International Journal of Scientific and Management Research
Volume 5 Issue 3 (March) 2022
ISSN: 2581-6888
Page: 42-54

A machine learning-based QSAR model for predicting phenols cytotoxicity

Latifa Douali[1]*
[1]Department of computer sciences, Regional Centre of Training and Education (CRMEF) Marrakech-Safi, Marrakech – **Morocco**

DOI - http://doi.org/10.37502/IJSMR.2022.5305

Abstract

Cytotoxicity is a very important aspect that gains big research interest in toxicology and pharmacology as it is related to whether a compound may cause cell damage, necrosis or apoptosis and this contributes effectively to determine the toxicity potential of a compound and to cancer treatment studies as well. The complexity and the sensitivity of cytotoxicity assays and the involvement of animal tests in many instances increase the need for rapid and reliable alternative methods. Quantitative structure-activity relationships (QSAR) are relevant techniques that provide mathematical models and help in chemicals screening and in predicting biological activities and eventually cytotoxicity. Because of their capacity to handle complex problems, machine learning contributed substantially to the QSAR field's evolvement. In this study, we established a predictive QSAR model based on machine learning, namely deep neural network to predict the cytotoxicity of phenols. The model exhibited high performances in predicting new compounds. It was proved that hydrophobic, steric and electronic effects are relevant in determining the cytotoxicity variability of phenols against Tetrahymena pyriformis.

Keywords: Cytotoxicity, Machine learning, Phenols, QSAR, Tetrahymena pyriformis, Risk assessment.

1. Introduction

Phenols are chemical compounds that can both be produced naturally and manufactured. They ubiquitously exist in the environment as the secondary metabolites of plants. They are present in fruits, cereals, vegetables. They are well-known by their anti-oxidant effect [1]–[4] and anti-cancer benefits as natural or synthetic products. It was suggested that they suppress oxidative stress by scavenging peroxy radicals [5]–[7]. Some phenols were reported as natural antimicrobial products that protect the plant from pathogens such as lignin, flavones, stilbenes and salicylic acid [8]. They might also be produced as a result of using pesticides in agriculture like the pentachlorophenol.

Phenolic compounds are also synthesised to be used in many industries such as preservatives, food processing, leather, plastic, resins. Phenols are known as industrial waste products present in soil, air and wastewater. In this sense, they represent big environmental and toxicological risks. Because of their high solubility and toxicity, they are considered as dangerous pollutants. The study of their toxicity against aquatic organisms is revealed to be very important.

Dangerous risks to humans may occur through multiple way of exposure to phenols, especially by inhalation, ingestion or dermal contact with these compounds or contaminated air or water [7], [9]–[11]. In fact, phenols irritate skin and cause its necrosis. More, many phenolic compounds that can be found in wastewater, e.g. bisphenol A, Butylated hydroxyanisole (BHA), some alkylphenols and nitrophenols can cause estrogenic disruption or teratogenic effects [12]–[14]

The antioxidant effect of phenols is suggested to be related to their ability to scavenge endogenous reactive oxygen species (ROS) according to equation (1).

$$ROO^{\bullet} + ArOH \xrightarrow{k_1} ROOH + ArO^{\bullet}$$

Equation (1)

The existence of the aryloxyl radical, under some circumstances, may trigger different kinds of secondary reactions and there is some chance to conduct to display pro-oxidant activity (Equation (3)). This pro-oxidant effect is pretended to involve interaction of the phenols with transition metals [15].

$$ROO^{\bullet} + X - H \xrightarrow{k_2} ROOH + X^{\bullet}$$

Equation (2)

$$ArO^{\bullet} + ROOH \xrightarrow{k_3} ArOH + ROO^{\bullet}$$

Equation (3)

Either the anti-oxidant mechanism of phenols involves a Hydrogen atom transfer or a single electron transfer, it is admitted that the anti-oxidant activity is attributed to some characteristics: O—H bond strength, hydrophobicity and hindrance in the anti-oxidant. It is noteworthy to mention that cytotoxicity may occur according the process of attacking ROS or by causing damage to cellular systems or DNA. The dual behavior of phenols raises big questions about their toxicity mechanism [6], [15], [16]. Cytotoxicity and anti-oxidant activity for many phenolic compounds should be determined. Many computational methods have been developed to relate chemical structure to biological activities. Quantitative Structure-Activity Relationships (QSAR) are relevant techniques that develop theoretical models and help to understand such mechanisms [6], [17]–[19]. QSAR explore and model the relationship between the chemical structure or physicochemical properties of a set of molecules and a target endpoint. The latter can be a feature, biological activity or physicochemical property. They require considerable and meticulous attention in all model development steps. Since their first development by Hansch et al. [20], they were successful in fulfilling the need to a knowledge-guided synthesis of prospective compounds and to speed up virtual screening and drug design process. They play major role in determining the factors enhancing or decreasing a given activity and hence understanding the mechanisms governing the activity and in predicting new congeners of synthesized drugs, more effective and more active.

QSAR models were very useful in several domains such as combinatorial chemistry, toxicology, high-throughput screening, drug development and drug design. They exhibited high performances in predicting biological activities that were experimentally confirmed afterward [6], [18], [21], [22]. In toxicology, QSAR approach may constitute an alternative to in vivo and animal tests. They proved to be time and costs saving.

Many statistical methods were used to establish QSAR models. Among these methods, machine learning (ML) proved to be highly pertinent and successful in developing accurate models especially with complex nonlinear data. Artificial neural networks (ANN) were first used in QSAR modelling in 1973 by Hiller et al. to distinguish between active and non-active compounds [23]

They have proven to be powerful for nonlinear optimization problems frequently encountered in biological and chemical processes. Artificial neural networks (ANN), support vector machines (SVM), random forest (RF) were among machine learning methods that were used as a QSAR model-building in numerous studies and for many chemical compounds.

Deep learning is a ML algorithm based on ANN that proves a high-performance in handling QSAR modeling. Owing to its ability to detect meaningful patterns in the data, it outperforms many statistical methods in inferring the relationships between the endpoints and the inputs and in generating accurate predictive models. ANN with backpropagation learning algorithm are the most popular ML methods used in many domains. They are based on several nodes arranged in multiple layers and can map any complex function. Each node can be activated by previous activated nodes via weighted interconnections. Many transformations and approximations are applied to the nodes' status and the weights until the production of the desired behavior.

Unlike shallow ANN, the process of producing outputs from inputs, with the deep neural networks (DNN) involves a multi-level learning strategy leading to more accurate data processing. Through the different layers, the algorithm goes hierarchically from low level pattern extraction to a higher level, in other words, to a higher level of data abstraction. Thus, the network optimization is faster and the problem of overfitting, from which suffer mostly the shallow ANN, is avoided [24]. Handling large datasets and implementing multiple layers becomes thus possible with DNN and helps approximating complex function. DNN gain now big interest and their usage opens new era in QSAR models establishment. Interestingly, the molecular descriptors that constitute the inputs of the DNN can be implemented without many restrictions on the dataset size [25]. One can describe the prototypes/molecules as precisely as possible using 1-D, 2-D and 3-D parameters. The selection of the inputs obeys to other considerations related to the relevance and the reliability of these descriptors.

It is interesting to note that DNN were primarily used in many classification and transcription problems. There are only few investigations using DNN in establishing regression models [26], [27]. As in many domains, they provide good predictive models for biological and pharmaceutical properties [26]–[29].

Many QSAR models studying different datasets of phenolic compounds were described in the literature [6], [18], [30]–[34]. Most of them have used multiple linear regression (MLR) and provide several linear models. Other models were also developed by partial least squares (PLS) and ANN.

For this investigation, we used DNN to develop a regression-based QSAR model to predict cytotoxicity of phenols to Tetrahymena pyriformis. For this purpose, we calculated several parameters to describe the molecular structure and the physico-chemical properties of the studied phenolic compounds and to serve as network inputs.

2. Material and Methods

2.1 Datasets

In this study, 250 phenolic compounds with different types of substituents on the aromatic ring were investigated. Substituents in ortho, meta and para positions were implemented in the same dataset. These compounds were widely studied by many authors [35], [36]. The data contain mono-, bi-, and tri-substituted phenols. In a previous work, Selassi et al. [32], [37], [38] studied separately different datasets of phenols with electron-releasing and electron-withdrawing substituents.

The target activity is the cytotoxicity of the 250 phenols to the Tetrahymena pyriformis ss expressed by IGC50; the 50% growth inhibitory concentration (mmol/L) of a compound to Tetrahymena pyriformis. Different modes of action (MOA) were considered. For calculation conformity, the values of log (1/IGC50) were considered as endpoints. It is noteworthy that the dataset contains much diversified endpoints. The values of log (1/IGC50) ranges from -1.5 to 2.71 with a mean value of 0.739 and a standard deviation of 0.828.

To be able to assess the model, the initial dataset was divided into a training and a predictive datasets. Thus, 80% randomly selected phenols were used for the model development and the remaining 20 % served as external dataset to assess the predictive capability of the DNN model. Both subsets contain compounds with electron-withdrawing and electron-releasing substituents.

2.1 Molecular Descriptors

It was emphasized in many instances that the quality of molecular description inherently affects the quality of a QSAR model [19], [39]. In this investigation, we paid considerable attention to the molecular descriptors to implement in the model towards getting a reliable model. In this study, the endpoints were sparse and the substituents were significantly diversified and there was a need to generate molecular descriptors that describes well the features of the whole dataset and delineates its diversity. Hence, molecular descriptors that describe the entire molecule have been chosen. Many structural parameters that give details on the molecular connectivity were implemented, namely Kier parameters [40] indices and structural fingerprints [41]. They were generated by the QSARIN software [42]. Actually, fingerprints are numerical values that encode fragments or subgroups in a molecule. To take into account the electronic interaction of the phenols, we introduced molecular descriptors such as the molar refractivity (MR) and the Mc Gowan volume (McVol) [43]. Those descriptors were calculated using the CLogP program. The hydrophobic character of the compounds was introduced via logP parameter, the octanol-water partition coefficient of the whole compound, it was generated by the CLogP program. Electronic aspect of the substituents plays a major role in the cytotoxicity of phenols. Many electronic parameters, such as LUMO (Lower Unoccupied Molecular Orbital) and HOMO (Highest Occupied Molecular Orbital), Pka (acid dissociation constant) and Ip (ionization potential) were implemented. These molecular descriptors were calculated after a geometry optimization of the molecules using the PM6 semi-empirical quantum method implemented in MOPAC 7 program [44], [45]. The implementation of these electronic descriptors helped to handle a dataset of phenols with electron-releasing and electron-withdrawing substituents. A total number of 118 molecular descriptors were then generated for the whole dataset.

2.3 Model development

For the purpose of this study, we developed a neural network based on Keras library and using Tensorflow framework [46]–[48]. In order to construct models with high predictive performance, many hyperparameters were optimized including the learning rate, the bias and batch size. The DNN architecture, consisting of many interconnected layers, was also monitored. DNN architectures consisted of two hidden layers, one input and one output layer. Although there is common consensus about the DNN capability of handling big data, the most important in the context of QSAR model establishment is the significance of the inputs and the information they provide. On this regard only relevant molecular descriptors were implemented as network inputs in this work. Hyperbolic tangent function was used as activation function for hidden nodes and the stochastic gradient descent algorithm was adopted.

Two regression models were developed as well using MLR and shallow ANN methods to compare with DNN. The ANN was constructed with the same inputs in the input layer, one hidden layer and one output layer. The learning rate was set to 0.1.

3. Results and Discussion

3.1 Model fitting

The dataset was randomly split into a training dataset (200 samples) and a predictive dataset (50 samples). The distributions of the cytotoxicity for both datasets (training and predictive) were quasi similar. The training stage afforded an opportunity to fine-tune the internal networks hyperparameters

To evaluate the accuracy of the established models, and to be able to compare it with other models, two metrics were adopted; the statistical root-mean-square deviation (RMSD) and the correlation coefficient R^2.

A DNN model with two layers, 20 and 14 nodes each respectively exhibited the best statistical metrics. The correlation coefficient was high and it was equal to 0.943, the RMSD was very low and it was equal to 0.194. These results were significantly better than the results of the MLR and ANN models. In fact, the developed MLR model implementing all molecular descriptors resulted in an R^2 of 0.67 and an RMSD of 0.51. The linear model generated one outlier. A constructed ANN model with 10 nodes in the hidden layer and with the sigmoid as an activation function, gave a relatively high R^2 of 0.74 and an RMSD of 0.60 adding more nodes to the hidden layer pushed the network into over fitting.

From the above-mentioned metrics, one can conclude that the DNN outperformed MLR and ANN. It accomplished a perfect fit of the data. It was able to extract the molecular features that govern the phenols cytotoxicity. It could extract information provided by the molecular descriptors that fed the DNN; Namely the physicochemical parameters augmented by the topological parameters. The DNN performances can be perceived by further examination of Fig. 1 which represents the cytotoxicity values calculated by the DNN model versus the experimental values.

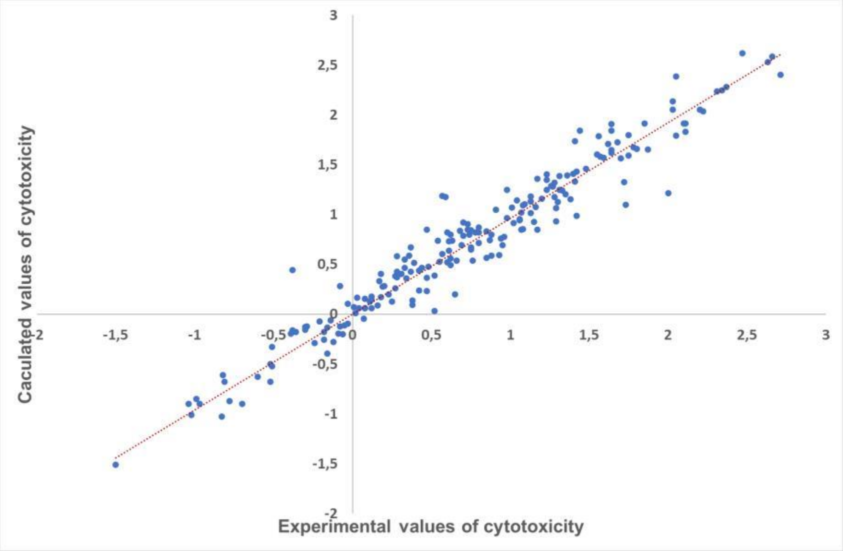

Figure 1: Calculated versus experimental cytotoxicity of phenols

3.2 Applicability domain

The applicability domain (AD) is one of the pillars to ensure the reliability of the model and its accuracy in predicting new compounds. [49].

AD is defined as a region in the chemical space containing the molecules used to train and validate the model. Distance to model allows to determine outliers and shows applicability of the constructed model to other similar molecules.

While there are several methods to determine the applicability domain, we used leverages to determine the distance to the model, the resulting AD is represented in Figure 2

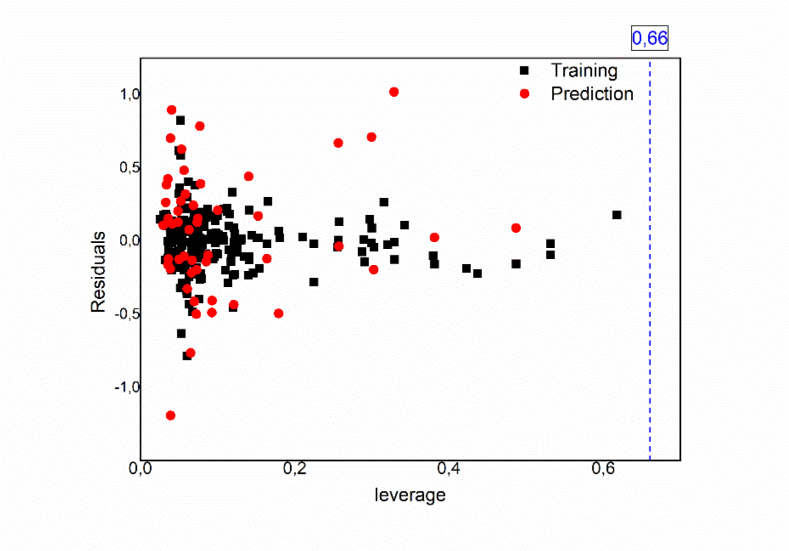

Figure 2: Applicability domain established for the model No outliers were detected

3.3 Residual analysis

To assess the accuracy of the established model, a residual analysis has been carried out. It allowed a closer scrutiny of the QSAR model and its capability to predict new compounds, external to the training dataset. The results are depicted in Figure 3

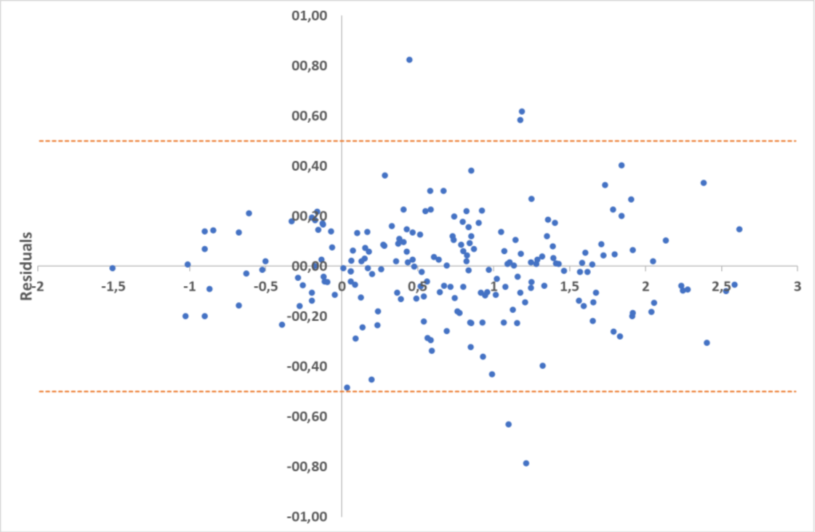

Figure 3: Residual plot

In figure 3, the values are evenly distributed along the X axis. The majority of the residuals lie in the interval [-0.5 – 0.5]. This proves the adaptability of the model and that the model is reliably applicable to predict the cytotoxicity of phenols to Tetrahymena pyriformis.

3.4 The leave one out validation

A Leave-One-Out validation procedure was carried out to evaluate the performance of the constructed model and to ensure that the network built was able to process new data and to predict the cytotoxicity of new compounds. The results were promising and exhibited a leave-one-out cross validated R^2 (q^2) of 0.941 and an RMSD of 0.203. These results led to adopt the constructed model to predict new activities.

3.5 The prediction of new compounds

In this investigation, 50 phenols were used as an external prediction dataset. The applicability domain showed that this data is structurally similar to the training dataset (Figure 2). To evaluate the DNN prediction capability, we used the coefficient of determination and the standard deviation of prediction (SDTP) metrics. Indeed, it have been indicated that the SDTP correlates with the prediction accuracy [34]. This stage resulted in a high R^2 that equals 0.739 and an SDTP equals 0.434. The predicted values versus the true values of cytotoxicity are reported in Figure 4. It shows that the established DNN model could predict correctly the cytotoxicity of almost all new phenols in the prediction dataset.

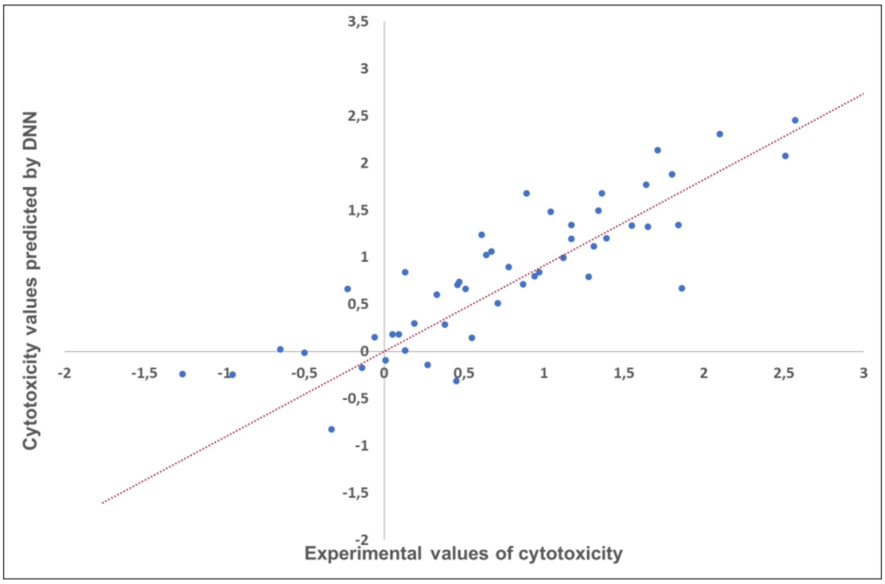

Figure 4: Cytotoxicity predicted by the DNN versus the experimental values

In comparison to other QSAR models in the literature established by different statistical methods, namely MLR and partial least squares (PLS), the present model showed better prediction capabilities. While in some models [34], a large number of outliers (80 outliers) was excluded, the present model was able to fit the whole dataset. The selected molecular descriptors provided the DNN with precise information. The DNN was successful to extract the features required to accurately predict the cytotoxicity.

3.6 Variation of cytotoxicity with respect to ClogP

Hydrophobicity plays a major role in the cytotoxicity of many chemical compounds and phenols as well [50], [51] and one of the major objectives of QSAR is to shed light on the factors influencing the biological activity/cytotoxicity of the studied compounds. In previous work, we demonstrated a parabolic distribution of the anti-HIV activity with regard to hydrophobic parameter. Here we report the variation of the cytotoxicity of the phenols studied versus the hydrophobic parameter ClogP, as it was depicted by the present model. The resulting plot is reported in Figure 5. It is in accordance with the fact that phenols have both lipophilic and hydrophilic character [52].

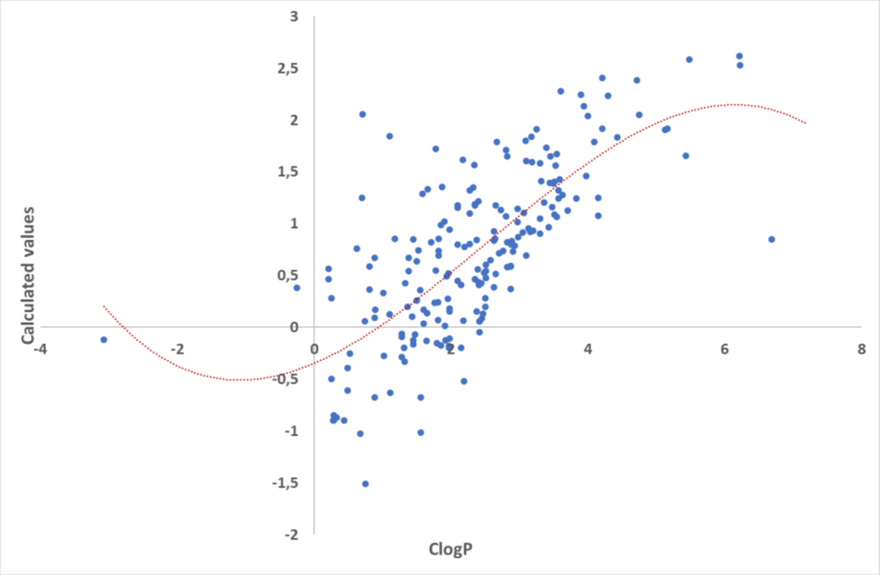

Fig. 5: Calculated values of cytotoxicity versus the ClogP parameter

4. Conclusion

In the present study, a DNN model was successfully developed to predict the phenols cytotoxicity to Tetrahymena pyriformis. As proved by the statistical metrics and the performed residual analysis, the DNN model succeeded in mapping the molecular features of the phenols to their cytotoxicity. Unlike other modeling approaches, reported in the literature and in this study, where several input outliers had to be excluded, the present DNN model fit well all the samples and predicted accurately all the proposed new compounds. Both the fitting quality and the predictability accuracy were significantly high.

Actually, two main factors played a part in this success:

- The use of DNN ensuring an automatic feature extraction capability and a non-linear transformation function involved to learn chemical patterns.
- The multi-component representation of the inputs. The present description of the compounds consisted on parameters that described precisely the molecular graphs of the compounds and hence simulating the spatial images of the molecules, and parameters that described the physicochemical characteristics of the molecules, caused mainly by the different substituents on the aromatic ring. This provided the networks with informative chemical features. Indeed, DNN excel in image recognition. Bearing in mind that a molecule is far from being a static image/graph, we proposed to add physicochemical parameters. They convey a valuable information on the intrinsic features of a molecule. Specifically, electronic and hydrophobic characteristics play a key role in a compound-biological system interaction. By instance, electron-releasing and electron-withdrawing groups affect differently the reactivity of a phenolic compound and hence affect their biological responses, and the hydrophobic character manages the penetration of a chemical compound into the biological system.

Basically, the parameters provided to the networks should be sufficiently precise and diverse to ensure reliable predictions.

Certainly, our main objective in this QSAR investigation is to build an accurate predictive model. Thus, an external dataset of 50 phenols, as sparse as the training dataset and with structural features close to the fitting dataset (containing electron-donor, electron-attracting and mono, bi-, tree- substituents) served as predictive dataset. In contrast to the models reported in the literature, the present DNN predicted the cytotoxicity of new compounds at about 74% of precision. All the proposed compounds were well predicted as it was asserted by the statistical metrics and the residual analysis. Cytotoxicity assays may benefit largely from deep learning and rigorously established models. This would be of considerable help to evolve animal-free assays.

References

[1] L. Zhao *et al.*, "Nutshell Extracts of Xanthoceras sorbifolia: A New Potential Source of Bioactive Phenolic Compounds as a Natural Antioxidant and Immunomodulator.," *J Agric Food Chem*, vol. 66, no. 15, pp. 3783–3792, Apr. 2018, doi: 10.1021/acs.jafc.7b05590.

[2] G. Liu *et al.*, "Antioxidant capacity of phenolic compounds separated from tea seed oil in vitro and in vivo," *Food Chem.*, vol. 371, p. 131122, Mar. 2022, doi: 10.1016/j.foodchem.2021.131122.

[3] B. Singh, J. P. Singh, A. Kaur, and N. Singh, "Phenolic composition, antioxidant potential and health benefits of citrus peel," *Food Res Int*, vol. 132, p. 109114, Jun. 2020, doi: 10.1016/j.foodres.2020.109114.

[4] Z. Rappoport, Ed., *The chemistry of phenols*. Hoboken, NJ: Wiley, 2003.

[5] N. R. Gassman, "Induction of oxidative stress by bisphenol A and its pleiotropic effects," *Environ. Mol. Mutagen.*, vol. 58, no. 2, pp. 60–71, 2017, doi: 10.1002/em.22072.

[6] R. Garg, S. Kapur, and C. Hansch, "Radical toxicity of phenols: a reference point for obtaining perspective in the formulation of QSAR," *Med. Res. Rev.*, vol. 21, no. 1, pp. 73–82, 2001.

[7] I.-H. Acir and K. Guenther, "Endocrine-disrupting metabolites of alkylphenol ethoxylates - A critical review of analytical methods, environmental occurrences, toxicity, and regulation," *Sci. Total Environ.*, vol. 635, pp. 1530–1546, Sep. 2018, doi: 10.1016/j.scitotenv.2018.04.079.

[8] W. Vermerris and R. Nicholson, *Phenolic Compound Biochemistry*. Springer Netherlands, 2006.

[9] F. Bajot, M. T. D. Cronin, D. W. Roberts, and T. W. Schultz, "Reactivity and aquatic toxicity of aromatic compounds transformable to quinone-type Michael acceptors," *SAR QSAR Environ Res*, vol. 22, no. 1–2, pp. 51–65, Mar. 2011, doi: 10.1080/1062936X.2010.528449.

[10] S. Gautam, Samiksha, S. S. Chimni, S. Arora, and S. K. Sohal, "Toxic effects of purified phenolic compounds from Acacia nilotica against common cutworm," *Toxicon*, vol. 203, pp. 22–29, Nov. 2021, doi: 10.1016/j.toxicon.2021.09.017.

[11] W. Wang, P. Xiong, H. Zhang, Q. Zhu, C. Liao, and G. Jiang, "Analysis, occurrence, toxicity and environmental health risks of synthetic phenolic antioxidants: A review," *Environmental Research*, vol. 201, p. 111531, Oct. 2021, doi: 10.1016/j.envres.2021.111531.

[12] Y. Ma *et al.*, "The adverse health effects of bisphenol A and related toxicity mechanisms," *Environ Res*, vol. 176, p. 108575, Sep. 2019, doi: 10.1016/j.envres.2019.108575.

[13] W. A. Yehye *et al.*, “Understanding the chemistry behind the antioxidant activities of butylated hydroxytoluene (BHT): a review.,” *Eur J Med Chem*, vol. 101, pp. 295–312, Aug. 2015, doi: 10.1016/j.ejmech.2015.06.026.

[14] N. Arnich *et al.*, “Conclusions of the French Food Safety Agency on the toxicity of bisphenol A,” *Int J Hyg Environ Health*, vol. 214, no. 3, pp. 271–275, Jun. 2011, doi: 10.1016/j.ijheh.2010.12.002.

[15] K. Jomová *et al.*, “A Switch between Antioxidant and Prooxidant Properties of the Phenolic Compounds Myricetin, Morin, 3′,4′-Dihydroxyflavone, Taxifolin and 4-Hydroxy-Coumarin in the Presence of Copper(II) Ions: A Spectroscopic, Absorption Titration and DNA Damage Study,” *Molecules*, vol. 24, no. 23, p. 4335, Jan. 2019, doi: 10.3390/molecules24234335.

[16] E. Papadaki, M. Z. Tsimidou, and F. T. Mantzouridou, “Changes in Phenolic Compounds and Phytotoxicity of the Spanish-Style Green Olive Processing Wastewaters by Aspergillus niger B60.,” *J Agric Food Chem*, May 2018, doi: 10.1021/acs.jafc.8b00918.

[17] A. Cherkasov *et al.*, “QSAR modeling: where have you been? Where are you going to?,” *Journal of medicinal chemistry*, vol. 57, no. 12, pp. 4977–5010, Jun. 2014, doi: 10.1021/jm4004285.

[18] C. Hansch, A. Jazirehi, S. B. Mekapati, R. Garg, and B. Bonavida, “QSAR of apoptosis induction in various cancer cells,” *Bioorg. Med. Chem.*, vol. 11, no. 13, pp. 3015–3019, Jul. 2003.

[19] C. D. Selassie, S. B. Mekapati, and R. P. Verma, “QSAR: then and now,” *Curr Top Med Chem*, vol. 2, no. 12, pp. 1357–1379, Dec. 2002, doi: 10.2174/1568026023392823.

[20] C. Hansch, P. Maloney, T. Fujita, and R. M. Muir, “Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients,” vol. 194, no. 4824, pp. 178-180., 1962.

[21] Y. K. Peterson, X. S. Wang, P. J. Casey, and A. Tropsha, “The Discovery of Geranylgeranyltransferase-I Inhibitors with Novel Scaffolds by the Means of Quantitative Structure-Activity Relationship Modeling, Virtual Screening, and Experimental Validation,” *J Med Chem*, vol. 52, no. 14, pp. 4210–4220, Jul. 2009, doi: 10.1021/jm8013772.

[22] L. Douali and D. Cherqaoui, “QSAR Studies of Non-Nucleoside Reverse Transcriptase Inhibitors: The Hydrophobic Effect,” *Current Computer - Aided Drug Design*, vol. 2, pp. 21–29, 2006, doi: 10.2174/157340906776056446.

[23] S. A. Hiller, V. E. Golender, A. B. Rosenblit, L. A. Rastrigin, and A. B. Glaz, “Cybernetic methods of drug design. I. Statement of the problem--the perceptron approach,” vol. 6, no. 5, pp. 411–421, 1973.

[24] S. Cohen, “Chapter 2 - The basics of machine learning: strategies and techniques,” in *Artificial Intelligence and Deep Learning in Pathology*, S. Cohen, Ed. Elsevier, 2021, pp. 13–40. doi: 10.1016/B978-0-323-67538-3.00002-6.

[25] L. Douali, D. Villemin, and D. Cherqaoui, “Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives.,” *Curr Pharm Des*, vol. 9, no. 22, pp. 1817–1826, 2003.

[26] Y. Y. Yanga, Z. Ye, Y. Su, Q. Zhao, X. Li, and D. Ouyang, “Deep learning for in vitro prediction of pharmaceutical formulations,” *Acta Pharmaceutica Sinica B*, vol. 9, no. 1, pp. 177–185, 2019.

[27] T. B. Hughes, G. P. Miller, and S. J. Swamidass, “Modeling epoxidation of drug-like molecules with a deep machine learning network,” *Cent. Sci. 1*, pp. 168–180, 2015.

[28] J. Cotterill, N. Price, E. Rorije, and A. Peijnenburg, “Development of a QSAR model to predict hepatic steatosis using freely available machine learning tools,” *Food and Chemical Toxicology*, vol. 142, p. 111494, Aug. 2020, doi: 10.1016/j.fct.2020.111494.

[29] F. Ghasemi, A. Mehridehnavi, A. Fassihi, and H. Pérez-Sánchez, “Deep neural network in QSAR studies using deep belief network,” *Applied Soft Computing*, vol. 62, pp. 251–258, 2018, doi: 10.1016/j.asoc.2017.09.040.

[30] J. A. Castillo-Garit, G. M. Casañola-Martin, S. J. Barigye, H. Pham-The, F. Torrens, and A. Torreblanca, “Machine learning-based models to predict modes of toxic action of phenols to Tetrahymena pyriformis,” *SAR QSAR Environ Res*, vol. 28, no. 9, pp. 735–747, Sep. 2017, doi: 10.1080/1062936X.2017.1376705.

[31] M. T. D. Cronin and T. W. Schultz, “Structure-toxicity relationships for phenols to Tetrahymena pyriformis,” *Chemosphere*, vol. 32, no. 8, pp. 1453–1468, 1996.

[32] C. Selassie and R. P. Verma, “QSAR of toxicology of substituted phenols,” *Journal of Pesticide Science*, vol. 40, no. 1, pp. 1–12, 2015, doi: 10.1584/jpestics.D14-097.

[33] C. D. Selassie *et al.*, “Comparative QSAR and the radical toxicity of various functional groups,” *Chem. Rev.*, vol. 102, no. 7, pp. 2585–2605, Jul. 2002.

[34] I. V. Tetko *et al.*, “Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection,” *J Chem Inf Model*, vol. 48, no. 9, Art. no. 9, Sep. 2008, doi: 10.1021/ci800151m.

[35] M. T. D. Cronin *et al.*, “Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis.,” *Chemosphere*, vol. 49, no. 10, pp. 1201–1221, Dec. 2002.

[36] V. Ruusmann, S. Sild, and U. Maran, “QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models,” *J. Cheminform.*, vol. 7, no. 1, p. 32, Jun. 2015, doi: 10.1186/s13321-015-0082-6.

[37] C. Hansch, S. C. McKarns, C. J. Smith, and D. J. Doolittle, “Comparative QSAR evidence for a free-radical mechanism of phenol-induced toxicity,” *Chem Biol Interact*, vol. 127, no. 1, pp. 61–72, 2000.

[38] C. D. Selassie, T. V. DeSoyza, M. Rosario, H. Gao, and C. Hansch, “Phenol toxicity in leukemia cells: a radical process?,” *Chem. Biol. Interact.*, vol. 113, no. 3, pp. 175–190, Jun. 1998.

[39] L. Douali, D. Villemin, and D. Cherqaoui, “Neural networks: Accurate nonlinear QSAR model for HEPT derivatives.,” *J Chem Inf Comput Sci*, vol. 43, no. 4, pp. 1200–1207, Aug. 2003, doi: 10.1021/ci034047q.

[40] L. B. Kier and L. H. Hall, *Molecular connectivity in chemistry and drug research*. New York: Academic Press, 1976.

[41] L. H. Hall and L. B. Kier, “Issues in representation of molecular structure: The development of molecular connectivity,” *J. Mol. Graph. Model.*, vol. 20, no. 1, pp. 4–18, Dec. 2001, doi: 10.1016/S1093-3263(01)00097-3.

[42] P. Gramatica, N. Chirico, E. Papa, S. Cassani, and S. Kovarich, “QSARINS: A new software for the development, analysis, and validation of QSAR MLR models,” *J. Comput. Chem.*, vol. 34, no. 24, pp. 2121–2132, 2013, doi: 10.1002/jcc.23361.

[43] J. C. McGowan, “Molecular volumes and structural chemistry,” *Recueil des Travaux Chimiques des Pays-Bas*, vol. 75, no. 2, pp. 193–208, 1956, doi: https://doi.org/10.1002/recl.19560750208.

[44] J. J. Stewart, “Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements,” *J. Mol. Model.*, vol. 13, no. 12, pp. 1173–1213, 2007.

[45] J. J. Stewart, “Application of the PM6 method to modeling proteins,” *J. Mol. Model.*, vol. 15, no. 7, pp. 765–805, 2009.

[46] F. Chollet, *Keras*. 2015. [Online]. Available: https://keras.io

[47] S.-C. Huang and T.-H. Le, “Chapter 1 - Introduction to TensorFlow 2,” in *Principles and Labs for Deep Learning*, S.-C. Huang and T.-H. Le, Eds. Academic Press, 2021, pp. 1–26. doi: 10.1016/B978-0-323-90198-7.00014-8.

[48] M. Abadi *et al.*, “TensorFlow: A system for large-scale machine learning,” *Google AI*, 2016. https://ai.google/research/pubs/pub45381 (accessed Sep. 15, 2018).

[49] R. Liu, H. Wang, K. P. Glover, M. G. Feasel, and A. Wallqvist, “Dissecting Machine-Learning Prediction of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure-Activity Relationship Models Based on Deep Neural Networks?,” *J Chem Inf Model*, vol. 59, no. 1, pp. 117–126, Jan. 2019, doi: 10.1021/acs.jcim.8b00348.

[50] S. Fujii, Y. Miyajima, H. Masuno, and H. Kagechika, “Increased Hydrophobicity and Estrogenic Activity of Simple Phenols with Silicon and Germanium-Containing Substituents,” *J. Med. Chem.*, vol. 56, no. 1, pp. 160–166, Jan. 2013, doi: 10.1021/jm3013757.

[51] C. Hansch, K. Kiehs, and G. L. Lawrence, “The Role of Substituents in the Hydrophobic Bonding of Phenols by Serum and Mitochondrial Proteins,” *J. Am. Chem. Soc.*, vol. 87, no. 24, pp. 5770–5773, Dec. 1965, doi: 10.1021/ja00952a044.

[52] M. Sobiesiak, “Chemical Structure of Phenols and Its Consequence for Sorption Processes,” in *Phenolic Compounds*, M. Soto-Hernandez, M. Palma-Tenango, and M. del R. Garcia-Mateos, Eds. Rijeka: IntechOpen, 2017. doi: 10.5772/66537.