

# A Molecular Image-Based Novel Quantitative Structure-Activity Relationship Approach, DeepSnap-Deep Learning and Machine Learning

Yasunari Matsuzaka and Yoshihiro Uesawa\*

Department of Medical Molecular Informatics, Meiji Pharmaceutical University, 204-8588 Tokyo, Japan

\* Correspondence: uesawa@my-pharm.ac.jp

DOI: <https://doi.org/10.21775/cimb.042.455>

## Abstract

The quantitative structure-activity relationship (QSAR) approach has been used in numerous chemical compounds as *in silico* computational assessment for a long time. Further, owing to the high-performance modeling of QSAR, machine learning methods have been developed and upgraded. Particularly, the three-dimensional structure of chemical compounds has been gaining increasing attention owing to the representation of a large amount of information. However, only many of feature extraction is impossible to build models with the high-ability of the prediction. Thus, suitable extraction and effective selection of features are essential for models with excellent performance. Recently, the deep learning method has been employed to construct prediction models with very high performance using big data, especially, in the field of classification. Therefore, in this study, we developed a molecular image-based novel QSAR approach, called DeepSnap-Deep learning approach for designing high-performance models. In addition, this DeepSnap-Deep learning approach outperformed the conventional machine learnings when they are compared. Herein, we discuss the advantage and disadvantages of the machine learnings as well as the availability of the DeepSnap-Deep learning approach.

## Introduction

Quantitative structure-activity relationship (QSAR) is a well-established *in silico* approach that can predict pharmacological and toxicological effects of chemical compounds with similar structures. This can be achieved by demonstrating linear

or nonlinear relationships between structural features of chemicals and their biological activities measured experimentally using chemical descriptors, representing values of characteristics of chemicals with the same basic skeleton (Muratov et al., 2020; Achary, 2020; Toropov and Toropova, 2020). Moreover, these descriptors are categorized from zero dimension (0D) to four dimension (4D) based on the compound space considered in the calculation (González-Díaz et al., 2007; Yap, 2011; Kurgan and Disfani, 2011; Comelli et al., 2014, Yuan et al., 2018; Schneider et al., 2019; Ginex et al., 2019; Tangadpalliwar et al., 2019). First, a 0D-molecular descriptor is a configuration descriptor that represents the molecular weight of the compound and the numbers of rotatable bonds as well as double or triple bonds, and a count descriptor that indicates the number of atoms and rings, as well as the total heavy atoms (Kramer and Gedeck, 2011; Damião et al., 2014). Secondly, a 1D-molecular descriptor is physical property value, such as the numbers of specific functional groups, partial fragment structures, hydrogen bond donor, and acceptor atoms, fingerprint that can be represented by the presence or absence of the partial structure, and various LogPs including AlogP, ClogP, MlogP, SlogP, XlogP, and so on (Moriguchi et al., 1992; Xue et al., 2000; Abreu et al., 2011; dos Reis et al., 2014; McDonagh et al., 2015). Third, a 2D-molecular descriptor is a topological or connectivity index which is a quantitative variable that characterizes topological features as an invariant for a molecular graph, for example, the topological polar surface area indicating the polar part of the surface of the molecules, Wiener index showing the sum of the shortest distances between the atoms in the molecules, and the Balaban J index exhibiting the average total bond distance in the molecules (Prasanna et al., 2005; Prasanna and Doerksen, 2009; Poša, 2011). Fourth, a 3D-molecular descriptor is a geometrical descriptors that shows 3D information of molecules, such as molecular size, molecular structure, symmetry, and atomic distribution, for example, highest occupied molecular orbital (HOMO)/lowest unoccupied molecular orbital (LUMO) energy levels calculated from quantum chemical calculations, and a weighted holistic invariant molecular (WHIM) descriptor that is an eigenvalue calculated from a molecular matrix corresponding to a molecular graph in which 3D-coordinates are weighted based on the characteristics of each atom (Chen, 2008; Uesawa and Mohri, 2012; Marunna et al., 2017; Elrhayam and Elharfi, 2019). Fifth, a 4D-molecular descriptor is calculated through the interaction with other compounds, for example, the interaction energy resulting from molecular dynamics simulation (Kumar and Kulkarni, 2017, Ataide Martins et al., 2018; Ma et al., 2019). Since the affinity of chemical substances to protein enzymes and receptors strongly depends on intermolecular interactions, such as hydrogen bonding, electrostatic interactions, and hydrophobic interactions, it has been observed that the complementary positional relationship between the base groups of the chemicals and their binding site of the enzyme or receptor required for the interactions of chemical compounds is important (Yang et al., 2017; Kimani et al., 2018; Pan et al., 2019; Garcia et al., 2020). Therefore, besides molecular descriptors that show physiochemical properties related to intermolecular interactions, QSAR analysis in a 3D space that utilizes descriptors representing the 3D structure of chemical substances has been recently attracting research attention (Chen et al., 2020; Hadni and Elhallaoui, 2020; Kumar et al., 2020; Liu et al., 2020; Zhang et al., 2020). Additionally, prediction models in classical QSAR

are constructed using descriptors as explanatory variables. The objective variables have various types and strengths, namely reactivity and affinity, of different chemical substances, including hormones and inhibitors that maintain homeostasis of biological functions, ligands act on energy metabolism, and agonist/antagonist for vitamins involved in signal transduction (Hsu et al., 2014; Santos et al., 2017; Nagamani et al., 2018; Sakkiah et al., 2018; Agrawal et al., 2019; Kato 2020). Moreover, these chemical substances are involved in a process called key event (KE), which explains an unexpected in this pathway, that induces the expression of physiological activity by interacting with target molecules, such as receptors (Bal-Price and Meek, 2017; Scotti et al., 2017; Brüggemann et al., 2018; Ehlert, 2018; Naqvi et al., 2018). Further, the type of physiological activity depends on the type of target molecule, and the strength of the physiological activity is determined by the strength of interaction with the target molecule (Devidas and Ranawat, 2019). In particular, QSAR analysis can be employed to identify or assess the potential of chemical compounds in toxicity that is initiated by a molecular initiating event through the KE, finally causing an adverse outcome, in which numerous exogenous substances interfere with hormonal systems to produce a range of developmental, reproductive, neurologic, immune, or metabolic diseases (Vinken, 2016; Ciallella and Zhu, 2019; Schneider et al., 2019).

In classical QSAR derived from a set of small molecules with similar target-specific biological activity, using various descriptors related to the characteristic of physiological activity, mathematical correlations and patterns have been analyzed to date by applying multivariate analysis, multiple linear regression, partial least squares (PLS) regression, and so on (Stanton, 2012; Varmuza et al., 2013; Martinez-Lopez et al., 2017; Baskin II, 2018; Cruz et al., 2018; Nnadi et al., 2018). In these analyses, increasing the number of descriptors that are explanatory variables increases the coefficient of determination,  $R^2$ , and improves the prediction accuracy. However, when there is a strong linear dependence among the explanatory variables, the regression coefficient increases, and the estimation of the partial regression coefficient becomes unstable (Venkatraman et al., 2004; Basak et al., 2006). Thus, the influence of each explanatory variable on the objective variable may be unclear, and reliability may decrease (multicollinearity). In addition, an orthogonal transformation that aggregates and uses explanatory variables with strong correlation is required by applying feature selection, that is, selecting appropriate subsets of explanatory variables and using them in regression equations, such as principal component regression, PLS regression, and regression with penalties, to bring partial regression coefficient values closer to zero, such as ridge, lasso, and elastic net regressions (Hemmateenejad and Yazdani, 2009; Eklund et al., 2014; Tsiliki et al., 2015; Al-Sha'er et al., 2016). In these cases, owing to the occurrence of bias in the estimation of the regression coefficient value and the aggregation of variables, the original effect of the explanatory variable may become unclear, and the predictability of the prediction result may deteriorate (Fearn et al., 2008). Furthermore, as a solution to the variable selection problems that combine variables among those that accurately explain the target variable of interest from numerous candidate explanatory variables, a variable specification method that specifies the optimal explanatory

variable is utilized based on scientific theory or knowledge (George, 2000). However, it is rare to specify the optimum explanatory variable in advance. Further, in the case of the best subset selection or round-robin method, in which analysis is performed by utilizing all the combinations of the explanatory variables, there are  $2^k - 1$  combinations of  $k$  explanatory variables, resulting in a huge calculation cost (Noon et al., 2011; Saavedra et al., 2020). Additionally, based on the usefulness of each univariate regression coefficient, there are some sequential selection methods, which include forward-backward stepwise selection method, forward stepwise selection method, backward stepwise selection method, and backward-forward stepwise selection method, that sequentially increase or decrease the explanatory variables individually (Goodarzi et al., 2012; Fatima et al., 2018; Fatima et al., 2019; Hryniewicz et al., 2019; Fatima and Agarwal, 2020; McCann et al., 2020). However, the effect of unselected explanatory variables is not corrected, so to make variable selections, it is necessary to include technical and academic knowledge to the variable selection criteria.

### **Machine learning**

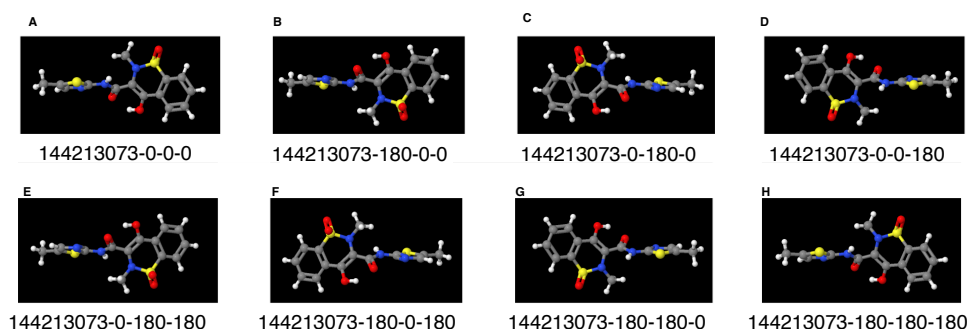
Moreover, to confound complex factors that contain numerous explanatory variables and construct nonlinear predictive models, it is often difficult to perform accurate modeling using conventional statistical methods. However, machine learning that can learn large amounts of data and automatically build models to perform classification and regression is suitable (Idakwo et al., 2018; Achary, 2020; Lin, et al., 2020; Muratov et al., 2020; Xiao et al., 2020; Yang et al., 2020). In machine learning algorithms, regularization to reduce extra features is internally performed so that the regression coefficient is stable. Furthermore, it is possible to reduce the data reading as well as the preprocessing and calculation costs, and by ameliorating the algorithm speed and clarifying the feature amount that has the most information, it is expected that the interpretation of the prediction result and the explanation can be improved. To date, machine learning algorithms, such as random forest (RF), support vector machine (SVM), eXtreme Gradient Boosting (Xgboost), Light Gradient Boosting Machine (LightGBM), Category Boosting (CATBoost), and neural network (NN), have been used to perform QSAR analysis (Zhang et al., 2018; Maltarollo et al., 2019; Sidorov et al. 2019; Matsuzaka et al., 2020; Matsuzaka and Uesawa, 2019a, 2020). The RF is an ensemble learning method based on bagging in which many decision trees that divide a group by conditional branching are connected and learning is performed in parallel for each model (Ghosh et al., 2020). In the case of classification in RF, the majority is output; the average value is finally output in the case of regression. Explaining the output results is relatively easy, and the effect of overfitting, which is a character of decision trees, can be reduced. Moreover, noise in the input data can easily cause over-learning, which leads to a reduction in generalization performance and an increase in calculation cost for complex datasets. Besides, an SVM performs two-class classification by constructing a classification boundary to increase the margin, which is the distance between the support vectors, which are the learning data located near the classification boundary, and the boundary (Cho et al., 2019). Thus, the classification accuracy does not depend on the dimension increase of the feature amount, so the

parameters can be easily optimized. In addition, since the calculation cost increases based on the learning data size, it is difficult to apply it to large-scale data, and it is also difficult to optimize the discriminant function owing to multiple classes. Further, Xgboost is a classifier configured plurality of classifiers gradient by collecting boosting, which ensemble learning combining by the plurality of weak classifiers, and RF (Zopluoglu, 2019). When constructing a new weak learner in the Xgboost, the weight is reduced for the data correctly classified by the constructed weak learners, whereas the weight is increased for the data not correctly classified, whereby update the construction of the weak learner serially in sequence. As described above, since the model is designed by aggregating the weightings based on the accuracy, highly accurate modeling can be expected. Additionally, many parameters need to be tuned to improve accuracy. In boosting, the learning cost increases because the learning is performed in series. LightGBM is also an ensemble learning method that applies gradient boosting based on a decision tree algorithm, but this method adopts the "Leaf-wise" method, which when training a decision tree, it grows based on the leaf of the decision tree (Zhang et al., 2019). This method has a shorter training time than the "Level-wise" method, in which the level of the decision tree grows; it is expected to improve the prediction accuracy since the structure of the decision tree generated by the "Leaf-wise" method is more complicated than that of the "Level-wise" method. Besides, overfitting is likely to occur, whereby hyperparameter adjustment to avoid this issue becomes complicated. Furthermore, in conventional search for branching points in a decision tree, it was necessary to read all the data points; in LightGBM, the features of the training data are divided into classes and made into histograms, reducing the computational cost even for large-scale data. Additionally, Catboost is also an ensemble learning method that uses a gradient boosting decision tree, but preprocessing is unnecessary because it can absorb categorical variables, namely qualitative variables, internally and process them (Romagnoni et al., 2019). Thus, it is expected that overfitting, such as target leakage, can be reduced by randomly selecting a dataset, repeating the calculation of statistics from the dataset, and calculating a substitute value. Moreover, the calculation cost is high, and it is difficult to output the prediction accuracy higher than the learning data. Furthermore, a simple perceptron, which is a type of NN, comprises two layers: input and output layers. A unit called neuron of the input data that corresponds to the type of explanatory variables is multiplied by a weight indicating a coupling strength, which is input to the output layer (Laudani et al., 2015; Roudi and Taylor, 2015). The sum of the values in each output layer and the bias for adding the variation to the data are added and output by a step function, which is an activation function. When the input value is less than zero, the output value is always zero, while when the input value is greater than or equal to zero, the output value is always 1, and finally the result is output. In learning the NN, the correct answer rate can be expected to be improved through parameter optimization by repeatedly adjusting the weight and bias. Moreover, this simple perceptron can be applied to a linearly separable problem in which two sets in an N-dimensional space can be separated by (N - 1)-dimensional space, but there was a problem in which non-linearly separable problems could not be handled (Van Calster et al., 2006; Martinetz et al., 2009). In the multi-layer

perceptron that adds a layer, also called intermediate layer or hidden layer, between the input and output layers based on the concept that complicated function approximation is possible by combining numerous simple neurons, it was possible to transform the nonlinear separation into a possible form. This was achieved by applying the activation function that performed the nonlinear transformation after performing the linear transformation in the intermediate layer. Furthermore, although the inability to update the coupling coefficient from the input layer to the intermediate layer was the cause of the nonlinear separation, the error backpropagation method corrects the weights of the intermediate and output layers based on the error between the output value from the output layer and the correct answer; this method further corrects the weight between the input and intermediate layers based on this corrected value that modified the entire NN each time, thereby enabling highly accurate nonlinear separation (Adigun and Kosko, 2019; Li, 2020; Ko and Lee, 2020). In a multi-layered NN, the process of reducing the error, which is the difference between the predicted and actual values, becomes complicated. Therefore, by applying the error backpropagation method and the stochastic gradient descent method, this problem can be avoided (Elfwing et al., 2015; Stromatias et al., 2017). In the case of a sigmoid function or its like, since there is a region where the amount of change called gradient in error is close to zero, the weight is hardly modified. The optimal solution was not obtained because all the gradients downstream from that region approach zero, and there was a problem (gradient disappearance problem) in which learning did not progress (Zhang, 2017). Besides, since the gradient of each layer is propagated by the product, it significantly disappears or increases (gradient explosion problem) as the number of layers increases, and the gradient of the NN becomes unstable (instability gradient problem) (Zhao et al., 2012; Sun et al., 2020). Consequently, training error is significantly reduced in the error regarding the training data, but generalization error cannot be reduced in the error regarding the test data, resulting in overfitting. In addition, there are problems in which the true optimal solution cannot be derived because the system converges to a local optimal solution, which is a partial optimal solution (local solution problem), and the calculation cost increases (Fu et al., 2018). Therefore, by utilizing a stacked autoencoder that reduces the number of neurons in the intermediate layer to be smaller than that in the input/output layer and performs repeated dimension compression or feature extraction by extracting very important information and deleting the rest, a deep neural network (DNN) composed of a multi-layered structured by preliminarily estimating an optimal value (preliminary learning) for the initial value of the weight of the neural network through learning using the same data as the output has enabled deep learning (Bengio et al., 2013).

### **Deep learning**

In conventional machine learning, the feature values of interest in the dataset were manually determined (feature extraction) based on knowledge and technology, and the model was designed by fitting the space (feature space) represented by the feature value into the data distribution. In deep learning, using a convolutional neural network (CNN), it has become possible to automate a series of processes of highly versatile feature extraction and model construction (Zhao et al., 2019; Heo et al., 2020). In addition, conventional machine learning

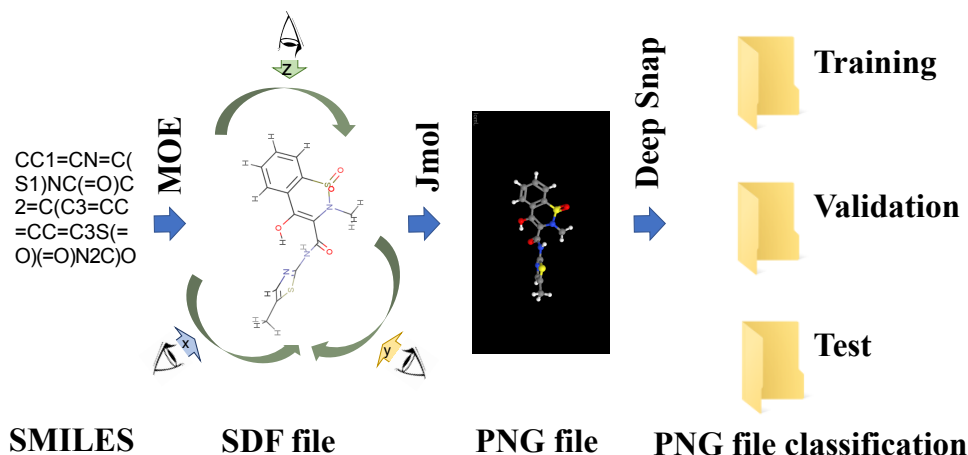


**Figure 1. Representative images obtained by rotating the 3D structure in 180° increments on DeepSnap.** The numbers below the images are the substance identification numbers (SID) provided in the PubChem database and increments of the viewing direction on the x-, y-, and z-axes. In addition, red, yellow, blue, white, and gray colors in the molecular structures indicate oxygen, sulfur, nitrogen, hydrogen, and carbon atoms, respectively.

has made it possible to process large-scale labeled training data, but when building a complex model, there was a problem in which the relationship between data size and generalization performance eventually becomes a plateau. Moreover, in deep learning, it was demonstrated that the relationship between the data size and the generalization performance is proportionally related, thereby enabling high accuracy modeling from big data (Gao et al., 2020; Lee and Chen, 2020). In addition, the learning result of this deep learning depends on the weight of the node, but it is difficult to theoretically interpret the accuracy or regularity of this result (black box problem) (Ishida et al., 2019; Leming et al., 2020). Further, to obtain a highly accurate result, there are still many unclear issues concerning the construction and learning method of the CNN. It is also relatively difficult to obtain large-scale labeled learning data with little noise. Therefore, in the case of image data, modeling with high accuracy was reported by employing feature extraction performed using CNN for a variant of the dataset with an increased amount of data, which were created by rotation, stretching, and so.

### DeepSnap-Deep learning method

The performance of the deep learning method may depend on the quantity and quality of the input datasets used in training and prediction. However, the deep learning approach has not established the preparation of suitable input data of three-dimensional (3D)-chemical structures. Therefore, as a new QSAR analysis method based on the present deep learning, DeepSnap-Deep learning method was developed by Uesawa at the Meiji Pharmaceutical University in 2018 (Uesawa, 2018). In this DeepSnap-Deep learning approach, the 3D-optimized molecular structures, which can be rotated at any arbitrary angle on the x-, y-, and z-axes, were photographed as a ball-and-stick model with different colors to represent the corresponding atoms to automatically input as much structural information as possible into the DL models. (Figure 1). Using this snapshot image as input data for deep learning, feature extraction through DNN and model



**Figure 2. A schematic of the DeepSnap-Deep learning procedure.** 3D chemical structures generated by molecular operating environment (MOE) software from a simplified molecular-input line-entry system (SMILES) format are captured as image pictures on arbitrary angles on the x-, y-, and z-axes, and saved as portable network graphics (PNG) files. These obtained images were split into the following three datasets for the input to deep learning: training, validation, and test.

construction are automatically performed (Figure 2). For example, when the 3D molecular structure is rotated in 45° increments on the x-, y-, and z-axes and photographed, a total of 512 images are captured for each molecule and saved in the portable network graphics (PNG) format. This allows for combining digital information regarding the 2D plane location of the atoms with pixel-level data representing the three primary colors (RGB). Therefore, when compared with the deep learning through DNN based on a 2D-graph structure, extraction of a larger amount of information regarding chemical structure can be expected. Moreover, this method can be used to predict the potential activity of many different chemicals to various receptors without the extraction of descriptors. In the process of drawing the 3D-chemical structure of this DeepSnap-Deep learning method, it is possible to adjust parameters, such as the atoms, color of atoms, bond radius, and pixel size (Matsuzaka and Uesawa, 2019b). Besides, by decreasing the drawing angle, the number of generated images per molecule can be increased, but no proportional and discontinuous relationship is observed between this number of generated images and the performance of the model. Additionally, in the DeepSnap-Deep learning method, a proportional relationship is observed between the number of generated images and the calculation cost. This demonstrated that the optimization of the number of generated images may minimize the calculation cost. In addition, the effects of protonation, stable structure selection in the 3D-process of the compound, and the background color of the generated image on model performance were also illustrated (Matsuzaka and Uesawa, 2019a,b, 2020). Further, when compared with the conventional machine learning methods, such as RF, Xgboost, LightGBM, Catboost, and NN,



that use molecular descriptors extracted through calculation software, Mordred (Moriwaki et al. 2018), it was observed that this DeepSnap-Deep learning method outperformed (Matsuzaka and Uesawa, 2019a,b, 2020; Matsuzaka et al. 2020). In this DeepSnap-Deep learning method, the classification performance was evaluated using information retrieved from a confusion matrix by cut-off points calculated with Youden index. Based on the sensitivity, which is a true positive rate identified as positive for all the positive samples including true and false positives, and the specificity, which is a true negative rate identified as negative for all the negative samples including true and false negatives, a confusion matrix regarding the predicted and experimentally defined labels was used to make the Receiver Operating Characteristic (ROC) curve and calculate the Area Under Curve (AUC), balanced accuracy, *F* value, and Matthews correlation coefficient. Utilizing proper statistical values can construct the prediction models with high performance. Thus, this DeepSnap-Deep learning method has the following advantages. First, the feature(s) in the molecular images can be automatically extracted by CNNs and can visualize the feature(s) by predicting the convolutional areas with NN. Second, high prediction performance can be expected as more detailed information of the chemical structure can be captured from different viewing directions along the x-, y-, and z-axes. Third, the determination and visualization of the conformer that is docked in the ligand-binding domain of protein may reveal the critical conformation of chemicals and protein domains related to the biological activity.

In the future, we hope that improvement and understanding of the interpretation and explanation of the prediction results from the DeepSnap-Deep learning method will yield modeling with high performance and the elucidation of the molecular mechanism exerted by chemical substances at the cellular, organ, and organism levels.

## References

- Abreu, R.M., Ferreira, I.C., Calhelha, R.C., Lima, R.T., Vasconcelos, M.H., Adegas, F., Chaves, R., and Queiroz, M.J. (2011). Anti-hepatocellular carcinoma activity using human HepG2 cells and hepatotoxicity of 6-substituted methyl 3-aminothieno[3,2-b]pyridine-2-carboxylate derivatives: in vitro evaluation, cell cycle analysis and QSAR studies. *Eur. J. Med. Chem.* **46**, 5800-5806. <https://doi.org/10.1016/j.ejmech.2011.09.029>.
- Achary, P.G.R. (2020). Applications Quantitative Structure-Activity Relationships (QSAR) based Virtual Screening in drug design: A Review. *Mini. Rev. Med. Chem.* in press. <https://doi.org/10.2174/1389557520666200429102334>.
- Adigun, O., and Kosko, B. (2019). Noise-boosted bidirectional backpropagation and adversarial learning. *Neural. Netw.* **120**, 9-31. <https://doi.org/10.1016/j.neunet.2019.09.016>.
- Agrawal, N., Chandrasekaran, B., and Al-Aboudi, A. (2019). Recent Advances in the In-silico Structure-based and Ligand-based Approaches for the Design and Discovery of Agonists and Antagonists of A2A Adenosine Receptor. *Curr. Pharm. Des.* **25**, 774-782. <https://doi.org/10.2174/1381612825666190306162006>.

- Al-Sha'er, M.A., Mansi, I., Khanfar, M., and Abudayyeh, A. (2016). Discovery of new heat shock protein 90 inhibitors using virtual co-crystallized pharmacophore generation. *J. Enzyme Inhib. Med. Chem.* *31*, 64-77. <https://doi.org/10.1080/14756366.2016.1218485>.
- Ataide Martins, J.P., Rougth de Oliveira, M.A., and Oliveira de Queiroz, M.S. (2018). Web-4D-QSAR: A web-based application to generate 4D-QSAR descriptors. *J. Comput. Chem.* *39*, 917-924. <https://doi.org/10.1002/jcc.25166>.
- Bal-Price, A., and Meek, M.E.B. (2017). Adverse outcome pathways: Application to enhance mechanistic understanding of neurotoxicity. *Pharmacol. Ther.* *179*, 84-95. <https://doi.org/10.1016/j.pharmthera.2017.05.006>.
- Basak, S.C., Natarajan, R., Mills, D., Hawkins, D.M., and Kraker, J.J. (2006). Quantitative structure-activity relationship modeling of juvenile hormone mimetic compounds for *Culex pipiens* larvae, with a discussion of descriptor-thinning methods. *J. Chem. Inf. Model.* *46*, 65-77. <https://doi.org/10.1021/ci050215y>.
- Baskin, II. (2018). Machine Learning Methods in Computational Toxicology. *Methods Mol. Biol.* *1800*, 119-139. [https://doi.org/10.1007/978-1-4939-7899-1\\_5](https://doi.org/10.1007/978-1-4939-7899-1_5).
- Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* *35*, 1798-828. <https://doi.org/10.1109/TPAMI.2013.50>.
- Brüggemann, M., Licht, O., Fetter, É., Teigeler, M., Schäfers, C., and Eilebrecht, E. (2018). Knotting nets: Molecular junctions of interconnecting endocrine axes identified by application of the adverse outcome pathway concept. *Environ. Toxicol. Chem.* *37*, 318-328. <https://doi.org/10.1002/etc.3995>.
- Chen, H.F. (2008). Computational study of histamine H3-receptor antagonist with support vector machines and three dimension quantitative structure activity relationship methods. *Anal. Chim. Acta.* *624*, 203-209. <https://doi.org/10.1016/j.aca.2008.06.048>.
- Chen, Q., Wang, F., and Zhou, B. (2020). Investigations of retinoic acid receptor-related orphan receptor-gamma t (ROR $\gamma$ t) agonists: a combination of 3D-QSAR, molecular docking and molecular dynamics. *J. Biomol. Struct. Dyn.* in press. <https://doi.org/10.1080/07391102.2020.1765873>.
- Cho, G., Yim, J., Choi, Y., Ko, J., and Lee, S.H. (2019). Review of Machine Learning Algorithms for Diagnosing Mental Illness. *Psychiatry Investig.* *16*, 262-269. <https://doi.org/10.30773/pi.2018.12.21.2>.
- Ciallella, H.L., and Zhu, H. (2019). Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. *Chem. Res. Toxicol.* *32*, 536-547. <https://doi.org/10.1021/acs.chemrestox.8b00393>.
- Cruz, J.V., Serafim, R.B., da Silva, G.M., Giuliatti, S., Rosa, J.M.C., Araújo Neto, M.F., Leite, F.H.A., Taft, C.A., da Silva, C.H.T.P., and Santos, C.B.R. (2018). Computational design of new protein kinase 2 inhibitors for the treatment of inflammatory diseases using QSAR, pharmacophore-structure-based virtual screening, and molecular dynamics. *J. Mol. Model.* *24*, 225. <https://doi.org/10.1007/s00894-018-3756-y>.
- Comelli, N.C., Duchowicz, P.R., and Castro, E.A. (2014). QSAR models for thiophene and imidazopyridine derivatives inhibitors of the Polo-Like Kinase 1. *Eur. J. Pharm. Sci.* *62*, 171-179. <https://doi.org/10.1016/j.ejps.2014.05.029>.

- Damião, M.C., Pasqualoto, K.F., Polli, M.C., and Parise Filho, R. (2014). To be drug or prodrug: structure-property exploratory approach regarding oral bioavailability. *J. Pharm. Pharm. Sci.* *17*, 532-540. <https://doi.org/10.18433/j3bs4h>.
- dos Reis, R.R., Sampaio, S.C., and de Melo, E.B. (2014). An alternative approach for the use of water solubility of nonionic pesticides in the modeling of the soil sorption coefficients. *Water Res.* *53*, 191-199. <https://doi.org/10.1016/j.watres.2014.01.023>.
- Ehlert, F.J. (2018). Analysis of Biased Agonism. *Prog. Mol. Biol. Transl. Sci.* *160*, 63-104. <https://doi.org/10.1016/bs.pmbts.2018.08.001>.
- Eklund, M., Norinder, U., Boyer, S., and Carlsson, L. (2014). Choosing feature selection and learning algorithms in QSAR. *J. Chem. Inf. Model.* *54*, 837-843. <https://doi.org/10.1021/ci400573c>.
- Elfving, S., Uchibe, E., and Doya, K. (2015). Expected energy-based restricted Boltzmann machine for classification. *Neural Netw.* *64*, 29-38. <https://doi.org/10.1016/j.neunet.2014.09.006>.
- Elrhayam, Y., and Elharfi, A. (2019). 3D-QSAR studies of the chemical modification of hydroxyl groups of biomass (cellulose, hemicelluloses and lignin) using quantum chemical descriptors. *Heliyon* *5*, e02173. <https://doi.org/10.1016/j.heliyon.2019.e02173>.
- Fatima, S., Gupta, P., and Agarwal, S.M. (2018). Insight into structural requirements of antiamebic flavonoids: 3D-QSAR and G-QSAR studies. *Chem. Biol. Drug Des.* *92*, 1743-1749. <https://doi.org/10.1111/cbdd.13343>.
- Fatima, S., Pal, D., and Agarwal, S.M. (2019). QSAR of clinically important EGFR mutant L858R/T790M pyridinylimidazole inhibitors. *Chem. Biol. Drug Des.* *94*, 1306-1315. <https://doi.org/10.1111/cbdd.13505>.
- Fatima, S., and Agarwal, S.M. (2020). Structure-activity Relationship Study on Therapeutically Relevant EGFR Double Mutant Inhibitors. *Med. Chem.* *16*, 52-62. <https://doi.org/10.2174/1573406415666190206204853>.
- Fearn, T., Hill, D.C., and Darby, S.C. (2008). Measurement error in the explanatory variable of a binary regression: regression calibration and integrated conditional likelihood in studies of residential radon and lung cancer. *Stat. Med.* *27*, 2159-2176. <https://doi.org/10.1002/sim.3163>.
- Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). Deep Ordinal Regression Network for Monocular Depth Estimation. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2018*, 2002-2011. <https://doi.org/10.1109/CVPR.2018.00214>.
- Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A Survey on Deep Learning for Multimodal Data Fusion. *Neural Comput.* *32*, 829-864. [https://doi.org/10.1162/neco\\_a\\_01273](https://doi.org/10.1162/neco_a_01273).
- George, E.I. (2000). The Variable Selection Problem. *J. Am. Stat. Assoc.* *95*, 1304-1308. <https://doi.org/10.1080/01621459.2000.10474336>
- Ghosh, T., Zhang, W., Ghosh, D., and Kechris, K. (2020). Predictive Modeling for Metabolomics Data. *Methods Mol. Biol.* *2104*, 313-336. [https://doi.org/10.1007/978-1-0716-0239-3\\_16](https://doi.org/10.1007/978-1-0716-0239-3_16).
- Ginex, T., Vazquez, J., Gilbert, E., Herrero, E., and Luque, F.J. (2019). Lipophilicity in Drug Design: An Overview of Lipophilicity Descriptors in 3D-

- QSAR Studies. *Future Med. Chem.* *11*, 1177-1193. <https://doi.org/10.4155/fmc-2018-0435>.
- González-Díaz, H., Olazábal, E., Santana, L., Uriarte, E., González-Díaz, Y., and Castañedo, N. (2007). QSAR Study of Anticoccidial Activity for Diverse Chemical Compounds: Prediction and Experimental Assay of trans-2-(2-nitrovinyl) furan. *Bioorg. Med. Chem.* *15*, 962-968. <https://doi.org/10.1016/j.bmc.2006.10.032>
- Garcia, M.L., de Oliveira, A.A., Bueno, R.V., Nogueira, V.H.R., de Souza, G.E., and Guido, R.V.C. (2020). QSAR studies on benzothiophene derivatives as Plasmodium falciparum N-myristoyltransferase inhibitors: Molecular insights into affinity and selectivity. *Drug Dev. Res.* in press. <https://doi.org/10.1002/ddr.21646>.
- Goodarzi, M., Dejaegher, B., Vander and Heyden, Y. (2012). Feature selection methods in QSAR studies. *J. AOAC. Int.* *95*, 636-651. [https://doi.org/10.5740/jaoacint.sge\\_goodarzi](https://doi.org/10.5740/jaoacint.sge_goodarzi).
- Hadni, H., and Elhallaoui, M. (2020). 3D-QSAR, docking and ADMET properties of aurone analogues as antimalarial agents. *Heliyon* *6*, e03580. <https://doi.org/10.1016/j.heliyon.2020.e03580>.
- Hemmateenejad, B., and Yazdani, M. (2009). QSPR models for half-wave reduction potential of steroids: a comparative study between feature selection and feature extraction from subsets of or entire set of descriptors. *Anal. Chim. Acta.* *634*, 27-35. <https://doi.org/10.1016/j.aca.2008.11.062>.
- Heo, T.Y., Kim, K.M., Min, H.K., Gu, S.M., Kim, J.H., Yun, J., and Min, J.K. (2020). Development of a Deep-Learning-Based Artificial Intelligence Tool for Differential Diagnosis between Dry and Neovascular Age-Related Macular Degeneration. *Diagnostics (Basel)* *10*, E261. <https://doi.org/10.3390/diagnostics10050261>.
- Hsu, C.W., Zhao, J., Huang, R., Hsieh, J.H., Hamm, J., Chang, X., Houck, K., and Xia, M. (2014). Quantitative high-throughput profiling of environmental chemicals and drugs that modulate farnesoid X receptor. *Sci. Rep.* *4*, 6437. <https://doi.org/10.1038/srep06437>.
- Hryniewicz, M., Iwaniak, A., Bucholska, J., Minkiewicz, P., and Darewicz, M. (2019). Structure-Activity Prediction of ACE Inhibitory/Bitter Dipeptides-A Chemometric Approach Based on Stepwise Regression. *Molecules* *24*, 950. <https://doi.org/10.3390/molecules24050950>.
- Idakwo, G., Luttrell, J., Chen, M., Hong, H., Zhou, Z., Gong, P., and Zhang, C. (2018). A review on machine learning methods for in silico toxicity prediction. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* *36*, 169-191. <https://doi.org/10.1080/10590501.2018.1537118>.
- Ishida, S., Terayama, K., Kojima, R., Takasu, K., and Okuno, Y. (2019). Prediction and Interpretable Visualization of Retrosynthetic Reactions Using Graph Convolutional Networks. *J. Chem. Inf. Model.* *59*, 5026-5033. <https://doi.org/10.1021/acs.jcim.9b00538>.
- Kato, H. (2020). Computational prediction of cytochrome P450 inhibition and induction. *Drug Metab. Pharmacokinet.* *35*, 30-44. <https://doi.org/10.1016/j.dmpk.2019.11.006>.

- Ko, T.Y., and Lee, S.H. (2020). Novel Method of Semantic Segmentation Applicable to Augmented Reality. *Sensors (Basel)* 20, 1737. <https://doi.org/10.3390/s20061737>.
- Kimani, N.M., Matasyoh, J.C., Kaiser, M., Nogueira, M.S., Trossini, G.H.G., and Schmidt, T.J. (2018). Complementary Quantitative Structure-Activity Relationship Models for the Antitrypanosomal Activity of Sesquiterpene Lactones. *Int. J. Mol. Sci.* 19, 3721. <https://doi.org/10.3390/ijms19123721>. <https://doi.org/10.1016/j.bmc.2006.10.032>.
- Kramer, C., and Gedeck, P. (2011). Three descriptor model sets a high standard for the CSAR-NRC HiQ benchmark. *J. Chem. Inf. Model.* 51, 2139-45. <https://doi.org/10.1021/ci200030h>.
- Kumar, A., Rai, S., Rathi, E., Agarwal, P., and Kini, S.G. (2020). Pharmacophore-guided fragment-based design of novel mammalian target of rapamycin inhibitors: extra precision docking, fingerprint-based 2D and atom-based 3D-QSAR modelling. *J. Biomol. Struct. Dyn.* 19, 1-19. <https://doi.org/10.1080/07391102.2020.1726816>.
- Kurgan, L., and Disfani, F.M. (2011). Structural protein descriptors in 1-dimension and their sequence-based predictions. *Curr. Protein Pept. Sci.* 12, 470-489. <https://doi.org/10.2174/138920311796957711>.
- Laudani, A., Lozito, G.M., Riganti Fulginei, F., and Salvini, A. (2015). On Training Efficiency and Computational Costs of a Feed Forward Neural Network: A Review. *Comput. Intell. Neurosci.* 2015, 818243. <https://doi.org/10.1155/2015/818243>.
- Lee, C.Y., and Chen, Y.P. (2020). Prediction of drug adverse events using deep learning in pharmaceutical discovery. *Brief Bioinform.* in press. <https://doi.org/10.1093/bib/bbaa040>.
- Leming, M., Górriz, J.M., and Suckling, J. (2020). Ensemble Deep Learning on Large, Mixed-Site fMRI Datasets in Autism and Other Tasks. *Int. J. Neural Syst.* in press. <https://doi.org/10.1142/S0129065720500124>.
- Li, H., Wang, Y., Zhen, Z.Z., Tan, Y., Chen, Z., Wang, X., Pei, T., and Wang, L. (2020). Identifying Microbe-Disease Association Based on a Novel Back-Propagation Neural Network Model. *IEEE/ACM Trans. Comput. Biol. Bioinform.* in press. <https://doi.org/10.1109/TCBB.2020.2986459>.
- Lin, X., Li, X., and Lin, X. (2020). A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules* 25, 1375. <https://doi.org/10.3390/molecules25061375>.
- Liu, Y.X., Gao, S., Ye, T., Li, J.Z., Ye, F., and Fu, Y. (2020). Combined 3D-quantitative structure-activity relationships and topomer technology-based molecular design of human 4-hydroxyphenylpyruvate dioxygenase inhibitors. *Future Med. Chem.* 12, 795-811. <https://doi.org/10.4155/fmc-2019-0349>.
- Ma, W., Wang, Y., Chu, D., and Yan, H. (2019). 4D-QSAR and MIA-QSAR study on the Bruton's tyrosine kinase (Btk) inhibitors. *J. Mol. Graph Model.* 92, 357-362. <https://doi.org/10.1016/j.jmkgm.2019.08.009>.
- Marunnan, S.M., Pulikkal, B.P., Jabamalaairaj, A., Bandaru, S., Yadav, M., Nayariseri, A., and Doss, V.A. (2017). Development of MLR and SVM Aided QSAR Models to Identify Common SAR of GABA Uptake Herbal Inhibitors used

- in the Treatment of Schizophrenia. *Curr. Neuropharmacol.* *15*, 1085-1092. <https://doi.org/10.2174/1567201814666161205131745>.
- Martinetz, T., Labusch, K., and Schneegass, D. (2009). SoftDoubleMaxMinOver: perceptron-like training of support vector machines. *IEEE Trans. Neural Netw.* *20*, 1061-1072. <https://doi.org/10.1109/TNN.2009.2016717>.
- Martinez-Lopez, Y., Caballero, Y., Barigye, S.J., Marrero-Ponce, Y., Millan-Cabrera, R., Madera, J., Torrens, F., and Castillo-Garit, J.A. (2017). State of the Art Review and Report of New Tool for Drug Discovery. *Curr. Top Med. Chem.* *17*, 2957-2976. <https://doi.org/10.2174/1568026617666170821123856>.
- Matsuzaka, Y., Hosaka, T., Ogaito, A., Yoshinari, K., and Uesawa, Y. (2020). Prediction Model of Aryl Hydrocarbon Receptor Activation by a Novel QSAR Approach, DeepSnap-Deep Learning. *Molecules* *25*, 1317. <https://doi.org/10.3390/molecules25061317>.
- Matsuzaka, Y., and Uesawa, Y. (2020). DeepSnap-Deep Learning Approach Predicts Progesterone Receptor Antagonist Activity With High Performance. *Front. Bioeng. Biotechnol.* *7*, 485. <https://doi.org/10.3389/fbioe.2019.00485>. eCollection 2019.
- Matsuzaka, Y., and Uesawa, Y. (2019a). Prediction Model with High-Performance Constitutive Androstane Receptor (CAR) Using DeepSnap-Deep Learning Approach from the Tox21 10K Compound Library. *Int. J. Mol. Sci.* *20*, 4855. <https://doi.org/10.3390/ijms20194855>.
- Matsuzaka, Y., and Uesawa, Y. (2019b). Optimization of a Deep-Learning Method Based on the Classification of Images Generated by Parameterized Deep Snap a Novel Molecular-Image-Input Technique for Quantitative Structure-Activity Relationship (QSAR) Analysis. *Front. Bioeng. Biotechnol.* *7*, 65. <https://doi.org/10.3389/fbioe.2019.00065>.
- McCann, M.R., McHugh, C.E., Kirby, M., Jennaro, T.S., Jones, A.E., Stringer, K.A., and Puskarich, M.A. (2020). A Multivariate Metabolomics Method for Estimating Platelet Mitochondrial Oxygen Consumption Rates in Patients with Sepsis. *Metabolites* *10*, E139. <https://doi.org/10.3390/metabo10040139>.
- McDonagh, J.L., van Mourik, T., and Mitchell, J.B. (2015). Predicting Melting Points of Organic Molecules: Applications to Aqueous Solubility Prediction Using the General Solubility Equation. *Mol. Inform.* *34*, 715-724. <https://doi.org/10.1002/minf.201500052>.
- Moriguchi, I., Hirano, S., Liu, Q., Nakagome, I., and Matsushita, Y. (1992). Simple Method of Calculating Octanol/Water Partition Coefficient. *Chem. Pharm. Bull.* *40*, 127-130. <https://doi.org/10.1248/cpb.40.127>
- Moriwaki, H., Tian, Y.S., Kawashita, N., and Takagi, T. (2018). Mordred: a molecular descriptor calculator. *J. Cheminf.* *10*, 4. <https://doi.org/10.1186/s13321-018-0258-y>
- Muratov, E.N., Bajorath, J., Sheridan, R.P., Tetko, I.V., Filimonov, D., Poroikov, V., Oprea, T.I., Baskin, II., Varnek, A., Roitberg, A., et al., (2020). QSAR without borders. *Chem. Soc. Rev.* in press. <https://doi.org/10.1039/d0cs00098a>.
- Nagamani, S., Kesavan, C., and Muthusamy, K. (2018). Atom-based and Pharmacophore-based 3D - QSAR Studies on Vitamin D Receptor (VDR). *Comb. Chem. High Throughput Screen* *21*, 329-343. <https://doi.org/10.2174/1386207321666180607101720>.

- Naqvi, A.A.T., Mohammad, T., Hasan, G.M., and Hassan, M.I. (2018). Advancements in Docking and Molecular Dynamics Simulations Towards Ligand-receptor Interactions and Structure-function Relationships. *Curr. Top Med. Chem.* 18, 1755-1768. <https://doi.org/10.2174/1568026618666181025114157>.
- Noon, A., Kalakech, A., and Kadry, S. (2011). A New Round Robin Based Scheduling Algorithm for Operating Systems: Dynamic Quantum Using the Mean Average. <https://arxiv.org/abs/1111.5348>.
- Nnadi, C.O., Althaus, J.B., Nwodo, N.J., and Schmidt, T.J. (2018). A 3D-QSAR Study on the Antitrypanosomal and Cytotoxic Activities of Steroid Alkaloids by Comparative Molecular Field Analysis. *Molecules* 23, 1113. <https://doi.org/10.3390/molecules23051113>.
- Pan, C., Meng, H., Zhang, S., Zuo, Z., Shen, Y., Wang, L., and Chang, K.J. (2019). Homology modeling and 3D-QSAR study of benzhydrylpiperazine  $\delta$  opioid receptor agonists. *Comput. Biol. Chem.* 83, 107109. <https://doi.org/10.1016/j.compbiolchem.2019.107109>.
- Poša, M. (2011). QSPR Study of the Effect of Steroidal Hydroxy and Oxo Substituents on the Critical Micellar Concentration of Bile Acids. *Steroids* 76, 85-93. <https://doi.org/10.1016/j.steroids.2010.09.003>.
- Prasanna, S., Manivannan, E., and Chaturvedi, S.C. (2005). QSAR analyses of conformationally restricted 1,5-diaryl pyrazoles as selective COX-2 inhibitors: application of connection table representation of ligands. *Bioorg. Med. Chem. Lett.* 15, 2097-2102. <https://doi.org/10.1016/j.bmcl.2005.02.035>.
- Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., and Hugot, J.P. (2019). Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). *Sci. Rep.* 9, 10351. <https://doi.org/10.1038/s41598-019-46649-z>.
- Roudi, Y., and Taylor, G. (2015). Learning with hidden variables. *Curr. Opin. Neurobiol.* 35, 110-118. <https://doi.org/10.1016/j.conb.2015.07.006>.
- Saavedra, L.M., Romanelli, G.P., and Duchowicz, P.R. (2020). A non-conformational QSAR study for plant-derived larvicides against Zika *Aedes aegypti* L. vector. *Environ. Sci. Pollut. Res. Int.* 27, 6205-6214. <https://doi.org/10.1007/s11356-019-06630-9>.
- Sakkiah, S., Guo, W., Pan, B., Kusko, R., Tong, W., and Hong, H. (2018). Computational prediction models for assessing endocrine disrupting potential of chemicals. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* 36, 192-218. <https://doi.org/10.1080/10590501.2018.1537132>.
- Santos, P., López-Vallejo, F., and Soto, C.Y. (2017). In silico approaches and chemical space of anti-P-type ATPase compounds for discovering new antituberculous drugs. *Chem. Biol. Drug Des.* 90, 175-187. <https://doi.org/10.1111/cbdd.12950>.
- Schneider, M., Pons, J.L., Labesse, G., and Bourguet, W. (2019). In Silico Predictions of Endocrine Disruptors Properties. *Endocrinology* 160, 2709-2716. <https://doi.org/10.1210/en.2019-00382>.
- Scotti, L., Mendonca, J.F.J., Ishiki, H.M., Ribeiro, F.F., Singla, R.K., Barbosa Filho, J.M., Da Silva, M.S., and Scotti, M.T. (2017). Docking Studies for Multi-Target

- Drugs. *Curr. Drug. Targets* 18, 592-604. <https://doi.org/10.2174/1389450116666150825111818>.
- Sidorov, P., Naulaerts, S., Arieu-Bonnet, J., Pasquier, E., and Ballester, P.J. (2019). Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data. *Front. Chem.* 7, 509. <https://doi.org/10.3389/fchem.2019.00509>. eCollection 2019.
- Stanton, D.T. (2012). QSAR and QSPR model interpretation using partial least squares (PLS) analysis. *Curr. Comput. Aided Drug Des.* 8, 107-127. <https://doi.org/10.2174/157340912800492357>.
- Stromatias, E., Soto, M., Serrano-Gotarredona, T., and Linares-Barranco, B. (2017). An Event-Driven Classifier for Spiking Neural Networks Fed with Synthetic or Dynamic Vision Sensor Data. *Front. Neurosci.* 11, 350. <https://doi.org/10.3389/fnins.2017.00350>. eCollection 2017.
- Sun, L., Wang, Y., He, J., Li, H., Peng, D., and Wang, Y. (2020). A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs. *Health Inf. Sci. Syst.* 8, 19. <https://doi.org/10.1007/s13755-020-00103-x>.
- Tangadpalliar, S.R., Vishwakarma, S., Nimbalkar, R., and Garg, P. (2019). ChemSuite: A package for chemoinformatics calculations and machine learning. *Chem. Biol. Drug Des.* 93, 960-964. <https://doi.org/10.1111/cbdd.13479>.
- Tsiliki, G., Munteanu, C.R., Seoane, J.A., Fernandez-Lozano, C., Sarimveis, H., and Willighagen, E.L. (2015). RRegrs: an R package for computer-aided model selection with multiple regression models. *J. Cheminform.* 7, 46. <https://doi.org/10.1186/s13321-015-0094-2>.
- Toropov, A.A., and Toropova, A.P. (2020). QSPR/QSAR: State-of-Art, Weirdness, the Future. *Molecules* 25, pii: E1292. <https://doi.org/10.3390/molecules25061292>.
- Uesawa, Y., and Mohri, K. (2012). Quantitative structure-interaction relationship analysis of 1,4-dihydropyridine drugs in concomitant administration with grapefruit juice. *Pharmazie* 67, 195-201. <https://doi.org/10.1691/ph.2012.1101>
- Uesawa, Y. (2018). Quantitative structure-activity relationship analysis using deep learning based on a novel molecular image input technique. *Bioorg. Med. Chem. Lett.* 28, 3400-3403. <https://doi.org/10.1016/j.bmcl.2018.08.032>. E
- Van Calster, B., Timmerman, D., Nabney, I.T., Valentin, L., Van Holsbeke, C., and Van Huffel, S. (2006). Classifying ovarian tumors using Bayesian Multi-Layer Perceptrons and Automatic Relevance Determination: a multi-center study. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2006, 5342-5345. <https://doi.org/10.1109/IEMBS.2006.260118>.
- Varmuza, K., Filzmoser, P., and Dehmer, M. (2013). Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS. *Comput. Struct. Biotechnol. J.* 5, e201302007. <https://doi.org/10.5936/csbj.201302007>.
- Venkatraman, V., Dalby, A.R., and Yang, Z.R. (2004). Evaluation of mutual information and genetic programming for feature selection in QSAR. *J. Chem. Inf. Comput. Sci.* 44, 1686-1692. <https://doi.org/10.1021/ci049933v>.
- Vinken, M. (2013). The Adverse Outcome Pathway Concept: A Pragmatic Tool in Toxicology. *Toxicology* 312, 158-165. <https://doi.org/10.1016/j.tox.2013.08.011>.
- Xiao, X., Fu, D., Shi, Y., and Wen, J. (2020). Optimized Mahalanobis-Taguchi System for High-Dimensional Small Sample Data Classification. *Comput. Intell.*



- Neurosci. 2020, 4609423. <https://doi.org/10.1155/2020/4609423>. eCollection 2020.
- Xue, L., Godden, J.W., and Bajorath, J. (2000). Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* 40, 1227-1234. <https://doi.org/10.1021/ci000327j>.
- Yang, H., Huang, Y., Liu, J., Tang, P., Sun, Q., Xiong, X., Tang, B., He, J., and Li, H. (2017). Binding modes of environmental endocrine disruptors to human serum albumin: insights from STD-NMR, ITC, spectroscopic and molecular docking studies. *Sci. Rep.* 7, 11126. <https://doi.org/10.1038/s41598-017-11604-3>.
- Yang, H., Lou, C., Li, W., Liu, G., and Tang, Y. (2020). Computational Approaches to Identify Structural Alerts and Their Applications in Environmental Toxicology and Drug Discovery. *Chem. Res. Toxicol.* in press. <https://doi.org/10.1021/acs.chemrestox.0c00006>.
- Yao, G., Wang, G., Wang, D., and Su, G. (2019). Identification of a novel mutation of FGFR3 gene in a large Chinese pedigree with hypochondroplasia by next-generation sequencing: A case report and brief literature review. *Medicine (Baltimore)* 98, e14157. <https://doi.org/10.1097/MD.00000000000014157>.
- Yap, C.W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32, 1466-1474. <https://doi.org/10.1002/jcc.21707>.
- Zhang, C., Li, Q., Ren, Y., and Liu, F. (2020). Molecular modeling studies of benzothiophene-containing derivatives as promising selective estrogen receptor downregulators: a combination of 3D-QSAR, molecular docking and molecular dynamics simulations. *J. Biomol. Struct. Dyn.* 20, 1-22. <https://doi.org/10.1080/07391102.2020.1751717>.
- Zhang, Y., Xie, R., Wang, J., Leier, A., Marquez-Lago, T.T., Akutsu, T., Webb, G.I., Chou, K.C., and Song, J. (2019). Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform.* 20, 2185-2199. <https://doi.org/10.1093/bib/bby079>.
- Zhang, X. (2017). Melanoma segmentation based on deep learning. *Comput. Assist. Surg (Abingdon)* 22, 267-277. <https://doi.org/10.1080/24699322.2017.1389405>.
- Zhao, T., Hachiya, H., Niu, G., and Sugiyama, M. (2012). Analysis and improvement of policy gradient estimation. *Neural Netw.* 26, 118-129. doi: 10.1016/j.neunet.2011.09.005.
- Zhao, Z., Deng, Y., Zhang, Y., Zhang, Y., Zhang, X., and Shao, L. (2019). DeepFHR: intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network. *BMC Med. Inform. Decis. Mak.* 19, 286. <https://doi.org/10.1186/s12911-019-1007-5>.
- Zopluoglu, C. (2019). Detecting Examinees With Item Preknowledge in Large-Scale Testing Using Extreme Gradient Boosting (XGBoost). *Educ. Psychol. Meas.* 79, 931-961. <https://doi.org/10.1177/0013164419839439>.

