A Binary Classification Model for Toxicity Prediction in Drug Design

Génesis Varela-Salinas[1], Hugo E. Camacho-Cruz[2](✉), Alfredo Juaŕez Saldivar[4](✉), Jose L. Martinez-Rodriguez[1], Josue Rodriguez-Rodriguez[1], and Carlos Garcia-Perez[3]

[1] Universidad Autónoma de Tamaulipas - UAM Reynosa Rodhe,
Carretera Reynosa - San Fernando, 88779 Reynosa, Tamaulipas, Mexico
[2] Universidad Autónoma de Tamaulipas - FMeISC de Matamoros,
Sendero Nacional Km. 3. H, Matamoros, Tamaulipas, Mexico
[3] Information and Communication Technology Department (ICT),
Helmholtz Zentrum München, Ingolstädter Landstrasse 1, 85764 München, Germany
[4] Laboratorio de Biotecnología Farmacéutica, Centro de Biotecnología Genómica,
Instituto Politécnico Nacional, 88710 Reynosa, Mexico

Abstract. Toxicity in drug design is a very important step prior to human or animal evaluation phases. Establishing drug toxicity involves the modification or redesign of the drug into an analog to suppress or reduce the toxicity. In this work, two different deep neural networks architectures and a proposed model to classify drug toxicity were evaluated. Three datasets of molecular descriptors were build based on SMILES from the Tox21 database and the AhR protein to test the accuracy prediction of the models. All models were tested with different sets of hyperparameters. The proposed model showed higher accuracy and lower loss compared to the other architectures. The number of descriptors played a key roll in the accuracy of the proposed model along with the Adam optimizer.

Keywords: Toxicity · Tox21 · Deep learning · Drug design

1 Introduction

Drug development is performed through a series of complex processes. One of the first steps is defining an enzyme as the drug target. Enzymes are proteins that act as drug targets for diseases in the drug design. Then small molecules are identified as active compounds that bind strongly with a protein target. The active compounds are subjected to various experimental evaluations involving cell line assays, animal assays, and human clinical trials [17]. In this regard, during the last decade computational techniques have improved the drug development. Among theses improvements, we can mention the prediction of synergistic drug combinations to avoid drug resistance or increase treatment efficiency and thus, reduce the drug dose to avoid toxicity [22]. Moreover, the use of large

H. Sanjurjo González et al. (Eds.): HAIS 2021, LNAI 12886, pp. 149–157, 2021.
https://doi.org/10.1007/978-3-030-86271-8_13

volume datasets have led drug research to apply complex calculations, where graphical processing units (GPUs) are used for data processing. Therefore, modern drug development has entered the era of big data [18] and new techniques are required. Nowadays, Deep Learning (DL) is a highly demanded technique to promote drug development in the area of artificial intelligence [23].

Deep Learning is of great interest in the process of drug design, in particular for toxicity prediction. Toxicity (or toxic action) is understood as the ability of a substance to cause a harmful response or severe damage to the body functions at cellular or molecular level, and in some cases death [14]. However, some active compounds can present toxicity in high doses but be harmless and even beneficial in small quantities, thus, failing in the latest phases of the development, even if they have obtained satisfactory results in vivo assays [17,21].

In drug design, toxicity evaluation plays a key role for further phases or the approval for human consumption. Nevertheless, the methods used to determine toxicity are slow, tedious, and expensive, not to mention that some of them raise ethical concerns due to the testing of the active compounds in animals [1,5]. For these reasons, predicting toxicity through computational techniques is convenient to accelerate the development of drugs and thus avoid the use of animals in the process.

Encouraged by these reasons, we decided to apply a DL model to toxicity prediction and contribute it to solve this type of problem. Our proposal uses a binary classification model and a dataset with molecular descriptors as feature elements of the AhR molecule from the Tox21 project [19]. The rest of this paper is organized as follows: Sect. 2 includes the data description and the step methods. Section 3 presents and analyzes the results, and finally, in Sect. 4 we present the discussion and conclusions.

2 Data and Methods

Diverse approaches have been proposed for addressing the toxicity prediction problem through machine learning strategies. For example, multiple heterogeneous neural network types and data representations of chemical compounds as SMILE strings [8] have been introduced. Other approaches are shallow networks via 2D features using PADEL descriptors [7] or Deep Neural Networks (DNN) with static and dynamic features [12].

Deep Learning models can be train to learn and recognize molecular descriptors that are active or toxic to a given type of chemical structures. Therefore, to evaluate the toxicity of drugs and reduce tests, a binary classification model was considered to predict if a drug is toxic or not. Figure 1 shows the proposed pipeline. These steps are described in the following subsections.

2.1 Dataset Creation

Tox21 [4,9,20] is a collaboration program of the NIH's NCATS and the National Toxicology Program at the National Institute of Environmental Health Sciences,

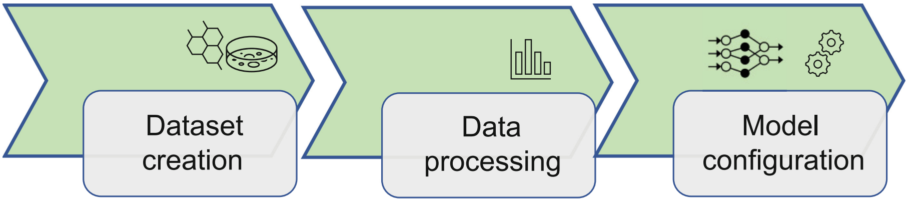

Fig. 1. Proposed method for toxicity prediction

the Environmental Protection Agency, and the Food and Drug Administration. In 2014, the Tox21 members set a Machine Learning challenge to predict the toxicity over 10,000 chemical compounds. Tox21 is divided in 12 assays giving priority to the toxicological evaluation of drugs. We select one assay to predict the toxicity through a Deep Learning model. In the following we will describe the steps to prepare the dataset to train the model. We selected the Aryl hydrocarbon Receptor (AhR) as the target and proceeded to download the list of drugs from the assay in SMILES format (https://tripod.nih.gov/tox21/assays/). The simplified molecular input line entry specification or SMILES is a specification in form of a line notation for describing the structure of a small chemical molecule. It was introduced by Arthur Weininger and David Weininger in the late 1980s. The list contains 8,170 drugs in total, and is divided into active (toxic) or non active (non toxic) regarding the AhR target. Next, we calculate molecular descriptors associated to the AhR target for each drug as shown in Table 1. The molecular descriptors were calculated with Pybel [15].

2.2 Data Processing

Due to the difference in range values between molecular descriptors, we preprocessed the data to ensure a better learning of the features. Data normalization is a recurring technique in Machine Learning for preprocessing data. This type of technique normalizes the data in a range between 0 and 1 for each column. The standard deviation help to avoid differences in values or information loss. For this work, we employed the Normalizer function from the Scikit-learn library [16].

2.3 Proposed Model Architecture

We tried three hyperparameter configurations and architectures designs, as shown in Table 2. We also set a different number of molecular descriptors for the input data (i.e., 4, 8 and 15). All models were run for 64 epochs and with a batch size of 128. We used the Adam [10] and SGD [3] optimizers for the experiments. And we run the training with 10%, 20% and 50% dropout in different models, as is also shown in Table 2.

Having the results of the experiments, we decided to set the proposed model as follows: the input layer as fully connected, 10 nodes in the first hidden layer with sigmoid activation function, and with a dropout at 10%; the second hidden

Table 1. Molecular descriptors.

Id descriptor	Molecular descriptor	Description	Data type
1	atoms	Number of atoms	Discrete
2	bonds	Number of bonds	Discrete
3	HBD	Number of Hydrogen Bond Donors	Discrete
4	HBA1	Number of Hydrogen Bond Acceptors 1	Discrete
5	HBA2	Number of Hydrogen Bond Acceptors 2	Discrete
6	nF	Number of Fluorine Atoms	Discrete
7	logP	Octanol/Water Partition Coefficient	Continuous
8	MW	Molecular Weight Filter	Continuous
9	tbonds	Number of triple bonds	Discrete
10	MR	Molar Refractivity	Continuous
11	abonds	Number of aromatic bonds	Discrete
12	sbonds	Number of single bonds	Discrete
13	dbonds	Number of double bonds	Discrete
14	rotors	Rotatable bonds filter	Discrete
15	MP	Melting point	Continuous

layer with 10 nodes and the RELU activation function, because this avoids gradient fading and saturation. RELU is a rectified function which means that a node will be only activated if the input is above a threshold. Therefore, it rectifies the input values between 0 and 1 regardless of whether they are positive or negative values. The output layer was set to one node to make a binary prediction with sigmoid activation function. We use the Adam [10] optimizer and the Binary Cross-Entropy as loss function. The proposed model is shown in Fig. 2.

3 Experiments and Results

This section presents the experiments used to evaluate the performance of the proposed model. We performed a comparison between the models shown in Table 2 by running the toxicity classification models in a local machine with the following characteristics: 1 node with Intel Core i7 processors at 4.3 Ghz, 8 GB of DDR4 memory, SATA III SSD at 1 TB at 6 GB/, a GPU GEFORCE GTX 1660Ti at 1770 Mhz and 6 GB DDR6. The operating system was LinuxMint version 19.7. Additionally, we applied several libraries such as Pybel [15] for molecular descriptors, Scikit-learn [16] and Pandas [13] for data preprocessing and for creating the input data set. First, the metrics used in the experiments are explained in the following Sect. 3.1. Next, in Sect. 3.2 we will show the scenarios of the experiments. Finally, in Sect. 3.3 we will present the results obtained.

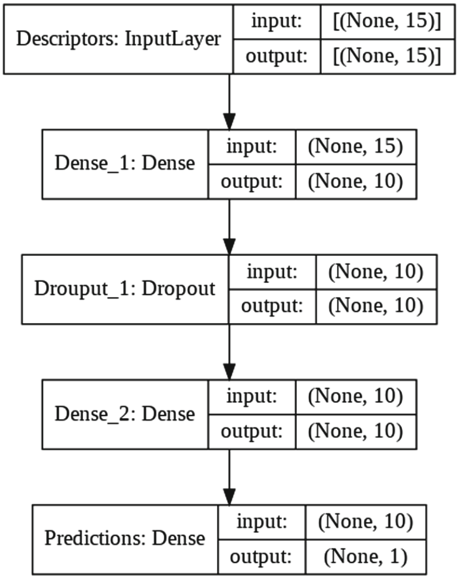

Fig. 2. Binary model architecture

3.1 Metrics

Having two classes denoted as positive and negative causes a binary classification problem. To measure the performance of the trained models, we used the Recall, Precision, and F-1 scores [2,6]. Then,

- Precision. It is the number of correctly classified positive examples divided by the total number that are classified positive. That is,

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

  where TP is the number of true positives, FP the number of false positives, and P precision.
- Recall. Measures the number of how many of the actual positives (true positives and false negatives) were predicted correctly as positives (true positives),

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

  where FN is the number of false negatives and R is the recall.
- F-1 is an harmonic measure that combines precision (P) and recall (R) as shown in

$$\text{F-1} = 2\frac{P \cdot R}{P + R}$$

F-1 is 1 when there is no FP, and FN and 0 when there is no TP. F-1 is particularly useful when the number of positive and negative classes are substantially different or imbalanced in the data.

3.2 Scenarios

As we mentioned in the previous Sect. 2.1, we calculated molecular descriptors to train the models. Pybel is limited to 15 molecular descriptors. Therefore, we calculated the maximum of 15 molecular descriptors, then we tried with a different set of molecular descriptors. After a statistical analysis of the molecular descriptors (not shown here), we made three datasets with the following number of descriptors: 4, 8 and 15. All models use standard deviation as normalization, and the 10 k-fold cross-validation.

Table 2. Model configuration.

Model	Layers	Optimization
M1	1^{st}: RELU - input 2^{nd}: RELU 6 nodes 3^{rd}: Sigmoid– output	Adam
M2	1^{st}: input 2^{nd}: Simoid 10 nodes 3^{rd}: RELU 10 nodes 4^{th}: Sigmoid– output	
	Dropout: 10% 1^{st} hidden layer	
M3	1^{st}: RELU – input 2^{nd}: RELU 16 nodes 3^{rd}: RELU 6 nodes 4^{th}: RELU 64 nodes 5^{th}: sigmoidal– output	SGD
	Dropout: 20% input layer 50% 4^{th} layer	

For each of the three datasets we used three different models (Table 2). We must highlight that the datasets were heavily unbalanced where the toxic samples were the lower class with 950 samples against 7,219 non toxic samples for the AhR receptor. To overcome this, we applied the under-sampling technique provided by the imblearn library [11].

We train the three models with the balanced training set in order to observe the performance according the settings of the hyperparameters. To validate and see if the models generalize well, we wanted to see if any of the models were able to adapt properly to unseen data and classify samples correctly. We used the k-fold cross-validation technique with 10 splits with the Scikit-learn library. We filtered the descriptors to keep only the highest correlated ones to improve the performance of the models.

3.3 Results

The metrics from the three models are shown in Table 3. Model 2 (M2) shows a high F-1 score of 0.89 using the 15 molecular descriptors while model 1 (M1) is the second best (F-1 of 0.88) also using the 15 molecular descriptors. Finally, in model 3 (M3) the F-1 score is 0.84. Evidently, when using 15 molecular descriptors the performance is better for all three models. It is also interesting that M2 and M1 showed a very close F1 score with 8 molecular descriptors. Finally, we can say that M3 performs lower than models M2 and M1. Additionally, we run a Support Vector Machine (SVM) and a Gaussian Naive Bayes (GaussianNB) algorithm from the Scikit-learn library in order to compare traditional ML methods against the three proposed models. Table 3 summarizes that both methods reach a F-1 score of 0.86 with 15 molecular descriptors while M2 has a F-1 score 0.86 with 8 molecular descriptors. It is clear that M2 with 15 molecular descriptors overpasses classical ML approaches. Although the idea of keeping only the descriptors with the highest correlation was supposed to improve the performance of the models (4 and 8 molecular descriptors), the results show that using all the descriptors provides better results in the F-1 score.

Table 3. Metrics for the five models.

Model	Desc	Precision	Recall	F-1
M1	4	0.961	0.776	0.859
	8	0.941	0.806	0.867
	15	0.913	0.854	0.881
M2	4	0.958	0.744	0.837
	8	0.948	0.788	0.860
	15	0.947	0.839	**0.890**
M3	4	0.833	0.737	0.781
	8	0.907	0.764	0.828
	15	0.907	0.791	0.844
SVM	4	0.943	0.755	0.838
	8	0.948	0.782	0.856
	15	0.938	0.798	0.862
GaussianNB	4	0.773	0.823	0.796
	8	0.832	0.843	0.837
	15	0.881	0.851	0.865

4 Conclusions

This paper presents a strategy to develop a binary classifier for toxicity prediction in the drug design pipeline. The dataset from the AhR Tox21 assay was used to calculate molecular descriptors, and it was used as input data to train a set of Machine Learning models. On one side, the experiments showed that the proposed model (M2) achieves promising results shown in the F-1 score when using 15 molecular descriptors and multiple hidden layers with the RELU and sigmoid activation functions. On the other side, M2 performs better than classical ML algorithms as shown in Table 3. In conclusion, the results show that more than 15 molecular descriptors could improve the F1-score for the SVM and the Gaussian Naive Bayes algorithm, and therefore the F1-score from M2.

References

1. Atkinson Jr., A.J., Markey, S.P.: Biochemical mechanisms of drug toxicity. In: Principles of Clinical Pharmacology, pp. 249–271. Elsevier (2007)
2. Bania, R.K.: COVID-19 public tweets sentiment analysis using TF-IDF and inductive learning models. INFOCOMP J. Comput. Sci. **19**(2), 23–41 (2020)
3. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Lechevallier, Y., Saporta, G. (eds.) Proceedings of COMPSTAT 2010, pp. 177–186. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-7908-2604-3_16
4. Collins, F.S., Gray, G.M., Bucher, J.R.: Transforming environmental health protection. Science **319**(5865), 906–907 (2008). https://doi.org/10.1126/science.1154619, https://science.sciencemag.org/content/319/5865/906
5. Dearden, J.C.: In silico prediction of drug toxicity. J. Comput. Aided Mol. Des. **17**(2–4), 119–127 (2003). https://doi.org/10.1023/A:1025361621494
6. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945). https://doi.org/10.2307/1932409, https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1932409
7. Karim, A., Mishra, A., Newton, M.H., Sattar, A.: Efficient toxicity prediction via simple features using shallow neural networks and decision trees. ACS Omega **4**(1), 1874–1888 (2019)
8. Karim, A., Singh, J., Mishra, A., Dehzangi, A., Newton, M.A.H., Sattar, A.: Toxicity prediction by multimodal deep learning. In: Ohara, K., Bai, Q. (eds.) PKAW 2019. LNCS (LNAI), vol. 11669, pp. 142–152. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30639-7_12
9. Kavlock, R.J., Austin, C.P., Tice, R.R.: Toxicity testing in the 21st century: implications for human health risk assessment. Risk Anal. **29**(4), 485–487 (2009). https://doi.org/10.1111/j.1539-6924.2008.01168.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2008.01168.x
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017)
11. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **18**(17), 1–5 (2017)
12. Mayr, A., Klambauer, G., Unterthiner, T., Hochreiter, S.: DeepTox: toxicity prediction using deep learning. Front. Environ. Sci. **3**, 80 (2016)

13. McKinney, W., et al.: Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, vol. 445, pp. 51–56 (2010)
14. Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Müller, L., Pähler, A.: Computational toxicology in drug development. Drug Discovery Today **13**(7–8), 303–310 (2008)
15. O'Boyle, N.M., Morley, C., Hutchison, G.R.: Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. Chem. Cent. J. **2**(1), 1–7 (2008)
16. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
17. Saldívar-González, F., Prieto-Martínez, F.D., Medina-Franco, J.L.: Descubrimiento y desarrollo de fármacos: un enfoque computacional. Educación química **28**(1), 51–58 (2017)
18. Sid, K., Batouche, M.C.: Big data analytics techniques in virtual screening for drug discovery. In: Lazaar, M., Tabii, Y., Chrayah, M., Achhab, M.A. (eds.) Proceedings of the 2nd International Conference on Big Data, Cloud and Applications, BDCA 2017, Tetouan, Morocco, 29–30 March 2017, pp. 9:1–9:7. ACM (2017). https://doi.org/10.1145/3090354.3090363
19. Thomas, R.S., et al.: The US Federal Tox21 program: a strategic and operational plan for continued leadership. ALTEX - Altern. Anim. Exp. **35**(2), 163–168 (2018)
20. Tice, R.R., Austin, C.P., Kavlock, R.J., Bucher, J.R.: Improving the human hazard characterization of chemicals: a Tox21 update. Environ. Health Perspect. **121**(7), 756–765 (2013). https://doi.org/10.1289/ehp.1205784, https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.1205784
21. Verbist, B., et al.: Using transcriptomics to guide lead optimization in drug discovery projects: lessons learned from the QSTAR project. Drug Discovery Today **20**(5), 505–513 (2015)
22. Wang, X., Song, K., Li, L., Chen, L.: Structure-based drug design strategies and challenges. Curr. Top. Med. Chem. **18**(12), 998–1006 (2018)
23. Zhang, L., Tan, J., Han, D., Zhu, H.: From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discovery Today **22**(11), 1680–1685 (2017)