Bioinformatics

http://dx.doi.org/10.7124/bc.000B18
UDC 577.004.9

T.G. Maiula, S.M. Yarmoliuk, I.I. Konvaliuk, V.A. Kunakh
Institute of Molecular Biology and Genetics, NAS of Ukraine
150, Akademika Zabolotnoho Str., Kyiv, Ukraine, 03143
t.h.maiula@imbg.org.ua

IN SILICO PREDICTION OF NEUROPROTECTIVE PROPERTIES OF NATURAL COMPOUNDS USING *SCUTELLARIA BAICALENSIS* AS AN EXAMPLE

Aim. *To develop, optimize, and evaluate effective in silico models for predicting the neuroprotective and anxiolytic properties of natural compounds using Scutellaria baicalensis as a case study.* ***Methods.*** *Construction and validation of machine learning models.* ***Results.*** *Three machine learning models, constructed using the Random Forest, XGBoost, and LightGBM algorithms, were developed for in silico prediction of the neuroprotective and anxiolytic activity of natural compounds. The classifiers achieved an accuracy of 75–78%. A binary classification approach was proposed, incorporating molecular descriptors and structural fingerprints, which, after preprocessing and optimization, enabled the identification of compounds with potential neuroprotective activity. The study confirms the effectiveness of these modeling approaches in predicting the neuroprotective, anxiolytic properties of S. baicalensis compounds. Application of the models to known phytochemicals from this plant verified previously reported bioactive substances: 46 out of 78 analyzed compounds were predicted to be potentially active.* ***Conclusions.*** *The in silico prediction of neuroprotective properties of bioactive compounds shows promise for screening and identifying phytocomplexes, particularly for applications in modern medicine such as the prevention and management of PTSD and other neurological disorders.*

Keywords: *in silico, machine learning, molecular descriptors, Scutellaria baicalensis, neuroprotective properties.*

Introduction

Modern pharmacology is rapidly evolving, employing advanced methods for the analysis and prediction of the biological activity of both synthetic drugs and natural compounds. The traditional experimental techniques, such as *in vitro* and *in vivo* testing, despite their high accuracy, are associated with substantial resource, time, and financial costs, as well as bioethical concerns related to the use of laboratory animals and challenges in reproducibility. These limitations have stimulated

Citation: Maiula T.G., Yarmoliuk S.M., Konvaliuk I.I., Kunakh V.A. (2025) *In silico* prediction of neuroprotective properties of natural compounds using *Scutellaria baicalensis* as an example. *Biopolymers & Cell*, 2(41), 139—149. http://dx.doi.org/10.7124/bc.000B18

the development of alternative research approaches, among which *in silico* methods have emerged as leading tools. These approaches allow us to effectively reduce experimental costs, accelerate the acquisition of results, and to eliminate the ethical issues associated with animal testing. *In silico* methods include virtual screening, molecular docking, quantitative structure-activity relationship (QSAR) analysis, and the prediction of ADMET properties [1]. These techniques are increasingly applied in the identification of bioactive compounds, including novel antibiotics in response to resistance, protein kinase inhibitors, and other therapeutics. In contemporary pharmacological research, *in silico* tools such as molecular docking and virtual screening are widely used to explore ligand-target interactions, thereby reducing the need for complex *in vitro* and *in vivo* experiments [2]. These technologies are also crucial in the study of natural products, especially when plant-based raw materials are scarce or difficult to access. Despite the existence of many herbal preparations with proven efficacy, their large-scale application is often limited by restricted availability of source material. In such cases, the application of *in silico* methods enables the rapid identification of promising bioactive molecules for further experimental validation [3].

One of the most promising medicinal plants with a broad therapeutic profile is *Scutellaria baicalensis*. This perennial herb from the Lamiaceae family has long been used in traditional medicine in East Asian countries and is known for its pronounced neuroprotective and anxiolytic effects, as well as antioxidant, anti-inflammatory, antiviral, antibacterial, and anticancer activities [4—6]. Among its key constituents, the flavonoids baicalin and baicalein have been shown to exert neuroprotective and anxiolytic effects through interaction with $GABA_A$ receptors (gamma-aminobutyric acid type A). Baicalin functions as a positive allosteric modulator at the benzodiazepine site $GABA_A$ receptors, selectively binding to subunits containing α2 and α3, thereby providing anxiolytic effects without significant sedation or muscle relaxation [7]. Baicalein, on the other hand, primarily targets non-benzodiazepine sites of the $GABA_A$ receptor and demonstrates similar effects, with additional neuroprotective properties and minimal impact on the serotonergic system [8]. Furthermore, baicalein has been shown to activate the TrkB/AKT signaling pathway, promoting synaptogenesis and neuroprotection [9]. Considering the current public health challenges in Ukraine — particularly the rise in cases of post-traumatic stress disorder (PTSD) and neurological disorders among military personnel and civilians affected by the ongoing war — the therapeutic potential of *S. baicalensis* bioactive compounds is of particular interest. The previous research has highlighted the plant's neuroprotective and anxiolytic mechanisms [6, 10—12]. However, due to the limited availability of raw plant material — *S. baicalensis* being naturally distributed primarily in East Asia (China, Korea, Japan) and partially in Russia [4] — the practical application and study of this plant remain uncommon in Ukraine. Therefore, to predict the biological activity, this study applied machine learning models: RandomForest, XGBoost, and LightGBM. These *in silico* models are highly effective for classifying complex pharmacological datasets, as they are capable of capturing nonlinear relationships between molecular descriptors and biological activity, while maintaining robustness against overfitting [13, 14].

RandomForest is an ensemble method based on constructing a large number of decision trees and averaging their outputs to achieve strong generalization. XGBoost is a gradient boosting algorithm that sequentially builds decision trees, where each subsequent tree corrects the errors of the previous one through gradient descent optimization [13]. LightGBM is a fast and efficient gradient boosting framework that uses a leaf-wise tree growth strategy, providing high computational speed and accuracy when working with large volumes of data. It is considered one of the most powerful machine learning tools for biomedical research [14].

The use of these models enables rapid and effective identification of promising compounds with

neuroprotective and anxiolytic properties, which was verified using a test set of known natural compounds derived from *S. baicalensis* [15]. The further application of *in silico* approaches will facilitate the discovery of additional active candidates, ensure more rational use of natural plant resources, and enhance the efficiency of future experimental research.

Therefore, the aim of this study was to develop and optimize *in silico* models for predicting the neuroprotective and anxiolytic properties of natural compounds, using *Scutellaria baicalensis* as a representative example.

Materials and Methods

To construct machine learning models for predicting neuroprotective and anxiolytic activity, the data were obtained from the ChEMBL database — an open-access repository of bioactive small molecules with known activity against various biological targets such as receptors and enzymes [1]. Access to the database was provided via the ChEMBL API. The responses from the API were returned in JSON (JavaScript Object Notation) format, a widely used data exchange standard suitable for parsing in Python. Computational operations were carried out using the cloud-based platform Google Colab, which supports integration with all the Python libraries used in this study and enables efficient calculations and predictions on server infrastructure, eliminating the need for local computational resources.

A list of mechanisms of action pharmacologically associated with neuroprotective and anxiolytic activity was compiled (including, for example, "Dopamine D2 receptor antagonist", "GABA receptor agonist", "Serotonin transporter inhibitor", "Serotonin 1a (5-HT1a) receptor agonist", "Muscarinic acetylcholine receptor M1 agonist", *etc.*). Using the ChEMBL API, a dataset of compounds matching these mechanisms was retrieved. A total of 104 such mechanisms were selected, and compounds acting as agonists, antagonists, or inhibitors of the respective targets were collected based on their ChEMBL IDs. For each selected molecule, IC50 values were extracted. After filtering and deduplication, the final dataset consisted of 687 compounds. An example of the results is presented in the Table 1.

To facilitate further analysis, IC50 values were converted to pIC50 using Equation 1.

$$pIC50 = -\log_{10}(IC50 \times 10^{-9}). \quad (1)$$

To facilitate classification, the pIC50 value was used to define bioactivity thresholds: molecules with pIC50 ≥ 6 were considered active (class 1), while all others were assigned to the inactive class (class 0). Due to class imbalance, oversampling was applied to the active class by duplicating those molecules in order to reduce model bias toward the majority class.

For each molecule, the corresponding SMILES notation was retrieved using its ChEMBL ID and used to calculate molecular descriptors [16]. The «RDKit» library, an open-source cheminformatics framework for Python, was employed to generate these descriptors based on SMILES structures. The following descriptors were computed: MW (molecular weight), TPSA (topological polar surface area), LogP (lipophilicity), and the number of hydrogen bond donors and acceptors (NumDonors / NumAcceptors). Additionally, molecular fingerprints were generated using both the Molecular ACCess System (MACCS keys, 167 bits) and Mor-

Table 1. **Molecules retrieved via ChEMBL API according to mechanisms of action**

molecule_chembl_id	mechanism_of_action
CHEMBL2359670	Dopamine D2 receptor antagonist
CHEMBL1201003	Serotonin 1b (5-HT1b) receptor agonist
CHEMBL3989558	Serotonin 1a (5-HT1a) receptor partial agonist
CHEMBL1214124	Glutamate receptor ionotropic AMPA antagonist
CHEMBL972	Monoamine oxidase B inhibitor

gan fingerprints (ECFP4, 2048 bits). MACCS is a predefined set of structural keys, where each bit indicates the presence or absence of specific chemical substructures. Morgan fingerprints are circular fingerprints that capture the atomic environment within a specified radius. In this case, binary vectors of length 2048 bits were generated using ECFP4 (Extended-Connectivity Fingerprints with radius = 2), a widely accepted standard in QSAR modeling and bioactivity prediction due to their sensitivity to molecular structure. The use of both molecular descriptors and structural fingerprints as input features is a common approach for predicting toxicity, bioactivity, and other physicochemical or pharmacological properties.

A total of 2220 descriptors were generated per molecule. Prior to model construction, data preprocessing included the removal of entries with invalid SMILES or errors during descriptor generation. Such preprocessing is recommended to improve model generalization and reduce the risk of overfitting [16].

Machine learning models based on the Random Forest algorithm were built with the following parameter settings: «n_estimators» = 100 (i.e., 100 decision trees), and «random_state» = 42. This configuration provides a robust baseline for the current dataset. While increasing the «n_estimators» parameter can enhance model stability, it does not always lead to significant performance improvements and may considerably increase computation time and memory usage. The XGBoost-based model was configured with «n_estimators» = 100, «learning_rate» = 0.1 (initial learning rate), and «max_depth» = = 5 (maximum depth of each tree). The LightGBM-based model was assigned the same hyperparameters as XGBoost. These parameter values are commonly recommended as initial settings in official documentation and are widely used in practice.

All models were implemented in Python using the «scikit-learn», «xgboost», and «lightgbm» libraries, and executed within the Google Colab environment.

For *in silico* modeling, the dataset was split into training (80%) and test (20%) subsets using the «stratify» = y parameter to preserve class balance — an important consideration in binary classification problems involving active (1) and inactive (0) classes. In addition, 5-fold cross-validation was performed to minimize the influence of random variations and improve the reliability of performance estimates.

The performance of the models was evaluated using the following metrics:

Accuracy — the proportion of correctly classified observations (Equation 2):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Precision — the proportion of correctly classified positive observations (Equation 3):

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall — the ability of the model to identify all positive observations (Equation 4):

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1-score — the harmonic mean of precision and recall (Equation 5):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

TP — true positives, TN — true negatives, FP — false positives, FN — false negatives predictions.

The average activity score (Average_Prediction) was calculated as the simple mean of predictions from all three models using Equation 6. The input values were either probabilities or binary outputs, representing the predicted potential activity of the compounds with respect to the studied biological property.

$$Average\ Prediction = \frac{Prf + Pxgb + Plgbm}{3} \quad (6)$$

Prf — prediction of the Random Forest model (0 or 1), Pxgb — prediction of the XGBoost model (0 or 1), Plgbm — prediction of the LightGBM model (0 or 1).

Results and Discussion

In the initial stage of the study, the three aforementioned models based on Random Forest, XGBoost, and LightGBM were evaluated. All models demonstrated comparable accuracy values. The classification results obtained on the test dataset are presented in the Table 2.

After performing cross-validation, the accuracy of the XGBoost- and LightGBM-based models increased to 77.03 and 77.21%, respectively.

All models demonstrated better precision for class 0 (inactive) and higher recall for class 1 (active). The differences between the models were statistically insignificant, allowing the selection of a primary model to be based on feature importance analysis. The minor variation in accuracy indicates consistency and stability across the models, which supports the reliability of the applied approach.

Feature importance analysis was conducted by identifying the top descriptors for each model individually. This type of analysis is commonly used in research to compare the relevance of features derived from Morgan fingerprints [16] or to select the most influential descriptors from the full feature set.

This analysis led to several important observations. The Random Forest-based model assigned the greatest weight to the LogP descriptor, reflecting compound hydrophobicity, along with molecular weight (MolWt) and a number of MACCS fingerprints. In contrast, XGBoost identified Morgan fingerprints (particularly Morgan_1536 and Morgan_428), which describe molecular fragments based on local structures, as the most informative features. LightGBM demonstrated a hybrid approach: its top features included both general structural descriptors (MolWt, LogP, TPSA) and structural fingerprints (MACCS and Morgan). These results confirm that each model interprets the feature importance differently, offering opportunities for their complementary use. A combined approach employing both LightGBM and XGBoost algorithms appears logical, since despite their nearly identical classification accuracy, the models rely on different groups of features. LightGBM prioritized global physicochemical properties — MolWt, LogP, TPSA, NumHAcceptors — while XGBoost based its predictions primarily on structural fingerprints such as Morgan_1536, Morgan_428, and MACCS_101, which capture local molecular fragments. This complementarity enhances the potential of integrated (ensemble) strategies, enabling broader coverage of structural-functional compound characteristics.

On the other hand, as illustrated in Fig. 1, the top descriptors identified by the Random Forest

Table 2. **Classification results for the test dataset of the models**

Metric	Random Forest		XGBoost		LightGBM	
Class	0	1	0	1	0	1
Precision	0,87	0,66	0,87	0,65	0,87	0,65
Recall	0,68	0,86	0,66	0,86	0,66	0,86
f1-score	0,76	0,75	0,75	0,74	0,75	0,74
Accuracy	75,36%		74,64%		74,64%	

Table 3. **List of descriptors with significant influence on the predicted variable**

No.	Random Forest	XGBoost	LightGBM
1	LogP	Morgan_1536	MolWt
2	MACCS_125	MACCS_101	LogP
3	MACCS_144	Morgan_428	TPSA
4	MolWt	MACCS_126	MACCS_99
5	MACCS_103	MACCS_116	MACCS_109
6	MACCS_82	MACCS_105	NumHAcceptors
7	MACCS_134	MACCS_54	MACCS_127
8	MACCS_145	Morgan_625	Morgan_322
9	MACCS_81	Morgan_52	MACCS_166
10	TPSA	Morgan_1	Morgan_799

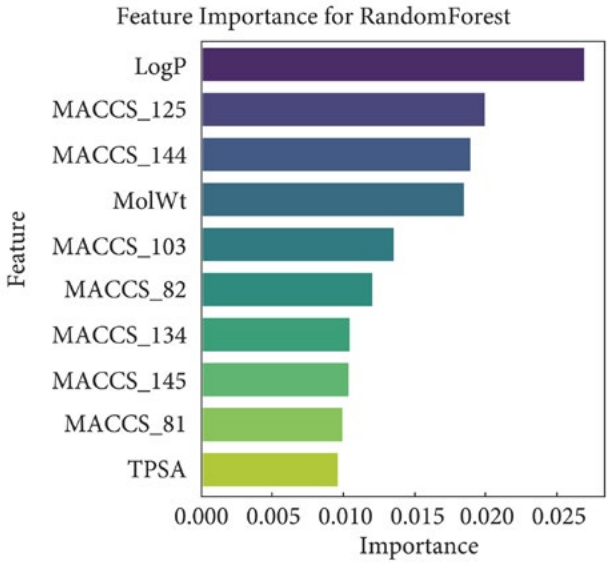

Fig. 1. Histogram of the most influential descriptors for Random Forest

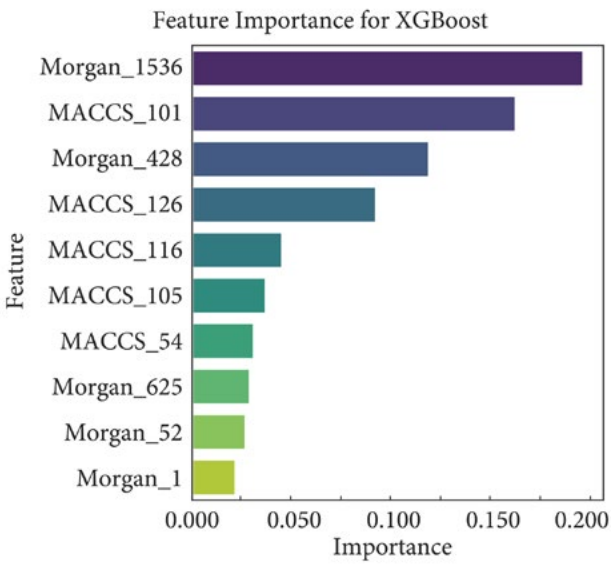

Fig. 2. Histogram of the most influential descriptors for XGBoost

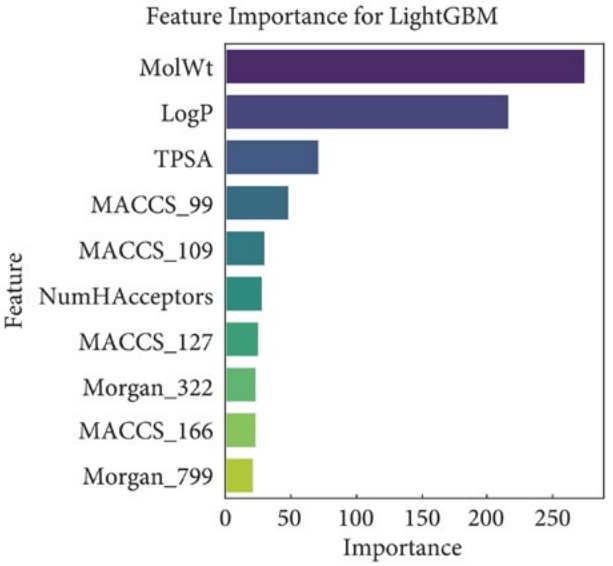

Fig. 3. Histogram of the most influential descriptors for LightGBM

model demonstrated more pronounced individual importance. For this reason, all three trained models were preserved. The «joblib» library was used to serialize the trained Python model objects into «.pkl» files, allowing future reuse without retraining [17].

The next stage of the study aimed to verify the predictive capacity of the constructed models in identifying neuroprotective and anxiolytic properties of natural compounds. For this purpose, a test set was compiled based on the phytochemical composition of *S. baicalensis*, according to previously published data [4]. The test set included natural compounds characteristic of *S. baicalensis.*

Based on available chemical names and structural formulas [4], 78 out of 126 compounds were described using SMILES notation. SMILES were constructed using BIOVIA Draw 2018 and DataWarrior (v06.04.02), which allow for the creation, visualization, and export of 2D molecules, as well as calculation of basic molecular properties. The remaining 48 compounds could not be transformed into SMILES format due to overly generic names or ambiguous structural representations in the source material.

For each of the 78 compounds, molecular descriptors were calculated using the same pipeline as in the training set — excluding IC50, pIC50, and activity class labels. Calculations were performed using the «RDKit» library. Prior to prediction, the structure of the new data (78 compounds)

Table 4. **Compounds of *Scutellaria baicalensis* predicted to be active by at least one of the trained models (Random Forest, XGBoost, LightGBM)**

No.	Name	Molecular formula	MW	Plant part	Prf	Pxgb	Plgbm	Average_ Prediction
1	4'-Hydroxyacetophenone	$C_8H_8O_2$	136	Root	0	1	1	0,6666667
2	4'-Hydroxywogonin (5,7,4'-Trihydroxy-8-methoxyflavone)	$C_{16}H_{12}O_6$	300	Root	0	1	1	0,6666667
3	Scutevulin (5,7,2'-Trihydroxy-8-methoxyflavone)	$C_{16}H_{12}O_6$	300	Root	0	1	1	0,6666667
4	(2S)-5,7,4'-Trihydroxy-6-methoxyflavanone	$C_{16}H_{14}O_6$	302	Root	0	1	1	0,6666667
5	Dihydrooroxylin A ((2S)-5,7-Dihydroxy-6-methoxyflavanone)	$C_{16}H_{14}O_5$	286	Root	0	1	1	0,6666667
6	5,7,4'-Trihydroxy-6-methoxyflavone	$C_{16}H_{12}O_6$	300	Aerial part	0	1	1	0,6666667
7	Tenaxin II (5,7,2'-Trihydroxy-6-methoxyflavone)	$C_{16}H_{12}O_6$	300	Root	0	1	1	0,6666667
8	Syringaldehyde (4-Hydroxy-3,5-dimethoxybenzaldehyde)	$C_9H_{10}O_4$	182	Root	0	1	1	0,6666667
9	Vanillin (4-Hydroxy-3-methoxybenzaldehyde)	$C_8H_8O_3$	152	Root	0	1	1	0,6666667
10	Acetosyringone (4'-Hydroxy-3',5'-dimethoxyacetophenone)	$C_{10}H_{12}O_4$	196	Root	1	0	1	0,6666667
11	7-Methoxychrysin (5-Hydroxy-7-methoxyflavone)	$C_{16}H_{12}O_4$	268	Aerial part	0	1	1	0,6666667
12	5,8,2'-Trihydroxy-7-methoxyflavone	$C_{16}H_{12}O_6$	300	Root	0	1	1	0,6666667
13	7-O-Methylwogonin (5-Hydroxy-7,8-dimethoxyflavone)	$C_{17}H_{14}O_5$	298	Root	0	1	1	0,6666667
14	(2S)-5,4'-Dihydroxy-7-methoxyflavanone	$C_{16}H_{14}O_5$	286	Aerial part	0	1	1	0,6666667
15	(2S)-7,2',6'-Trihydroxy-5-methoxyflavanone	$C_{16}H_{14}O_6$	302	Root	0	1	1	0,6666667
16	(2S)-7-Hydroxy-5-methoxyflavanone	$C_{16}H_{14}O_4$	270	Root	0	1	1	0,6666667
17	5,7,6'-Trihydroxy-2'-methoxyflavone	$C_{16}H_{12}O_6$	300	Root	0	1	1	0,6666667
18	(2R,3R)-3,5,7,2',6'-Pentahydroxyflavanone	$C_{15}H_{12}O_7$	304	Root	0	1	1	0,6666667
19	Protocatechuic acid (3,4-Dihydroxybenzoic acid)	$C_7H_6O_4$	154	Root	0	1	1	0,6666667
20	p-Hydroxybenzoic acid	$C_7H_6O_3$	138	Root	0	1	1	0,6666667
21	Viscidulin I (5,7,2',6'-Tetrahydroxyflavonol)	$C_{15}H_{10}O_7$	302	Root	0	1	1	0,6666667
22	(2S)-5,7,2',6'-Tetrahydroxyflavanone	$C_{15}H_{12}O_6$	288	Root	0	1	1	0,6666667
23	(+)-Eriodictyol ((2S)-5,7,3',4'-Tetrahydroxyflavanone)	$C_{15}H_{12}O_6$	288	Root	0	1	1	0,6666667

End of the Table 4

No.	Name	Molecular formula	MW	Plant part	Prf	Pxgb	Plgbm	Average_ Prediction
24	p-Hydroxybenzaldehyde	$C_7H_6O_2$	122	Root	0	1	1	0,6666667
25	Protocatechuic aldehyde (3,4-Dihydroxybenzaldehyde)	$C_7H_6O_3$	138	Root	0	1	1	0,6666667
26	Isocarthamidin ((2S)-5,7,8,4'-Tetrahydroxyflavanone)	$C_{15}H_{12}O_6$	288	Leaf; Root	0	1	1	0,6666667
27	Carthamidin ((2S)-5,6,7,4'-Tetrahydroxyflavanone)	$C_{15}H_{12}O_6$	288	Leaf; Root	0	1	1	0,6666667
28	5,7-Dihydroxy-6,8-dimethoxyflavone	$C_{17}H_{14}O_6$	314	Root	0	1	0	0,3333333
29	5,7,2'-Trihydroxy-6,8-dimethoxyflavone	$C_{17}H_{14}O_7$	330	Root	0	1	0	0,3333333
30	Tenaxin I (5,2'-Dihydroxy-6,7,8-trimethoxyflavone)	$C_{18}H_{16}O_7$	344	Root	0	1	0	0,3333333
31	5,8-Dihydroxy-6,7-dimethoxyflavone	$C_{17}H_{14}O_6$	314	Root	0	1	0	0,3333333
32	5,8,2'-Trihydroxy-6,7-dimethoxyflavone	$C_{17}H_{14}O_7$	330	Root	0	1	0	0,3333333
33	Viscidulin III (5,7,3',6'-Tetrahydroxy-8,2'-dimethoxyflavone)	$C_{17}H_{14}O_8$	346	Root	0	1	0	0,3333333
34	Wogonin (5,7-Dihydroxy-8-methoxyflavone)	$C_{16}H_{12}O_5$	284	Root; Aerial part; Hairy root	0	1	0	0,3333333
35	Oroxylin A (5,7-Dihydroxy-6-methoxyflavone)	$C_{16}H_{12}O_5$	284	Root	0	1	0	0,3333333
36	Syringic acid (4-Hydroxy-3,5-dimethoxybenzoic acid)	$C_9H_{10}O_5$	198	Root	0	0	1	0,3333333
37	Vanillic acid (4-Hydroxy-3-methoxybenzoic acid)	$C_8H_8O_4$	168	Root	0	0	1	0,3333333
38	Genkwanin (5,4'-Dihydroxy-7-methoxyflavone)	$C_{16}H_{12}O_5$	284	Aerial part	0	1	0	0,3333333
39	5,8-Dihydroxy-7-methoxyflavone	$C_{16}H_{12}O_5$	284	Root	0	1	0	0,3333333
40	Viscidulin II (5,2',6'-Trihydroxy-7,8-dimethoxyflavone)	$C_{17}H_{14}O_7$	330	Root	0	1	0	0,3333333
41	Rivularin (5,6'-Dihydroxy-7,8,2'-trimethoxyflavone)	$C_{18}H_{16}O_7$	344	Root; Hairy root	0	1	0	0,3333333
42	Skullcapflavone I (5,2'-Dihydroxy-7,8-dimethoxyflavone)	$C_{17}H_{14}O_6$	314	Root; Hairy root	0	1	0	0,3333333
43	5,6,7-Trihydroxy-4'-methoxyflavone	$C_{16}H_{12}O_6$	300	Root	0	0	1	0,3333333
44	5,7,6'-Trihydroxy-2'-methoxyflavonol	$C_{16}H_{12}O_7$	316	Root	0	1	0	0,3333333
45	5,7,6'-Trihydroxy-8,2'-dimethoxyflavone	$C_{17}H_{14}O_7$	330	Root	0	1	0	0,3333333
46	Baicalein 7-O-β-D-glucoside	$C_{21}H_{20}O_{10}$	432	Root; Aerial part	0	0	1	0,3333333

was aligned to the format of the training data to ensure compatibility with the models. Specifically:

• columns were ordered identically to the training dataset;

• SMILES structures were validated for correctness using «RDKit».

The saved models files — «random_forest_model.pkl», «xgb_model.pkl», and «lgb_model.pkl» — were loaded into the Google Colab environment using the «joblib» library [17]. Each of the 78 compounds was individually evaluated by all three models. The Average_Prediction score was calculated as the mean of model predictions (Equation 6).

In Table 4, the compounds are listed in a simple numerical sequence, while their names, descriptions, and the corresponding plant parts are retained as reported in the source [4], ensuring traceability and consistency with the phytochemical data of *S. baicalensis*.

Based on the data presented in Table 4, 27 compounds received a positive activity prediction from two models simultaneously (predominantly XGBoost and LightGBM), indicating high prediction consistency and increased confidence in their biological relevance. An additional 19 compounds were predicted as active by only one model, which may suggest marginal activity or descriptor values near classification thresholds. Only one compound was predicted as active by the Random Forest-based model, which may be attributed to the model's lower sensitivity to weak activity signals or its reliance on a different subset of descriptors. This distribution of results confirms the sensitivity and realism of the models and highlights that the combined use of two independent, yet high-performing algorithms — XGBoost and LightGBM — which rely on different groups of descriptors, provides a more robust identification of potentially active compounds. Notably, 42 out of the 46 predicted active compounds were found specifically in the root of *S. baicalensis* — the plant part traditionally considered the most pharmacologically valuable due to its high concentration of bioactive flavonoids. Although from a chemical perspective some of the compounds that received a positive Average_Prediction are not typical flavonoids or their glycosides, their inclusion among the potentially active candidates can be explained by several factors. Many simple phenolic compounds, such as vanillin or p-hydroxybenzoic acid, are intermediates in flavonoid biosynthesis, exhibit their own pharmacological activity (e.g., antioxidant, anti-inflammatory), and share structural features with bioactive fragments present in the training datasets. The machine learning models likely identified similarities in their molecular descriptors to those of active compounds in the training set, which led to a positive prediction.

These findings demonstrate the potential of the proposed models as effective *in silico* tools for preliminary screening of compounds with targeted pharmacological activity in the chemical profiles of underexplored yet promising medicinal plants. This approach offers the advantage of significantly reducing the scope of experimental validation by prioritizing compounds with the highest predicted likelihood of biological activity.

Conclusions

The application of the developed machine learning models to the analysis of 78 components of *S. baicalensis* enabled the identification of several compounds with a high probability of exhibiting significant biological activity. Given the multicomponent nature of this plant's phytocomplexes, the potential synergistic effects of individual constituents cannot be excluded, as they may enhance the biological impact even of compounds with moderate standalone activity. This property makes *S. baicalensis* particularly promising for further investigation in pharmacology and biomedicine, especially in the development of multicomponent phytopharmaceuticals. In this context, the use of network pharmacology approaches is advisable, as they allow the study of interactions between multiple plant-derived components, the prediction of synergistic effects, and the construction of personalized combinations of active substances targeting specific neuropharmacological pathways [3].

Considering the *in silico* prediction results obtained, the proposed models may be useful for identifying promising bioactive compounds with potential anxiolytic and neuroprotective activity in less-studied medicinal plant species. Such an approach is especially relevant in the current context of Ukraine, where the consequences of war have led to a significant increase in the prevalence of post-traumatic stress disorder (PTSD), sleep disturbances, and other neurological conditions. According to the data from Ukrainian psychiatric hospitals, the proportion of hospitalizations due to war-related psychological trauma rose from 12.2% in January 2022 to 17.3% in April 2024, reflecting the growing need to support the patients suffering from trauma-induced mental disorders, including PTSD [18].

In this setting, the use of safe and plant-based agents — particularly phytocompounds with confirmed *in silico* activity — may serve as an effective alternative to traditional pharmacological therapies, which are often associated with undesirable side effects. Furthermore, the application of *in silico* methods, including the construction and use of machine learning models, represents a cost-effective solution for rapidly and efficiently identifying potentially active compounds, with minimal resource expenditure, for subsequent experimental validation.

REFERENCES

1. *Zagrychuk OG, Matiashchuk YO, Korzhovska VV, et al., and Zagrychuk HY.* Use of in silico research methods for predicting pharmacokinetic properties and searching for biologically active compounds. *Pharm Rev.* 2024; **3**:53—67.
2. *Volynets GP, Iungin OS, Gudzera OI, et al., and Tukalo MA.* Identification of novel antistaphylococcal hit compounds. *J Antibiot (Tokyo).* 2024; **77**(10):665—78.
3. *Li L, Yang L, Yang L, et al., and Li P.* Network pharmacology: a bright guiding light on the way to explore the personalized precise medication of traditional Chinese medicine. *Chin Med.* 2023; **18**(1):146.
4. *Wang ZL, Wang S, Kuang Y, et al., and Ye M.* A comprehensive review on phytochemistry, pharmacology, and flavonoid biosynthesis of *Scutellaria baicalensis*. *Pharm Biol.* 2018; **56**(1):465—84.
5. *Tao F, Cai Y, Deng C, et al., and Sun H.* A narrative review on traditional Chinese medicine prescriptions and bioactive components in epilepsy treatment. *Ann Transl Med.* 2023; **11**(2):129.
6. *Zhang K, Pan X, Wang F, et al., and Wu C.* Baicalin promotes hippocampal neurogenesis via SGK1- and FKBP5-mediated glucocorticoid receptor phosphorylation in a neuroendocrine mouse model of anxiety/depression. *Sci Rep.* 2016; **6**:30951.
7. *Wang F, Xu Z, Ren L, et al., and Xue H.* GABA A receptor subtype selectivity underlying selective anxiolytic effect of baicalin. *Neuropharmacology.* 2008; **55**(7):1231—7.
8. *de Carvalho RS, Duarte FS, de Lima TC.* Involvement of GABAergic non-benzodiazepine sites in the anxiolytic-like and sedative effects of the flavonoid baicalein in mice. *Behav Brain Res.* 2011; **221**(1):75—82.
9. *Ding S, Zhuge W, Hu J, et al., and Zhuge Q.* Baicalin reverses the impairment of synaptogenesis induced by dopamine burden via the stimulation of GABAAR-TrkB interaction in minimal hepatic encephalopathy. *Psychopharmacology (Berl).* 2018; **235**(4):1163—78.
10. *Ma Y, Zhou X, Zhang F, et al., and Tao X.* The effect of *Scutellaria baicalensis* and its active ingredients on major depressive disorder: a systematic review and meta-analysis of literature in pre-clinical research. *Front Pharmacol.* 2024; **15**:1313871.
11. *EghbaliFeriz S, Taleghani A, Tayarani-Najaran Z.* Central nervous system diseases and Scutellaria: a review of current mechanism studies. *Biomed Pharmacother.* 2018; **102**:185—95.
12. *Limanaqi F, Biagioni F, Busceti CL, et al., and Fornai F.* Potential Antidepressant Effects of *Scutellaria baicalensis*, Hericium erinaceus and Rhodiola rosea. *Antioxidants (Basel).* 2020; **9**(3):234.
13. *Taha K.* Machine learning in biomedical and health big data: a comprehensive survey with empirical and experimental insights. *J Big Data.* 2025; **12**(1):61.
14. *Kanber BM, Al Smadi A, Noaman NF, et al., and Alsmadi MK.* LightGBM: a leading force in breast cancer diagnosis through machine learning and image processing. *IEEE Access.* 2024; **12**:39811—32.

15. Zhang L, Tian Y, Wang J, *et al., and* Fan H. Network pharmacology-based research on the effect of *Scutellaria baicalensis* on osteosarcoma and the underlying mechanism. *Medicine (Baltimore).* 2023; **102**(46):e35957.
16. *Nguyen-Vo TH, Nguyen L, Do N, et al., and Le L.* Predicting Drug-Induced Liver Injury Using Convolutional Neural Network and Molecular Fingerprint-Embedded Features. *ACS Omega.* 2020; **5**(39):25432—9.
17. *Polishchuk MM, Tsyben DV, Kapliuk YI.* Information processing using machine learning tools in Python. *Computer-Integrated Technologies: Education, Science, Production.* 2023; **53**:205—9.
18. *Pinchuk I, Yachnik Y, Goto R, Skokauskas N.* Mental health services during the war in Ukraine: 2-years follow up study. *Int J Ment Health Syst.* 2025; **19**(1):11.

Received 21.04.2025

Т.Г. Маюла, С.М. Ярмолюк, І.І. Конвалюк, В.А. Кунах
Інститут молекулярної біології і генетики НАН України
вул. Академіка Заболотного, 150, Київ, Україна, 03143
t.h.maiula@imbg.org.ua

IN SILICO ПРОГНОЗУВАННЯ НЕЙРОПРОТЕКТОРНИХ ВЛАСТИВОСТЕЙ СПОЛУК ПРИРОДНОГО ПОХОДЖЕННЯ НА ПРИКЛАДІ ШОЛОМНИЦІ БАЙКАЛЬСЬКОЇ (*SCUTELLARIA BAICALENSIS*)

Мета. Розробка, оптимізація та апробація ефективних in silico моделей прогнозування нейропротекторних та анксіолітичних властивостей сполук природного походження на прикладі шоломниці байкальської (*Scutellaria baicalensis*). ***Методи.*** Побудова моделей машинного навчання. ***Результати.*** Розроблено три моделі машинного навчання, побудовані із застосуванням алгоритмів Random Forest, XGBoost та LightGBM, для прогнозування нейропротекторної та анксіолітичної активностей природних сполук. Побудовані класифікатори досягли точності на рівні 75—78%. Запропоновано підхід бінарної класифікації із залученням молекулярних дескрипторів і структурних фінгерпринтів, який після обробки та оптимізації дозволяє виявляти сполуки з потенційною нейропротекторною активністю. Обґрунтовано ефективність застосування методів in silico моделювання для прогнозування нейропротекторних та анксіолітичних властивостей сполук *S. baicalensis*. Застосування моделей до компонентів цього виду засвідчило їхню здатність верифікувати вже відомі біологічно активні речовини**:** з 78 досліджених сполук 46 були ідентифіковані як потенційно активні. ***Висновки.*** Застосування *in silico* прогнозування нейропротекторних властивостей біоактивних сполук є перспективним для скринінгу фітокомплексів, зокрема у фармації та медицині — для профілактики та підтримки при ПТСР і нервових розладах.

Ключові слова. *in silico*, машинне навчання, молекулярні дескриптори, *Scutellaria baicalensis*, нейропротекторні властивості.