A structure-activity relationship modelling of opioid compounds by using machine learning

A structure-activity relationship modelling of opioid compounds by using machine learning

Fatima Sapundzhi[1], **Meglena Lazarova**[2], **Tatyana Dzimbova**[1,3], **Slavi Georgiev**[4,5], **Antonina Ivanova**[6]

[1] Department of Communication and Computer Engineering, Faculty of Engineering, South-West University "Neofit Rilski", 66 Ivan Myhailov Str., 2700 Blagoevgrad, Bulgaria
[2] Faculty of Applied Mathematics and Informatics, Technical University of Sofia, 8 "St. Kliment Ohridski", Blvd., 1000 Sofia, Bulgaria
[3] Institute of Molecular Biology, Bulgarian Academy of Sciences, Acad. G.Bonchev Str, bl. 21, 1113 Sofia, Bulgaria,
[4] Department of Information Modeling, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str., bl. 8, 1113 Sofia, Bulgaria
[5] Department of Applied Mathematics and Statistics, Faculty of Natural Sciences and Education, University of Ruse, 8 Studentska Str., 7004 Ruse, Bulgaria
[6] Department of Computer Science, Faculty of International Economics and Administration, Varna Free University "Chernorizets Hrabar", 84 Yanko Slavchev Str., Chaika Resort, 9007 Varna, Bulgaria

sapundzhi@swu.bg

Abstract. Opiates are among the oldest drugs that are used to treat many medical problems. They are analgesic and sedative drugs that contain opium. The morphine is its most active ingredient and it is a widely used pain reliever despite its side effects. The main objective of this study is to construct a model which gives the structure-activity relationship among a series of mu-opioid ligands and molecular docking results. For this purpose, a model of mu-opioid receptors using machine learning is introduced. By obtaining a relationship between the docking results and the in vivo test, we could predict the biological effect of the newly synthesized ligands.

1. Introduction

Opioids are pain relievers that are extracted from the papaver somniferum plant [1]. They are widely used as analgesic drugs in the treatment of pain in humans. Opioids bind to three brain receptors: the mu-opioid receptor (MOR), the kappa-opioid receptor (KOR), and the delta-opioid receptor (DOR) in the central nervous system (CNS) and the peripheral organs [2,3]. In contrast to mu-opioid and kappa-opioid agonists, delta-opioid agonists have limited antinociceptive properties which are measured by morphine-sensitive antinociceptive assays as a result of DOR activation. The DOR system may play an essential role in regulating mood and emotional states [4]. Although there are powerful clinically available analgesics (morphine, oxycodone, and fentanyl), they are highly addictive. It is important for

the treatment of pain to develop new opioid drugs that produce analgesia without causing dependence [5,6,7].

With the help of computer-aided drug design or CADD, the drug developers could create more drugs faster by using the state-of-the-art technology. Molecular docking is one of the fundamentals of CADD that analyzes the binding interaction between the target and the small molecules called ligands. These ligands are potential drug candidates for the development of phenotypic and therapeutic models targeted by the CADD proteins [8, 9,10].

In our previous work [11] we have investigated the relationship between the values of the biological activity of delta-opioid analogues that were previously synthesized and biologically tested and have made some investigations over the results of the *in silico* docking. Moreover, the calculation of the minimal energy conformation for each obtained ligand-receptor complex after the docking procedure was investigated.

The main purpose of this study is to create a model for the structure-activity relationship of a series of delta-opioid ligands and molecular docking results by using machine learning where the Delta-opioid receptor (DOR) has a crystal structure (PDBid: 4ej4). This is a continuation of the recent study [12] where the relationship had been sought between the scores and it is obtained by another optimization algorithm where the biological activity of the studied compounds is used. The current paper briefly presents the data while the proposed fitting algorithm as well as the machine learning methods are explained in detail. The paper concludes with a result analysis and with an outline for future research.

2. Experimental remarks

2.1. *Docking procedure*

The docking procedure was performed by using the software for molecular docking GOLD and the ChemScore algorithms, see [11, 13, 14]. The scoring results give information on how good the pose is. Precisely, the goodness of the pose is sketched by the scale of the score.

2.2. *Receptor*

We use the delta-opioid receptor's model with a crystal structure which is published in the RCSB Protein DataBase (PDBid: 4ej4), (http://www.rcsb.org).

2.3. *Ligands*

The Delta-opioid ligands that were previously synthesized by our colleagues were also used in this study, see [11, 15, 16], Table 1. Ligand preparation was performed by the Molegro Molecular Viewer 2.5 (MMV) program, (www.clcbio.com). The prepared structures were utilized for molecular docking by GOLD software. The total energies of the formed ligand–receptor complex after docking were computed by MMV 2.5 where the MolDock scoring algorithm is used, see [17].

3. Results and discussions

In order to establish a relationship between the biological activity of the studied delta-opioid compounds and the results from molecular docking (the values of the scoring functions) the Surface Curve Fitting Toolbox in MATLAB was applied [11].

In our previous research, see [11, 18-20] the experimental data fitting for DOR was carried out by using a polynomial function $z = f(x, y)$ where the values $(z_1, z_2, \dots, z_n)$ of the dependent variable z represent the values of the biological activity of the mu-opioid ligands. The values $(x_1, x_2, \dots, x_n)$ of the independent variable x represent the results from the docking – these are the values of ChemScore function (calculated by GOLD). The values $(y_1, y_2, \dots, y_n)$ of the independent variable y represent the total energies for the ligand-receptor complex which are formed after the docking – the values of MolDock function (calculated by MMV).

Table 1. Data for the biological activity of ligands and docking studies, see [11].

Ligand	*ChemScore*	*Total energy*	*Ligand efficacy*
[Cys(Bzl)^2-Leu^5]-enk	38.91	-170.657	9.3
[Cys(Bzl)^2-Met^5]-enk	35.19	-125.108	3.5
[Cys(O_2NH_2)^2-Leu^5]-enk	28.48	-118.805	29.2
[Cys(O_2NH_2)^2-Met^5]-enk	25.82	-87.343	7.3
[DCys(O_2NH_2)^2-Leu^5]-enk	31.84	-136.187	7.4
[DCys(O_2NH_2)^2-Met^5]-enk	31.55	-139.449	7.1
[HCys(O_2NH_2)^2-Leu^5]-enk	32.75	-100.702	30.2
[HCys(O_2NH_2)^2-Met^5]-enk	26.55	-112.164	3.4
[D-Pen^{2,5}]- enkephalin (DPDPE)	29.23	896.877	4.5
[Leu^5]-enkephalin	31.62	-119.009	5.8
[Met^5]-enkephalin	32.22	-106.792	3.6

The obtained polynomial model of 3rd degree [11] could be interpreted as a surface-fitting function and it analyses the experimental data using a least squares method.

The Surface Fitting Toolbox of MATLAB (http://www.mathworks.com/products/matlab) was applied for analysing the behaviour of one variable that depends on multiple independent variables.

To assess the goodness of fit the following statistical measures were employed: SSE (Sum of squares due to error), $R-Square$ (R^2), $Adjusted\ R^2$, $RMSE$ (*Root Mean Squared Error*).

The results are as follows:

Table 2. Goodness of fit for the polynomial models obtained by the least squares method.

Models Poly (x,y)	Degree of x	Degree of y	SSE	R^2	Adj R^2	RMSE
Poly11	1	1	6988	0.7369	0.6784	27.86
Poly22	2	2	5456	0.7946	0.6234	30.15
Poly33	3	3	0.6290	1.000	0.9999	0.5608

The obtained model shows good fitting properties and significant predictive ability. and was suitable for determining the structure-biological activity relationship: $R^2 = 1.0\ SSE = 0.6290,\ adj\ R^2 = 0.9999,\ RMSE = 0.5608$. Similar studies have been conducted with other compounds, see [11, 18-21].

In this research our aim is to identify a relationship of the from $z = f(x, y)$ where z represents the values of the biological activity of the delta-opioid ligands. The function z is a function of two variables where the independent variables x and y are the values of the GoldScore function and the MolDock function, respectively. Our objective is to discover a nonlinear correlation that utilizes machine learning methodologies, see [22].

The regression models that are employed encompass the k-Nearest Neighbors, Gradient Boosting, Random Forest, and Extra Trees. The latter trio is classified as an ensemble method and is expected to outperform the former, the traditional method which is incorporated primarily for comparison purposes, see [23, 24].

The **k-Nearest Neighbors** (k-NN) algorithm is a multifunctional and easily understood approach that is used across various machine learning applications like regression or classification tasks. In particular, k-NN regression, a derivative of the method, serves as a beneficial instrument for predicting continuous data points.

The core concept of the k-NN regression is founded on the presumption that the data points in the dataset are close in the feature space and they are likely to show analogous outcome values. This algorithm functions by pinpointing 'k' data points in the training dataset that are closest to a new,

unobserved data point employing a certain distance metric, generally the Euclidean distance in the multi-dimensional feature space.

Once the 'k' nearest data points are ascertained, the algorithm creates a forecast for the new data point. For the regression tasks, this prediction is typically the mean of the dependent variable for the 'k' closest data points. This basically suggests that the outcome for an unobserved data point is the average of the outcomes of its neighboring points. A chief benefit of the k-NN regression is its straightforwardness and logical nature. It does not make any explicit suppositions about the underlying functional form of the data, categorizing it as a non-parametric approach. This adaptability allows the k-NN adequately to fit a sophisticated and non-linear data. Nevertheless, the k-NN regression also presents its own set of difficulties. Selecting an ideal 'k' can be a complicated task – a smaller 'k' may result in a model that is excessively sensitive to noise, whereas a larger 'k' may overgeneralize the model, disregarding crucial trends in the data. In addition, the k-NN regression can find its challenging to deal with high-dimensional datasets which is a problem that is frequently referred to as the "curse of dimensionality".

Gradient Boosting is an influential machine learning method utilized in an array of regression and classification applications. Basically the gradient boosting regression is an ensemble algorithm that constructs a predictive model by sequentially adjusting a collection of weak learners to the data, each striving to correct the errors committed by the one before it. The weak learners are generally - decision trees, although different base models may be employed. The fundamental concept is to amalgamate the outputs of numerous simple models to generate a single very precise forecast. This idea of "boosting" arises from the proposition that a combination of weak learners when they are aptly merged, could evolve into a strong learner.

In gradient boosting, the sequential inclusion of weak learners effectively operates as a method of steepest descent, therefore the term "gradient". The method estimates the gradient of the loss function (the measure that gauges the accuracy of the model's forecasts, compared to the actual values) relative to the model parameters, and incorporates new models that orient in the direction that reduces the loss.

For regression purposes, the objective is to predict a continuous outcome variable. In the context of Gradient Boosting Regression, the ensemble of trees is trained to forecast the residuals or errors of the preceding trees. Thus, each subsequent tree is effectively drawing nearer to the true, unidentified function that we are aiming to approximate.

In general, Gradient Boosting has its advantages and disadvantages. It is an exceptionally potent method, capable of fitting complex, non-linear data, and it often performs well even on datasets composed of a mix of categorical and numerical features. Nevertheless, gradient-boosting models can be susceptible to overtraining, particularly if the data is riddled with noise. They could also be computationally challenging and necessitate meticulous tuning of various hyperparameters. For example the number of estimators, the depth of the trees, and the learning rate.

Random Forest is a renowned machine learning technique, regularly employed in an array of regression and classification tasks. As indicated by its name, the Random Forest model consists of an assembly of individual decision trees. These trees are structured in a way that ensures variety and thus resilience in the collective predictive strength.

In the terms of regression, the Random Forest Regressor operates by producing numerous individual decision trees during training, each tree trained on a random data subset and making its distinct prediction. When a new prediction is necessitated, each tree in the forest generates its prediction and the final output is the mean of these individual forecasts. This procedure enables the Random Forest Regressor effectively to capture the intricate, non-linear associations in data.

The core principle of the Random Forest is the notion that an assembly of "weak learners" could amalgamate to compose a "strong learner". Each decision tree in the forest is a weak learner, trained on a random data subset using a random feature subset at each split. This technique is known as bootstrap aggregating or "bagging" coupled with feature randomness. The strength of the Random Forest lies in

its straightforwardness and adaptability. It is relatively unsusceptible to overtraining, due to the randomness incorporated in its construction. It effectively handles both numerical and categorical data and manages missing data efficiently. Moreover, it offers an inherent method for feature importance estimation which can be advantageous for comprehending the model. However, in regard to all models, the Random Forest has its limitations. It can be computationally demanding and slower in both training and forecasting, especially when the tree count gets large. Additionally, it might not perform as well with very high-dimensional sparse data, such as text data and could be less interpretable compared to single decision trees.

Extra Trees Regressor, abbreviated from "Extremely Randomized Trees" is another machine learning method used for regression tasks that belongs to the ensemble learning category. Like the Random Forest the Extra Trees constructs multiple decision trees during the training stage. It introduces an extra degree of randomness, rendering it more resilient and less susceptible to overfitting. In a regular decision tree the optimal split amongst a random subset of features in the node is chosen during the tree growth process. However, in Extra Trees, for each feature being considered, a random value is chosen as the split point, as opposed to the optimal split. Therefore, the Extra Trees Regressor brings in randomness not just at the sample level but also at the level of each decision tree's individual split.

The forecasting process for regression tasks is identical to that of the Random Forest Regressor. Every tree in the ensemble generates a forecast and the final result is the average of these individual forecasts. The randomness in the Extra Trees Regressor aids in capturing complex, non-linear relationships making it an extremely effective tool for regression.

Nevertheless, it is crucial to highlight the distinctions between the Extra Trees and the Random Forest techniques. While they both are using bagging and random feature subsets, the main difference resides in how they split the nodes. The Extra Trees technique is faster due to its randomness at each split, which leads to a quicker training process. However, this could sometimes result in a minor performance decline, thus presenting a trade-off to consider.

The Extra Trees is less likely to overtrain. It handles numerical and categorical data well and could effectively manage missing data. Like Random Forest, it offers an inherent method for feature importance assessment like the Random Forest. The drawback is similar to the other tree-based methods. It might not perform optimally with extremely high-dimensional sparse data such as text data. Its ensemble nature makes it less interpretable compared to the single decision trees.

We have applied the four machine learning methods to build a multivariate regression. The metrics for each model were computed using 10-fold cross-validation and a hyperparameter tuning process was employed to find the optimal parameters. The best hyperparameters can be found in Table 3. Fig. 1 provides the flowchart of the algorithm.

Cross-validation is a tool used to measure how a machine learning model would perform on a novel data set, i. e. data has not been seen during the training. It helps to mitigate the overtraining - a scenario where the model overly memorizes the training data, which usually leads to a poor performance on the unseen data. The most frequently used form of the cross-validation is the k-fold cross-validation where the initial data set is randomly divided into k equally sized subsets. One of these k subsets is set aside for the model validation and the remaining k-1 subsets are utilized for training. This cross-validation procedure is iterated k times ensuring that each subset serves as validation data once. The results from the k folds can then be combined (typically by averaging) to provide a unified measure of the model performance.

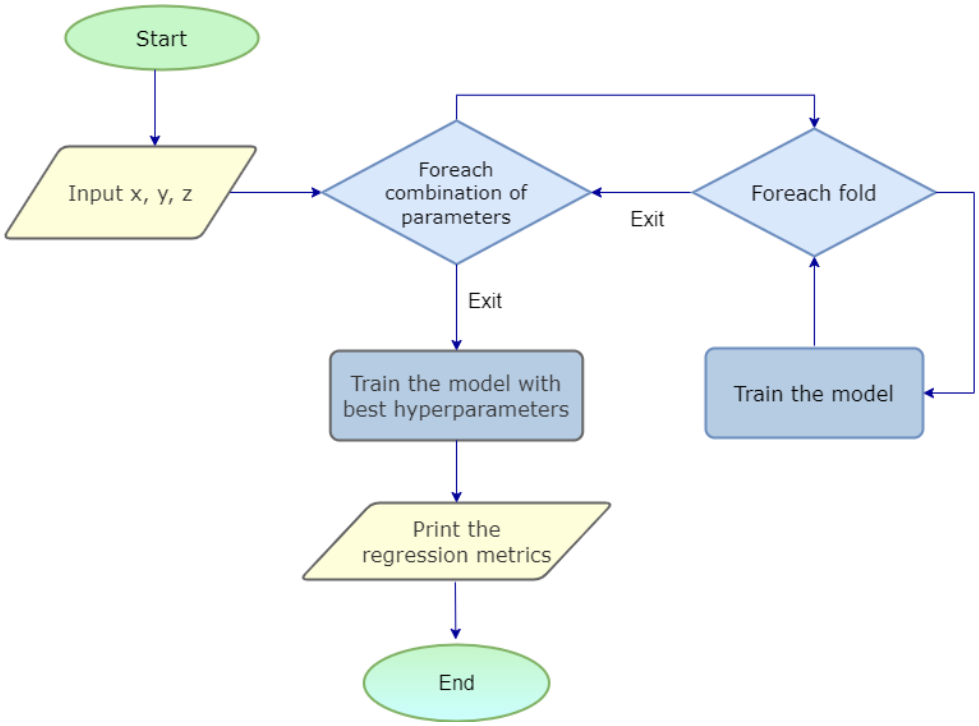

Fig. 1. Flowchart of the regression algorithm (https://app.diagrams.net/).

However, hyperparameter optimization refers to the process of finding the best set of hyperparameters for a machine learning model. Hyperparameters, unlike model parameters, are preset before training and do not change during the learning process. For example, in a Random Forest model, hyperparameters might include the number of trees in the forest or the number of features that each tree evaluates when splitting a node. Hyperparameters significantly affect model performance, making their appropriate selection vital. For the hyperparameter optimization, we apply a 'grid search' strategy, where we define a range of potential values for each hyperparameter and then assess the performance of every possible combination.

Table 3: Modeling by using machine learning regressors.

Regressor	**Best hyperparameters**	**SSE**	**R-square**	**RMSE**
k-Nearest Neighbors	{'n_neighbors': 5, 'weights': 'uniform'}	1223.4328	-0.2559	10.5461
Gradient Boosting	{'max_depth': 2, 'n_estimators': 50}	6.6080	0.9932	0.7751
Random Forest	{'max_depth': 4, 'n_estimators': 200}	228.5116	0.7654	4.5578
Extra Trees	{'max_depth': 2, 'n_estimators': 200}	763.3891	0.2163	8.3306

The optimal hyperparameter combination is presented in the second column of Table 3. The other columns contain the measures, see [14] which establish the comparison between the both approaches. Apparently from the results, the **Gradient Boosting** method indicates an optimal performance. The coefficient of determination reaches one, while the error metrics are almost negligible. In comparison to Table 2, the Gradient Boosting rates as the third-order polynomial approximation Poly33. These findings will agree with the results in [12].

The method can be employed to predict the biological activity of some new ligands via the optimization function ChemScore. The forecast surface is given in Fig. 2, where the white dots are the given data in Table 1. The result could be compared with Fig. 3 in [11].

4. Conclusions

We conducted some docking studies in order to understand the interactions between delta-opioid ligands and the delta-opioid receptor (DOR). These studies allowed us to anticipate the biological activity of newly engineered analogs, which manifested a significantly higher activity compared to the other compounds in the series tested, by creating a correlation between the docking outcomes and in vitro tests.

From this evaluation, we have improved our comprehension of how the biological impacts of compounds correspond to *in silico* experiments. Further, this research opens the avenue to determine if the biological macromolecule models (DOR) are in harmony with the actual three-dimensional structures of the molecules.

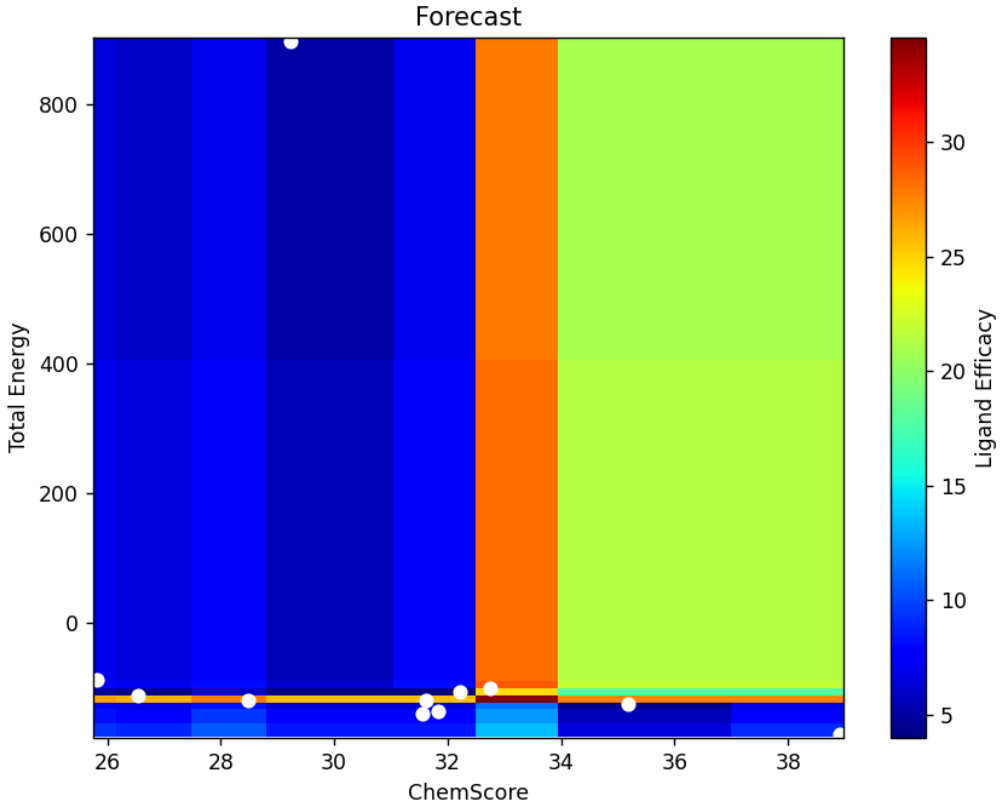

Fig. 2. Forecast surface.

The insights derived from these studies could be instrumental in predicting the potential effectiveness of compounds with known docking scores and total energy calculations. We aim to apply these findings to future research and compound analysis, thereby advancing our understanding of molecular docking

and its implications for drug design and development. In essence, our work provides a powerful analytical tool that bridges the gap between in silico predictions and biological effects, accelerating the process of identifying promising compounds for further investigation.

Acknowledgments

This research is supported by the Bulgarian National Science Fund under Project KP-06-M62/1 "Numerical deterministic, stochastic, machine and deep learning methods with applications in computational, quantitative, algorithmic finance, biomathematics, ecology and algebra" from 2022.

References

[1] Ehrlich A., Kieffer B, Darcq E. 2019 *Expert opinion on therapeutic targets* **23(4)** 315–326 https://doi.org/10.1080/14728222.2019.1586882
[2] Valentino R, Volkow N 2018 *Neuropsychopharmacology* **43(13)** 2514-2520 https://doi.org/10.1038/s41386-018-0225-3
[3] Darcq E, Kieffer B. 2018 *Nat Rev Neurosci* **19(8)** 499–514 https://doi.org/10.1038/s41583-018-0028-x
[4] Jutkiewicz E. 2006 *Molecular interventions* **6(3)** 162–169. https://doi.org/10.1124/mi.6.3.7
[5] Morgan M, Christie M 2011 *British journal of pharmacology* **164(4)** 1322–1334 https://doi.org/10.1111/j.1476-5381.2011.01335.x
[6] Reinscheid R, Nothacker H, Bourson A, 1995 *Science* **270** 792-794 https://doi.org/10.1126/science.270.5237.79
[7] Le Merrer J, Becker J, Befort K, Kieffer L 2009 *Physiol Rev*. **89(4)** 1379–1412 https://doi.org/10.1152/physrev.00005.2009
[8] Agarwal R., Singh A., Sen S. 2016 *Applied Case Studies and Solutions in Molecular Docking-Based Drug Design,* 1-28 IGI Global. https://doi.org/10.4018/978-1-5225-0362-0.ch001
[9] Sapundzhi F., Dzimbova T. 2022 *Bulgarian Chemical Communications* **54 (B1)**, 97-102
[10] Sapundzhi F., Popstoilov M., Lazarova M. 2023 *Numerical Methods and Applications. NMA 2022. Lecture Notes in Computer Science (LNCS*) 13858. Springer, Cham.
[11] Sapundzhi F., Dzimbova T., Pencheva N., Milanov P. 2017 *Bulgarian Chemical Communications* **49 (4)** 768 – 774.
[12] Sapundzi F, Lazarova M, Dzimbova T and Georgiev S 2023 Application of machine learning for modelling of biological data *Submitted to J. Phys. Conf. Proc.*
[13] Jones G, Willett P, Glen R, Leach A, Taylor R 1997 *J. Mol. Biol.* **267**, 727-748.
[14] Sapundzhi F, Prodanova K, Lazarova M 2019 *AIP Conference Proceedings*, 2172, 100008 1-6
[15] Pencheva N., Bocheva A, Dimitrov E., Ivancheva C 1996 *Eur. J. Pharmacol*. **304** 99-108.
[16] Pencheva N., Milanov P., Vezenkov L., Pajpanova T., Naydenova E. 2004 *Eur. J. Pharmacol.* **498** 249-256.
[17] R. Thomsen, M. Christensen, J. Med. Chem., 49, 3315(2006).
[18] Sapundzhi F, Dzimbova T. 2019 *International Journal of Online and Biomedical Engineering* **15(15)** 39–56 https://doi.org/10.3991/ijoe.v15i15.11566
[19] Sapundzhi F., Dzimbova T., Pencheva N., Milanov P. 2016 *Bulgarian Chemical Communications* **49 (E)**, 23-30.
[20] Sapundzhi F 2019 *Bulgarian Chemical Communications* **51 (4),** 569-579.
[21] Traykov M, Trenchev I, 2016 *Genetika* **52(9)**, 1089–1096, https://doi.org/10.7868/s0016675816080130
[22] Russel SJ, Norvig P 2020 *Artificial Intelligence: A Modern Approach* 4th Ed Pearson USA
[23] Kelleher JD, Namee BM, D'Arcy A 2020 *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* 2nd Ed The MIT Press USA
[24] Nedyalkov I, Stefanov A, Georgiev G, 2018 *IX National Conference with International Participation (ELECTRONICA)*, Sofia, Bulgaria, 1-4,