INTERPRETABLE QSAR MODEL FOR HEALTH RISK ASSESSMENT OF HAZARDOUS CHEMICAL BASED ON STRUCTURE-TO-TOXICITY TRANSFORMER

SangYoun Kim, Shahzeb Tariq, SungKu Heo, ChanHyeok Jeong, MinHyeok Shin,
TaeYong Woo, and ChangKyoo Yoo[*]
Integrated Engineering Major, Dept. of Environmental Science and Engineering,
College of Engineering, Kyung Hee University, Seocheon-dong 1, Giheung-gu,
Yongin-Si, Gyeonggi-Do 446-701, Republic of Korea

Abstract

After the twentieth century, over a million chemicals were utilized without testing their toxicity. Quantitative Structure–Activity Relationships (QSAR), one of the *in silico* testing, can reduce time and cost comparing with traditional *in vivo* testing. In recent years, advances in machine learning have enabled the analysis of a broad variety of molecular descriptors for QSAR studies. However, conventional molecular descriptors based QSAR models are unable to directly exploit the three-dimensional information included in a molecule's structure. Additionally, the machine learning algorithm has a disadvantage of less interpretability. In this context, this study proposes a development of a Structure-to-Toxicity (S2T)-transformer QSAR model based on Simplified Molecular Input Line Entry System (SMILES) to predict toxicity index of PCBs. S2T transformer trained directly SMILES which indicates the molecular structure of PCBs. Moreover, attention mechanism in transformer architecture allows S2T transformer to interpret the correlation between molecular structure and K_{OW} to overcome the disadvantage of machine learning algorithm. The proposed SMILES based S2T-transformer QSAR model indicated the higher predictive performance of 0.83 R^2 score. Also, S2T-transformer QSAR model interpreted the atom contribution of the molecular structure by attention weights in the transformer architecture.

Keywords

Quantitative structure-activity relationships (QSAR), Transformer architecture, Attention mechanism.

Introduction

After the turn of the twentieth century, over a million chemical compounds exhibiting unique physiochemical properties were utilized to suit the demands of consumer goods, food production, and the pharmaceutical industry (Judson et al., 2009). Additionally, more than 190 million chemical compounds have been registered by Chemical Abstract Services (CAS) as of July 1, 2022. It is evident that as the worldwide market for newly produced chemical goods expands, the hazardous risks to human health and ecosystems will rise. Comparatively, the chemical compounds that have been identified, only a tiny fraction have been tested for their toxicity (Judson et al., 2009). This data gap introduces a great concern to human society as consumers are exposed to thousands of chemicals via various routes (Judson et al., 2009). Therefore, European commission has urged to find innovative ways to provide safe and sustainable chemicals for toxic free environment (Tang et al., 2018).

[*] ckyoo@khu.ac.kr

The National Research Council of the United States (NRC) proposes *in silico* testing as an alternative to animal testing for toxicity assessment (Tang et al., 2018). *In silico* testing is predicated on the idea that the toxicity of a chemical substance is determined by its intrinsic property (Rim, 2020). Consequently, *in silico* testing have become the most prevalent method for providing meaningful and reliable toxicity data (Rim, 2020). In comparison to *in vivo* procedures, these methods are advantageous since they reduce time and money and give vital insight into the mechanisms that cause toxicity.

Among *in silico* approaches, Puzyn et al., (2011) proposed Quantitative Structure–Activity Relationships (QSAR) to predict the toxicity properties of chemical compounds (Puzyn et al., 2011). A QSAR model can estimate the toxicity of novel compounds based on their molecular structure and the toxicity of known chemicals with comparable molecular structures (Tang et al., 2018). In recent years, advances in machine learning have enabled the analysis of a broad variety of chemical descriptors for QSAR studies (Tang et al., 2018). These QSAR models may be used for chemical risk assessment, chemical property prediction, drug development, and design of novel chemicals.

Huang et al., (2021) developed the QSAR model for both the octanol/water partition coefficient (K_{OW}) and LC50 using MLR. This research investigated which substructures had a strong relationship with two hazardous characteristics. Heo et al., (2019) developed prediction and classification model for the sex-hormone binding globulin (SHBG) and estrogen receptor (ER) simultaneously using deep neural network (DNN). However, conventional QSAR algorithms are unable to directly exploit the three-dimensional information included in a molecule's structure.

Alternatively, researchers have proposed the utilization of simplified molecular input line entry system (SMILES) string instead of the conventional molecular descriptors. Sabando et al., (2022) predicted several physicochemical properties and compared the model performance of SMILES string based QSAR and the model performance of the conventional molecular descriptor based QSAR using bidirectional long short-term memory (Bi-LSTM). Chen et al., (2020) identified the compound-protein interaction (CPI) and highlighted the import interacting regions of protein sequences and compound atoms.

According to literature review, QSAR studies can be divided into two groups i.e., based on molecular descriptor and SMILES. To summarize, in the first method toxicological information is estimated using molecular descriptors, physicochemical properties, and thermodynamic properties. While on the other side, the SMILES string is employed, which depicts the molecular structure in a single text using alphabet, number, and parenthesis. These studies either predicted the toxicity qualities or assisted in the identification of new compounds via the use of linear or nonlinear algorithms by constructing the QSAR model.

Although the transformer technique, which is a state-of-the-art approach in machine learning, was used in the majority of prior research based on SMILES strings, only one study employed Bi-LSTM. These studies proposed a variation of the original transformer algorithm to compensate for the SMILES string. The highlight of these investigations, however, is the examination of attention mechanisms by which non-linear attention weights establish a clear link between the structure and target property of a chemical molecule. Additionally, the attention mechanism enables us to overcome the disadvantage of AI algorithms that have less interpretability.

In this context, this study proposes a development of a Structure-to-Toxicity (S2T)-transformer QSAR model based on SMILES to predict toxicity index of PCBs. The results will be compared with the transformer-QSAR model based on conventional molecular descriptors.

Materials and methods

Proposed methodology

Figure 1 presents the proposed framework for developing S2T-transformer QSAR model. The development of S2T-transformer QSAR model can be divided into four main stages: 1) the molecular structure information and toxicity indices of the PCBs were collected to predict the toxicity property. 2) the representation of molecular structure was converted into molecular descriptors and SMILES. 3) the S2T-transformer QSAR model was developed to predict the toxicity index. The performances were evaluated. 4) SMILES based S2T-transformer QSAR model can provide the attention contribution which indicates significant part of the molecular structure to toxicity indices. This analysis is a crucial process to predict the toxicity properties of unknown chemicals.

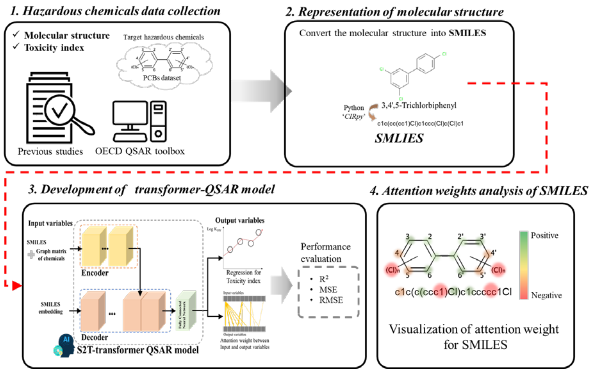

1. Hazardous chemicals data collection
✓ Molecular structure
✓ Toxicity index
Target hazardous chemicals
PCBs dataset
Previous studies
OECD QSAR toolbox
2. Representation of molecular structure
Convert the molecular structure into SMILES
Python 'CIRpy'
3,4',5-Trichlorbiphenyl
c1c(cc(cc1)Cl)c1ccc(Cl)c(Cl)c1
SMLIES
3. Development of transformer-QSAR model
Input variables
SMILES
Graph matrix of chemicals
SMILES embedding
Encoder
Decoder
Fully Connected Neural Network
S2T-transformer QSAR model
Output variables
Log K_OW
Regression for Toxicity index
Input variables
Output variables
Attention weight between Input and output variables
Performance evaluation
R²
MSE
RMSE
4. Attention weights analysis of SMILES
3
2
2'
3'
4
4'
(Cl)n
5
6
6'
5'
(Cl)n
c1c(c(ccc1)Cl)c1ccccc1Cl
Positive
Negative
Visualization of attention weight for SMILES

Figure 1. Schematic representation of proposed framework for developing transformer-QSAR model based on SMILES combined with molecular descriptors

Polychlorinated biphenyl (PCBs)

PCBs are persistent organic pollutants which have high toxicity, environmental persistence, and bioaccumulation. PCBs are comprised by 209 structurally similar compounds varied by location of chlorine atom on the biphenyl ring. In general, the location of chlorine atom relates the toxic activity of the compound (D. Kim et al., 2016). The PCB dataset containing the toxicity index i.e. K_{OW} was collected from Organization for Economic Co-operation and Development (OECD) QSAR Toolbox and related studies (D. Kim et al., 2016).

The partition coefficients K_{OW} indicate the lipophilicity of a PCB, which is defined as the ratio of its concentration in the octanol phase to its concentration in the aqueous phase at equilibrium. While the log K_{OW} shows the likelihood of a substance to MOVE from the aqueous phase to the lipid phase (Zhu et al., 2022). The number of PCBs which have K_{OW} was 139.

Representation of molecular structure

Molecular descriptors

Molecular descriptors are mathematical representation of a chemical that encode significant structural features and physicochemical properties of chemicals (Todeschini & Consonni, 2010). Molecular descriptors of PCBs were employed to predict and to classify the toxicity indices. DRAGON 6 software was utilized to calculate molecular descriptors of the PCBs. The 4,885 molecular descriptors in all 29 different groups were employed to describe the structural diversity of chemicals (Mauri et al., 2006). Then, for pre-processed, the molecular descriptors were normalized. Afterward, not a number and constant molecular descriptors which had standard deviations of less than 0.01 were excluded.

Key molecular descriptors selection

Only the 139 number of K_{OW} of PCBs were collected while the dimension of molecular descriptors was 4,885. The greater number of molecular descriptors than the number of chemicals can cause overfitting problems (J. Huang & Fan, 2013). Previous study demonstrated that key molecular descriptors selection process could handle the overfitting problem and improve the QSAR model performance (Algamal et al., 2015). In this study, two key molecular descriptor selection methods are employed to prevent overfitting problem and to improve the predictive performance of QSAR modeling (Algamal et al., 2015). The variable importance in partial least square projection (VIP) and the elastic-net regularization are applied after the molecular descriptors are pre-processed.

VIP represents the importance of variables according to weights of output variables (Mehmood et al., 2012). The VIP score can be formulated as Eq. (1).

$$VIP_j = \sqrt{\frac{p\sum_{k=1}^{l}(q_k^2 t_k^{'} t_k)(w_{jk}/\|w_{jk}\|^2)}{\sum_{k=1}^{l}(q_k^2 t_k^{'} t_k)}} \tag{1}$$

where p is the number of input variables, w is the loading weights, l is the number of selected latent variables. $q_k^2 t_k^{'} t_k$ represents the variance explained by each component. If VIP scores of input variables is larger than 1, the input variables are selected as the important variable. Previous studies suggested the threshold between 0.83 and 1.21 can yield more relevant variables (Mehmood et al., 2012). Therefore, input variables which have VIP score more than 1.21 were selected as important variables.

The elastic-net regularization optimizes the β with a penalized least squares method, as shown in Eq. (2).

$$\hat{\beta} = \arg\min_{\beta} |\mathrm{y} - \mathrm{X}\beta|^2$$
$$subject\ to\ (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leq \mathrm{s\ for\ some\ s} \tag{2}$$

where α is equal to $\lambda_2/(\lambda_1 + \lambda_2)$. The function $(1 - \alpha)|\beta|_1 + \alpha|\beta|^2$ is the elastic-net penalty, which is convex combination of the lasso and ridge penalty. The lasso penalty regularizes the coefficients of unimportant variables to be zero and the ridge penalty regularizes the coefficients of inter-correlated variables to be zero (Ogutu et al., 2012). In this manner, the molecular descriptors which have the coefficients of less than 0.01 were excluded. The key molecular descriptors were employed as input variables for S2T-transformer QSAR model.

Simplified molecular input line entry system (SMILES)

Simplified molecular input line entry system (SMILES) represents molecules structure as a one-dimensional text (Weininger, 1988). An atom is described by the alphabet of the element symbol, and a bond is represented by a single bond '-', and double bonds '='. The branches in molecular structure are specified by parentheses '()'. The aromatic rings are represented by numerical order in each ring. The atoms present in the aromatic ring are written in lowercase, while the others are capitalized (H. Kim et al., 2021). The encoding rule of SMILES is graphically shown in Figure 2.

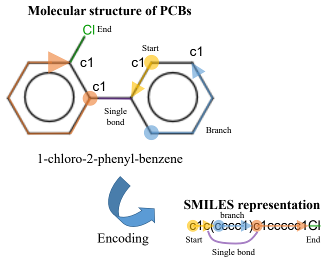


Figure 2. The graphical diagram of SMILES rule to represent the molecular structure of PCB: 1-chloro-2-phenyl-benzene

SMILES contains the information to describe the 3-D molecular structure of the target chemicals. Also, each symbol of SMILES has correlation between the molecular descriptors, it improves on model performance by greater speed and better usage of computational costs (Weininger, 1988). It is very useful language for QSAR modeling as a one-SMILES is including one exact molecular structure of a chemical. Comparing to the molecular descriptor, SMILES can avoid overfitting issues due to less number of features. Therefore, SMILES overcomes the drawback of the molecular descriptor which requires feature selection process. Also, SMILES is one-dimensional text, therefore, it can be easily handled by state-of-art machine learning technique for natural language processing such as attention mechanism. The CAS registry number of PCBs was converted into SMILES using 'CIRpy' library in python. The SMILES was tokenized and employed as input variables for S2T-transformer QSAR model.

Development transformer based QSAR model

QSAR model is the mathematical formulation of the correlationship between the molecular structure and the measured activities of the chemical compounds, as shown in Eq. (3) (D. Kim et al., 2016).

$$\text{Properties} = f(MolecularStructure) + E_r \qquad (3)$$

where, f represents the function between the molecular structure and the activity of chemicals, E_r represents an error between the predicted value and the measured value. The activities of a chemical can be calculated by f and E_r. By selecting the notation of molecular structure the procedure to develop QSAR model can be distinguished. In this study, molecular descriptors and SMILES were employed to represent the molecular structure. Depending on the definition of the function f, various QSAR models can be developed (D. Kim et al., 2016). QSAR model can predict the activities of unknown chemicals utilizing its molecular structure.

Transformer model and attention mechanism

In this study, transformer model was employed as the function of QSAR. The transformer model was proposed by (Vaswani et al., 2017), which was originally developed for natural language processing (NLP) tasks. Transformer is an autoencoder model consisting of only multi-head attention layer not recurrent network or convolutional network to solve NLP tasks. Recently, transformer model indicated high performance in NLP tasks, and many novel models have been established.

The key technique in the transformer framework is multi headed self-attention layer. A multi headed self-attention layer consists of several scaled dot attention layers to extract interaction information between the encoder and the decoder. The scaled dot attention layer takes three inputs, the keys, K, the values, V and the queries, Q, and calculates the attention as Eq. (4) (Vaswani et al., 2017):

$$attention(Q,K,V) = softmmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (4)$$

where d_k is a scaling factor depending on the layer size. A multi headed self-attention can be formulated by parallel of scaled dot attention layers as Eq. (5) (Vaswani et al., 2017).

$$\begin{aligned} Multihead(Q,K,V) &= Concat(head_1, \ldots, head_h)W^O \\ head_i &= attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \qquad (5)$$

where W^O, W_i^Q, W_i^K, and W_i^V are the parameter matrices. A multi headed self-attention allows the transformer to focus on some important features from the input variables dynamically, which directly captures the interaction features of the given two sequences (Vaswani et al., 2017). In addition, the original transformer was designed to solve sequence prediction tasks and utilized mask operation to cover the downstream context of each word. whereas, this study modified the mask operation of the decoder to ensure that the model is accessible to whole sequence of SMILES, which is one of the most crucial modification to transformer architecture (Chen et al., 2020).

In this study, Transformer-CPI (Compound-protein interaction) model which is proposed by Chen et al., (2020) was employed to develop S2T-transformer QSAR model, as shown in Figure 3. Transformer-CPI model replaced the conventional self-attention layers in the encoder part with a gated convolutional network (GCN). The encoder of transformer-CPI was modified to extract hidden features of chemicals and to provide SMILES of the chemicals as encoder output. As well as binary classification task of transformer-CPI was modified to predict toxicity properties as regression task.

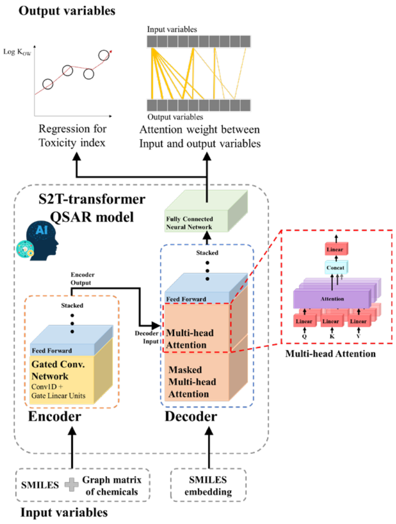


Figure 3. The architecture of the S2T-transformer QSAR model

In order to evaluate the predictive performance of S2T-transformer QSAR models, three statistical parameters were employed, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE).

Results and discussion

Key molecular descriptors for PCBs

VIP in partial least square and elastic-net regularization were utilized sequentially to select the key molecular descriptors. The 92 molecular descriptors in 12 groups were selected by VIP score and elastic-net regularization as shown in Table 1. The relative ratio of groups and the coefficients were indicated in Figure 4. 2D matrix-based descriptors, Functional group descriptors, and 2D Atom Pairs groups had high portion of key molecular descriptors. MPC10 descriptor, which is 2D matrix-based descriptors, had the highest coefficient value as 1.22.

Table 1. Categories of key molecular descriptors for log K_{OW} of PCBs

Group	Original	Selected number
Topological indices	75	2
2D matrix-based descriptors	550	26
3D autocorrelations	80	3
GETAWAY descriptors	273	4
Functional group counts	154	19
Atom-type E-state indices	170	2
2D Atom Pairs	1596	33
2D autocorrelations	213	1
WHIM descriptors	114	1
RDF descriptors	210	1
Geometrical descriptors	38	0
3D-MoRES descriptors	224	0
Total	4885	92

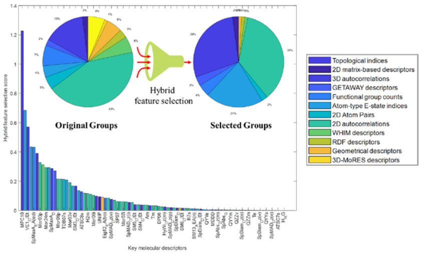


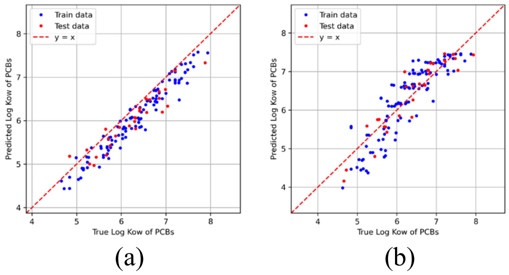
Figure 4. Coefficient of key molecular descriptors for log K_{OW} by VIP and elastic-net regularization

Prediction performance of S2T-transformer model

Table 2 indicates the predictive performance of key molecular descriptors based and SMILES based S2T-transformer QSAR model on the test dataset. SMILES based S2T-transformer QSAR model had the higher predictive performance of 0.83 in terms of R^2. Predictive results of key molecular descriptors based S2T-transformer QSAR model had been biased compared to predictive results of SMILES based S2T-transformer QSAR model, as shown in Figure 5.

Table 2. The predictive performance of key molecular descriptors based and SMILES based S2T-transformer QSAR model on test dataset

Input variables	R^2	RMSE	MAE
Key molecular descriptors	0.76	0.34	0.29
SMILES	0.83	0.34	0.30



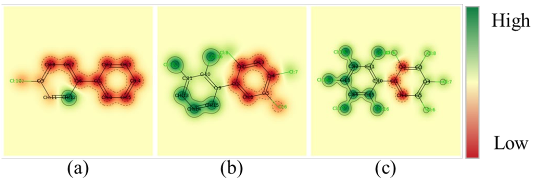
(a)

(b)

Figure 5. Prediction results for log K_{OW} of PCBs; (a) based on key molecular descriptors, and (b) based on SMILES

Attention weight analysis for SMILES of PCBs

S2T-transformer QSAR model based on SMILES calculated the attention weight between the SMILES and K_{OW} as shown in Figure 6. The number and location of chlorine atom contributes to the toxicity of the PCBs. Several chlorine atoms had positive contributions. 2,2',3,3',4,4',5,5',6-PCB has higher K_{OW} than 4-PCB due to the number of chlorine atom in the molecule. Thus the overall prediction results made sense, and the S2T-transformer QSAR model based on SMILES found the features of the molecular structure and could be utilized to interpret the correlation between toxicity activity and the molecular structure.



(a)

(b)

(c)

Figure 6. Attention weight analysis for log K_{OW} of (a) 4-PCB, (b) 2,2',3,4',5'-PCB, (c) 2,2',3,3',4,4',5,5',6-PCB

Conclusion

This study developed the S2T-transformer QSAR model based on SMILES. The proposed SMILES based S2T-transformer QSAR model indicated the higher predictive performance of 0.83 R^2 score than conventional key molecular descriptors based S2T-transformer QSAR model. Also, SMILES based S2T-transformer QSAR model interpreted the atom contribution of the molecular structure by attention weights in the transformer architecture.

Acknowledgments

This research is funded by the BK21 FOUR program of national Research Foundation of Korea, the National Research Foundation (NRF) grant funded by the South Korean government (MSIT) (No. 2021R1A2C2007838), Korea Ministry of Environment (MOE) as Graduate School specialized in Climate Change and Project for Collabo R&D between Industry, Academy, and Research Institute funded by Korea Ministry of SMEs and Start-ups in 2022 (Project No.S3301144).

References

Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. (2015). High-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive LASSO. *Journal of Chemometrics*, *29*(10), 547–556.

Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., Luo, X., Chen, K., Jiang, H., & Zheng, M. (2020). TransformerCPI: Improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, *36*(16), 4406–4414.

Heo, S. K., Safder, U., & Yoo, C. K. (2019). Deep learning driven QSAR model for environmental toxicology: Effects of endocrine disrupting chemicals on human health. *Environmental Pollution*.

Huang, J., & Fan, X. (2013). Reliably assessing prediction reliability for high dimensional QSAR data. *Molecular Diversity*.

Huang, T., Sun, G., Zhao, L., Zhang, N., Zhong, R., & Peng, Y. (2021). Quantitative structure-activity relationship (QSAR) studies on the toxic effects of nitroaromatic compounds (NACs): A systematic review. *International Journal of Molecular Sciences*, *22*(16).

Judson, R., Richard, A., Dix, D. J., Houck, K., Martin, M., Kavlock, R., Dellarco, V., Henry, T., Holderman, T., Sayre, P., Tan, S., Carpenter, T., & Smith, E. (2009). The toxicity data landscape for environmental chemicals. *Environmental Health Perspectives*, *117*(5), 685–695.

Kim, D., Lee, S., Kim, M., Lee, E., & Yoo, C. K. (2016). Development of QSAR model based on the key molecular descriptors selection and computational toxicology for prediction of toxicity of PCBs. *Korean Chemical Engineering Research*.

Kim, H., Na, J., & Lee, W. B. (2021). Generative Chemical Transformer: Neural Machine Learning of Molecular Geometric Structures from Chemical Language via Attention. *Journal of Chemical Information and Modeling*, *61*(12), 5804–5814.

Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2006). DRAGON software: An easy approach to molecular descriptor calculations. *Match*, *56*(2), 237–248.

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62–69.

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. (2012). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, *6*(SUPPL. 2).

Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T. P., Michalkova, A., Hwang, H. M., Toropov, A., Leszczynska, D., & Leszczynski, J. (2011). Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature Nanotechnology*.

Rim, K. T. (2020). In silico prediction of toxicity and its applications for chemicals at work. *Toxicology and Environmental Health Sciences*, *12*(3), 191–202.

Sabando, M. V., Ponzoni, I., Milios, E. E., & Soto, A. J. (2022). Using molecular embeddings in QSAR modeling: Does it make a difference? *Briefings in Bioinformatics*, *23*(1), 1–21.

Tang, W., Chen, J., Wang, Z., Xie, H., & Hong, H. (2018). Deep learning for predicting toxicity of chemicals: a mini review. *Journal of Environmental Science and Health - Part C Environmental Carcinogenesis and Ecotoxicology Reviews*, *36*(4), 252–271.

Todeschini, R., & Consonni, V. (2010). Molecular Descriptors for Chemoinformatics. In *Molecular Descriptors for Chemoinformatics*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Weininger, D. (1988). SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, *28*(1), 31–36.

Zhu, M., Su, H., Bao, Y., Li, J., & Su, G. (2022). Experimental determination of octanol-water partition coefficient (KOW) of 39 liquid crystal monomers (LCMs) by use of the shake-flask method. *Chemosphere*.