applied sciences

MDPI

Article

A Quantitative Structure-Property Relationship Model Based on Chaos-Enhanced Accelerated Particle Swarm Optimization Algorithm and Back Propagation Artificial Neural Network

Mengshan Li *, Huaijin Zhang, Liang Liu, Bingsheng Chen, Lixin Guan and Yan Wu

College of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000, China; hjz_gnnu@163.com (H.Z.); liangliu_gnnu@163.com (L.L.); gzcbs101@163.com (B.C.); glxszu@126.com (L.G.); jci_wu@163.com (Y.W.)

* Correspondence: jcimsli@163.com; Tel.: +86-797-839-3668

Received: 24 May 2018; Accepted: 5 July 2018; Published: 11 July 2018

Featured Application: The proposed hybrid intelligent model can be applied in engineering design, material performance prediction, numerical calculation, and the prediction of physical and chemical properties.

Abstract: A quantitative structure-property relationship (QSPR) model is proposed to explore the relationship between the pKa of various compounds and their structures. Through QSPR studies, the relationship between the structure and properties can be obtained. In this study, a novel chaos-enhanced accelerated particle swarm algorithm (CAPSO) is adopted to screen molecular descriptors and optimize the weights of back propagation artificial neural network (BP ANN). Then, the QSPR model based on CAPSO and BP ANN is proposed and named the CAPSO BP ANN model. The prediction experiment showed that the CAPSO algorithm was a reliable method for screening molecular descriptors. The five molecular descriptors obtained by the CAPSO algorithm could well characterize the molecular structure of each compound in pKa prediction. The experimental results also showed that the CAPSO BP ANN model exhibited good performance in predicting the pKa values of various compounds. The absolute mean relative error, root mean square error, and square correlation coefficient are respectively 0.5364, 0.0632, and 0.9438, indicating the high prediction accuracy. The proposed hybrid intelligent model can be applied in engineering design and the prediction of physical and chemical properties.

Keywords: quantitative structure-property relationship; hybrid intelligence; artificial neural network; particle swarm optimization

1. Introduction

In quantitative structure-property relationship (QSPR) modeling, some mathematical and artificial intelligence methods are used to explore the chemical and physical properties of various substances. These methods, including mathematical statistics, machine learning methods, and artificial intelligence methods, can reflect the relationship between the activity and structure of compounds. Through QSPR studies, the relationship between the structure and activity of compounds can be mined [1,2]. The QSPR model can be used to predict the activity of unknown materials and discover key influencing factors of the activity of related substances, such as groups or substituents determining the activity of the molecular structure [3,4]. Nowadays, QSPR has been applied in the fields of computer science, chemistry, materials science, medicine science, and life sciences [5,6].

The establishment of the QSPR model mainly involves the following steps: acquisition of experimental data, construction and optimization of the molecular structure, calculation and screening of molecular descriptors, establishment and verification of the model, etc. First of all, the variable selection is important in many fields, such as spectroscopy [7,8], QSPR [9,10], and other fields [11,12]. The selection of molecular descriptors largely determines the quality of the QSPR model [13–15]. The step of molecular descriptor screening aims to reflect more structural information so that there is no noise in the descriptors. Many methods have been developed to screen molecular descriptors and can be mainly divided into two categories [16–18]. The first category includes the common methods, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and forward/backward/bi-directional stepwise multiple linear regression (MLR). The second includes the modern search algorithms, such as genetic algorithm (GA), simulated annealing algorithm (SA), ant colony algorithm (AC), particle swarm optimization (PSO), and other swarm intelligence algorithms [7,11,19–21]. The common methods are the most simple and efficient, but their overall performances are low in complex nonlinear problems. The modern search algorithms based on the optimization strategy have obvious advantages and can search for optimal variables and deal with complex large data points. The model establishment is important in the QSPR study and commonly used QSPR models include two-dimensional (2d), three-dimensional (3d), and four-dimensional (4d) models [22–24]. According to the modeling ideas, these methods can be divided into linear and nonlinear QSPR methods. Linear methods mainly include multiple regression methods (MLR), partial least squares (PLS), and principal component regression (PCR) [25]. Nonlinear methods include support vector regression (SVR) and artificial neural network (ANN) methods [26–30].

However, the QSPR study based on various artificial intelligence algorithms also has some shortcomings, such as high computational cost [31]. Therefore, it is necessary to develop a QSPR model with high accuracy, high efficiency, and good stability.

The pKa value is a key parameter of some compounds, but its determination experiments are cumbersome. Therefore, it is important to develop a pKa prediction model with high accuracy. Polanski et al. [32] proposed a model based on ANN and PLS to predict the pKa of aromatic acids and alkyl acids. Luan et al. [33] developed a model with radial basis function artificial neural network (RBF ANN) and the heuristic method (HM) and obtained the better performance in pKa prediction. These studies showed that ANN has outstanding performance in pKa prediction. However, the performance of ANN is sensitive to its parameters and training algorithm. Many artificial intelligence algorithms, including various evolutionary algorithms, are applied in ANN training. However, the evolutionary algorithm also has its own shortcomings, such as the tendency to fall into the local extreme value and a slow convergence rate in the later stage, which lead to unsatisfactory results of QSPR modeling based on the evolutionary algorithm [34–37]. In this paper, a novel QSPR model is proposed based on BP ANN and the chaos-enhanced accelerated particle swarm algorithm (APSO) reported in recent years [38]. An improved APSO is applied in the screening of molecular descriptors and the optimization of the weights of BP ANN. Then, combined with other artificial neural networks, the QSPR model is used to predict the pKa values of various compounds.

2. Modeling Theory and Methods

2.1. Chaos-Enhanced Accelerated Particle Swarm Optimization Algorithm

Particle swarm optimization (PSO) was proposed by Eberhart and Kennedy in 1995 [39], but the performance of the standard PSO algorithm was not high enough and showed some defects, such as parameter sensitivity, premature convergence, and slow local search. In recent years, a variant PSO called accelerated PSO (APSO) has attracted wide attention from scholars [38,40–42]. Although the APSO improves the convergence speed, it may also lead to premature convergence and omit some extreme values. Therefore, in this study we propose a new chaos-enhanced accelerated particle

swarm optimization algorithm (CAPSO) by integrating chaos theory into the improvement of the APSO algorithm.

In the APSO algorithm, the influence of the inertial weight factor or cognitive factor on the particle is not considered and the algorithm is only improved by the global exploration factor [43]. The main idea of the algorithm is to fully attribute the power to the variable that is responsible for global search and to consider the update of the particle with the exploration factor. In the whole search process, the particle is only constrained by the global extreme value. The position update formula is:

$$X_{i,d}^{K+1} = (1 - C_2)X_{i,d}^{K} + C_2 p_{g,d}^{K} + C_1 r \quad (1)$$

where C_1 and C_2 are learning factors; r is the random number between 0 and 1; $X_{i,d}^{K+1}$ is the position of particle i in d-dimensional k-th iteration; and $p_{g,d}^{K}$ is the position of the global extremum of the whole population in the d-th dimension.

Compared with the standard PSO algorithm, APSO adds two parameters, C_1 and C_2, to reduce the randomness in the iterative process. In this paper, C_1 represents the monotonically decreasing function: $C_1 = \delta^t$, where $0 < \delta < 1$ and t is the current iteration number. Therefore, the performance of the APSO algorithm is mainly affected by parameter C_2. For common problems, the value is (0.2,0.7). When C_2 is 1, the particle can converge at any time to the current global value and does not change any more. Moreover, the global value may not be the real global value at all. When C_2 is 0, the global search speed of the algorithm is extremely slow. Therefore, the optimization of C_2 is important in the APSO algorithm.

A chaotic system refers to a deterministic system involving random irregular movements, whose behavior is uncertain, unrepeatable, and unpredictable. In a chaotic system, when the initial conditions are slightly changed, the system will be greatly different after continuous amplification. In the process of the APSO algorithm, the value of learning factor C_2 is uncertain and unpredictable and has partial chaotic characteristics. Therefore, in order to simulate the chaotic characteristics of C_2, the classical logistic equation is used to realize the evolution of chaotic variables and optimize the parameters in this paper. The iterative formula is provided as follows:

$$X_i^{K+1} = 4X_i^K(1 - X_i^K) \quad (2)$$

when $0 < X_i^K < 1$, the logistic equation is in a completely chaotic state.

The CAPSO algorithm involves the following steps:

Step 1: To initialize the particle group. The particles in the PSO algorithm are initialized. The optimal value of the individual extremum is selected as the global optimal value to generate chaotic values;

Step 2: To calculate the adaptive value of group particles;

Step 3: The adaptive value of each particle is compared with that of the particle at the best position. If the adaptive value is better, the best position is updated;

Step 4: The learning factor C_2 is obtained from the chaotic sequence (generated by Equation (2)) and the position of the particle is updated with Equation (1);

Step 5: If the end condition of the algorithm is satisfied, the global optimum position is the optimal solution. The result is saved and the algorithm is completed. Otherwise, return to Step 2.

2.2. QSPR Model Based on the Hybrid Intelligent Method

The back propagation artificial neural network (BP ANN) is one of the most important network models. It generally consists of an input layer, hidden layer, and output layer [44–46]. The implementation of BP ANN mainly consists of two processes: a learning process and a working process [47,48].

In a three-layer BP ANN, each layer consists of several nodes. The input layer receives the input information of the network. Then, the input information is processed and sent to the hidden layer.

The relationship between the input and output can be expressed as:

$$\text{Input}: \ net = x_1 w_1 + x_2 w_2 + \ldots x_n w_n \tag{3}$$

$$\text{Output}: \ y = f(net) = \frac{1}{1 + e^{-net}} \tag{4}$$

where $x_i, x_2, \ldots x_n$ are the input vectors of the network; $w_i, w_2, \ldots w_n$ are the connection weights for each input vector; and y is the output of the network.

In the BP ANN model, the nonlinear relationship between input and output is established by determining the weight and deviation between each layer. Structurally, the nonlinear relationship between the input and output can be understood as: output $y = f(w_{ih}, w_{ho}, b_o)$, where w_{ih}, w_{ho}, b_o are, respectively, the weight vector between the input layer and the hidden layer, the weight vector between the hidden layer and the output layer, and the deviation vector of the hidden layer. The performance of the network depends on the three main parameters of the network (w_{ih}, w_{ho}, b_o).

To improve the BP algorithm, a prediction model based on CAPSO and BP ANN, called CAPSO BP ANN, is proposed based on the optimization of BP ANN parameters with the CAPSO algorithm. The CAPSO BP ANN model makes full use of the strong global search capability of the PSO algorithm and the fast local search capability of the BP algorithm, thus improving the prediction speed and accuracy of the model. In CAPSO BP ANN, the PSO algorithm is proposed to optimize BP ANN parameters (w_{ih}, w_{ho}, b_o). Therefore, in the PSO optimization algorithm, the particle is designed as the structure with weight vector w_{ih}, weight vector w_{ho}, and deviation vector b_o:

$$particle(i) = [w_{ih}, w_{ho}, b_o] \tag{5}$$

The implementation of the CAPSO BP ANN model can be simply described as follows:

Step 1: To initialize the model. The connection weights, deviations, and population parameters of the model are initialized by the random method;

Step 2: Model training. The CAPSO algorithm is used to optimize the parameters of BP ANN and the particle structure is designed.

Step 3: Parameter adjustment. Based on the error of the output, the parameters are adjusted until the number of execution times reaches the set value or the error satisfies the setting condition.

Step 4: Output. After training, the model outputs each parameter and then the trained model is tested.

2.3. Model Evaluation

The evaluation of the model is mainly based on the stability and reliability of the model [49]. In this paper, the evaluation indices of prediction accuracy including the absolute average relative deviation (*AARD*) and the root mean square error of prediction (*RMSEP*) are defined as follows:

$$AARD = \frac{1}{N} \sum_{i=1}^{N} \frac{|\overline{y}_i - y_i|}{y_i} \tag{6}$$

$$RMSEP = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \overline{y}_i)^2} \tag{7}$$

The squared correlation coefficient (R^2) reflects the correlation between predicted values and experimental values and is defined as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{N}(y_i - y_{ave})(\overline{y}_i - \overline{y}_{ave})\right]^2}{\sum_{i=1}^{N}(y_i - y_{ave})^2 \sum_{i=1}^{N}(\overline{y}_i - \overline{y}_{ave})^2} \quad (8)$$

In these formulas, N is the number of samples; y_i is the predicted or calculated value of the model; y_i is the actual value obtained in experiments; y_{ave} is the average of actual values of the samples; and $\overline{y}_{ave}$ is the average of predicted values.

3. Experimental Study

3.1. Experimental Data

The comprehensive performance of the model was verified by the prediction experiments of the pKa values of various compounds. The experimental database was obtained from References [50–52] and is shown in Table 1. Table S1 lists the compound families used for the QSPR modeling in this paper. The database consists of 268 records of data. The largest organic molecules contain up to 50 non-hydrogen atoms, eight aromatic rings, and 11 heteroatoms. In order to obtain a more reasonable prediction model, the database is randomly divided into three subsets: training set, verification set, and testing set [53]. The training set is used to establish the model. The verification set is used to optimize and validate the model. The testing set is used to test the performance of the model and the tested performance can directly reflect the comprehensive performance of the model.

Table 1. Statistical table of experimental data.

Number of Compounds	Experimental pKa	References
31	0.70–4.99	[50,51]
34	5.00–6.99	[50–52]
16	7.00–7.99	[50–52]
46	8.00–8.99	[50–52]
80	9.00–9.99	[50–52]
45	10.00–10.99	[50–52]
16	11.00–13.80	[50,51]

In this paper, 70% of the data are used for training. Both the verification set and testing set account for 15%. The numbers of the experimental data in the training set, validation set, and testing set are 188, 40, and 40, respectively.

3.2. Screening of Molecular Descriptors

The molecular descriptors are generated by the following methods:

- Construction of molecular structure. This is performed using Chemdraw UItra 7.0 software.
- Optimization of molecular structure. The molecular structure is further optimized in Hyper Chem 7.5 software.
- Calculation of molecular descriptors. The optimized molecular structure is imported into CODESSA software and the corresponding molecular descriptors are obtained by calculation.

Through molecular descriptor computing software, 733 molecular descriptors are generated and some of the molecular descriptors are closely related to each other. When modeling, it is necessary to filter a large number of calculated molecular descriptors in order to select the descriptors which are the most closely related to the research questions. The quality of the QSPR model depends on the way to determine molecular descriptors to a large extent.

In this study, the CAPSO algorithm is used to screen a large number of calculated molecular descriptors. The implementation process of filtering molecular descriptors with CAPSO is described as follows:

Step 1. Population initialization. To set the population size and initialize the population individual as a molecular descriptor; to set the number of iterations and the maximum number of iterations.

Step 2. Adaptive evaluation. To calculate the fitness of all the molecular descriptors of a population.

Step 3. Molecular descriptor selection. To select the next generation of molecular descriptors based on individual fitness values.

Step 4. Population renewal. To iterate the molecular descriptors in the population and obtain the next generation of molecular descriptor population.

Step 5. Re-evaluation of individual adaptive values. To calculate the fitness of all of the molecular descriptors of the population through iterative evolution and re-evaluate the merits and demerits of the individuals.

Step 6. Iteration. To judge whether the iteration condition is satisfied. If it is satisfied, the evolution is ended, otherwise turn to Step 3 and continue to perform the iteration.

Finally, five molecular descriptors were selected through CAPSO's search for molecular descriptors (Table 2).

Table 2. Molecular descriptors selected by the chaos-enhanced accelerated particle swarm optimization algorithm (CAPSO) algorithm.

No.	Molecular Descriptors	Descriptor Types
MD1	Relative number of N atoms	Constitutional descriptors
MD2	Randic index (order 3)	Topological descriptors
MD3	RNCG relative negative charged (QMNEG/QTMINUS) (Quantum-Chemical PC)	Electrostatic descriptors
MD4	RNCS Relative negative charged SA (SAMNEG * RNCG) (Zefirov's PC)	Electrostatic descriptors
MD5	Maximum net atomic charge	Quantum descriptors

Five molecular descriptors belonging to four types were selected by CAPSO: constitutional descriptors, topological descriptors, electrostatic descriptors, and quantum descriptors.

The relative number of N atoms (molecular descriptor 1 (MD1)) is a constitutional descriptor and usually proportional to the density of the electron cloud. When the polarity of the positive and negative charge of a molecule increases, its pKa value decreases. The relative number of N atoms can be used to characterize the composition of the molecular structure.

The Randic index (order 3) (MD2) is a topological descriptor for molecular size, shape, branching degree, and dispersion force. As the molecular dispersion increases, the molecular volume increases, leading to the decrease of pKa value. The Randic index (order 3) can represent the topological structure of molecules.

RNCG relative negative (QMNEG/QTMINUS) (quantum-chemical PC) (MD3) and RNCS relative negative charged SA (SAMNEG * RNCG) (Zefirov's PC) (MD4) are electrostatic descriptors, which depend on the distribution of the charges on the molecule. The negative coefficient of the relative negative charge is inversely proportional to the pKa value and the probability that positive ions replace protons is inversely proportional to the contact area and the pKa value of the negative atomic solvent. The relative negative charge and its surface area can be used to characterize the electrostatic parameters of molecules.

The maximum net atomic charge (No. MD5) is a quantum chemical descriptor, which is proportional to pKa and related to the largest net atom. It can be used to characterize the quantum chemical structure of molecules.

In conclusion, the molecular descriptors selected by the CAPSO algorithm can objectively characterize the molecular structure theoretically and reflect the relationship between the pKa value and the molecular structure. The CAPSO algorithm can provide a reference for the selection of molecular descriptors in all methods of QSPR modeling.

3.3. Model Structure

The CAPSO BP ANN model was established with the molecular descriptors selected by CAPSO. The CAPSO BP ANN model adopted the three-layer structure composed of the input layer, the hidden layer, and the output layer. The input layer includes five input parameters representing the selected five molecular descriptors. The input parameters are: relative number of N atoms, Randic index (order 3), RNCG relative negative charged (QMNEG/QTMINUS) (Quantum-Chemical PC), RNCS relative negative charged SA (SAMNEG * RNCG) (Zefirov's PC), and maximum net atomic charge. The output layer has one output parameter representing the corresponding pKa value.

In this paper, the number of hidden layers is estimated with the formula: $(2 \times \text{sqrt}(m \times n) + 1$, where m and n are the numbers of the nodes of the input and output layers), and then the number of optimal hidden layer neurons is determined by the heuristic method. The model in this paper contains five input nodes and one output node, so the number of hidden layer neurons is estimated to be 5. Then, we assumed that the number of the neurons of the hidden layer was tested from 3 to 15, respectively. Figure 1 shows the comparison diagram of predicted errors and the number of hidden layer neurons.

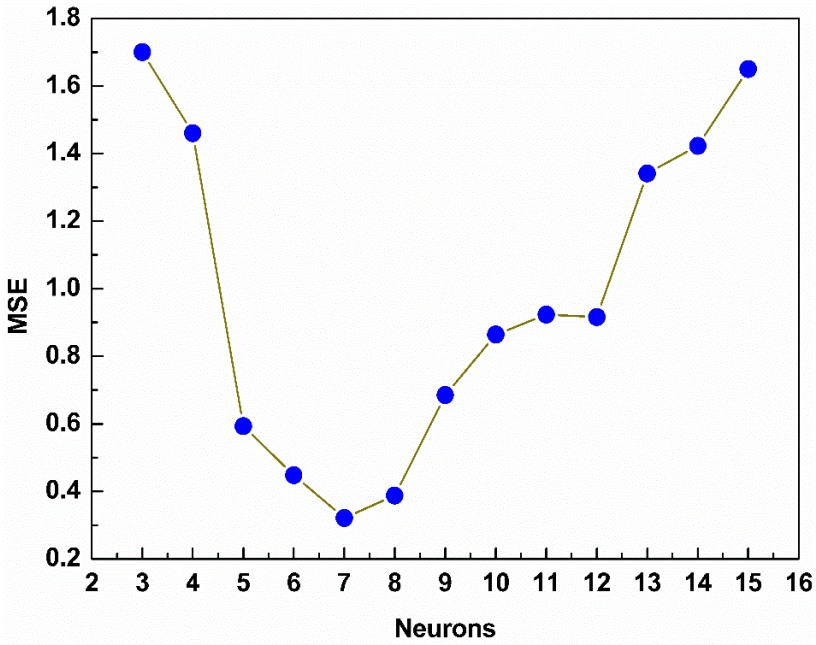

Figure 1. Optimization comparison diagram of the number of hidden layer neurons. MSE: Mean square error.

As shown in Figure 1, with the increase in the neurons of the hidden layer, the mean square error (MSE) decreases first and then increases. When the number is 7, the training MSE is the lowest and the structure of the prediction model is optimal. The model structure is 5-7-1.

4. Results and Discussion

A three-layer (5-7-1) CAPSO BP ANN prediction model was established to predict the pKa values of the compounds. MSE values were adopted as performance metrics for the model. To ensure the

generalization ability, the model was run 10 times. The optimized CAPSO BP ANN parameters used in this paper are summarized in Table 3.

Table 3. Optimized model parameters.

Parameters	Values
Training data proportion	70%
Validation data proportion	15%
Testing data proportion	15%
Training algorithm	CAPSO
Number of input neurons	5
Number of hidden neurons	7
Number of output neuron	1
Number of particles in CAPSO	50
Maximum iteration times	1000

First, 188 sets of data from the training set and 40 sets of data from the validation set were respectively used for model training and validation. Figure 2 shows the comparison between the experimental value and the predicted value in the training set and validation set, respectively. The circle and square respectively represent the predicted values of the model in the training set and the validation set. The vertical distances between the predicted data points and lines represent the absolute error of predicted values and experimental values.

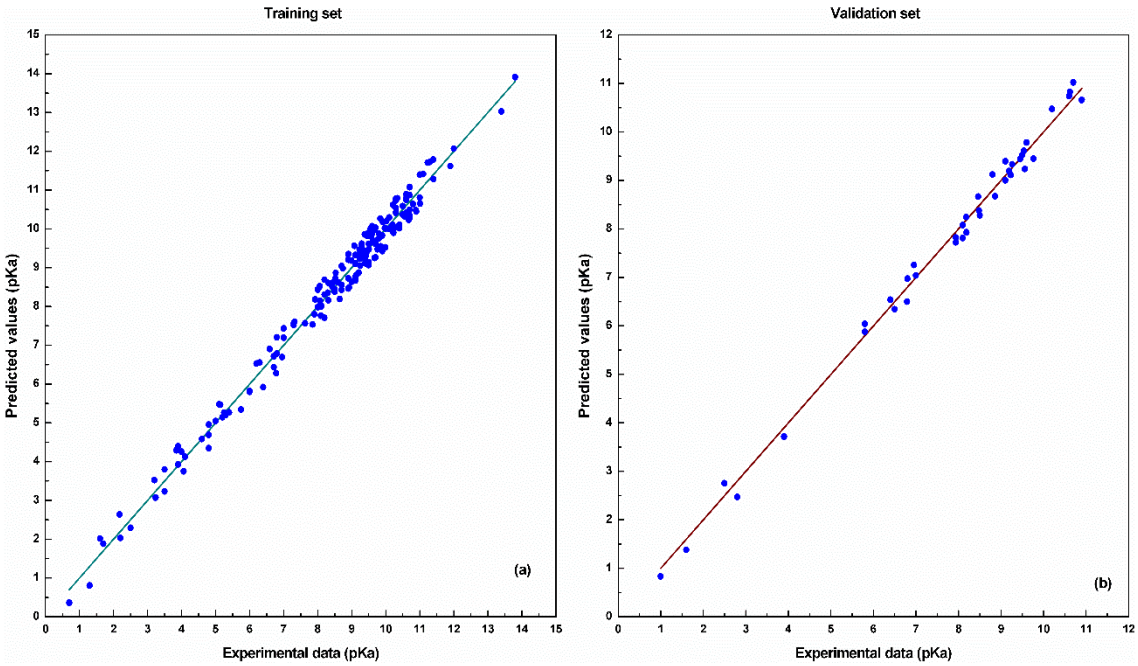

Figure 2. Comparison between the predicted and experimental values in the training and validation sets.

In the training set, the predicted value of the model training is distributed around the actual value, indicating the high coincidence degree. From the vertical distance between the prediction data points and the line, we can see that the prediction error of the model is small and that the prediction accuracy is high. In the validation set, the prediction results are significantly better than those in the training set, indicating that the training effect of the model is good.

Figure 3 shows the correlation between the actual value and the predicted value of the model in the testing set. In the testing set, the predicted value of the model is also consistent with the actual value. Table 4 shows the results of the model in the training set, validation set, and testing set.

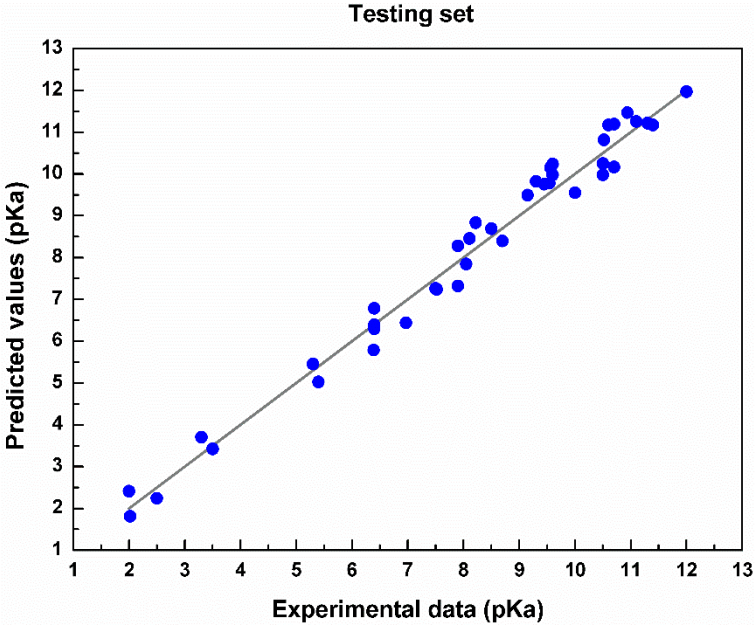

Figure 3. Comparison between the predicted and experimental values in the testing set.

Table 4. Statistics of the model prediction performance. *AARD*: Absolute average relative deviation. *RMSEP*: Root mean square error of prediction. R^2: Squared correlation coefficient.

Sets	*AARD*	*RMSEP*	R^2
Training	0.3436	0.0335	0.9771
Validation	0.3101	0.0211	0.9886
Testing	0.5364	0.0632	0.9438

The prediction results of the model in each subset are good and the prediction error is small, indicating the better comprehensive performance. The prediction performance of the model is better in terms of prediction accuracy and correlation. The above results prove that the prediction performance of the model is excellent.

In this paper, the partial derivative (PaD) method [13,54] was adopted to assess the sensitivity of the output against slight changes of the five molecular descriptors in the inputs. Figure 4 shows the contributions of the five input variables (five molecular descriptors).

Quantitatively, the Randic index (order 3) (MD2) contributes the most; the relative number of N atoms (MD1) and maximum net atomic charge (MD5) contribute roughly the same proportion (about 20%). The contributions of RNCG relative negative charged (QMNEG/QTMINUS) (Quantum-Chemical PC) (MD3) and RNCS relative negative charged SA (SAMNEG * RNCG) (Zefirov's PC) (MD4) are relatively small, but they all belong to the electrostatic descriptors. Among the four types of descriptors, electrostatic descriptors contribute the most, followed by topological descriptors and constitutional descriptors, and the quantum descriptors contribute the least (Figure 4).

Moreover, three artificial intelligence models, BP ANN, SVM, and PSO BP ANN, were selected as the comparison models. In addition, Jensen et al. [50] used PM6, PM7, PM3, AM1, and DFTB3 methods to predict the pKa values of some amine groups and indicated that PM3/COSMO was the best pKa prediction method. Therefore, in order to verify the performance of the model, the PM3/COSMO model [50] was selected as the comparison model in the study. Figure 5 shows the correlation and residual curve between the experimental values and the predicted values of each model in the testing set.

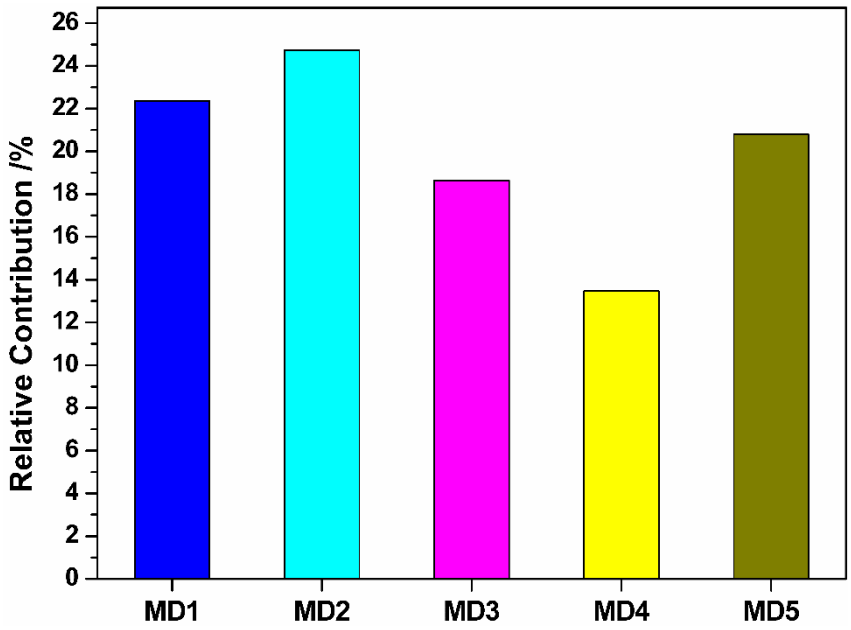

Figure 4. Contributions of the five molecular descriptors.

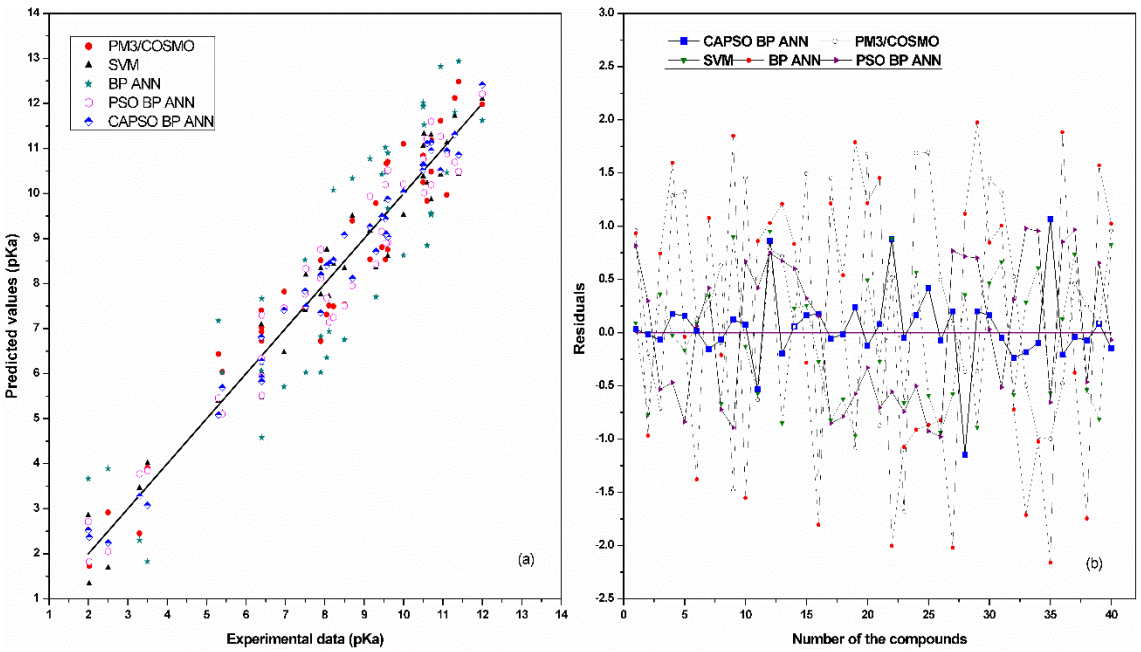

Figure 5. Comparison of the testing results of each model.

As shown in Figure 5a, the vertical distance between the prediction data points and the experimental data indicates that the prediction data of CAPSO BP ANN model are close to the experimental values. The prediction performance of the method proposed in this study is better than that of other methods. It can be seen from the residual curve that the error of the model proposed in this study is close to 0 (Figure 5b). Apart from some prediction points that have large errors, the prediction errors are generally smaller than those of other comparison models. Table 5 shows the evaluation results of each model.

Table 5. Statistical results of each model. *AARD*: Absolute average relative deviation. *RMSEP*: Root mean square error of prediction. R^2: Squared correlation coefficient.

Models	*AARD*	*RMSEP*	R^2
PM3/COSMO	0.8724	0.1439	0.8346
SVM	0.7333	0.1038	0.8863
BP ANN	1.2134	0.5354	0.6958
PSO BP ANN	0.7229	0.1029	0.8872
CAPSO BP ANN	0.5364	0.0632	0.9438

In order to verify the performance of each comparison model, the confidence interval (*C.I.*) of RMSEP in the testing set was calculated [49–55] (Table 6).

Table 6. Confidence intervals of *RMSEP* for each model. *C.I.*: Confidence interval.

Models	*C.I.* (90%)	*C.I.* (95%)	*C.I.* (99%)
PM3/COSMO	(0.04721, 0.24059)	(0.02904, 0.25876)	(0.00067, 0.29450)
SVM	(0.03368, 0.17393)	(0.02050, 0.18710)	(0.00054, 0.21303)
BP ANN	(0.39482, 0.67598)	(0.36841, 0.70240)	(0.31644, 0.75437)
PSO BP ANN	(0.03327, 0.17253)	(0.02019, 0.18561)	(0.00056, 0.21135)
CAPSO BP ANN	(0.05066, 0.07574)	(0.04830, 0.07810)	(0.04367, 0.08273)

Table 5 shows that the accuracy and relevance of the CAPSO BP ANN model have obvious advantages, including the lowest *RMSEP* and the highest R^2. The performances of PM3/COSMO and SVM are equivalent to that of the PSO BP ANN model. From Table 6, it can be observed that the CAPSO BP ANN model has the narrowest *C.I.*, 90%, 95%, or 99%. From the tables, we can see that the CAPSO BP ANN model with the lowest *RMSEP* and the narrowest *C.I.* is superior to other models.

To verify the stability and robustness of the models, an applicability domain study was proposed, as shown in Figure 6. The critical leverage is 0.213. The CAPSO BP ANN model has eight outliers (four outliers from the training set, two outliers from validation set, and two outliers from the testing set) and six influential values. The PSO BP ANN model has 11 outliers, including six outliers from training sets, three outliers from the verification set, and one outlier from the testing set. The SVM model has nine outliers and seven influential values, while the BP ANN model has 10 outliers and six influential values. All of the other values are within the applicability domain. Although all models show good performance, the CAPSO BP ANN proved its superiority, and the highest number of its respective observations was found to be within the warning limits of the defined applicability domain.

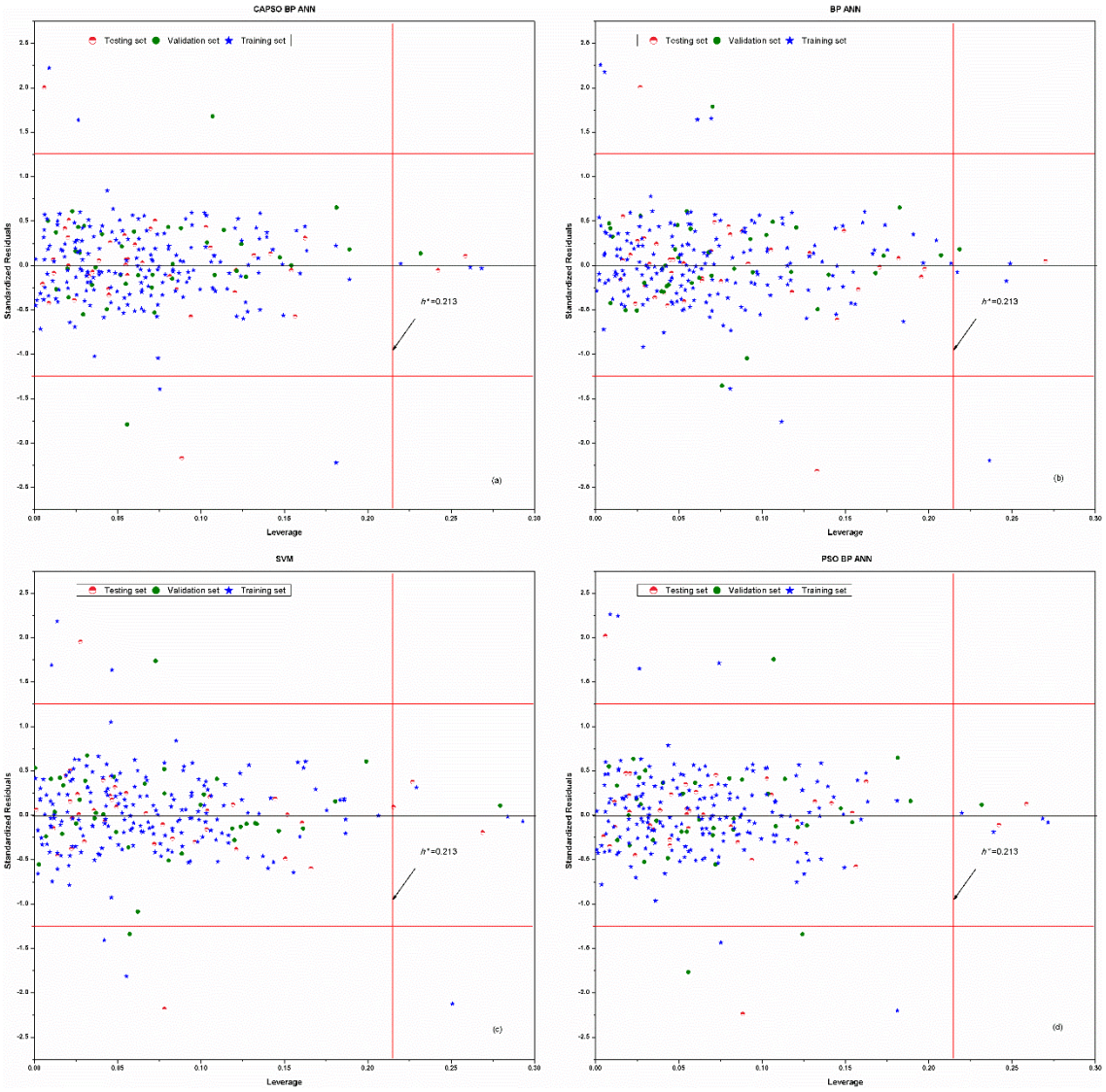

Figure 6. Applicability domain of each model.

5. Conclusions

In this study, in order to solve the problem of molecular descriptor selection and model establishment in QSPR research, a novel chaos-enhanced accelerated particle swarm optimization algorithm (CAPSO) was proposed. The algorithm was applied in the selection of molecular descriptors and QSPR modeling, and a prediction model called CAPSO BP ANN was obtained. Through the prediction experiment of the pKa values of compounds, the conclusions are drawn as follows:

The CAPSO algorithm could be applied in the selection of molecular descriptors. Prediction experiments showed that the five molecular descriptors selected by the CAPSO algorithm could well represent the molecular structures of various compounds in the prediction of the pKa value and provide the basis for the selection of molecular descriptors.

The CAPSO BP ANN model based on the PSO algorithm and BP ANN exhibited good performance in the prediction experiment of the pKa values of various compounds and achieved a higher prediction accuracy and correlation. The experimental results showed that the CAPSO BP ANN model could provide the basis for QSPR modeling.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-3417/8/7/1121/s1, Table S1: The experimental compounds in this paper.

Author Contributions: M.L. conceived and designed the experiments. M.L., B.C., and H.Z. wrote the main manuscript text. Y.W., L.L., and L.G. analyzed the data. All authors read and approved the final manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors gratefully acknowledge the support from the National Natural Science Foundation of China (grant numbers: 51663001, 51463015, and 61741103).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cumming, J.G.; Davis, A.M.; Muresan, S.; Haeberlein, M.; Chen, H.M. Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discov.* **2013**, *12*, 948–962. [CrossRef] [PubMed]
2. Rojas, C.; Ballabio, D.; Consonni, V.; Tripaldi, P.; Mauri, A.; Todeschini, R. Quantitative structure-activity relationships to predict sweet and non-sweet tastes. *Theor. Chem. Acc.* **2016**, *135*, 1–13. [CrossRef]
3. Patel, M.; Chilton, M.L.; Sartini, A.; Gibson, L.; Barber, C.; Covey-Crump, L.; Przybylak, K.R.; Cronin, M.T.D.; Madden, J.C. Assessment and reproducibility of quantitative structure-activity relationship models by the nonexpert. *J. Chem. Inf. Model.* **2018**, *58*, 673–682. [CrossRef] [PubMed]
4. Liu, C.M.; Dou, X.W.; Zhang, L.; Kong, W.J.; Wu, L.; Duan, Y.P.; Yang, M.H. Development of a broad-specificity antibody-based immunoassay for triazines in ginger and the quantitative structure-activity relationship study of cross-reactive molecules by molecular modeling. *Anal. Chim. Acta* **2018**, *1012*, 90–99. [CrossRef] [PubMed]
5. Gebreyohannes, S.; Dadmohammadi, Y.; Neely, B.J.; Gasem, K.A.M. A comparative study of QSPR generalized activity coefficient model parameters for vapor-liquid equilibrium mixtures. *Ind. Eng. Chem. Res.* **2016**, *55*, 1102–1116. [CrossRef]
6. Dardonville, C.; Caine, B.A.; de la Fuente, M.N.; Herranz, G.M.; Mariblanca, B.C.; Popelier, P.L.A. Substituent effects on the basicity (pKa) of aryl guanidines and 2-(arylimino) imidazolidines: Correlations of pH-metric and UV-metric values with predictions from gas-phase ab initio bond lengths. *New J. Chem.* **2017**, *41*, 11016–11028. [CrossRef]
7. Balabin, R.M.; Smirnov, S.V. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. *Anal. Chim. Acta* **2011**, *692*, 63–72. [CrossRef] [PubMed]
8. Xiaobo, Z.; Jiewen, Z.; Povey, M.J.W.; Holmes, M.; Hanpin, M. Variables selection methods in near-infrared spectroscopy. *Anal. Chim. Acta* **2010**, *667*, 14–32. [CrossRef] [PubMed]
9. Goodarzi, M.; Dejaegher, B.; Vander Heyden, Y. Feature selection methods in QSAR studies. *J. AOAC Int.* **2012**, *95*, 636–651. [CrossRef] [PubMed]
10. Zuvela, P.; Liu, J.J.; Macur, K.; Baczek, T. Molecular descriptor subset selection in theoretical peptide quantitative structure-retention relationship model development using nature-inspired optimization algorithms. *Anal. Chem.* **2015**, *87*, 9876–9883. [CrossRef] [PubMed]
11. Zuvela, P.; Liu, J.J. On feature selection for supervised learning problems involving high-dimensional analytical information. *RSC Adv.* **2016**, *6*, 82801–82809. [CrossRef]
12. Heberger, K.; Skrbic, B. Ranking and similarity for quantitative structure-retention relationship models in predicting lee retention indices of polycyclic aromatic hydrocarbons. *Anal. Chim. Acta* **2012**, *716*, 92–100. [CrossRef] [PubMed]
13. Zuvela, P.; David, J.; Wong, M.W. Interpretation of ANN-based QSAR models for prediction of antioxidant activity of flavonoids. *J. Comput. Chem.* **2018**, *39*, 953–963. [CrossRef] [PubMed]
14. Zuvela, P.; Macur, K.; Liu, J.J.; Baczek, T. Exploiting non-linear relationships between retention time and molecular structure of peptides originating from proteomes and comparing three multivariate approaches. *J. Pharm. Biomed.* **2016**, *127*, 94–100. [CrossRef] [PubMed]
15. Shi, J.; Hu, X.; Zou, X.; Zhao, J.; Zhang, W.; Huang, X.; Zhu, Y.; Li, Z.; Xu, Y. A heuristic and parallel simulated annealing algorithm for variable selection in near-infrared spectroscopy analysis. *J. Chemom.* **2016**, *30*, 442–450. [CrossRef]
16. Pandit, A.; Sengupta, S.; Krishnan, M.A.; Reddy, R.B.; Sharma, R.; Venkatesh, C. First report on 3D-QSAR and molecular dynamics based docking studies of GCPII inhibitors for targeted drug delivery applications. *J. Mol. Struct.* **2018**, *1159*, 179–192. [CrossRef]

17. Shahlaei, M. Descriptor selection methods in quantitative structure-activity relationship studies: A review study. *Chem. Rev.* **2013**, *113*, 8093–8103. [CrossRef] [PubMed]
18. Barley, M.H.; Turner, N.J.; Goodacre, R. Improved descriptors for the quantitative structure-activity relationship modeling of peptides and proteins. *J. Chem. Inf. Model.* **2018**, *58*, 234–243. [CrossRef] [PubMed]
19. Soper-Hopper, M.T.; Petrov, A.S.; Howard, J.N.; Yu, S.S.; Forsythe, J.G.; Grover, M.A.; Fernandez, F.M. Collision cross section predictions using 2-dimensional molecular descriptors. *Chem. Commun.* **2017**, *53*, 7624–7627. [CrossRef] [PubMed]
20. Khajeh, A.; Modarress, H.; Zeinoddini-Meymand, H. Application of modified particle swarm optimization as an efficient variable selection strategy in QSAR/QSPR studies. *J. Chemom.* **2012**, *26*, 598–603. [CrossRef]
21. Li, M.S.; Liu, L.; Huang, X.Y.; Liu, H.S.; Chen, B.S.; Guan, L.X.; Wu, Y. Prediction of supercritical carbon dioxide solubility in polymers based on hybrid artificial intelligence method integrated with the diffusion theory. *RSC Adv.* **2017**, *7*, 49817–49827. [CrossRef]
22. Liu, Q.Z.; Wang, S.S.; Li, X.; Zhao, X.Y.; Li, K.; Lv, G.C.; Qiu, L.; Lin, J.G. 3D-QSAR, molecular docking, and oniom studies on the structure-activity relationships and action mechanism of nitrogen-containing bisphosphonates. *Chem. Biol. Drug Des.* **2018**, *91*, 735–746. [CrossRef] [PubMed]
23. Wang, N.N.; Dong, J.; Deng, Y.H.; Zhu, M.F.; Wen, M.; Yao, Z.J.; Lu, A.P.; Wang, J.B.; Cao, D.S. ADME properties evaluation in drug discovery: Prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting. *J. Chem. Inf. Model.* **2016**, *56*, 763–773. [CrossRef] [PubMed]
24. Fujita, T.; Winkler, D.A. Understanding the roles of the "two QSARs". *J. Chem. Inf. Model.* **2016**, *56*, 269–274. [CrossRef] [PubMed]
25. Borisek, J.; Drgan, V.; Minovski, N.; Novic, M. Mechanistic interpretation of artificial neural network-based QSAR model for prediction of cathepsin K inhibition potency. *J. Chemom.* **2014**, *28*, 272–281. [CrossRef]
26. Du, X.J.; Wang, J.L.; Jegatheesan, V.; Shi, G.H. Dissolved oxygen control in activated sludge process using a neural network-based adaptive PID algorithm. *Appl. Sci.* **2018**, *8*, 261. [CrossRef]
27. Verma, R.P.; Matthews, E.J. Estimation of the chemical-induced eye injury using a weight-of-evidence (WoE) battery of 21 artificial neural network (ANN) c-QSAR models (QSAR-21): Part I: Irritation potential. *Regul. Toxicol. Pharm.* **2015**, *71*, 318–330. [CrossRef] [PubMed]
28. Yasrab, R.; Gu, N.J.; Zhang, X.C. An encoder-decoder based convolution neural network (CNN) for future advanced driver assistance system (ADAS). *Appl. Sci.* **2017**, *7*, 312. [CrossRef]
29. Selzer, D.; Neumann, D.; Schaefer, U.F. Mathematical models for dermal drug absorption. *Expert Opin. Drug Metab. Toxicol.* **2015**, *11*, 1567–1583. [CrossRef] [PubMed]
30. Hassanzadeh, Z.; Kompany-Zareh, M.; Ghavami, R.; Gholami, S.; Malek-Khatabi, A. Combining radial basis function neural network with genetic algorithm to QSPR modeling of adsorption on multi-walled carbon nanotubes surface. *J. Mol. Struct.* **2015**, *1098*, 191–198. [CrossRef]
31. Dolara, A.; Grimaccia, F.; Leva, S.; Mussetta, M.; Ogliari, E. Comparison of training approaches for photovoltaic forecasts by means of machine learning. *Appl. Sci.* **2018**, *8*, 228. [CrossRef]
32. Polanski, J.; Walczak, B. The comparative molecular surface analysis (COMSA): A novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615–625. [CrossRef]
33. Luan, F.; Xue, C.X.; Zhang, R.S.; Zhao, C.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Prediction of retention time of a variety of volatile organic compounds based on the heuristic method and support vector machine. *Anal. Chim. Acta* **2005**, *537*, 101–110. [CrossRef]
34. Li, M.; Huang, X.; Liu, H.; Liu, B.; Wu, Y.; Wang, L. Solubility prediction of supercritical carbon dioxide in 10 polymers using radial basis function artificial neural network based on chaotic self-adaptive particle swarm optimization and k-harmonic means. *RSC Adv.* **2015**, *5*, 45520–45527. [CrossRef]
35. Li, M.S.; Huang, X.Y.; Liu, H.S.; Liu, B.X.; Wu, Y.; Xiong, A.H.; Dong, T.W. Prediction of gas solubility in polymers by back propagation artificial neural network based on self-adaptive particle swarm optimization algorithm and chaos theory. *Fluid Phase Equilibr.* **2013**, *356*, 11–17. [CrossRef]
36. Azad, F.N.; Ghaedi, M.; Asfaram, A.; Jamshidi, A.; Hassani, G.; Goudarzi, A.; Azqhandi, M.H.A.; Ghaedi, A. Optimization of the process parameters for the adsorption of ternary dyes by ni doped FEO(OH)-NWs-AC using response surface methodology and an artificial neural network. *RSC Adv.* **2016**, *6*, 19768–19779. [CrossRef]
37. Li, M.; Wu, W.; Chen, B.; Wu, Y.; Huang, X. Solubility prediction of gases in polymers based on an artificial neural network: A review. *RSC Adv.* **2017**, *7*, 35274–35282. [CrossRef]

38. Gandomi, A.H.; Yun, G.J.; Yang, X.S.; Talatahari, S. Chaos-enhanced accelerated particle swarm optimization. *Commun. Nonlinear Sci.* **2013**, *18*, 327–340. [CrossRef]
39. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the 1995 IEEE International Conference on Neural Networks—ICNN'95, Perth, Western Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948.
40. Wang, F.; Zhou, L.D.; Wang, B.; Wang, Z.; Shafie-Khah, M.; Catalao, J.P.S. Modified chaos particle swarm optimization-based optimized operation model for stand-alone CCHP microgrid. *Appl. Sci.* **2017**, *7*, 754. [CrossRef]
41. Liang, C.H.; Tong, X.M.; Lei, T.Y.; Li, Z.X.; Wu, G.S. Optimal design of an air-to-air heat exchanger with cross-corrugated triangular ducts by using a particle swarm optimization algorithm. *Appl. Sci.* **2017**, *7*, 554. [CrossRef]
42. Jiang, G.W.; Luo, M.Z.; Bai, K.Q.; Chen, S.X. A precise positioning method for a puncture robot based on a PSO-optimized BP neural network algorithm. *Appl. Sci.* **2017**, *7*, 969. [CrossRef]
43. Yang, X.S.; Deb, S.; Fong, S. Accelerated particle swarm optimization and support vector machine for business optimization and applications. In Proceedings of the Networked Digital Technologies, Macau, China, 11–13 July 2011; Volume 136, pp. 53–66.
44. Han, F.; Zhu, J.S. Improved particle swarm optimization combined with backpropagation for feedforward neural networks. *Int. J. Intell. Syst.* **2013**, *28*, 271–288. [CrossRef]
45. Li, M.S.; Zhang, H.J.; Chen, B.S.; Wu, Y.; Guan, L.X. Prediction of pKa values for neutral and basic drugs based on hybrid artificial intelligence methods. *Sci. Rep.* **2018**, *8*. [CrossRef] [PubMed]
46. Zolfaghari, S.; Noor, S.B.M.; Mehrjou, M.R.; Marhaban, M.H.; Mariun, N. Broken rotor bar fault detection and classification using wavelet packet signature analysis based on fourier transform and multi-layer perceptron neural network. *Appl. Sci.* **2018**, *8*, 25. [CrossRef]
47. Valdez, F.; Melin, P.; Castillo, O. Modular neural networks architecture optimization with a new nature inspired method using a fuzzy combination of particle swarm optimization and genetic algorithms. *Inf. Sci.* **2014**, *270*, 143–153. [CrossRef]
48. Li, N.J.; Wang, W.J.; Hsu, C.C.J.; Chang, W.; Chou, H.G.; Chang, J.W. Enhanced particle swarm optimizer incorporating a weighted particle. *Neurocomputing* **2014**, *124*, 218–227. [CrossRef]
49. SAMUELS, M.L. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **1981**, *68*, 589–599.
50. Jensen, J.H.; Swain, C.J.; Olsen, L. Prediction of pKa values for druglike molecules using semiempirical quantum chemical methods. *J. Phys. Chem. A* **2017**, *121*, 699–707. [CrossRef] [PubMed]
51. Eckert, F.; Klamt, A. Accurate prediction of basicity in aqueous solution with COSMO-RS. *J. Comput. Chem.* **2006**, *27*, 11–19. [CrossRef] [PubMed]
52. Luan, F.; Ma, W.P.; Zhang, H.X.; Zhang, X.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Prediction of pKa for neutral and basic drugs based on radial basis function neural networks and the heuristic method. *Pharm. Res.* **2005**, *22*, 1454–1460. [CrossRef] [PubMed]
53. Kennard, R.W.; Stone, L.A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137–148. [CrossRef]
54. Gevrey, M.; Dimopoulos, I.; Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.* **2003**, *160*, 249–264. [CrossRef]
55. Leifeld, P.; Cranmer, S.J.; Desmarais, B.A. Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. *J. Stat. Softw.* **2018**, *83*. [CrossRef]