 2017;**20**:136–44.
3. Danon JJ, Reekie TA, Kassiou M. Challenges and opportunities in central nervous system drug discovery. *Trends Chem* 2019;**1**:612–24.
4. Makhouri FR, Ghasemi JB. In silico studies in drug research against neurodegenerative diseases. *Curr Neuropharmacol* 2018;**16**:664–725.
5. Hyman SE. A glimmer of light for neuropsychiatric disorders. *Nature* 2008;**455**:890–3.
6. Menken M, Munsat TL, Toole JF. The global burden of disease study: implications for neurology. *Arch Neurol* 2000;**57**:418–20.
7. Doniger S, Hofmann T, Yeh J. Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J Comput Biol* 2002;**9**:849–64.
8. Lanevskij K, Japertas P, Didziapetris R. Improving the prediction of drug disposition in the brain. *Expert Opin Drug Metab Toxicol* 2013;**9**:473–86.
9. Wager TT, Hou X, Verhoest PR, *et al*. Central nervous system multiparameter optimization desirability: application in drug discovery. *ACS Chem Nerosci* 2016;**7**:767–75.
10. Vastag M, Keseru GM. Current in vitro and in silico models of blood-brain barrier penetration: a practical view. *Curr Opin Drug Discov Devel* 2009;**12**:115–24.

11. Zhang YY, Liu H, Summerfield SG, *et al*. Integrating in silico and in vitro approaches to predict drug accessibility to the central nervous system. *Mol Pharm* 2016;**13**: 1540–50.
12. Shi J, Zhao G, Wei Y. Computational QSAR model combined molecular descriptors and fingerprints to predict HDAC1 inhibitors. *M S-Med Sci* 2018;**34**(Focus issue F1):52–8.
13. Goodwin JT, Clark DE. In silico predictions of blood-brain barrier penetration: considerations to "keep in mind". *J Pharmacol Exp Ther* 2005;**315**:477–83.
14. Pajouhesh H, Lenz GR. Medicinal chemical properties of successful central nervous system drugs. *NeuroRx* 2005;**2**:541–53.
15. Chen H, Winiwarter S, Engkvist O. *In Silico Tools for Predicting Brain Exposure of Drugs. Blood-Brain Barrier in Drug Discovery*. Hoboken, NJ: John Wiley & Sons, Inc., 2015.
16. Fischer H, Gottschlich R, Seelig A. Blood-brain barrier permeation: molecular parameters governing passive diffusion. *J Membr Biol* 1998;**165**:201–11.
17. van de Waterbeemd HC, Camenisch G, Folkers G, *et al*. Estimation of blood-brain barrier crossing of rrugs using molecular size and shape. And H-bonding descriptors. *J Drug Target* 1998;**6**:151–65.
18. Shen MY, Su BH, Esposito EX, *et al*. A comprehensive support vector machine binary hERG classification model based on extensive but biased end point hERG data sets. *Chem Res Toxicol* 2011;**24**:934–49.
19. Chang CY, Hsu MT, Esposito EX, *et al*. Oversampling to overcome overfitting: exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods. *J Chem Inf Model* 2013;**53**:958–71.
20. Shao CY, Su BH, Tu YS, *et al*. CypRules: a rule-based P450 inhibition prediction server. *Bioinformatics* 2015;**31**: 1869–71.
21. Wang PH, Tu YS, Tseng YJ. PgpRules: a decision tree based prediction server for P-glycoprotein substrates and inhibitors. *Bioinformatics* 2019;**35**:4193–5.
22. Sherer EC, Verras A, Madeira M, *et al*. QSAR prediction of passive permeability in the LLC-PK1 cell line: trends in molecular properties and cross-prediction of Caco-2 permeabilities. *Mol Inform* 2012;**31**:231–45.
23. Svetnik V, Liaw A, Tong C, *et al*. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 2003;**43**:1947–58.
24. Vapnik VN. An overview of statistical learning theory. *IEEE Trans Neural Netw Learn Syst* 1999;**10**:988–99.
25. Vapnik VN. *The Nature of Statistical Learning Theory*. New York: Springer, 2000.
26. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. *Adv Neural Inf Process Syst* 2016;**29**:3844–52.
27. Renn A, Su BH, Liu H, *et al*. Advances in the prediction of mouse liver microsomal studies: from machine learning to deep learning. *Wiley Interdiscip Rev Comput Mol Sci* 2021;**11**:e1479.
28. Lo YC, Rensi SE, Torng W, *et al*. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 2018;**23**:1538–46.
29. Krizhevsky A, Sutskever I, Hinton GE. Image net classification with deep convolutional neural networks. *Commun ACM* 2017;**60**:84–90.
30. Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process Mag* 2012;**29**:82–97.
31. Arulkumaran K, Deisenroth MP, Brundage M, *et al*. Deep reinforcement learning: a brief survey. *IEEE Signal Process Mag* 2017;**34**:26–38.
32. Su BL, Lu SJ. Accurate recognition of words in scenes without character segmentation using recurrent neural network. *Pattern Recognit* 2017;**63**:397–405.
33. Miao R, Xia LY, Chen HH, *et al*. Improved classification of blood-brain-barrier drugs using deep learning. *Sci Rep* 2019;**9**:8802.
34. Elangovan A, He J, Verspoor K. *Memorization vs. generalization: quantifying data leakage in NLP performance evaluation*. Association for Computational Linguistics, 2021;1325–1335. https://aclanthology.org/2021.eacl-main.113.
35. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. *J Chem Inf Model* 2018;**58**:916–32.
36. Xie LX, Xu L, Kong R, *et al*. Improvement of prediction performance with conjoint molecular fingerprint in deep learning. *Front Pharmacol* 2020;**11**:606668.
37. Caceres EL, Tudor M, Cheng AC. Deep learning approaches in predicting ADMET properties. *Future Med Chem* 2020;**12**:1995–9.
38. Ghose AK, Herbertz T, Hudkins RL, *et al*. Knowledge-based, central nervous system (CNS) lead selection and lead optimization for CNS drug discovery. *ACS Chem Nerosci* 2012;**3**:50–68.
39. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 2011;**32**:1466–74.
40. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
41. *Deepchem: democratizing deep-learning for drug discovery*. https://deepchem.io/ (7 April 2021, date last accessed).
42. Probst D, Reymond JL. A probabilistic molecular fingerprint for big data settings. *J Chem* 2018;**10**:66.
43. Awale M, Probst D, Reymond JL. WebMolCS: a web-based Interface for visualizing molecules in three-dimensional chemical spaces. *J Chem Inf Model* 2017;**57**:643–9.
44. Oh J, Laubach M, Luczak A. Estimating neuronal variable importance with random forest. In: *Conference Proceedings of the IEEE Engineering in Medicine and Biology Society*. Newark, NJ, USA: IEEE (Institute of Electrical and Electronics Engineers), 2003. pp. 33–4.
45. Kingma DP, Ba LJ. *Adam: a method for stochastic optimization*. San Diego: Ithaca, NY. 2015: arXiv:1412.6980.
46. Fu XC, Wang GP, Liang WQ, *et al*. Predicting blood-brain barrier penetration of drugs using an artificial neural network. *Pharmazie* 2004;**59**:126–30.
47. Sun S, Adejare A. Fluorinated molecules as drugs and imaging agents in the CNS. *Curr Top Med Chem* 2006;**6**:1457–64.
48. Hsu KW. A theoretical analysis of why hybrid ensembles work. *Comput Intell Neurosci* 2017;**2017**:1930702.
49. Khagi B, Kwon GR, Lama R. Comparative analysis of Alzheimer's disease classification by CDR level using CNN, feature selection, and machine-learning techniques. *Int J Imag Syst Tech* 2019;**29**:297–310.
50. Heravi MM, Zadsirjan V. Prescribed drugs containing nitrogen heterocycles: an overview. *RSC Adv* 2020;**10**:44247–311.
51. Gomtsyan A. Heterocycles in drugs and drug discovery. *Chem Heterocycl Compd* 2012;**48**:7–10.
52. Liljebris C, Larsen SD, Ogg D, *et al*. Investigation of potential bioisosteric replacements for the carboxyl groups of peptidomimetic inhibitors of protein tyrosine phosphatase

1B: identification of a tetrazole-containing inhibitor with cellular activity. *J Med Chem* 2002;**45**:1785–98.

53. Gupta M, Lee HJ, Barden CJ, *et al*. The blood-brain barrier (BBB) score. *J Med Chem* 2019;**62**:9824–36.
54. Ellis MJ, Lloyd D, Mcnab H, *et al*. Gas-phase pyrolysis of 2,3-Dihydro-1,4-Diazepines - involvement of the saturated portion of the ring in chemical-reactions and novel cis-trans isomerization of a fused-ring system. *J Chem Soc Chem Commun* 1995;2337–8. The Royal Society of Chemistry. http://dx.doi.org/10.1039/C39950002337.
55. Mikitsh JL, Chacko AM. Pathways for small molecule delivery to the central nervous system across the blood-brain barrier. *Perspect Medicin Chem* 2014;**6**:11–24.
56. Zhang YH, Xia ZN, Qin LT, *et al*. Prediction of blood-brain partitioning: a model based on molecular electronegativity distance vector descriptors. *J Mol Graph* 2010;**29**:214–20.
57. Aalizadeh R, von der Ohe PC, Thomaidis NS. Prediction of acute toxicity of emerging contaminants on the water flea Daphnia magna by ant Colony optimization - support vector machine QSTR models. *Environ Sci Process Impacts* 2017;**19**:438–48.
58. Umar AB, Uzairu A, Shallangwa GA, *et al*. QSAR modelling and molecular docking studies for anti-cancer compounds against melanoma cell line SK-MEL-2. *Heliyon* 2020;**6**:e03640.
59. Ding Q, Hou S, Zu S, *et al*. VISAR: an interactive tool for dissecting chemical features learned by deep neural network QSAR models. *Bioinformatics* 2020;**36**(11):3610–2.
60. Farhid H, Khodkari V, Nazeri MT, *et al*. Multicomponent reactions as a potent tool for the synthesis of benzodiazepines. *Org Biomol Chem* 2021;**19**(15):3318–58.
61. National Center for Biotechnology Information. *Pub Chem Compound Summary for CID 60172, Adefovir*, 2021. Retrieved 27 July 2021. https://pubchem.ncbi.nlm.nih.gov/compound/Adefovir.
62. Shinfuku M, Kishimoto T, Uchida H, *et al*. Effectiveness and safety of long-term benzodiazepine use in anxiety disorders: a systematic review and meta-analysis. *Int Clin Psychopharmacol* 2019;**34**:211–21.
63. Olsen RW, McCabe RT, Wamsley JK. GABAA receptor subtypes: autoradiographic comparison of GABA, benzodiazepine, and convulsant binding sites in the rat central nervous system. *J Chem Neuroanat* 1990;**3**(1):59–76.