Machine learning for the prediction of phenols cytotoxicity

Latifa Douali [a],[1],[*]

[a] Regional Centre of Training and Education (CRMEF) Marrakech-Safi, Marrakech -Morocco, Department of computer sciences
[1] Email l_douali@yahoo.fr*
* corresponding author

ARTICLE INFO

Article history
Received September 23, 2021
Revised October 21, 2021
Accepted March 31, 2021
Available online March 31, 2021

Keywords
Cytotoxicity
Deep neural networks
Phenols
QSAR
Tetrahymena pyriformis
Risk assessment

ABSTRACT

Quantitative structure-activity relationships (QSAR) are relevant techniques that assist biologists and chemists in accelerating the drug design process and help understanding many biological and chemical mechanisms. Using classical statistical methods may affect the accuracy and the reliability of the developed QSAR models. This work aims to use a machine learning approach to establish a QSAR model for phenols cytotoxicity prediction. This issue concern many chemists and biologists. In this investigation, the dataset is diverse, and the cytotoxicity data are sparse. Multi-component description of the compounds has then been considered. A set of molecular descriptors fed the deep neural network (DNN) and served to train the DNN. The established DNN model was able to predict the cytotoxicity of the phenols at high precision. The correlation coefficient at the fitting stage was higher than other statistical methods reported in the literature or developed in the present work, specifically multiple linear regression (MLR) and shallow artificial neural networks (ANN), and was equal to 0.943. The predictive capability of the model, as estimated by the coefficient of determination on an external predictive dataset, was significantly high and was about 0.739. This finding could help implement many molecular descriptors relevant to describing the compounds, representing the effects governing the phenols cytotoxicity toward Tetrahymena pyriformis, avoiding overfitting and outlier exclusion.

1. Introduction

Phenols are chemical compounds that are very abundant in nature and can be synthesized. They are largely used in agriculture, food processing, and many industries. Mainly, they are present in many synthetic products; dye, leather, polymers, pesticides, resin manufacturing, and wood preservatives. They are present in many kinds of cereal, fruits, and vegetables. They are well-known for their beneficial antioxidant effect [1]–[4]. They were suggested suppressing oxidative stress by scavenging peroxy radicals [5], [6]. However, it was also reported that they present fatal environmental and toxicological risks [7]. They are discharged in nature as a waste product of industrial units and cause substantial environmental damage to air, water, and soil [8], [9]. Their toxicity is very high. The Environmental and Protection Agency (EPA) and the European Union (EU) consider the phenols as pollutants at high risk and mandate the removal of phenols from wastewater during the treatment process. It was reported in many instances that some phenols cause serious health problems [10], [11]. They are harmful to humans and animals, even in small amounts [7], [12]–[14]. Besides the skin irritation, liver, and kidney damage that phenols can cause to humans, estrogenic and teratogenic problems may also be evolved [15], [16].

The described perplexing behavior of phenols raises important questions about their toxicity mechanism [5], [9], [17] and there is an excessive need to develop theoretical models describing these processes or models predicting the toxicity accurately. Quantitative Structure-Activity Relationships (QSAR) are relevant techniques that develop theoretical models and help to understand such mechanisms [18], [19]. QSAR models are of big interest to biologists and chemists in many research domains for modeling the relationships between key parameters and endpoints [20]. Endpoints may be a biological activity, toxicity, or physicochemical property. The objective of the development of QSAR models is mainly to shed light on the complex chemical and biological mechanisms and hence to accelerate the drug design process [18], [21], [22]. In toxicology, the QSAR approach is important in investigating the toxicity of new compounds. Cytotoxicity assays may benefit mainly from rigorously established models. While several statistical methods help establish QSAR models, machine learning (ML) proved to be highly pertinent and successful in developing accurate models. They have proven powerful, especially for non-linear optimization problems frequently encountered in biological and chemical processes. Artificial neural networks (ANN), support vector machines (SVM), and random forest (RF) was used as a QSAR model-building in numerous studies and for many chemical compounds.

Deep learning is an ML algorithm based on ANN. It is essentially based on emulating the biological neurons' behavior. Known for their ability to map complex functions, deep neural networks (DNN) represent an ideal tool for establishing non-linear QSAR and building highly predictive models. Using ML and ANN in QSAR has known many advances since its first application by Hiller et al. [23]. The authors used perceptron to classify 1,3-dioxanes as active or inactive concerning their physiological activity. This field has known a considerable growth, and it was developed in a well-established field with new approaches and methodologies. ANN is the most popular non-linear tool used in QSAR development. Genetic, Bayesian, K-nearest neighbors (KNN), and the back-propagation neural networks were used. They proved to be more efficient than other statistical methods and ML models. However, although the advantages ANN present, they suffer from serious problems: Actually, molecular descriptors constitute the inputs of the networks, and in many instances, there is a need to use many of them since they contribute to a better description of the molecules and their effects on the biological activity or property. Often, correlations between them added to the dataset size limit their implementation in the model and reduce the model interpretability and reliability [24]. In previous work, we were brought to determine a ratio ϱ to avoid this problem [25]. Very often, the application of ANN leads to the overfitting problem that significantly affects the model's prediction ability. Much attention was given to these problems, and the typical solutions proposed to deal with these issues were to reduce the number of inputs, which affects the chemical information provided to the model, to use few neurons in the hidden layers, or limit the number of samples used to train the network by removing the samples that the modeling technique cannot appropriately handle.

The immense progress provided by the DNN helps to circumvent these issues [26]–[29]. In fact, by effectively using a multi-level learning strategy, the DNN can process data more accurately. We go hierarchically through the different layers from low-level pattern extraction to a higher level. These different levels correspond to different levels of data abstraction. Processing these extracted patterns leads to high accuracy in detecting meaningful features and making accurate predictions. They are rapidly optimized, and they avoid the overfitting issue from which they suffer the application of shallow neural networks [19], [30], [31]. Using numerous descriptors, large datasets, and multiple hidden layers becomes possible. They now gain significant interest in almost all problems that need information extraction without human intervention and QSAR model development.

It is noteworthy that DNNs are intensively used in classification and transcription problems. Few works use DNN in building regression models [32], [33]. They achieve high performance in predicting biological and pharmaceutical properties [32]–[35]. Many QSAR models studying different datasets of phenols were established [17], [36]–[41]. Most of them have used multiple linear regression (MLR) and led to many linear models. Other models were also developed, namely by MLR and ANN.

In this study, we use DNN to develop a regression-based QSAR model to predict the cytotoxicity of phenols to Tetrahymena pyriformis. For this purpose, we generated several molecular descriptors to serve as network inputs.

2. Method

2.1. Dataset

Phenols are chemical compounds with a structure containing a hydroxy group connected to an aromatic ring. In the present work, we investigate a dataset containing 250 phenolic compounds with different types of substituents on the aromatic ring. Substituents in different positions (ortho, meta, and para) were considered. These compounds were studied earlier by Cronin et al. [42], and are available via the QDB Databank repository [42], [43]. The data contain mono-, bi-, and tri-substituted phenols. In previous work, Selassi *et al.* [39], [44], [45] studied different datasets of phenols. Datasets with electron-releasing and electron-withdrawing substituents were considered separately, and in the present work, both types of substituents were implemented in the same dataset. The structures of phenol and some studied phenolic compounds are represented in Fig. 1.

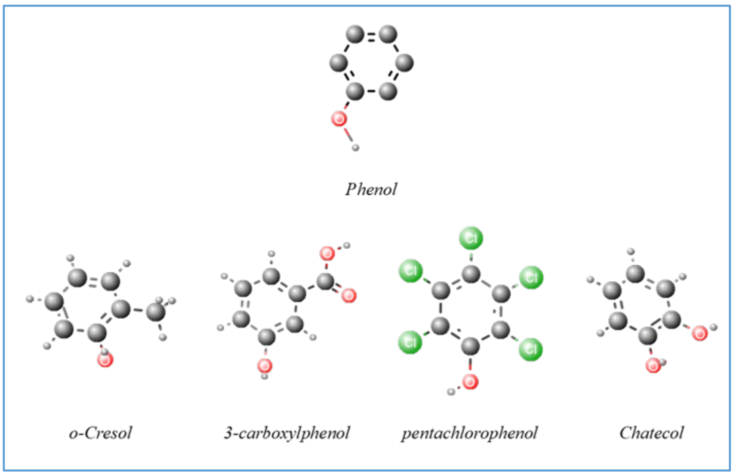

Fig. 1. The general structure of some phenols studies.

IGC50 expresses the cytotoxicity of the 250 phenols to the Tetrahymena pyriformis, the 50% growth inhibitory concentration (mmol/L) of a compound to Tetrahymena pyriformis, regardless of their mode of action (MOA). For calculation conformity, the log values (1/IGC$_{50}$) were considered as endpoints. It is noteworthy that the dataset contains very diversified endpoints. The log values (1/IGC$_{50}$) range from -1.5 to 2.71, with a mean value of 0.739 and a standard deviation of 0.828. The variability of the cytotoxicity in our data is shown in Fig. 2.

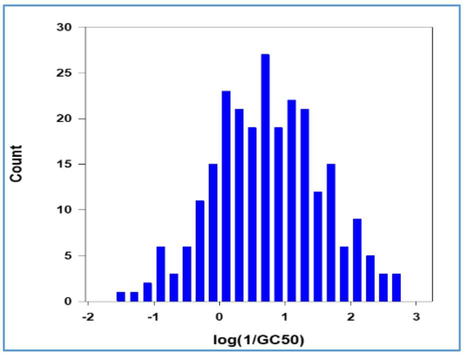

Fig. 2. The distribution of the cytotoxicity endpoints.

To build the model, the dataset was divided into two subsets. We used randomly selected 80% of phenols to perform the fitting stage, and the remaining 20 % served as an external predictive dataset. The datasets were carefully checked to ensure that compounds with electron-withdrawing and electron-releasing substituents were implemented in both subsets.

2.2. Generation of Molecular Descriptors

From our experience, a successful QSAR study relies largely on the set of molecular descriptors used and their ability to depict the biological effect meticulously [25], [46]. There are two levels of molecular description; the first level consists of describing only the substituents, and the second consists of describing the whole molecule. For this investigation, the endpoint values were sparse, and the substituents were significantly diversified, and there was a need to generate unswerving molecular descriptors that represent the features of the whole dataset and characterize its diversity. We choose then to use molecular descriptors that describe the entire molecule.

DNN accepts as inputs a large amount of information as precise as possible to extract features and achieve good results. In fact, a chemical compound is constituted of a connection of atoms of different nature. Each atom has its inherent characteristics, and it is connected to other atoms and so affects the behavior of the whole compound. A small change in this structure induces, in many instances, meaningful changes in the biological behavior of the compound. The molecular connectivity approach based on chemical graph theory [47] provides such precision and information. Many structural descriptors that give details on the molecular connectivity were implemented. They were represented by Kier [47], [48] indices a structural fingerprint [49]. They were generated by the QSARIN software [50]. Actually, fingerprints are numerical values that encode fragments or subgroups in a molecule.

Molecular reactivity depends not only on the graphical representation of a molecule but also on the atoms' intrinsic quantum and physicochemical properties. In addition, molecules interact with biological systems involving three-dimensional dynamic processes and mechanisms. To take into account these characteristics, we implemented molecular descriptors such as the molar refractivity (MR) and the Mc Gowan volume (McVol) [51]. Those descriptors were calculated using the CLogP program. The hydrophobic character of the compounds, which reflects their penetration mechanism through biological systems, is a determining factor. We introduced this characteristic via logP parameter, the octanol-water partition coefficient of the whole compound, calculated using the CLogP program.

The electronic aspect of the substituents was considered, and many electronic descriptors, such as HOMO (Highest Occupied Molecular Orbital) and LUMO (Lower Unoccupied Molecular Orbital), Ip (the ionization potential), Pka (the acid dissociation constant) were implemented. These parameters were calculated after a geometry optimization of the molecules using the PM6 semiempirical quantum method implemented in MOPAC 7 program [52], [53]. This implementation helped to investigate a dataset of phenolic compounds containing both electron-releasing and electron-withdrawing substituents. A total number of 118 molecular descriptors were then generated.

2.3. Generation of Molecular Descriptors

To establish the QSAR model, we developed our own program using the Keras library package [54] with the Tensorflow framework [28], [55]. DNNs are based on ANN concepts with many deep nodes [26]. Many hyperparameters had to be adjusted with attention to the calculation time optimization and the network accuracy. The inputs of the networks were carefully chosen. Although many molecular descriptors were generated, only the relevant ones were considered and served as inputs. A total number of 24 molecular descriptors were then implemented. The robustness of DNN lies in its non-linear units that process data features at one level into feature data at a higher level. The hyperbolic tangent function was used as the activation function of the hidden nodes. The stochastic gradient descent optimizer was adopted, and the learning rate was 0.01 [35] to update the network parameters. One node in the output layer represented the target activity, log(1/IGC50). The constructed networks were trained for 10000 epochs.

Two regression models were also developed using MLR and ANN methods to compare with DNN. The ANN was constructed with one hidden layer. The learning rate was set to 0.1.

3. Results and Discussion

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 assess the accuracy of the established models, two metrics were adopted; the statistical root-mean-square deviation (RMSD) and the correlation coefficient R². The analytical formulae of these parameters are given in (1) and (2) below.

$$RMSD = \sqrt{\frac{\sum_{t=1}^{t=N}(\hat{T}_t - T_t)}{N}} \tag{1}$$

$$R^2 = \frac{\sum_t (T_t - \hat{T})}{\sum_t (T_t - \bar{T}_t)} \tag{2}$$

where $\hat{T}$ is the calculated cytotoxicity, $\overline{T}$ is the mean of the observed cytotoxicity, and N is the number of studied compounds. A correlation coefficient and an RMSD as close as possible to 1 and 0, respectively, are anticipated. A developed MLR model implementing all molecular descriptors resulted in an R² of 0.67 and an RMSD of 0.51. The linear model generated one outlier. The established ANN resulted in a relatively high R² of 0.74 and an RMSD of 0.60 with ten nodes in the hidden layer with the sigmoid activation function. Adding more nodes to the hidden layer pushes the network into overfitting.

For the DNN model, the best achievement, revealed by the values of the statistical metrics R² and RMSD, was obtained by a DNN structure with two layers containing 20 and 14 nodes, respectively; While the correlation coefficient is high and it amounts 0.943, the RMSD is very low, and it amounts 0.194. It is significantly better than the results of the MLR and ANN models. Furthermore, the problem of overfitting was avoided. The previous works reported in the literature had to remove any outliers. However, no outliers were detected in the present study. Much information can be deduced from these compounds, and removing them from the model will lead to poor performance.

The metrics mentioned above can infer that the DNN outperformed MLR and ANN. It accomplished a perfect fit of the data. It extracted 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. 3. It represents the cytotoxicity values calculated by the DNN model versus the experimental values.

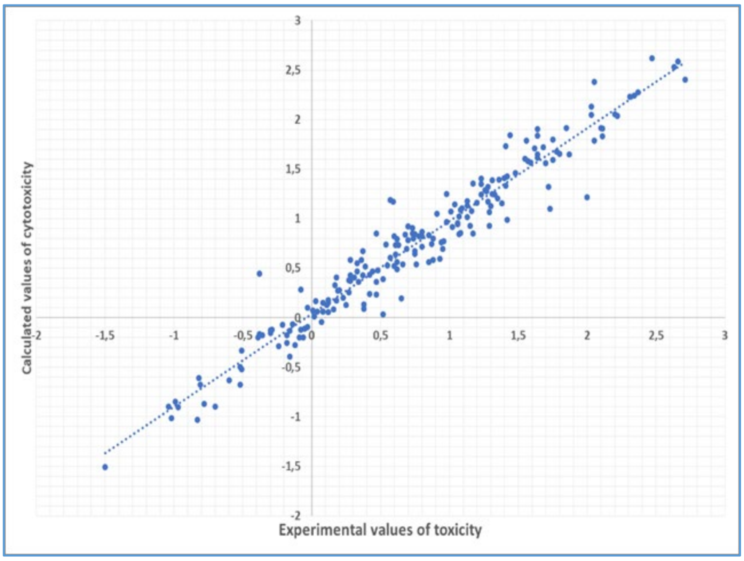

Fig. 3. Calculated versus experimental cytotoxicity of phenols.

After the fulfilling results of the fitting stage, we had to validate the model and ensure that the network built could process new data and predict the cytotoxicity of new compounds. A testing stage was then performed using a Leave-One-Out procedure. The promising results led to a leave-one-out cross-validated R^2 (q^2) of 0.941 and an RMSD of 0.203. These results led to adopting the constructed model to predict new activities.

In this investigation, our main focus was the model's predictive accuracy. Fifty phenolic compounds were used as an external prediction set. We ensured that this dataset comprised new phenolic compounds structurally close to the compounds in the fitting stage (Fig. 2), and the network had never been seen. To assess the DNN prediction capability, we used the coefficient of determination and the root-mean-square error of prediction (RMSEP) metrics. This stage resulted in a high R^2 that equals 0.739, and an RMSEP equals 0.434. The predicted values versus the valid cytotoxicity values are reported in Fig. 4. It shows that the established DNN model can predict the cytotoxicity of new phenols with high accuracy.

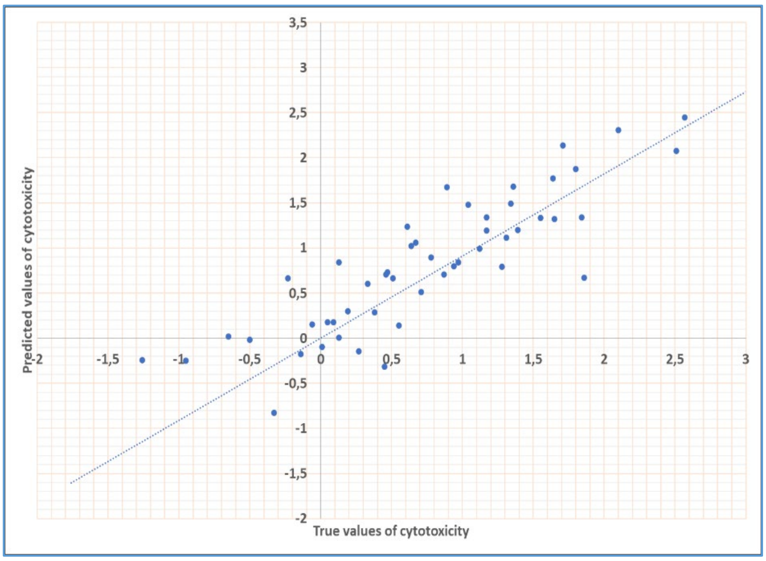

Fig. 4. Calculated versus experimental cytotoxicity of phenols.

Moreover, a residual analysis has been carried out to assess our model. It led to a closer examination of the established regression-model quality and its capability to predict the cytotoxicity of new phenols to Tetrahymena pyriformis. This analysis offers the opportunity to investigate each error made on the target outputs. The results are depicted in Fig. 5.

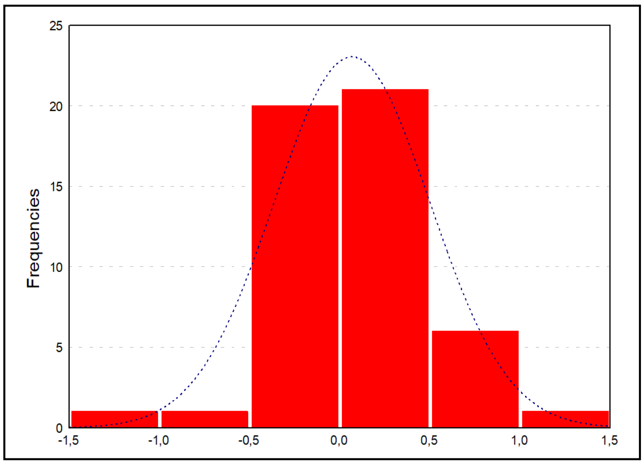

Fig. 5. The DNN generates prediction errors.

A gaussian-like distribution of the prediction residuals was obtained. As shown in the histogram of Fig. 5, most of the errors made on the predicted values of cytotoxicity were between -0.5 and +0.5. This proves the accuracy and the validity of the model and that the choice of the model is appropriate.

Outliers, the samples from the dataset that a developed model cannot fit correctly, represent a big challenge for QSAR modeling, yet the success of the QSAR model relies on its ability to predict the endpoints of new compounds. Removal of these prototypes is assumed to improve the accuracy of the model.

Interestingly, for almost all the established QSAR models for phenols cytotoxicity reported in the literature [39], [42] using other statistical methods (MLR, partial least squares (PLS)), there was a necessity to exclude large numbers of outliers to obtain good fitting models. Especially in [42], up to 80 outliers were excluded. Though, the fitting ability of the developed models was modest compared to the present DNN-developed model. The highest correlation coefficient reported in his work was 0.83. Models established with 200 prototypes led to poor statistical fit, and the correlation coefficients did not exceed 0.69. The big question was to determine the reason behind the existence of these outliers. As Maggiora suggested [56], this might be due to the existence of activity cliffs, which are samples with small differences in chemical structure that exhibit dramatic changes in the target activity. A tiny change in the structure causes a considerable effect on the target activity. Nevertheless, highly precise methods, such as DNN, may detect the subtle nuances and succeed to extract the features to determine precisely the endpoints. Furthermore, molecular recognition is a key ingredient to determine the biological activity. The implication of structural description along with physicochemical description of the compounds disclose tiny valuable details on the activity variability.

Similarly, for the prediction stage, outliers had to be excluded. Up to 4 outliers were prone to be excluded from all models established in [42]. A direct explanation is that the model could not predict at least these four compounds, which calls the model's reliability into question. For the model developed by the DNN, no outlier was excluded. The DNN could fit all the prototypes. The selected molecular descriptors provided the DNN with precise information. The DNN successfully extracted the features required to predict the cytotoxicity accurately.

A DNN model was successfully developed to predict the phenol's cytotoxicity to Tetrahymena pyriformis in the present study. 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 all the proposed new compounds accurately. Both the fitting quality and the predictability accuracy were significantly high. Two main factors played a part in this success: 1) the use of DNN ensuring an automatic feature extraction capability and non-linear transformation functions involved in learning chemical patterns, and 2) the multi-component representation of the inputs. The present description of the compounds consisted of parameters that precisely described the molecular graphs of the compounds and simulated the molecules' spatial images. Parameters that described the physicochemical characteristics of the molecules were caused mainly by the different substituents on the aromatic ring. This provided the networks with informative chemical features. Indeed, DNN excel in image recognition. Considering that a molecule is far from a static image/graph, we proposed adding physicochemical parameters. They convey valuable information on the intrinsic features of a molecule. Specifically, electronic and hydrophobic characteristics play a key role in a compound-biological system interaction. For instance, electron-releasing and electron-withdrawing groups affect the reactivity of a phenolic compound differently and hence affect their biological responses, and the hydrophobic character manages the penetration of a chemical compound into the biological system. The parameters provided to the networks should be sufficiently precise and diverse to ensure reliable predictions.

4. Conclusion

The research 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 a 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 in evolving animal-free assays.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] G. W. Burton, Y. Le Page, E. J. Gabe, and K. U. Ingold, "Antioxidant activity of vitamin E and related phenols. Importance of stereoelectronic factors," *J. Am. Chem. Soc.*, vol. 102, no. 26, pp. 7791–7792, Dec. 1980, doi: 10.1021/ja00546a032.

[2] W. M. El-Husseiny, M. A.-A. El-Sayed, N. I. Abdel-Aziz, A. S. El-Azab, E. R. Ahmed, and A. A.-M. Abdel-Aziz, "Synthesis, antitumour and antioxidant activities of novel α,β-unsaturated ketones and related heterocyclic analogues: EGFR inhibition and molecular modelling study," *J. Enzyme Inhib. Med. Chem.*, vol. 33, no. 1, pp. 507–518, Jan. 2018, doi: 10.1080/14756366.2018.1434519.

[3] 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.

[4] 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.

[5] 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, Nov. 2019, doi: 10.3390/molecules24234335.

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

[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. W. Anku, M. A. Mamo, and P. P. Govender, "Phenolic Compounds in Water: Sources, Reactivity, Toxicity and Treatment Methods," in *Phenolic Compounds - Natural Sources, Importance and Applications*, InTech, 2017, doi: 10.5772/66927.

[9] 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.*, vol. 66, no. 19, pp. 4891–4901, May 2018, doi: 10.1021/acs.jafc.8b00918.

[10] M. Khoshnamvand, Z. Hao, O. O. Fadare, P. Hanachi, Y. Chen, and J. Liu, "Toxicity of biosynthesized silver nanoparticles to aquatic organisms of different trophic levels," *Chemosphere*, vol. 258, p. 127346, Nov. 2020, doi: 10.1016/j.chemosphere.2020.127346.

[11] 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.

[12] 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, Jan. 2011, doi: 10.1080/1062936X.2010.528449.

[13] 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.

[14] 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," *Environ. Res.*, vol. 201, p. 111531, Oct. 2021, doi: 10.1016/j.envres.2021.111531.

[15] J. Moreman, O. Lee, M. Trznadel, A. David, T. Kudoh, and C. R. Tyler, "Acute Toxicity, Teratogenic, and Estrogenic Effects of Bisphenol A and Its Alternative Replacements Bisphenol S, Bisphenol F, and Bisphenol AF in Zebrafish Embryo-Larvae," *Environ. Sci. Technol.*, vol. 51, no. 21, pp. 12796–12805, Nov. 2017, doi: 10.1021/acs.est.7b03283.

[16] R. Chianese *et al.*, "Bisphenol A in Reproduction: Epigenetic Effects," *Curr. Med. Chem.*, vol. 25, no. 6, pp. 748–770, Feb. 2018, doi: 10.2174/0929867324666171009121001.

[17] 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, Jan. 2001, doi: 10.1002/1098-1128(200101)21:1<73::AID-MED3>3.0.CO;2-5.

[18] L. Douali and D. Cherqaoui, "QSAR Studies of Non-Nucleoside Reverse Transcriptase Inhibitors: The Hydrophobic Effect," *Curr. Comput. Aided-Drug Des.*, vol. 2, no. 1, pp. 21–29, Mar. 2006, doi: 10.2174/157340906776056446.

[19] F. Ghasemi, A. Mehridehnavi, A. Pérez-Garrido, and H. Pérez-Sánchez, "Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks," *Drug Discov. Today*, vol. 23, no. 10, pp. 1784–1790, Oct. 2018, doi: 10.1016/j.drudis.2018.06.016.

[20] C. HANSCH, P. P. MALONEY, T. FUJITA, and R. M. MUIR, "Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients," *Nature*, vol. 194, no. 4824, pp. 178–180, Apr. 1962, doi: 10.1038/194178b0.

[21] C. Hansch, J. P. Björkroth, and A. Leo, "Hydrophobicity and Central Nervous System Agents: On the Principle of Minimal Hydrophobicity in Drug Design," *J. Pharm. Sci.*, vol. 76, no. 9, pp. 663–687, Sep. 1987, doi: 10.1002/jps.2600760902.

[22] C. Hansch, D. Hoekman, A. Leo, D. Weininger, and C. D. Selassie, "Chem-Bioinformatics: Comparative QSAR at the Interface between Chemistry and Biology," *Chem. Rev.*, vol. 102, no. 3, pp. 783–812, Mar. 2002, doi: 10.1021/cr0102009.

[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," *Comput. Biomed. Res.*, vol. 6, no. 5, pp. 411–421, Oct. 1973, doi: 10.1016/0010-4809(73)90074-8.

[24] A. Cherkasov *et al.*, “QSAR Modeling: Where Have You Been? Where Are You Going To?," *J. Med. Chem.*, vol. 57, no. 12, pp. 4977–5010, Jun. 2014, doi: 10.1021/jm4004285.

[25] 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, Jul. 2003, doi: 10.1021/ci034047q.

[26] G. Gini, F. Zanoli, A. Gamba, G. Raitano, and E. Benfenati, "Could deep learning in neural networks improve the QSAR models?," *SAR QSAR Environ. Res.*, vol. 30, no. 9, pp. 617–642, Sep. 2019, doi: 10.1080/1062936X.2019.1650827.

[27] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[28] S.-C. Huang and T.-H. Le, "Introduction to TensorFlow 2," *Princ. Labs Deep Learn.*, pp. 1–26, 2021, doi: 10.1016/B978-0-323-90198-7.00014-8.

[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[30] G. B. Goh, N. O. Hodas, and A. Vishnu, "Deep learning for computational chemistry," *J. Comput. Chem.*, vol. 38, no. 16, pp. 1291–1307, Jun. 2017, doi: 10.1002/jcc.24764.

[31] S. Cohen, "The basics of machine learning: strategies and techniques," *Artif. Intell. Deep Learn. Pathol.*, pp. 13–40, 2021, doi: 10.1016/B978-0-323-67538-3.00002-6.

[32] Y. Yang, Z. Ye, Y. Su, Q. Zhao, X. Li, and D. Ouyang, "Deep learning for in vitro prediction of pharmaceutical formulations," *Acta Pharm. Sin. B*, vol. 9, no. 1, pp. 177–185, Jan. 2019, doi: 10.1016/j.apsb.2018.09.010.

[33] T. B. Hughes, G. P. Miller, and S. J. Swamidass, "Modeling Epoxidation of Drug-like Molecules with a Deep Machine Learning Network," *ACS Cent. Sci.*, vol. 1, no. 4, pp. 168–180, Jul. 2015, doi: 10.1021/acscentsci.5b00131.

[34] 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 Chem. Toxicol.*, vol. 142, p. 111494, Aug. 2020, doi: 10.1016/j.fct.2020.111494.

[35] F. Ghasemi, A. Mehridehnavi, A. Fassihi, and H. Pérez-Sánchez, "Deep neural network in QSAR studies using deep belief network," *Appl. Soft Comput.*, vol. 62, pp. 251–258, Jan. 2018, doi: 10.1016/j.asoc.2017.09.040.

[36] M. T. D. Cronin and T. W. Schultz, "Structure-toxicity relationships for phenols to Tetrahymena pyriformis," *Chemosphere*, vol. 32, no. 8, pp. 1453–1468, Apr. 1996, doi: 10.1016/0045-6535(96)00054-9.

[37] 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.

[38] 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, doi: 10.1016/S0968-0896(03)00184-6.

[39] C. Selassie and R. P. Verma, "QSAR of toxicology of substituted phenols," *J. Pestic. Sci.*, vol. 40, no. 1, pp. 1–12, 2015, doi: 10.1584/jpestics.D14-097.

[40] C. D. Selassie *et al.*, "Comparative QSAR and the Radical Toxicity of Various Functional Groups," *Chem. Rev.*, vol. 102, no. 7, pp. 2585–2606, Jul. 2002, doi: 10.1021/cr940024m.

[41] 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, pp. 1733–1746, Sep. 2008, doi: 10.1021/ci800151m.

[42] M. T. . 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, doi: 10.1016/S0045-6535(02)00508-8.

[43] 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, Dec. 2015, doi: 10.1186/s13321-015-0082-6.

[44] 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, Jun. 2000, doi: 10.1016/S0009-2797(00)00171-X.

[45] 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, doi: 10.1016/S0009-2797(98)00027-1.

[46] 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, Aug. 2003, doi: 10.2174/1381612033454423.

[47] L. B. Kier and L. H. Hal, *Molecular Connectivity in Chemistry and Drug Research*. New York: Academic Press, 1976. Available: https://www.elsevier.com/books/molecular-connectivity-in-chemistry-and-drug-research/kier/978-0-12-406560-4.

[48] L. B. Kier and L. . Hall, "An Electrotopological-State Index for Atoms in Molecules," *Pharm. Res.*, vol. 7, pp. 801–807, 1990, doi: 10.1023/A:1015952613760.

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

[50] 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, Sep. 2013, doi: 10.1002/jcc.23361.

[51] J. C. McGowan, "Molecular volumes and structural chemistry," *Recl. des Trav. Chim. des Pays-Bas*, vol. 75, no. 2, pp. 193–208, Sep. 2010, doi: 10.1002/recl.19560750208.

[52] J. J. P. 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, Dec. 2007, doi: 10.1007/s00894-007-0233-4.

[53] J. J. P. Stewart, "Application of the PM6 method to modeling proteins," *J. Mol. Model.*, vol. 15, no. 7, pp. 765–805, Jul. 2009, doi: 10.1007/s00894-008-0420-y.

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

[55] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283, [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

[56] G. M. Maggiora, "On Outliers and Activity Cliffs - Why QSAR Often Disappoints," *J. Chem. Inf. Model.*, vol. 46, no. 4, pp. 1535–1535, Jul. 2006, doi: 10.1021/ci060117s.