Chronic Diseases Prevention Review
22 (2022) 1-7
Available at http:// www.cancercellresearch.org
ISSN 2161-2609

QSAR model study on antibacterial activity of aromatic nucleus-linked symmetric membrane surfactants against Staphylococcus aureus

Guangran Fu[1], Anling Zhang[2] *
[1]Department of Oncology, The Affiliated Hospital of Qingdao University, Qingdao,Shandong 266000
[2]Modern Educational Technology Center, Qingdao University, Qingdao, 266071,Shandong, China

Abstract: Due to the serious resistance of bacteria such as Staphylococcus aureus, which have posed a greater threat to all aspects of human beings, it is urgent to develop new antibacterial agents. The aim of our research is to predict the MIC value of new antibacterial agents to establish a highly predicting quantitative structure activity relationship model. Genetic algorithms genetic programming (GEP) was used to establish a quantitative structure-activity relationship (QSAR) model of 33 aromatic nucleated symmetric membrane active agents. QSARs are an effective method for screening new structures and predicting the various properties of synthetic compounds. We used heuristic methods (HM) and gene expression programming (GEP) algorithms to establish linear and nonlinear models, respectively. The results showed that The square of the correlation coefficient of the heuristic method is 0.57, and the S2 is 0.11. In gene expression programming, the square of correlation coefficient and the mean square error for the training set are 0.68 and 0.14, respectively. The square of correlation coefficient and the mean square error for the test set are 0.64 and 0.11, respectively. The nonlinear models are more satisfaction. The study indicates that this kind of compound has the possibility of further optimization.

Keywords: Antibacterial activity; Staphylococcus aureus; Heuristic method; Quantitative structural activity relationships; Gene expression programming
Received 18 March 2022, Accepted 19 March 2022

1. Introduction

Since the discovery of the first antibiotic, antimicrobial resistance (AMR) has been observed, and antimicrobial abuse and misuse have led to increased levels of clinical resistance[1]. Antimicrobial resistance is a multi-layered problem with catastrophic impacts on humans, livestock, the environment and the biosphere[2,3]. This imminent international crisis requires global attention and commitment to envisage and implement solutions. The World Health Organization has released a report on the crisis of antibiotic resistance and encouraged global leaders to take action. But in recent years, various solutions have not achieved significant results. The frequency of infections caused by antibiotic-resistant bacteria are increasing and leading to significant morbidity and mortality. New antimicrobials are in great need to treat infections that are resistant to currently available agents[4, 5]. Staphylococcus aureus is the most common clinical pathogen, which is community-acquired or hospital-acquired infections[6]. Since the discovery of methoxycillin-resistant staphylococcus aureus, the infection caused by it has quickly spread throughout the sphere. Traditional drugs inhibit bacterial growth by interfering with metabolic pathways. However, these drugs appear to be susceptible to the development of bacterial resistance[7]. To overcome bacterial resistance, there is an urgent need for antimicrobial agents with new mechanisms of action or structures. Membrane-targeting compounds are considered promising future antibiotics. The bacterial cell membrane has become an attractive target since its essential and highly conservative structure have been key challenges to resistance mechanisms[8, 9].

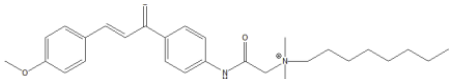
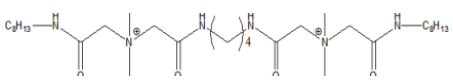
It is crucial to assess the activity of antimicrobial agents. But obtaining new antimicrobials through experiments and proving their activity and efficacy is a huge project that requires a lot of time, effort and money. Fortunately, advanced computer technology is a highly available strategy that can obtain new structures and predict security risks. Quantitative Structure-Activity Relationship (QSAR) has always been an effective method for screening new structures and predicting various properties of synthetic compounds. In recent decades, QSARs have been applied to database analysis to establish a quantitative relationship model between molecular multivariate structural parameters and experimental properties[10, 11]. In our study, we analyzed the 33 molecular structures in literature[12]. After that, we used the heuristic method (HM) to search for all computed molecular descriptors and remove unimportant variables. We used the stepwise regression method in HM, establishing a multivariate linear regression equation with selected descriptors. However, the MIC value of the antibacterial agents is influenced by a variety of factors, and most of the biological data is not linear, so we still need to build nonlinear models to predict more accurately.

2.Methods

2.1Date Set

We collected 33 compounds from Chu Wenchao' article and put them into training set and test set. The training set consists of 23 items, and the test set consists

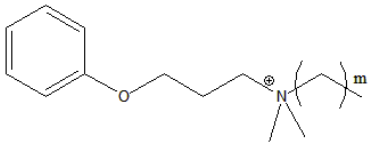
Table 1. Experimental and calculated log(MIC) of 33 compounds (heuristic method (HM) and GEP).

No.	Compound	Exp. log(MIC)	Calculate HM Pred.a	HM Resid.b	GEP Pred.a	GEP Resid.b
1		0	0.327	0.327	0.111	0.111
2		0.301	0.147	-0.154	0.140	0.161
3		0	0.279	0.279	-0.239	0.239

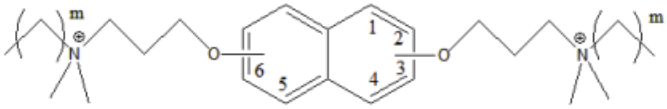
Compound No.	m	n	Exp. log(MIC)	Calculate HM Pred.a	HM Resid.b	GEP Pred.a	GEP Resid.b
4	5	2	1.505	0.886	-0.619	0.583	0.922
5	7	2	-0.301	0.039	0.340	-0.014	0.287
6	9	2	-0.301	-0.159	0.142	-0.118	0.183
7	11	2	0	-0.223	-0.223	-0.068	0.068
8	5	3	1.204	0.751	-0.453	0.463	0.741
9	7	3	0	0.027	0.027	0.307	0.307
10	9	3	-0.602	-0.310	0.292	0.021	0.623
11	11	3	0	-0.318	-0.318	0.012	0.012
12	5	4	0.301	0.730	0.429	0.556	0.255
13	7	4	0	-0.265	-0.265	-0.236	0.236
14	9	4	-0.301	-0.456	-0.155	-0.360	0.059
15	11	4	0	0.098	0.098	0.407	0.407
16	5	5	0.301	0.483	0.182	0.097	0.204
17	7	5	-0.602	-0.220	0.382	-0.085	0.517
18	9	5	-0.301	-0.336	-0.035	-0.116	0.185
19	11	5	0.301	-0.465	-0.766	0.047	0.254
20	8	4	-0.301	-0.305	-0.004	-0.265	0.036

22 (2022) 1-7
ISSN 2161-2609

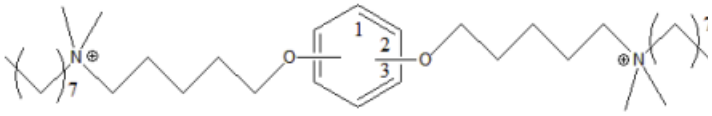
21	6	6	-0.301	-0.229	0.072	-0.158	0.143



Compound No.	m	Exp. log(MIC)	Calculate HM Pred.a	HM Resid.b	GEP Pred.a	GEP Resid.b
22	5	1.806	1.715	-0.091	1.560	0.246
23	7	1.204	0.795	-0.409	1.388	0.184
24	9	0.301	0.520	0.219	1.044	0.743
25	11	-0.301	0.151	0.452	0.357	0.658



Compound No.	diether	m	Exp. log(MIC)	Calculate HM Pred.a	HM Resid.b	GEP Pred.a	GEP Resid.b
26	1,6	5	0.903	0.856	-0.047	0.609	0.294
27	1,6	7	-0.301	-0.085	0.216	0.039	0.340
28	1,6	9	-0.301	0.168	0.469	0.261	0.562
29	2,3	7	0	-0.183	-0.183	-0.083	0.083
30	2,3	9	0	-0.293	-0.293	0.018	0.018
31	1,5	7	-0.301	-0.140	0.161	-0.150	0.151



Compound No.	diether	Exp. log(MIC)	Calculate HM Pred.a	HM Resid.b	GEP Pred.a	GEP Resid.b
32	1,3	-0.301	-0.393	-0.092	-0.307	0.006
33	1,2	0	0.023	0.023	0.097	0.097

of 10 items. The former set was used to build models, while the latter set was used to test models' reliability and stability (Table 1).

aThe predicted log（MIC）.

bResidue = Pred. - Exp.

2.2Calculation of the Descriptors

We plotted the structures in the ChemDraw application and optimized them in the Hyperchem program[13,14], using MM+ molecular mechanical force fields and semi-empirical AM1 methods[15]. Subsequently, we used MOPAC software to obtain three files formats.

2.3Development of Linear Model

By using CODESSA software, we got about 420 descriptors for each compound. However, before building a mathematical model, we must choose the appropriate descriptor. Screening descriptors should follow the principles[16]: (a) Remove the "0" variables, (b) remove highly correlated variables because mutual inclusion in the compounds leads to worse results, and (c) remove variables that are less relevant to other variables. By us-

22 (2022) 1-7
ISSN 2161-2609

ing the HM method, four descriptors are finally selected and subsequently used to construct an optimal multivariate linear regression model. The process is based on a gradual increase in descriptor, the accuracy of which is guaranteed by the maximum regression coefficient (R2) value, the interaction test coefficient (R2cv), and the F test (F) value[17].

2.4. Development of Nonlinear Model by the GEP Algorithm

GEP is an adaptive evolutionary algorithm based on the structure and function of genes[18]. Compared with other machine learning methods, it has significant improvements in data processing and generalization capabilities. The GEP algorithm consists of the following steps (Figure 1): Randomly generate a certain number of chromosomal individuals (initial population), express chromosomes, calculate fitness, select individuals for genetic manipulation and obtain offspring with new characteristics. The process should be repeated for several generations before finding the ideal solution.

The GEP algorithm consists of 5 processes:

(1) Initialization: Set the evolutionary algebra counter and the maximum evolutionary algebra and generate individuals as the initial population randomly.

(2) Computer fitness: The fitness of each individual in the initial population is calculated.

(3) Selection: Selection is used to determine which individuals are recombined or hybridized, and how many offspring the selected individuals will produce. According to the above adaptations, the individual choice of parents is made. The following algorithms can be selected: Roulette Selection, Random Traversal Sampling, Local Selection, Truncated Selection, Tournament choose.

(4)Hybridization: Genetic recombination is a combination of information from the parental mating population to produce new individuals. According to different expression methods, each code is divided into the following types: real-valued recombination, discrete recombination, intermediate recombination, linear recombination, extended linear recombination.

(5)Variation: The variation of the offspring after hybridization is actually a genetic change in the offspring caused by a small probability of perturbation. Different expressions of individual encodings use the following algorithms: real-value mutations and binary mutations.

The GEP algorithm consists of 5 steps: selection of coding scheme, fitness function and function set (+,-,*, /), etc., control parameter selection, genetic operator design and terminal standard settings. All of these processes are listed in Figure 1.

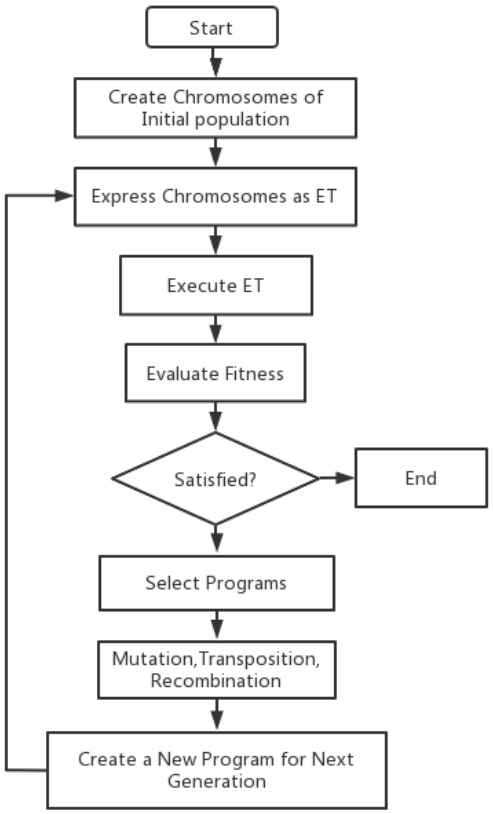


Figure 1. Flowchart of GEP algorithm.

3.Results

3.1. The Results of HM

The CODESSA software calculated about 420 descriptors for each compound. To find the most suitable combination of descriptors, a multivariate linear regression equation with numbers from 1 to 10 is calculated. The values for R2, R2cv, and S2 are 0.72, 0.58, and 0.11, respectively, and they assess the predictive power of the model, as shown in Figure 2. Finally, it was found that the model with 4 descriptors was the most suitable, as detailed in Table 2. The results show that the number of descriptors increases, R2, R2cv gradually increases, and S2 gradually decreases.

22 (2022) 1-7
ISSN 2161-2609

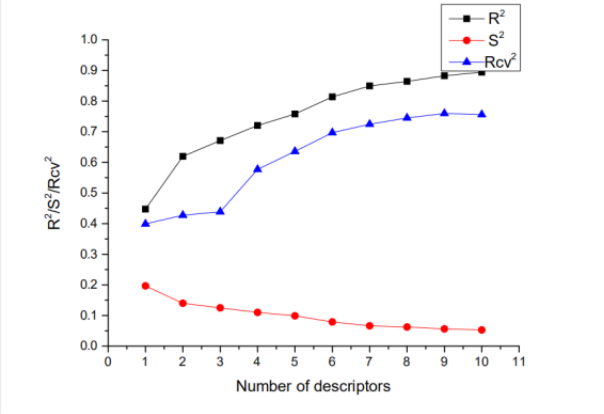


Figure 2. Influence of the number of descriptors on the correlation coefficient, the cross-validation correlation coefficient, and the square error.

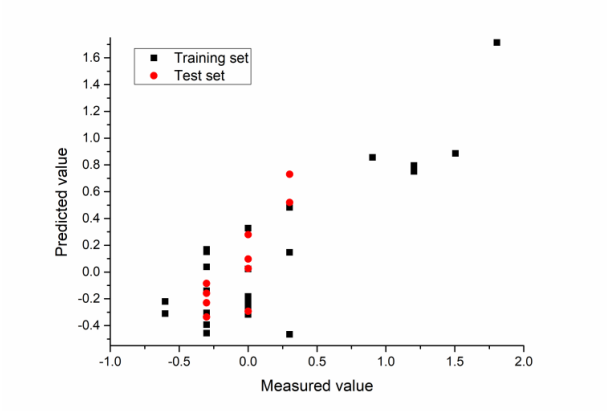


Figure 3.The plot of measured and calculated log(EC50) by HM

Table 2. The selected molecular descriptors, and their physical-chemical meaning, coefficient, and T-test.

Symbol	Physical-chemical Meaning	Coefficient	T-test
AERIC	Avg electroph. react. index for a C atom	1.78E+03	7.5137
MPCC	Min partial charge for a C atom [Zefirov's PC]	4.99E+04	4.8274
1XGP	1X GAMMA polarizability (DIP)	4.87E-06	2.3847
W3WP	WNSA-3 Weighted PNSA (PNSA3*TMSA/1000) [Zefirov's PC]	-2.64E-01	-4.364

Table 3. Correlation matrix of the 4 descriptors

-	AERIC	MPCC	1XGP	W3WP
AERIC	1	-0.4925	-0.0641	-0.1811
MPCC	-	1	-0.2039	0.0918
1XGP	-	-	1	-0.4074
W3WP	-		-	1

Since regression equations are obtained using stepwise regression, there is usually no multicollinearity between the obtained molecular descriptors. However, for the accuracy of the results, the correlation coefficient calculation between descriptors is shown in Table 3. Any two molecular descriptors do not exceed 0.8, indicating that all descriptors are independent. Therefore, the linear model built has strong statistical reliability. The diagram of the HM model is shown in Figure 3.

The linear model equation is as follows:

Log(MIC)=1780AERIC+49900MPCC+0.000004871XGP-0.264W3WP+1657.6

3.2. The results of GEP

After the CODESSA software selects the descriptors, we integrated their parameters and Log (MIC) values, and then inputted the files into the APS software to build a nonlinear model. The GEP algorithm consists of two processes: Dividing the dataset into random training sets and test sets, and building a nonlinear model using the structure and Log(MIC) values in the training set. To verify reliability, test set was calculated and analyzed at the same time. Table 4 lists the parameters set in the GEP algorithm. The training and test sets have R2 of 0.68 and 0.64, respectively, and the mean squared error is 0.14 and 0.11, respectively. In addition, the nonlinear model equations decoded by ET are as follows:

Log(MIC)=sin((1/((((x1*x1)+sin(x2))+((x2/x3)+x-1)))))+apsNop((apsNop((x3*x1))*x1));+x1 +(1/((((sin(x3)+x2)+(x4-x2))-x2)))+ sin(((sin(x4)/apsNop(x-4))*(x2+sin(x3))))

where x1, x2, x3 and x4 represented variables of AERIC,MPCC,1XGPandW3WP respectively.

The predicted values and measured values are listed in Fig. (4).

Table 4. Parameters for the simple symbolic regression problem

Number of Chromosomes	100
Generations without Change	200
Head Size	8

22 (2022) 1-7
ISSN 2161-2609

Number of Genes	5
Number of Tires	3
Max. Complexity	5
Mutation rate	0.044
Inversion rate	0.1
IS Transposition rate	0.1
RIS Transposition rate	0.1
1-Point Recombination rate	0.3
2-Point Recombination rate	0.3
Gene Recombination rate	0.1
Gene Transposition rate	0.1
Function sets	+,-,*,/,Inv,Nop,Sin

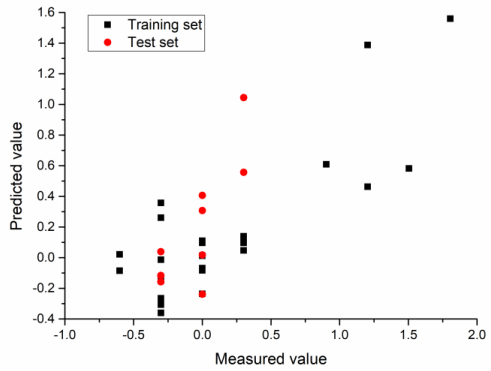


Figure 4. Plot of measured and calculated lg(IC50) by GEP.

4.Discussion

There were four descriptors selected by CODESSA software in the model: Avg electroph. react. index for a C atom, Min partial charge for a C atom [Zefirov's PC], 1X GAMMA polarizability (DIP),
WNSA-3 Weighted PNSA (PNSA3*TMSA/1000) [Zefirov's PC].

Avg electroph. react. index for a C atom[19]: Under the action of non-uniform electric field, the stress of the polarization of particles is unbalanced, so directional movement along the gradient of electric field intensity occurs. The positive or negative of the average electrophoretic force of the C atom can determine the polarity of the charge it carries, and the er the polarity, the higher the antibacterial activity of the compound.
Min partial charge for a C atom [Zefirov's PC][20]: The min partial charge of an atom represents the most basic chargability of the atom in the electric field, and the positive and negative electrical properties of the atom can be judged, from which the characteristics of the atom's gain and loss of electrons can be judged. The stronger the characteristics of the electron gain and loss, the greater the antibacterial activity changing of the compound.

1X GAMMA polarizability (DIP)[21]: The polarization rate, as a measure of the polarization strength of atoms, molecules, or ions under the action of an electric field, can reflect the difficulty of the polarization process. The reactive center embodying strong antibacterial activity in the compound needs to have a structural condition, that is, an electronic relay system composed of an electron containment center and an electron supply center. The higher the ability of the negatively charged central atom to supply electrons in the nucleophilic reaction, and the higher the ability of the positive central atom to accommodate electrons in the electrophilic reaction, the stronger the bacteriostatic activity of the compound. Therefore, 1XGP has a positive effect on enhancers.
WNSA-3 Weighted PNSA (PNSA3*TMSA/1000) [Zefirov's PC][22,23]: WNSA-3 is a quantum-chemical descriptor, which characterizes molecules by molecular shape and electron distribution and is defined in Equation: WNSA-3=PNSA3*TMSA/1000, where PNSA3 is the partial positively charged molecular surface area and the TMSA is the total molecular surface area. It indicated the influence of charge distribution on antibacterial activity. It can be seen from the descriptor that the surface area of the positive charge has a negative effect on the antibacterial activity of the antibacterial agent.
How association affects antibacterial activity needs to be further studied, which is of great significance for studying the mechanism of modulation.

5.Conclusion

In this work, a quantitative model has been developed to predict antibacterial activity of aromatic nucleus-linked symmetric membrane surfactants against taphylococcus aureus by GEP. Compared with the linear prediction model built by HM, the nonlinear model built by GEP has better predictive power and stability. As a result, GEP offers a new approach to solving more complex technical and scientific problems. This provides guidance for further research into the design and synthesis of antibacterial agent.

Acknowledgments
The authors are very grateful to the APS software providers.

References

[1] Christina T K. Entering a post-antibiotic era?[J]. Nature Reviews Microbiology, 2013, 11(3):146- 146.
[2] Chandra P, Mk U, Ke V, et al. Antimicrobial re-

22 (2022) 1-7
ISSN 2161-2609

sistance and the post antibiotic era: better late than never effort[J]. Expert opinion on drug safety, 2021, 20(11):1375-1390.
[3] Malak S A, Yasir A A, Abdulrahman W B, et al. Antimicrobial resistance in general ICUs in Saudi Arabia; a systematic review[J]. International Journal of Medicine in Developing Countries, 2020, 4(2):513-517.
[4] Boucher H W, Talbot G H, Benjamin D K, et al. 10 x '20 Progress-development of new drugs active against gram-negative bacilli: an update from the Infectious Diseases Society of America[J]. Clinical infectious diseases:an official publication of the Infectious Diseases Society of America, 2013, 56(12):1685-1694.
[5] Falagas M, Bliziotis I A. Pandrug-resistant Gram-negative bacteria: the dawn of the post-antibiotic era?[J]. International journal of antimicrobial agents, 2007, 29(6):630-636.
[6] Kareiviene V, Pavilonis A, Sinkute G, et al. Staphylococcus aureus resistance to antibiotics and spread of phage types[J]. Medicina (Kaunas, Lithuania), 2006, 42(4):332-339.
[7] Payne D J, Gwynn M N, Holmes D J, et al. Drugs for bad bugs: confronting the challenges of antibacterial discovery[J]. Nature reviews. Drug discovery, 2007, 6(1):29-40.
[8] Herzog I M, Fridman M. Design and synthesis of membrane-targeting antibiotics: from peptides- to aminosugar-based antimicrobial cationic amphiphiles[J]. MedChemComm, 2014, 5(8):1014-1026.
[9] Hurdle J G, O'Neill A J, Chopra I, et al. Targeting bacterial membrane function: an underexploited mechanism for treating persistent infections[J]. Nature reviews. Microbiology, 2011, 9(1):62-75.
[10] Neely W B, Branson D R, Blau G E, et al. Partition coefficient to measure bioconcentration potential of organic chemicals in fish[J]. Environmental Science & Technology, 1974, 8(13):1113-1115.
[11] Roy K, Sanyal I, Roy P P, et al. QSPR of the bioconcentration factors of non-ionic organic compounds in fish using extended topochemical atom (ETA) indices[J]. SAR and QSAR in environmental research, 2006, 17(6):563-582.
[12] Chu W, Yang Y, Cai J F, et al. Synthesis and Bioactivities of New Membrane-Active Agents with Aromatic Linker: High Selectivity and Broad-Spectrum Antibacterial Activity[J]. ACS infectious diseases, 2019, 5(9):1535-1545.
[13] Froimowitz M. HyperChem: a software package for computational chemistry and molecular modeling[J]. BioTechniques, 1993, 14(6):1010-1013.
[14] Yang B, Zhai H L, Si H Z, et al. QSAR Studies on the IC50 of a Class of Thiazolidinone/Thiazolide Based Hybrids as Antitrypanosomal Agents[J]. Letters in Drug Design & Discovery, 2021, 18(4):406-415.
[15] Gábor I, Csonka. Analysis of the core-repulsion functions used in AM1 and PM3 semiempirical calculations: Conformational analysis of ring systems. [J]. Journal of Computational Chemistry, 1993, 14(8):895-898.
[16] Si H Z, Zhao J A, Cui L H, et al. Study of human dopamine sulfotransferases based on gene expression programming[J]. Chemical biology & drug design, 2011, 78(3):370-377.
[17] Si Y R, Xu X Y, Hu Y F, et al. Novel QSAR model to predict Activity of natural Products against Covid-19[J]. Chemical biology & drug design, 2021, 97(4):978-983.
[18] Candida F. Genetic representation and genetic neutrality in gene expression programming[J]. Advances in Complex Systems, 2002, 5(4):389-408.
[19] Mehmet S K, Çiğdem Y, Mehmet Y, et al. Quantitative structure–activity relationship analysis of perfluoroiso-propyldinitrobenzene derivatives known as photosystem II electron transfer inhibitors[J]. BBA-Bioenergetics, 2012, 1817(8):1229-1236.
[20] Maslechko A, Verstraelen T, Van ETS, et al. Multiscale partial charge estimation on graphene for neutral, doped and charged flakes[J]. Physical chemistry chemical physics: PCCP, 2018, 20(31):20678-20687.
[21] Pérez-Garrido A, Helguera AM, Rodríguez FG, et al. QSAR models to predict mutagenicity of acrylates, methacrylates and α,β-unsaturated carbonyl compounds[J]. Dental materials, 2010, 26(5):397.
[22] Liao S L, Song J, Wang Z D, et al. Quantitative calculation of the effect of terpenoids association with carbon dioxide on their mosquito repellent activity[J]. Acta Entomol. Sinica, 2012, (09):58-65.
[23] Beteringhe A, Radutiu A C, Culita D C, et al. Quantitative Structure–Retention Relationship (QSRR) Study for Predicting Gas Chromatographic Retention Times for Some Stationary Phases[J]. QSAR & Combinatorial Science, 2008, 27(8):996-1005.