ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(8), DOI: 10.17485/ijst/2016/v9i8/87901, February 2016* ISSN (Online) : 0974-5645

Computer Assisted QSAR/QSPR Approaches – A Review

B. Firdaus Begam and J. Satheesh Kumar*

Department of Computer Applications, Bharathiar University, Coimbatore – 641046, Tamil Nadu, India;
firdh_2002@yahoo.com, jsathee@rediffmail.com

Abstract

Background/Objectives: Quantitative Structure–Activity Relationship (QSAR) / Quantitative Structure – Property Relationship (QSPR) model is based on changes in molecular structure that would reflect changes in observed biological activity or physico-chemical property. **Methods/Statistical analysis:** QSAR/QSPR involves chemistry, biology and statistics fields for analysis. It has been widely accepted model for predicting association between molecular structure and its activity. Over the years many algorithms have been proposed and applied in QSAR/QSPR studies. Framework of model involves molecular structure (graph) representation, calculation of molecular descriptors (graph invariants) and multiple linear regression method is applied for analysis. Model has been validated through statistical parameters (R and R2). **Findings:** Methods involved in model development were reviewed for QSAR/QSPR studies. Multiple Linear Regression is one of the best methods for developing QSAR/QSPR model. This work focuses on developing QSPR model for predicting boiling point of alkyl benzene molecules using Multiple Linear Regression method. Wiener index, Harary Index, Hyper Wiener Index, Hyper Harary Index, and Randic index are calculated for analysis. The model has been validated by calculating R and R2 value. Various models were developed based on different combinations of descriptors to analysis which contribute best in predicting boiling point. Best fit model has been identified by developing model with different combinations of descriptors and rank them based on highest R and R2 value. Model with highest value has been taken for prediction of boiling point as best fit model as n (number of molecules) =14, R= 0.9934 and R2=0.9968. **Applications/Improvements:** Review on methods involved in prediction analysis has enlightened that model with reduced molecular descriptor subset and outlier detection method shows better performance by improving quality of the dataset Main application of QSAR/QSPR analysis is in drug discovery process. As it has reduced the time taken for lead identification and optimization in drug discovery process.

Keywords: Descriptor, Descriptor Selection, Mathematical Model, Multiple Linear Regression, QSAR, QSPR

1. Introduction

Quantitative Structure – Property / Quantitative Structure – Activity Relationships are mathematical model which relates the physico-chemical property / biological activity of compounds to their chemical structures[1,2]. In other words, it provides a model to exploit or explore the relationship between chemical structure and physico-chemical / biological actions which results in development of Novel Chemical Entity (NCE) in biomolecular discovery[3]. International Union of Pure and Applied Chemistry define QSAR as "Quantitative Structure– Activity Relationships (QSAR) is mathematical relationships linking chemical structure and pharmacological activity in a quantitative manner for a series of compounds. Methods which can be used in QSAR include various regression and pattern recognition techniques"[4].

QSPR is analogous to QSAR. Quantitative Structure–Toxicity Relationship (QSTR) or Quantitative Structure–Pharmacokinetic Relationships (QSPkR) are similar models used on toxicological and pharmacokinetic system which are based chemical structure. Regardless of minor changes in the terminology they all follow same basic principles and protocols to build a model. It has been widely used as an important key field in drug discovery

* *Author for correspondence*

process for predicting physicochemical properties and biological activity of molecules. It expresses the characteristics of molecule through various molecular descriptors. It can be expressed as,

Property (P) = f(s) ***(1)***

where, P represents physicochemical and biological properties, f represents the relationship, and s represents description of molecular structure in terms of empirical, non-empirical or a combination of descriptors[5]. The term used to state computer assisted mathematical characterization of fundamental chemical structure is QSAR/QSPR model.

Main objective of the model is to select descriptors/ factors which are needed for particular property/activity of the molecule. And secondly model can be used to predict the property/activity of unknown compound. QSPR/QSAR model has shown its significance in drug discovery process[6]. The model has dedicated its role in the process of lead optimization as to improve or design a new chemical compound with well-defined property/ activity.

Need and application of QSAR/QSPR model are,

- Time and cost taken to understand physical, chemical and biological property of molecule through experiment are very expensive.
- Model gives better understanding about the interaction or reaction between molecules and its activity.
- It can provide useful information about biological effect of the compound which would help in drug research and testing ADMET.
- It can also be used to predict the property or activity of the compound before synthesis. It mainly helps in reducing or replacing the molecule taken for testing in wet lab.
- It is becoming more useful and reliable.

Computer based mathematical QSAR/QSPR model are based on chemical information extracted based on chemical structure not based on experimental values. The quality of the model depends on various factors like, quality of dataset, descriptor analysis, descriptor analysis, outlier detection, statistical methods and validation. Organization of the paper is as follows, steps involved in development of QSAR/QSPR model have been discussed in section 2. Detailed descriptions of predictive QSAR/ QSPR model has been discussed in section 3. QSPR model developed using multiple linear regression method in section 4 and result and discussion in section 5 and section 6 concludes the paper.

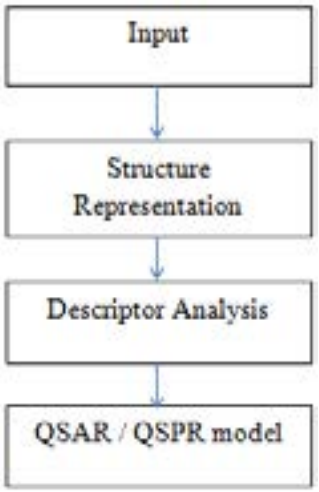

Figure 1. Frame work of QSAR / QSPR model.

2. QSAR/ QSPR Methodologies

QSAR/QSPR model consist of three steps namely, structure representation, descriptor analysis and model building as in Figure 1. The model starts with collection of data which are considered for prediction of property. Various representations of chemical data are available to store chemical information of data in computer systems. Some of them are chemical formula, IUPAC name, SMILES, MDL molfile, Tripols and others[7]. From the input data chemical structures are visualized in 1D, 2D or 3D form as in Figure 2. Chemical structural features called molecular descriptor has been found closely related to target property of the molecule. Thousands of descriptors are available for various applications and it can be divided into 0D, 1D, 2D and 3D descriptors (Figure 3), dimensionality refers to chemical representation from which they can be calculated[8]. Numerical invariants of chemical structures were calculated by analysing and manipulating structural information of the molecule is called as molecular descriptors. Randic[9] proposed set of requirements of molecular descriptors as shown below, Structural interpretation, Show good correlation with at least one property, Preferably allow for the discrimination of isomers, Applicable to local structure, Generalizable to "higher" descriptors, Independence, Simplicity, Not to be based on properties, Not to be trivially related to other descriptors, Allow for efficient construction, Use familiar structural concepts, Show the correct size dependence, Show gradual change with gradual change in structures.

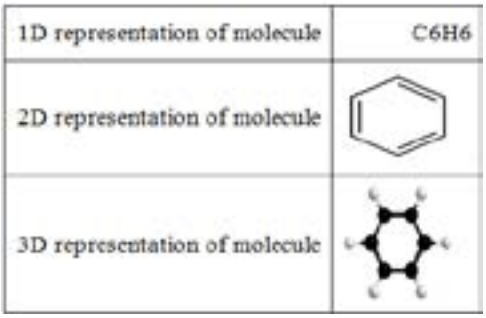

Figure 2. Various representation of chemical structure of Benzene.

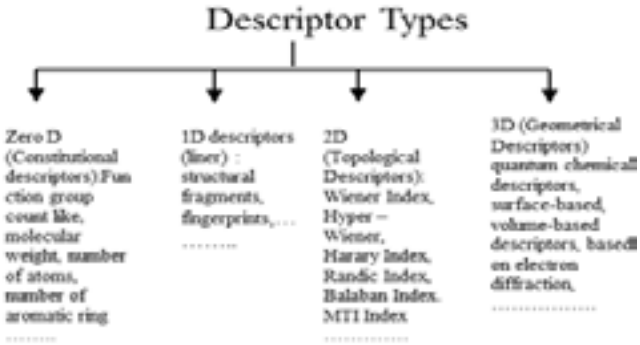

Figure 3. Descriptor types.

Chemical space is a used to store set of calculated descriptor for each molecule in multidimensional space[10]. Chemical space can be represented as in Figure 4, for each molecule (M_i), the target property (Y_i) and various descriptors (X_{1n}, X_{2n},….X_{nn}) are defined. Molecular properties can be shared by multiple molecules; can share locations in defined chemical space[11]. Chemical space can be used to visualize the variance or hidden patterns of molecules. Chemical space should not be used directly for creating QSAR/QSPR model as it may contain some unwanted data. Chemical space can be cured by means of pre-processing which improves the quality of the data. Pre-processing step in descriptor analysis helps in handling problems like[12],

- Different descriptors may relate to same structural information
- Descriptors that are not related to target property
- Large set of molecular descriptors.

The quality of the data can be analyzed through correlation coefficients and variance among descriptors. It is applied to identify descriptors which relate to same structural information. Principal Component Analysis (PCA)[13] method can be used to visualize chemical matrix which would help in creating better QSAR / QSPR model. Firdaus et al. (2014) visualized drug-likeness chemical space (based on Lipinski Rule of five) through PCA method and variance or hidden patterns in the dataset are interpreted through score plot, load plot and biplot[14].

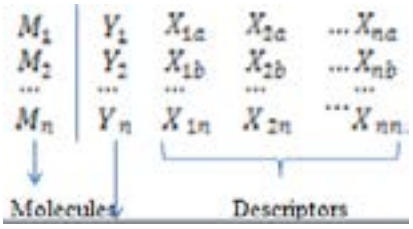

Figure 4. Chemical Space Data matrix.

Descriptor selection is aimed at getting rid of those calculated descriptors that are redundant, noisy, or irrelevant for the model building tasks[15]. Descriptors selection methods are applied to select set descriptors that relate to target property which results in compression of dataset. Feature selection aims to choose a small number of the most informative descriptors in a context-dependent way. Using descriptor selection results[16] in,

- Dimensionality reduction, as visualization and interpretation of data becomes more understandable
- Improves the performance of prediction process
- Reduces Storage space and measurement requirement
- Reduces time for manipulation and retrieval process

Descriptor selection process has been vital in developing QSAR/QSPR model as small set of descriptors can optimize predictivity of the model. Various descriptor selection methods are Forward selection method[17], backward elimination method, stepwise selection method[18], Leaps-and-bounds regression, successive projection algorithm[19], genetic algorithm method, Artificial neural network[20], Simulated Annealing (SA), Uninformative variable Elimination – Partial Least Square (UVE-PLS)[21], Robust principal components regression based on principal sensitivity vectors (RPPSV), Minimum Redundancy Maximum Relevance[22], Ant Colony optimization (ACO)[23], Particle Swarm Optimization (PSO) Fuzzy based Minimum Redundancy Maximum Relevance[24] and other related algorithms

The molecules which hold property that deviate or over represented from target property are called as outlier. It could influence the performance of remaining data. Two types of outliers available namely leverage outliers and activity outliers[25]. Leverage outlier is also known as structural outliers where similarities between compounds

are identified based on molecular descriptors. Activity outlier separates the compounds which deviate on different sides of activity. Sphere-Exclusion algorithm[26], Monte-Carlo cross-validation[27], M-estimators[28], least median of squares (LMS)[29], Least Trimmed Squares (LTS)[30], Robust Principal Component Regression (RPCR)[31], Robust Partial Least Squares (RPLS)[32] and robust principal components Regression based on Principal Sensitivity Vectors (RPPSV)[33] are some of the outlier detection methods. Outlier detection and removal process improves the quality of data set used for building QSAR/QSPR model.

3. Methods to Correlate Molecular Structure with Property/ Activity

QSAR/QSPR are influential mathematical model used for designing novel chemical molecules and predicting activity of molecules based on physical, chemical and biological properties. It relates the target property and set of molecular regression in linear manner and it is used to predict known/unknown target property of molecules. This approach is an application of data analysis methods and statistics to predict accurately the biological actions or property of the molecule based on its structure. Development of QSAR/QSPR model in solving various predictive chemical problems has found growing applications in chemical data mining and combinatorial library design.

Many approaches have been used in formulating mathematical model of QSAR/ QSPR but most widely used technique is linear regression. Regression analysis is a powerful means for establishing a correlation between independent variables and a dependent variable such as biological activity[34]. QSAR/QSPR studies include boiling point, partition coefficient[35], dissociation constant[36], thermodynamic behaviour[37], hydrophobicity[38], aqueous solubility[39], and biological properties include activity (IC50/EC50)[40], acute toxicity (LD50)[41], bioconcentration[42], carcinogenicity[43], inhibition constant[44], mutagenicity[45], penetration[46], pharmacokinetics[47], ADME studies[48], drug resistance[49], drug metabolism[50], toxicity prediction[51] and other related areas.

Various computational methods like multiple linear regression, principal component regression analysis[52], partial least square regression[53], Artificial Neural Networks[54], Support Vector Machine[55] were applied over various QSR/QSPR problems and the model which shows best are considered for further analysis. Determining the quality of the model is the crucial step of modelling. It establishes the reliability and relevance of the model developed for a specific purpose. The application of the QSAR model depends of predictive and significance ability of model which can be established by various validation methods. It helps to determine the complexity of an equation that the amount of data justifies. Validation methods are broadly classified as internal and external methods. The methods are correlation of determination (R), Squared correlation coefficient (R^2), cross validation (Q^2), adjusted R^2 (R^2adj), chi-squared test (χ^2), root mean-squared error (RMSE), bootstrapping and scrambling (Y-Randomization)[56,57].

4. Multiple Linear Regression Analysis to Predict Boiling Point (BP) of Alkyl Benzene Compounds

This research paper discusses classical QSPR approaches to predict boiling point of alkyl benzene molecules using Multiple Linear Regression (MLR) method. It is one of the most widely used methods for developing QSPR/QSAR model. MLR analysis is been applied to understand or identity certain structural features which influence a particular property. Regression analysis helps in developing a model which support in designing a new molecule with required property. This model depends on molecular descriptors which replicates molecular properties to provide meaningful relationship between structure and activity/property of molecule[58]. Chemical graph theory plays a vital role in forming mathematical model for QSAR/QSPR based on molecular descriptors. This work focuses on predicting boiling point of 14 alkyl benzene molecules. Structural molecular files were downloaded from NIST website. Firdaus and Satheesh (2014) have proposed sparse based method for structural representation of a chemical compound and chemical reaction[59,60]. Boiling point is an important and major property to identify and characterize a chemical compound. It is also used as an indicator of a molecule to represent volatility of molecule. Descriptors derived

Table 1. Boiling point, topological indices, calculated boiling point for alkyl benzene

Molecule	BP(·C)	W	H	WW	HH	R	Calculated BP(·C)	Residual
Benzene	80.1	27	10	42	8.9	3	75.6902	-4.4098
Toluene	110.6	42	12.916	71	11.26	3.3938	110.7996	0.1996
Ethyl benzene	136.2	64	15.78	122	13.5193	3.9319	138.997	2.797
O-xylene	144.4	60	16.1667	106	13.8403	3.8045	148.2419	3.8419
P-Xylene	138.4	62	16.033	115	13.7381	3.7877	139.2664	0.8664
Propyl benzene	159.2	94	18.6833	203	15.7662	4.4319	162.0505	2.8505
1ethyl-2-methlybenzene	165.2	86	19.2833	167	16.24	4.3425	163.5157	-1.6843
1ethyl-3-methlybenzene	161.3	88	19.15	176	16.1462	4.3257	159.5018	-1.7982
1ethyl-4-methlybenzene	162	90	19.0667	187	16.087	4.3257	160.5602	-1.4398
1,2,3-Trimethylbenzene	176.1	82	19.667	151	16.5697	4.2152	179.284	3.184
1,2,4-trimethylbenzene	169.4	84	19.53	160	16.4672	4.19	169.2999	-0.1001
1,3,5-trimethylbenzene	164.7	84	19.50	159	16.4375	4.1815	161.2743	-3.4257
1,2,3,4-Tetramethylbenzene	205	109	23.3667	211	19.4186	4.62	203.8427	-1.1573
n-Butyl benzene	183.3	133	21.6429	323	18.0358	4.9319	183.1	-0.2

through graph theory are called as topological descriptors. Wiener index (W), Harary Index (H), Hyper Wiener Index(WW), Hyper Harary Index (HH), and Randic index (R) are some of well-known structurally related indices were considered and calculated for developing QSPR model for alkyl benzene molecules (Table 1). As the data size is not complex dimensionality reduction approach has not be considered in this analysis.

Multiple Linear regression methods defines linear mathematical equation based on desired property and molecular descriptors as represented below in equation 1,

$$Y = \sum_{i=1}^{n} X_i b_i \quad (1)$$

where, Y represent desired property, X_1 to X_n represents specific molecular descriptor and to b_n represents coefficients of descriptors. Various models have been developed and analyzed as shown in Table 2. QSPR model has been validated through R which represents degree of correlation between all pairs of descriptors and R^2 coefficient which specifies the proportion of variation (Y) explained by means of regression analysis.

5. Result and Discussion

QSPR model for alkyl benzene molecules has been developed based on multiple linear regression method. The model is been developed for all different single and group of descriptors, as single and two variables (descriptors) does not shown good correlation, the combination of other descriptors taken for model development is been shown in Table 2. The best MLR equation obtained for BP of alkyl benzene molecules with descriptors W, H, WW, HH and R. The mathematical equation obtained,

BP = - 4.93 × W + 98.353 × H + 1.159 × WW + (-95.35) × HH + 11.353 × R

(n =14, R^2 = 0.9968, R= 0.9934)

Table 2. Possible descriptor set for calculating BP and quality validated by R and R^2 coefficients

Model No.	Descriptor	R	R^2
1	W, H, WW	0.9952	0.9905
2	W, H, HH	0.9959	0.9918
3	W, H, R	0.9959	0.9919
4	W, WW, HH	0.9920	0.9841
5	W, HH, R	0.9959	0.9918
6	H, WW, HH	0.9957	0.9915
7	H, WW, R	0.9959	0.9919
8	W, HH, R	0.9959	0.9918
9	WW, HH, R	0.9959	0.9918
10	W, H, WW, HH	0.9966	0.9932
11	W, H, WW, R	0.9961	0.9923
12	W, H, HH, R	0.9959	0.9919
13	W, WW, HH, R	0.9959	0.9919
14	H, WW, HH, R	0.9959	0.9919
15	W, H, WW, HH, R	0.9968	0.9934

According to Tropsha et al. (2003)[61] a model can be considered as predictive if $R^2 > 0.6$. QSPR model shows that all descriptors directly contribute in predicting the target

property of molecules. The experimental and calculated or predicted boiling point and their residual values are shown in Table 1. Correlation between the experimental and calculated boiling point is shown in Figure 5.

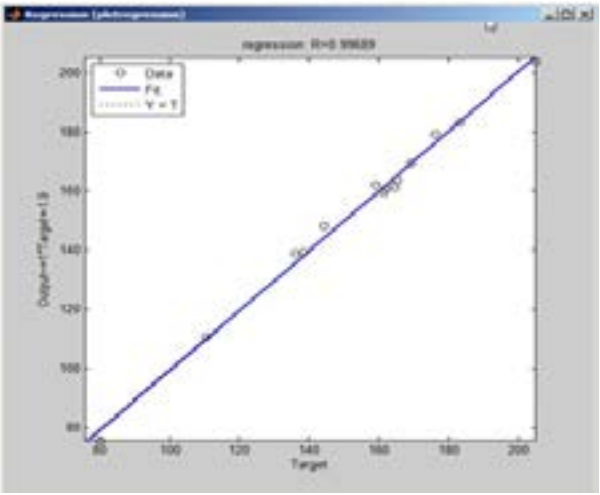

Figure 5. Regression plot for Target versus Calculated property of alkyl benzene molecules.

6. Conclusion

QSAR/ QSPR models are multidimensional functions which quantify the relation between a physicochemical and biological property of molecules based on the chemical information obtained from its structure. The relationship can be analyzed by encoding chemical structures with calculated molecular descriptors through regression analysis. The study has shown that the performance of the model can be improved by reducing number of molecular descriptors and identifying and removing outliers present in chemical space when the size of the dataset is complex. QSPR model to predict boiling of alkyl benzene molecules using multiple linear regression model has been developed. Model have been validated through and R^2 value. Best model with higher R=0.9934 and R^2 =0.9968 with Wiener index, Harary Index, Hyper Wiener Index, Hyper Harary Index, and Randic index are considered for predicting boiling point.

7. References

1. Faulon JL, Bender A. Handbook of cheminformatics algorithm; CRC press: New York; 2010.
2. Leach AR, Gillet VJ. An introduction to chemoinformatics; Springer Science Business Media Inc: India; 2007
3. Khajeh A, Modarress H, Zeinoddini-Meymand H. Application of modified particle swarm optimization as an efficient variable selection strategy in QSAR/QSPR studies. Journal of Chemometrics. 2012; 26(11–12):598–603.
4. Van de Waterbeemd H, Carter RE, Grassy G, Kubinyi H, Martin YC, Tute MS, Willett P. Glossary of terms used in computational drug design (IUPAC Recommendations 1997). Pure and Applied Chemistry. 1997; 69(5):1137–52.
5. Grover M, Singh B, Bakshi M, Singh S. Quantitative structure-property relationships in pharmaceutical research. Part 1. Pharmaceutical Science and Technology Today. 2000; 3(1):28–35.
6. Puzyn T, Leszczynski J, Cronin MTD. Recent advances in QSAR studies: methods and applications; Springer Science, Business Media B.V.: New York; 2010.
7. Gasteiger J, Engel T. Chemoinformatics: a text book; Wiley-VCH Verlag GmbH & Co.: Weinhaim; 2006.
8. Todeschini R, Consonni V. Handbook of Molecular descriptors; John Wiley and Sons: Weinhaim; 2008.
9. Randic M. Generalized molecular descriptors. Journal of Mathematical Chemistry. 1991, 7(1),155–68.
10. Oprea TI, Gottfries J. Chemography: the art of navigating in chemical space. Journal of Combinatorial Chemistry. 2001; 3(2):157–66.
11. Wegner JK, Frohlich H, Mielenz HM, Zell A. Data and graph mining in chemical space for adme and activity data sets. QSAR Combinatorial Science. 2006; 25(3):205–20.
12. Dudek AZ, Arodz T, Galvez J. Computational Methods in developing Quantitative Structure-Activity Relationships (QSAR): a review. Combinatorial Chemistry & High Throughput Screening. 2006; 9(3):213–28.
13. Khodadoust S. Application of artificial neural network for prediction of retention time for some pesticides in liquid chromatography. Indian Journal of Science and Technology. 2012 Feb; 5(2):2001–8 doi: 10.17485/ijst/2012/v5i2/30331.
14. Begam BF, Kumar JS. Visualization of chemical space through Principal Component Analysis. World Applied Sciences Journal, 29 (Data Mining and Soft Computing Techniques). 2014:53–59.
15. Goodarzi M, Dejaegher B, Heyden YV. Feature selection methods in QSAR studies. Journal of AOAC International. 2012; 95(3):636–51.
16. Xu L, Zhang WJ. Comparison of different methods for variable selection. Analytica Chimica Acta. 2001; 446(1):475–81.
17. Katritzky AR, Lobanov VS, Karelson M. QSPR: The Correlation and quantitative prediction of chemical and physical properties from structure. Chemical Society Reviews. 1995; 24(4):279–87.
18. Shahlaei M. Descriptor selection methods in quantitative structure–activity relationship studies: a review study. Chemical Reviews. 2013; 113(10):8093–103.
19. Hou TJ, Wang JM, Xu XJ. Applications of genetic algorithms on the structure–activity correlation study of a group of non-nucleoside HIV-1 inhibitors. Chemometrics and Intelligent Laboratory Systems. 1999; 45(1–2):303–10.
20. Daszykowski M, Stanimirova I, Walczak B, Daeyaert F, De

Jonge MR, Heeres J, Koymans LM H, Lewi PJ, Vinkers HM, Janssen PA, Massart DL. Improving QSAR models for the biological activity of HIV reverse transcriptase inhibitors: aspects of outlier detection and uninformative variable elimination. Talanta. 2005; 68(1):54–60.
21. Yang SS, Lu WC, Gu TH, Yan L-M, Li GZ. QSPR study of n-octanol/water partition coefficient of some aromatic compounds using support vector regression. QSAR and Combinatorial Science. 2009; 28(2):175–82.
22. Goodarzi M, Freitas MP, Jensen R. Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl) uracil derivatives using MLR, PLS and SVM regressions. Chemometrics and Intelligent Laboratory Systems. 2009; 98(2):123–29.
23. Reymond JL, Van Deursen R, Blum LC, Ruddigkeit L. Chemical space as a source for new drugs. MedChemComm. 2010; 1(1):30–38.
24. Huerta EB, Duval B, Hao J-K. Fuzzy logic for elimination of redundant information of microarray data. Genomics, Proteomics and Bioinformatics. 2008; 6(2):67–73.
25. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. Journal of Computer-aided Molecular Design. 2003; 17(2–4):241–53.
26. Konovalov DA, Sim N, Deconinck E, Heyden YV, Coomans D. Statistical confidence for variable selection in QSAR models via Monte Carlo cross-validation. Journal of Chemical Information and Modeling. 2008; 48(2):370–83.
27. Huber PJ. Robust statistics. International Encyclopedia of Statistical Science. 2011:1248–51.
28. Rousseeuw PJ. Least median of squares regression. Journal of the American Statistical Association. 1984; 79(388):871–80.
29. Agullo J, Croux C, Van Aelst S. The multivariate least-trimmed squares estimator. Journal of Multivariate Analysis. 2008; 99(3):311–38.
30. Walczak B, Massart DL. Robust principal components regression as a detection tool for Outliers. Chemometrics and Intelligent Laboratory Systems. 1995; 27(1):41–54.
31. Hubert M, Branden KV. Robust methods for partial least squares regression. Journal of Chemometrics. 2003; 17(10):537–49.
32. Zhang MH, Xu QS, Massart DL. Robust principal components regression based on principal sensitivity vectors. Chemometrics and Intelligent Laboratory Systems. 2003; 67(2):175–85.
33. Draper NR, Smith H. Applied Regression Analysis. 2nd ed.; John Wiley & Sons: New York; 1981.
34. Firpo M, Gavernet L, Castroa EA, Toropov AA. Maximum topological distances based indices as molecular descriptors for QSPR. Part 1. Application to alkyl benzenes boiling points. Journal of Molecular Structure: Theochem. 2010; 501–502:419–25.
35. Mercader AG, Goodarzi M, Duchowicz PR, Fernández FM, Castro EA. Predictive QSPR study of the dissociation constants of diverse pharmaceutical compounds. Chemical Biology & Drug Design. 2010; 76(5):433–40.
36. Mercader A, Castro EA, Toropov AA. Maximum topological distances based indices as molecular descriptors for QSPR. 4. Modeling the enthalpy of formation of hydrocarbons from elements. International Journal of Molecular Sciences. 2001; 2(2):121–32.
37. Tueros M, Castro EA, Toropov AA. Maximum topological distance-based indices as molecular descriptors for QSPR. 3–Calculation of the hydrophobicity of polyaromatic hydrocarbons. Molecular Modeling Annual. 2001; 7(6):178–83.
38. Delaney JS. Predicting aqueous solubility from structure. Drug Discovery Today. 2005; 10(4):289–95.
39. Jenssen H, Fjell CD, Cherkasov A, Hancock REW. QSAR modeling and computer-aided design of antimicrobial peptides. Journal of Peptide Science. 2008; 14(1):110–14.
40. He G, Feng L, Chen H. A QSAR study of the acute toxicity of halogenated phenols. Procedia Engineering. 2012; 43:204–09.
41. Gramatica P, Papa E. QSAR modeling of bioconcentration factor by theoretical molecular descriptors. QSAR and Combinatorial Science. 2003; 22(3):374–85.
42. Valerio LG, Arvidson KB, Chanderbhan RF, Contrera JF. Prediction of rodent carcinogenic potential of naturally occurring chemicals in the human diet using high-throughput QSAR predictive modelling. Toxicology and Applied Pharmacology. 2007; 222(1):1–16.
43. Eroglu E, Türkmen H, Guler S, Palaz S, Oltulu O. A DFT-based QSARs study of acetazolamide/sulfanilamide derivatives with carbonic anhydrase (CA-II) isozyme inhibitory activity. International Journal of Molecular Sciences. 2007; 8(2):145–55.
44. Ringeissen S, Note R, Dochez C, Flamand N, Ouedraogo-Arras G, Meunier JR. Evaluation of (Q) SAR models for the prediction of mutagenicity potential. AATEX. 2007; 14:469–73.
45. Chung KK, Do DQ. Modelling the effect of structural QSAR parameters on skin penetration using genetic programming. Advances in Natural Sciences: Nanoscience and Nanotechnology. 2010; 1(3):1–7.
46. Vieira JB, Braga FS, Lobato CC, Santos CF, Costa JS, Bittencourt JAH, Santos CBR. A QSAR, pharmacokinetic and toxicological study of new artemisinin compounds with anticancer activity. Molecules. 2014; 19(8):10670–697.
47. Hansch C, Leo A, Mekapati SB, Kurup A. QSAR and ADME. Bioorganic & Medicinal Chemistry. 2004; 12(12):3391–400.
48. Nandy A, Kar S, Roy K. Development and validation of regression-based QSAR models for quantification of contributions of molecular fragments to skin sensitization potency of diverse organic chemicals. SAR and QSAR in Environmental Research. 2013; 24(12):1009–23.
49. Bugrim A, Nikolskaya T, Nikolsky Y. Early prediction of drug metabolism and toxicity: systems biology approach and modelling. Drug Discovery Today. 2004; 9(3):127–35.

50. Dunn WJ. QSAR approaches to predicting toxicity. Toxicology Letters. 1998; 43(1):277–83.
51. Saghaie L, Sakhi H, Sabzyan H, Shahlaei M, Shamshirian D. Stepwise MLR and PCR QSAR study of the pharmaceutical activities of antimalarial 3-hydroxypyridinone agents using B3LYP/6-311++ G** descriptors. Medicinal Chemistry Research. 2013; 22(4):1679–88.
52. Shamshirian D, Natarajan R, Nirdosh I, Basak SC, Mills DR. QSAR modeling of flotation collectors using principal components extracted from topological indices. Journal of Chemical Information and Computer Sciences. 2002; 42(6):1425–30.
53. Cramer RD. Partial Least Squares (PLS): its strengths and limitations. Perspectives in Drug Discovery and Design. 1993; 1(2):269–78.
54. Espinosa G, Yaffe D, Cohen Y, Arenas A, Giralt F. Neural network based Quantitative Structural Property Relations (QSPRs) for predicting boiling points of aliphatic hydrocarbons. Journal of Chemical Information and Computer Sciences. 2000; 40(3):859–79.
55. Mei H, Zhou Y, Liang G, Li Z. Support vector machine applied in QSAR modelling. Chinese Science Bulletin. 2005; 50(20):2291–96.
56. Tropsha A. Best practices for QSAR model development, validation and exploitation. Molecular Informatics. 2010; 29(6–7):476–88.
57. Frimayanti N, Yam ML, Lee HB, Othman R, Zain SM, Rahman NA. Validation of Quantitative Structure-Activity Relationship (QSAR) model for photosensitizer activity prediction. International Journal of Molecular Science. 2011; 12(12):8626–44.
58. Maggiora GM. On outliers and activity cliffs why QSAR often disappoints. Journal of Chemical Information and Modeling. 2006; 46(4):1535–35.
59. Begam BF, Kumar JS. Topostructural view of chemical components using sparsity. IJRSCIT, 2014; 2(A):214–19.
60. Begam BF, Kumar JS. Representation and visualization of chemical reaction using graph theory. International Journal of Applied Engineering Research. 2014; 9(21):4933–38.
61. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation the absolute essential for successful application and interpretation of QSPR models. QSAR and Combinatorial Science. 2003; 22(1):69–77.