Computerized Detection of JWH Synthetic Cannabinoids Class Membership Based on Machine Learning Algorithms and Molecular Descriptors

Catalina Mercedes Burlacu
Department of Chemistry Physics and Environment
“Dunarea de Jos” University of Galati
Galati, Romania
catalina.burlacu@ugal.ro

Mirela Praisler
Department of Chemistry Physics and Environment
“Dunarea de Jos” University of Galati
Galati, Romania
Mirela.Praisler@ugal.ro

Adrian Constantin Burlacu
Department of Chemistry Physics and Environment
“Dunarea de Jos” University of Galati
Galati, Romania
ab273@student.ugal.ro

Abstract—An Artificial Neural Networks (ANN) model identifying JWH Synthetic Cannabinoids, that we have developed based on a combination of topological, 3D-MoRSE (Molecule Representation of Structure based on Electron diffraction) and ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) molecular descriptors, is described and analyzed. The validation results indicate that this computerized system has a very high potential for efficiently predicting the class membership of JWH and discriminating them from a large variety of (non-JWH) substances of forensic interest.

Keywords—JWH synthetic cannabinoids, artificial neural networks, molecular descriptors

I. INTRODUCTION

The emergence of synthetic cannabinoids such as the JWH series is an ongoing challenge for forensic, clinical analytical chemists and toxicologists. Because the JWH series of compounds produces euphoric and hallucinogenic effects, they cause serious social problems concerning their abuse liability. New analogs continuously emerge on the black market, with the aim to circumvent the legal consequences. JWH synthetic cannabinoids are known to be lipid soluble and nonpolar. They contain 22 to 26 carbon atoms, all sharing a common structural feature, i.e. a side chain composed of 4 to 9 saturated carbon atoms, which ensures an optimal activity in the cannabinoid receptor [1].

Detection plays a critical role in preventing substance abuse. In this paper, we are assessing the efficiency of a machine learning algorithm, i.e. an Artificial Neural Networks (ANN) model, designed to recognize JWH synthetic cannabinoids based on a combination of topological and 3D-MoRSE (Molecule Representation of Structure based on Electron diffraction) and ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) molecular descriptors. These descriptors allow the representation of the chemical characteristics of the targeted molecules in a numerical vector that is adequate for computer processing.

The proposed method consists of translating the model in a data-oriented manner on a large set of molecular structures. The obtained model can be provide meaningful molecular descriptors for performing structure searching without needing any retraining. In order to find which molecular descriptors are the most relevant for our purpose, we performed a variety of tests related to predictive QSAR (Quantitative Structure - Activity Relationships), toxicity and virtual screening tasks.

II. DATABASE, METHODS AND RESULTS

A. Database

The input database consists of 150 designer drugs including JWH synthetic cannabinoids, as well as non-JWH substances. The later class includes synthetic cannabinoids other than JWH and a variety of other illicit substances (e.g. doping drugs). The database was built and updated by performing automatic data extraction. For this process, we have used the KnowItAll Informatics System 2021 software package and the chemical abstracts service (CAS) number identifier from the Wiley Database Collection. The substances included in the database were divided into three classes, referred to as JWH, non-JWH Cannabinoids and Others. The group of positives contains 50 JWH synthetic cannabinoids, while the group of negatives includes 100 compounds (50 non-JWH cannabinoids and 50 others drugs). The list of these compounds was presented in a previous paper [2].

B. Methods

The 3D representation of the molecular structures was obtained, for all the 150 compounds forming the input database, with the HyperChem release 8.0.10 and Gaussian 16 software packages. The geometries of the molecular structures included in the database were fully optimized by using algorithms specific to the AM1 semi-empirical quantum method. The improved molecular geometry and associated parameters corresponding to the minimum energy of the molecular system were computed based on the Polak-Ribiere theory.

A number of 300 QSAR and 50 ADMET molecular descriptors have been selected from three blocks, i.e. topological, 3D-MoRSE, and toxicity. The molecular descriptors were computed for each molecular structure in the database by using the online modules and web applications of Swiss Institute of Bioinformatics scientific platforms. The initial 350 input molecular descriptors were evaluated from the point of view of their relative importance to the classification system. Taking into account its variation, we have selected as final input the first 150 most relevant descriptors. Their list was reported in a previous paper [3]. The detailed definitions, mathematical formulas and chemical significance of these topological and 3D-MoRSE descriptors are presented in detail in Todeschini et al [4].
C. Results

The ANN model designed to recognize JWH synthetic cannabinoids based on the 150 most relevant molecular descriptors was named nd150_DETECT_DNN and was built by using the Neural Designer 5.9.2 software. Fig.1 shows the architecture of this system.

ANN represents a predictive artificial intelligence tool. The Neural Designer software may build networks with deep architectures, which represent an efficient type of universal approximators [5]. In our case, the system contains three input nodes, corresponding to Topological, 3DMORSE and ADMET descriptors respectively. The size of the scaling layer is equal to the number of input nodes, i.e. 3.

The optimal number of neurons for the nd150_DETECT_DNN system was determined with the growing neurons algorithm. The following chart (Fig.2) presents the dynamics of the training error and the selection error, as determined for the different subsets while performing the above-mentioned algorithm.

We chose the growing neurons algorithm for selecting the optimal number of neurons in this application (see Table 2). This procedure starts with a minimum number of neurons and then incrementally increases their number in an iterative way. The growing inputs algorithm was used for selecting the optimal set of inputs (see Table 3). Hence, the inputs were added progressively according to their correlations with the targets.

Table 1 shows the results obtained for the selection of the neurons, as performed by the growing neurons algorithm. The model selection algorithms aim to determine which ANN has the topology that corresponds to the optimized error on new data. The most adequate model selection methods are the order selection algorithms and the input selection algorithms. The former determine the optimal number of hidden neurons in the architecture of the ANN. The later ones are designed to identify the optimal subset of input variables.

| Description                                      | Value     |
|--------------------------------------------------|-----------|
| Optimal order                                    | 3         |
| Optimum training order                          | 0.42421   |
| Optimum selection order                         | 0.48846   |
| Epochs number                                    | 10        |
| Stopping criterion                              | Maximum neurons |
| Elapsed time                                     | 00:00:40  |

Table II. The neurons selection algorithm

| Description                                      | Value     |
|--------------------------------------------------|-----------|
| Minimum neurons                                  | 1         |
| Maximum neurons                                  | 10        |
| Step                                             | 1         |
| Trials number                                    | 99        |
| Selection loss goal                              | 0         |
| Maximum selection failures                       | 100       |
| Maximum iterations number                        | 1000      |
| Maximum time                                     | 3600      |

Table III. The input selection algorithm

| Description                                      | Value     |
|--------------------------------------------------|-----------|
| Trials number                                    | 100       |
| Selection error goal                             | 0         |
| Maximum selection features                       | 100       |
| Maximum inputs number                            | 3         |
| Minimum correlation                              | 0         |
| Maximum correlations                             | 1         |
| Maximum iterations number                        | 1000      |
| Maximum time                                     | 3600      |

The following chart (Fig.3) presents the nature of the samples in the input data subsets. The 150 samples were divided into 90 training samples (60%), 30 selection samples (20%), and 30 testing samples (20%).
The loss index defines the task the ANN must perform and represents a measure of the quality of the possible. The loss index is often tested by performing a linear regression analysis regarding the scaled ANN outputs and the associated targets, as computed for an independent testing subset [9]. This method yields a number of parameters for each output variable. The first two parameters, and , are the y-intercept and the slope of the best linear regression correlating the outputs to the targets. The third parameter, , is the correlation coefficient which the selection error increases.

Table VII lists the linear regression parameters obtained for the output associated with JWH cannabinoids. The values of the intercept are very small (nearly equal to 0), while the slope and correlation are very close to 1, so we may conclude the regularization term measures the parametric values in the ANN system. If it is added to the error, then the ANN will be characterized by smaller weights and biases. In this case, its response will be smoother. In addition, overfitting may be easier avoided. In our case, we applied the L2 regularization method, which is based on the squared sum of all the ANN parameters.

Choosing the appropriate error method depends on the particular application. In our case, we selected the Normalized Squared Error (MSE). The normalized squared error is equal to 1 when the outputs from the ANN are equal to the mean values of the target variables. A normalized squared error equal to 0 indicates a perfect prediction of the data [7].

The quasi-Newton method, presented in Table VII, was used as optimization algorithm. Although it is based on Newton's method, this approach has the advantage of not requiring the calculation of second derivatives. Instead, it determines an approximation of the inverse Hessian at each iteration of the algorithm, based only on gradient information.

The loss of a model is often tested by performing a linear regression analysis regarding the scaled ANN outputs and the associated targets, as computed for an independent testing subset [9]. This method yields a number of parameters for each output variable. The first two parameters, and , are the y-intercept and the slope of the best linear regression correlating the outputs to the targets. The third parameter, , is the correlation coefficient which the selection error increases.

Table VIII lists the linear regression parameters obtained for the output associated with JWH cannabinoids. The values of the intercept are very small (nearly equal to 0), while the slope and correlation are very close to 1, so we may conclude

![Distribution of samples in the training, selection and testing sets.](Image)

Table IV depicts the values used for scaling the inputs. They include the minimum, maximum, mean, and standard deviation (the scalers are the Mean Standard Deviation).

| Name           | Minimum | Maximum | Mean   | Deviation |
|----------------|---------|---------|--------|-----------|
| Topological    | 4.3     | 7.9     | 5.843333| 0.828066  |
| 3DMORSE        | 2       | 4.4     | 3.054000| 0.433594  |
| ADMET          | 1       | 6.9     | 3.75867 | 1.764420  |

The nd150_DETECT_DNN system contains three perceptron layers. The parameters used for scaling the inputs are displayed in Table V. They include the minimum, maximum, mean, and standard deviation (the scalers are the Minimum and the Maximum).

| Name            | Scaling Input Class |
|-----------------|---------------------|
| JWH cannabinoids| Minimum 0 Maximum 1 | 0.333333 0.472984 |
| non-JWH cannabinoids | Minimum 0 Maximum 1 | 0.313333 0.465403 |
| Other           | Minimum 0 Maximum 1 | 0.353333 0.479606 |

Table VI indicates the size of each layer and the corresponding activation function. The size of the unscaling layer is three (equal to the number of outputs).

| Name         | Corresponding activation function |
|--------------|-----------------------------------|
| perceptron_layer_1 | Inputs number 3 Neuron number 1 Activation function Hyperbolic Tangent (tanh) |
| perceptron_layer_2 | Inputs number 4 Neuron number 1 Activation function Linear |
that the ANN is assigning the JWH class membership remarkably well.

### TABLE VIII. Linear Regression Parameters for the JWH Cannabinoids Output

| Value   |
|---------|
| Intercept | 0.0099 |
| Slope    | 0.9249 |
| Correlation | 0.9306 |

Fig. 4 illustrates the linear regression for the output corresponding to the class of JWH cannabinoids. Each circle represents a predicted value versus the actual one. The line indicates the best linear fit.

### TABLE IX. Linear Regression Parameters for the Non-JWH Cannabinoids Output

| Value   |
|---------|
| Intercept | 0.1150 |
| Slope    | 0.9328 |
| Correlation | 0.4234 |

Fig. 5 illustrates the linear regression for the non-JWH cannabinoids output.

### TABLE X. Linear Regression Parameters for the Output Corresponding to Drugs Other Than Cannabinoids

| Value   |
|---------|
| Intercept | -0.0911 |
| Slope    | 1.019  |
| Correlation | 0.7364 |

The confusion matrix computed for this system, presented in Table XII, shows that it is remarkably sensitive, as it detects all the analyzed JWH synthetic cannabinoids without exception. The ANN model is characterized by a rate of true positives (JWH and non-JWH cannabinoids) of 100% and a rate of false negatives (cannabinoids misclassified as Others) of 0%. As many of 148 samples were correctly classified and only two samples were incorrectly assigned the class identity.
III. CONCLUSIONS

The detection of the JWH synthetic cannabinoids, as well as of non-JWH cannabinoids and discriminating them from other compounds of forensic interest is extremely important. These compounds have very limited legal uses. Hence, there are very few studies about their pharmacological activity and toxicity.

The most important characteristic of any system screening for these drugs of abuse is its capacity to recognize correctly the class identity of the positive samples, which, according to the requirements for a forensic tool, should not be misclassified under any circumstances.

The results obtained for the nd150_DETECT_DNN model, which was designed to recognize the class identity of JWH synthetic cannabinoids and discriminate them from non-JWH cannabinoids and other substances of forensic interest, indicates that the model is exceptionally efficient. All the JWH or non-JWH cannabinoids have been recognized as such, and no substance belonging to the class of non-JWH cannabinoids or of the class of Others was misclassified as JWH cannabinoid. The two drugs that were misclassified belong to the class of Others and were incorrectly assigned the class identity of non-JWH cannabinoids.

ACKNOWLEDGMENT

An appreciation of the Neural Designer software packages, an important and useful tool used in the construction of the nd150_DETECT_DNN architecture system presented in this paper.

REFERENCES

[1] European Monitoring Centre for Drugs and Drug Addiction, Joint EMCDDA, “Understanding the Spice phenomenon”, Thematic papers. Luxembourg: Publications Office of the European Union, 2009.

[2] C. M. Burlacu, S. Gosav, B. A. Burlacu, and M. Praisler, “Convolutional Neural Network Detecting Synthetic Cannabinoids,” International Conference on e-Health and Bioengineering (EHB), Iași, România, 2021.

[3] C. M. Burlacu, A. C. Burlacu, and M. Praisler, “Physico-chemical analysis, systematic benchmarking, and toxicological aspects of the JWH aminooalkylindole class-derived synthetic JWH cannabinoids,” Annals of the “Dunarea de Jos” University of Galati. Fascicle II, Mathematics, Physics, Theoretical Mechanics, vol. 44, pp. 34-45, 2021.

[4] R. Todeschini and V. Consonni, Handbook of molecular descriptors. Weinheim: Wiley, 2000.

[5] S. Gosav, M. Praisler, J. Van Boeckxlaer, A.P. De Leenheer and D. L Massart, “Class identity assignment for amphetamines using neural networks and GC–FTIR data”, Spectrochim. Acta A Mol. Biomol. Spectrosc., vol. 64, pp. 1110–1117, 2006.

[6] S. Gosav, Studii privind identificarea amfetaminelor ilicite prin metode spectrale şi de inteligenţă artificială. Iași: Tehnopress, 2008.

[7] A. Ion, S. Gosav, M. Praisler, and T. Gosav, “Machine learning tool detecting NBOMe drugs of abuse based on geometrical descriptors,” International Conference on e-Health and Bioengineering (EHB), Iași, România, 2020.

[8] A. F. M. Agarap, “Deep Learning using Rectified Linear Units (ReLU),” Comput. Sci., Mathematics ArXiv, 4090379, pp. 1–7, 2018.

[9] A. Ion, S. Gosav, M. Praisler, and C. Negoiţa, “Artificial neural network designed to identify NBOMe hallucinogens based on 3D-MoRSE descriptors and topological descriptors,” in 23rd ICSTCC International Proceedings, R.-E Precup, Ed. IEEE, 2019, pp. 872-876.

[10] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, vol. 33, pp. 159–174, 1977.