UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

The Measurement of Molecular Biological Activity based on Quantitative Structure Activity Relationships

Hentabli Hamza, Maged Nasser, Naomie Salim & Faisal Saeed
Faculty of Computing
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia
Email: maged.m.nasser@gmail.com

Submitted: 12/01/2018. Revised edition: 31/05/2018. Accepted: 2/06/2018. Published online: 21 November 2018

***Abstract*—The urge of producing new chemical compounds under eco-friendly production restriction and with minimum side effects is significantly rising, considering the difficulties that conventional methods are dealing with, from financial investments and being time consuming to multi resistance microorganisms and untreatable diseases. Rational molecular design methods like Computer Aided Molecular Design (CAMD) and Quantitative Structure Activity Relationship (QSAR) allow the production of new substances with pre-decided properties and correlate the structure to biological activity, which influences the drug development process and minimizes the financial investment involved in the process. QSAR employs several descriptors to decode the molecular configuration of a compound, which facilitate the understanding of its physical and biological properties. There are various conditions such as the selection of compounds and descriptors that must be accomplished for developing an applicable model of QSAR. The underlining principles and steps of QSAR are explained in this review paper.**

***Keywords*—Computer aided molecular design, CAMD, Quantitative Structure Activity Relationship, QSAR, Drug design, Chemical dataset, molecular design**

I. Introduction

The survival of mankind is threatened by ever increasing multi-drug resistant micro-organisms, untreatable diseases like AIDS and the contamination of the food chain by chemicals used to kill pests, all these problems can be solved by development of new chemical compounds. So, it is imperative that simulation models are set up to study the various threats in front of mankind and more effective drugs with minimum side effects are developed and all the chemicals used in our daily lives like cosmetics are tested for safety. Also, it must be assessed how different chemicals contaminate the surrounding environment and stress should be placed on developing and adopting eco-friendly production processes. Public sentiment frowns upon animal testing and researchers must find a suitable alternative to conducting clinical trials on animals or at least try and modify the way animals are treated during the trials and must pressurize the regulatory bodies to include results obtained from non-animal testing in the drug development and approval process [1]. All these methods must be put to practical use in drug development, analysis of different experimental procedures and must dictate the choice of the correct substance in validation experiments.

Though a large number of possible drug candidates are entered in to the Chemical Abstracts Service annually a shortage of new drugs that can treat human diseases still remains. On the other hand, the financial investment needed to develop a new drug has increased by leaps and bounds. So, drug manufacturers have had to pump a huge amount of money and invest years of time in the drug development process [2]. Although the methods used in the entire process of drug development are cumbersome and expensive they can be used to study chemicals used in our daily lives or those that have definite scientific potential [2, 3].

Conventional methods may take a long time to deliver a compound suitable for clinical trials. Over the past ten years rational molecular design methods [4, 5] like computerized data compilation, detailed analysis of structural configuration, tabulation of information in a chemical data base and processing and recovery of the data as and when needed [6-7], molecular modeling, emphasis on study of structural configurations of new chemicals and correlating the structure to biological activity along with different statistical methods have influenced the drug development process to a great extent and cut down significantly the financial investment involved in the process.

When rational drug design focuses on the dangers to mankind and tries to study the effect of drugs on environmental conditions then many different statistical methods are used to get a cohesive picture from all the information related to biological and chemical aspects of the problem and then an effort is made to understand what kind of chemical composition can help in finding solutions to the problems looming large in front of us. The most important procedures involved in the process include grading of risks, assigning importance to different tasks, dividing in categories and ensuring proper labeling as per the regulations. But, it must be remembered that the complicated web of drug development and search of new compounds with biological activity requires interdisciplinary cooperation [8] between fields like computers and math. Every chemical compound that possesses some biological activity cannot act as a drug if it has certain adverse effects. Once a compound is developed and proven to have therapeutic action it must undergo a series of tests for safety purposes. So, computational chemistry alone does not define the drug development process, this process needs the synergy between experimental methods and computerized techniques.

If a structure of a chemical compound has to be analyzed, then both molecular modeling and statistical methods can be applied successfully. A compound usually derives its therapeutic effect from more than one chemical action so if a thorough analysis of the compound is performed a greater amount of information can be collected. The discipline of chemometrics has come up to help in the study of large number of data sets which are related to different types of biological effects and multiple descriptors, the field uses both sophisticated computer program and statistical methods. The definition of chemometrics as laid down in the 1970s is the field of science that utilizes different mathematical and statistical methods to derive the best possible process of analysis of large amount of information. Artificial intelligence and multivariate methods are used for quick tabulation and prediction of molecular and biological effects of a compound [9-13].

Drug manufacturers have reaped enormous benefits by the use of molecular techniques based on structural configuration. But, if proper information about the structure of the intended target is not available then the trial and error method has to be used in which different compounds are tested for their biological effects. The technique of introducing stepwise changes in the structure of a compound has sparked scientific interest in the past. It is not easy to decide which modification would result in the desired action and be applicable in all parameters of the process so property descriptors are used to select the desired properties and statistical methods are applied to ensure that the correct modification is chosen.

The modern era belongs to the field of combinatorial chemistry where actual synthesis of a compound is not required and a large number of structural configurations can be designed by combinations of different structural components and huge number of compounds with structural similarities can be stored in an electronic database [14]. Data mining methods and High Throughput Screening [HTS] can filter through the vast expanse of information looking for a compound with a definite activity; they usually use a part of the molecular structure to analyze up to a million compounds. If the correct pharmacophore is not available then the most commonly occurring pattern may be identified by these methods. The technique used by combinatorial chemistry ensures that the tabulated data is the widest possible so that it covers the widest possible range of structural configurations and the search for a biologically active compounds is a successful one [15]. With the use of the method of molecular diversity the possibility of discovery of a new drug has improved. Since, large amount of information about the chemical structure of different compounds is available at the click of a mouse the chances of the discovery of a new drug have increased manifold [16]. The use of combinatorial chemistry has led to an ever increasing number of compounds to be synthesized and tested and have reduced the costs involved in drug discovery as compared to the conventional methods of synthesizing a substances in a lab and then undertaking tests to determine its activity. Since only limited finance is available for the actual synthesis and clinical trials of compounds it is extremely important to stop the unregulated synthesis of compounds and ensure that only those compounds are synthesized which have some desired therapeutic effect. Proper study of the pathogenesis of a disease and analysis of the molecular target can help in synthesis of compounds which can have the desired action; such compounds can be synthesized by use of experimental methods based on statistical techniques.

There are 2 methods used in experimental design. The first method involves making the drug candidate as effective as possible, but this is applicable only in the latter steps of the optimization procedure. Such methods can be applied to a small number of variations of a known compound and does not cover a broad area. So, any compound that has the desired properties but lies outside the coverage area stands to be missed when this method is used. The second approach is to conduct a stepwise analysis of the entire database which can be performed at any stage of the drug development process. For these methods to work efficiently, it is essential that clearly defined relationships between the structural configuration of a compound and its properties exists both in quantitative and qualitative terms. The entire descriptor space can be thoroughly analyzed with the help of a small number of compounds with the help of these two methods and the biological activity of the drug candidate can be tested under the context of its structural arrangement.

Computer Aided Drug Design which is also known as Computer Aided Molecular design (CAMD) is a method that helps in studying the chemical properties of structural configurations which have been developed via software programs. The concept of rational molecular design uses synergy of chemical combinations and permutations via computer software and advanced computer technology so that new compounds can be developed. The important processes used in computational chemistry are molecular quantum mechanics, analysis on basis of structural configuration, thorough analysis of available molecular data and vast variety of large number of compounds, molecular graphics and illustration of data illustrating the binding of ligand to receptor, calculating the intermolecular bonds and Quantitative Structure Activity Relationships (QSAR). Many different sectors have benefitted from the use of CAMD like study of organic chemicals, development of new drugs, study of biochemical phenomenon occurring in nature, guest host chemistry which deals with structural analysis of inhibitors of enzymes, catalysts and solutions used in experiments. Other fields like agriculture, animal husbandry, medicine and material sciences (like study of compounds made up of different molecules, polymers, chemicals, semiconductors and nonlinear phenomenon) have also benefitted from these developments.

In the face of such positive results in favor of use of computers in drug development an important point that has to be kept in mind is that computers are not exchangeable for the human expertise in chemical analysis. The computer should be used only as a helping tool to understand the fine biological and chemical aspects of the molecule being analyzed. Extensive research has been conducted on the use of rational drug design in the field of drug development [17]. However, the initial stages of the research did not have the support of computer technology as this technology was still in its infancy and was not widely available. So, researchers used statistical methods to define the interrelationship between biological activity of a compound and its structural configuration. But, with advances in computer technology, availability of sophisticated gadgets with the capability of producing detailed graphics the field has developed into a much-sophisticated branch called rational drug design or computer aided drug design.

Work done in the field of linking structural configuration with physical activity of a compound has stressed on the need for further understanding of this complicated relationship. Some examples of this kind of work include Quantitative Structure-Activity relationship (QSAR) and Quantitative Structure Property relationship (QSPR) which helps in estimating the physical properties of compounds and Quantitative Structure Toxicity Relationship (QSTR) pertaining to medical advances and study of environmental contamination. In this paper, all these studies will be focus just on QSAR to simplify the process of study.

II. QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIPS

The basic purpose served by QSAR is that it serves to define the interrelationship between the structure of a molecule and its characteristic properties (physical or biological activity, side effects) and uses statistical methods to extend this relationship to all compounds with similar structure. So, the net result is depicted by a very elegant equation-Function = f (molecular structural configuration and its properties).

The QSAR method amalgamates measurement techniques (biological and chemical measurements), statistical methods and the final compilation of results.The QSAR method can be used for molecular design in various sectors like environmental sciences, predicting biological activity of related compounds, weeding out the best drug candidates, analyzing drug pharmaco dynamics and enhancing the drug development process by putting forward new structural configurations.

A. Objective of Quantitative Structure Activity Relationships

The analysis of the interrelationship between structure and activity is done to understand the factors that influence the activity of a system so that the behavior of that particular system can be understood and analyzed in detail. The approach used to serve this purpose is to create a mathematical simulation that is based on experimental data and molecular properties which are ascertained for a variety of molecules. The simulator should be as close to the system being studied so that new structural configurations can be discovered within the context of biological and chemical phenomenon occurring within the system. There are a number of factors that affect the biological activity and can be described in terms of affinity to water, electrical charge, and effect of structural arrangement on chemical activity and relationship of different parts with each other amongst many others. Previously these were calculated on basis of actual experimental results, but now sophisticated computer techniques are used.

According to Computer-Aided Drug Design the interrelationship between the structure of a molecule and its effect as a drug in the human body helps us to understand drug mechanism in a better way. A simulation model with valid and consistent chemical variables can help in understanding various biological processes and thus increase the number of new structures thus proportionately increasing the number of drug candidates. So, in a nut shell it can be said that QSAR helps in understanding the biological activity of compounds which are still in the virtual space as well as of those who have not undergone testing and also helps to understand which chemical variable is the defining factor that influences the biological activity of the compound.

B. Underlying Principles

The methodology of QSAR depends on the numerical demonstration of chemical structure. The objective of QSAR is to find out the method in which the biological activity of a specific group of compounds is affected by the changes in the chemical structure. It was found that the descriptors, who exemplify the compounds, represent the aspects that control the biological activities. However, the same mechanism is applicable for expressing the biological activity of these descriptors. Hence, QSAR tries in finding those molecular aspects that have an effect on its function and the changes required for bringing an enhancement to their features. Therefore, it is expected that a relative equivalent change in the biological response reflects any sort of methodical change in the chemical structure, in case of a chain of molecules activated biologically.

The model of QSAR assumes that between the structure of molecule and biological activity an essential association exists. The systematic change leads to this relationship. Moreover, these correlations are shown by the multivariate physicochemical depiction of the group of compounds. Using several descriptors, the numerical form of the molecular structure can be decoded and this helps to understand every possible property (physical, chemical, and biological) of a chemical substance.

Nevertheless, stating the dissimilarity between correlation and causation is very essential. A satisfactory QSAR correlation means an efficient action of a compound, which is not always because of a specific descriptor. Extra information about several mechanisms that causes the biological action can be balanced by the absence of proof related to causation.

According to the principle of similarity, alike compounds have alike actions and this principle forms the base of QSAR model. Moreover, linear relationships are based on both the principle of similarity and principle of neighborhood. According to neighborhood principle, alike action is presented by the molecules found in the same area of the descriptor space.

C. QSAR Model

A multivariate mathematical correlation is expressed by QSAR. This relation is between a set of physicochemical features or descriptors, represented by $\{x_{ij}\}$, and an experimental function or biological activity represented by $\{y_i\}$. The QSAR relationship is a mathematical model. As it is used for accounting the observed action, hence it is quantitative. In case of a compound i, the linear equation connecting properties of the molecule, x_1, x_2… x_n, with the desired action, Then

$$y_i = x_{i1}b_1 + x_{i2} + ... + x_{in}b_n + e_i$$

In case of n selected descriptors, xj and when the earlier equation is expressed in a compact form, then the QSAR equation is:

$$y_i = \sum_{j=1}^{n} x_{ij}b_j + e_i$$

Here the linear slopes b, expresses the relation between the specific feature of molecule x_{ij} and the action y_i, in the compound represented by i where a constant is represented by e_i. A regression analysis is used to calculate the slopes and the constant. The models having a single dependent variable or y observation is only considered in this particular study. However few models can function along with various biological events. This variable quality decides the QSAR model’s potential. Certain features of the chemical structure that might be determined experimentally or theoretically are explained by the independent variables or descriptors. These descriptors are also the physicochemical properties. Poor models of QSAR are the outcome of inappropriate selection of independent variables. A great number of descriptors could be utilized in a usual study of QSAR, but as there are a lot of factors that can make any model associate successfully hence over fitting must be given the priority. The smallest number of descriptors representing the compounds action in the study is searched by the final equation of QSAR. A single independent variable to five compounds represents the maximum ratio.

III. CONDITIONS FOR APPLICABILITY OF QSAR

There are several conditions associated with various the features matching the model and must be fulfilled for developing an applicable and dependable model of QSAR [18-19]. The biological action of the chain of compounds and the physicochemical features must be related. Moreover, it is very usual and important that the same aforementioned mechanism must be used to explain the chemical actions. It is prohibited to use a related chemical structure. Nevertheless, determining the action mechanism is difficult, if not impossible, whereas establishing similarity of chemical is often less difficult. Therefore, in case of congeneric molecular series, which are also sets of chemically similar compounds, QSARs are developed expecting that their action mechanism will be similar. The compounds having different action mechanism does not fit in the association properly and lays outsides. Thus, the main areas required for the development of such models are the major conditions that are considered for building a suitable model of QSAR.

A. Selection of Compounds

For achieving valid outcomes it is very important to select suitable chemical sets that help in the development of QSAR models [20]. Chemicals applying a given activity impact through a common mechanism forms a suitable set. The common mechanisms are those that can be controlled using a single equation of QSAR.

1) Homogeneity

According to the chemical similarity and homogeneity of compounds, the system being investigated should have the same influence mechanism and also that the variability and diversity of chemical structures and features have certain limits. Hence, strong clustering is expected due to influential outliers’ absence.When compounds of various classes are separated into different subgroups, there is a formation of clusters. Treating each class independently is also an option that should be considered in this case.

2) Representativeness

The chemical domain of interest has to be extended according to the definition of the chemical space. This extension is performed by the selection of the group of compounds. Thus, for assessing the utility a great number of pertinent chemicals need to be chosen. Multivariate design [21] known as Statistical Molecular Design (SMD) can be used [22] for this purpose. Non-statistical techniques [23], Factorial Design (FD)[24], Fractional Factorial Design (FFD) [25], Central Composite Design (CCD), D-Optimal Design (DOD) [26, 27] Principal Component Analysis (PCA) [28] and Cluster Analysis (CA) [29] are among several procedures of selection that have been suggested.

B. Selection of Descriptors

In case of the physicochemical or structural descriptors, there can be a problem of collinearity due to content superfluous information in some descriptors; this happens particularly when a large number of descriptors have to be analyzed. The factors used in QSAR must have appropriate meaning and in physical aspects should be interpretable easily. Therefore, valuable approach into the mechanism must be provided by the selected descriptors [30-32].

C. Biological Data

High quality and dependable biological data is needed in terms of experimental activity. Well standardized assays having a clear and explicit endpoint must be used to measure the biological activities and that too in a steady and reliable manner [31].

Preferably, the same protocol and possibly same laboratory must be used to obtain the source of data. It is suitable to consider only a single source when data has to be retrieved from a literary source. Moreover, for having a base for detecting and rejecting outliers of the model, mechanistic insights of the chemicals should be considered carefully. Lastly, it is important to consider that biological measurements are subjected to experimental mistakes.

The techniques used for studying biological and toxicological endpoints are found in increasing complexity order, i.e. in silico techniques that accounts for electronic and general molecular properties, in vivo techniques that suits to more detailed studies on particular organs and individuals, and the last in vitro methods that at cellular level gives a satisfactory explanation.

1) Types of Data

On a continuous scale there is a distribution of biological data, which by the help of correlation methods leads to the derivation of a quantitative QSAR equation. At times it is categorized into distinct groups. For instance, based on the activity's strength, a chemical can be categorized as active or inactive, or in various classes. In these cases, other statistical methods, for example categorization methods should be applied. Here the compounds' physicochemical features are used for distinguishing between activity and inactivity. In case more than two such properties are utilized, they could be joined into principal elements; the various classes can help in differentiating by a plot of two major principal components.

D. Some Instruction to Apply QSAR

The models of QSAR must be easy, apparent, and mechanistically understandable [30, 31]. Though later an in-depth study of statistical methods will be conducted, yet over-fitting and non-linearity of data must be avoided. However, clearness and justification of the model are recommended strongly.

IV.QSAR STEPS

Based on the multivariate data study and statistical experimental design, there are a number of iterative steps during the development of QSAR in drug design, which causes the compounds design having an action profile [30-32].

A. Formulation of Classes of Similar Compounds

Selecting the biological activities of interest, choosing structural features to be varied and selecting structural domain (structural class) are among the initial steps. QSARs being based on compounds that are too different in structure are not desirable because the biological action mechanism often varies between diverse compounds. Therefore, the perfect circumstances and classes of compounds that are same both chemically and biologically correspond. Every compound in each class is same structure-wise and its working depends on the similar action mode it does function. However, for causing certain systematic variation in the biological function the compounds should be sufficiently different.

The compounds, chain of compounds are classified based on their chemical structure; this is done to form classes of alike or analogous compounds. The compounds general strength of character, their reactivity, substituent, and information about their biological mechanism helps in achieving the purpose. New classes are formed in case it is found by the following data study that a homogeneous class is not formed by the compounds being studied.

B. Quantitative of Structural Variation and Choice of the QSAR Model

Generally, a number of descriptor variables needs to have adequate and significant data about the biological activities so that the changes in the structure can be described appropriately. Thus description of the structure is multivariate. However, predicting about the useful descriptor variables is very complicated.

Therefore, having an independent design variables group that influences the biological effect is well-situated. However, distinguishing design variables that change freely is not possible during the molecular optimization. In optimization of molecules substitution designs or the whole structure of molecule is changed.

C. Selection the Training Set of Compounds (Series Design)

In any model of QSAR, the chosen training set must standardize the model so that a well-balanced division is shown as well as representative compounds are present in it. Selecting the training set systematically allows the important features of structure to vary systematically and simultaneously, which is required to achieve the standardization mentioned above.

D. Synthesis and Biological Testing

Minimization of the biological testing is necessary; hence for obtaining a wide and constant picture of biological features it is important to subject just the representative training test to extensive testing. The biological variables cover several features of the biological profiles of compounds that are studied, and all these forms the response matrix. With more conduction of biological tests, the resulting QSAR model's stability becomes better leading to an improved predictive potential. Time and money are saved when only some representative compounds are tested, thus sticking onto the animal welfare principles. The association between the administered doses and the responses they draw out are shown by the dose-response curves; usually recorded in terms of biological dimensions.

E. QSAR Development: Data Analysis

Data relating to important aspects of the chemical and biological data structure is extracted for calculating the best mathematical expression; this expression connects the physicochemical descriptors and biological responses together. Changing few descriptors variables, deleting outliers, and showing different chemical and biological features might be necessary. Whether a descriptor variable is applicable to certain use or not is found out by analysis of QSAR

F. Validation and Predictions for Non-Tested Compounds

Lastly, predicting the biological activities of non-tested compounds is the final objective of a QSAR. These compounds belong to the class that is being researched. However, at first an experiment is conducted to verify the model's predictive potential. For this purpose, few additional compounds are tested biologically which is similar to that of the training set. Later the result of the experiment and the values predicted by the QSAR are compared. It might be used for a more broad prognostication in case the QSAR predicts within acceptable limits. The mistakes in prediction, the accuracy and the obtained biological measurements range must be compared.

G. Data Analysis and Interpretation of Results for the Proposal of New Compounds

Indeed, any QSAR development where the steps are severally repeated till the time adequate information related to a class of compounds is achieved is known as an iterative cycle. This method is performed for designing compounds having the activity profile that is desired or for making a conclusion that such a profile is not possible to be achieved.

V. Conclusion

Quantitative Structure Activity Relationship QSAR illustrates the interrelationship between the structure of a molecule and its biological properties and employs statistical methods to expand this relationship to all compounds with similar structure. QSAR is the most responsible computational technique used for years for well measuring on substituent's physicochemical property and biological activity of molecules. Currently the main objective of these researches is to guess biological activity of unknown molecules on the basis of previously synthesized compounds. Number of developments has been taken and different constitutional descriptors have been explored. QSAR model is based on the principle of similarity and one of QSAR's objectives is to determine the method that combines the biological activity and the changes in a chemical structure in a specific group of a compound. Some of the conditions of applying QSAR include the selection of a suitable chemical sets and utilizing different descriptors of molecular, these descriptors are frequently used to compute the basis of physicochemical properties of ligand molecule.

The constructed mathematical model which indicates the association between molecular descriptors and biological activity, is validated internally and externally in order to assess the predicative power of the QSAR model. The interpretations of these models are carried out by various methods like pattern recognition, machine and artificial intelligence. The constructed mathematical model which indicates the association between molecular descriptors and biological activity, is validated internally and externally in order to assess the predicative power of the QSAR model.

Since then a variety of quantitative structure activity relationship studies have been reported to predict cytotoxicities, depressant and antibacterial activity of chemical compounds. Thousands of QSAR equations have been formulated using the QSAR methodology to validate and elucidate the predicative power of QSAR hypothesis about the mechanism of action of drugs at the molecular level and a more complete understanding of physicochemical phenomena such as hydrophobicity. In 1962 Hansch and Muir published their brilliant study on the 2D structure-activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity.

The QSAR method is used in several sectors like environmental sciences, predicting biological activity of related compounds, selecting the best drug candidates. QSAR is considered to be an extremely efficient instrument in molecular design and accelerates the initial steps of drug development process. Furthermore, it enhances the effectiveness and reduces the cost of newly developed drugs.The QSAR is a knowledge-based method where a statistical prediction model is made about biological activity and the presence of molecular descriptor. The aim of carrying out a QSAR study is with the help of computational methods the QSAR model can help evaluate biological activity; this is mostly done to reduce failure rate in the drug development process.

Acknowledgment

This work is supported by the Ministry of Higher Education (MOHE) and the Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under the Research University Grant Category (R.J130000.7828.4F985).

References

[1] Combes, R., Barratt, M., & Balls, M. (2003). An Overall Strategy for the Testing of Chemicals for Human Hazard and Risk Assessment under the EU REACH System. *ATLA-NOTTINGHAM-*, 31(1), 7-20.

[2] Werth, B. (2013). The Billion-dollar *Molecule: The Quest for the Perfect Drug.* Simon and Schuster.

[3] Dickson, M., & Gagnon, J. P. (2004). Key Factors in the Rising Cost of New Drug Discovery and Development. *Nature Reviews Drug Discovery,* 3(5), 417-429.

[4] Gore, M., & Desai, N. S. (2014). Computer-aided Drug Designing. *Clinical Bioinformatics,* 313-321.

[5] Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational Methods in Drug Discovery. *Pharmacological Reviews,* 66(1), 334-395.

[6] Goto, S., Okuno, Y., Hattori, M., Nishioka, T., & Kanehisa, M. (2002). LIGAND: Database of Chemical Compounds and Reactions in Biological Pathways. *Nucleic Acids Research,* 30(1), 402-404.

[7] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., & Schomburg, D. (2004). BRENDA, the Enzyme Database: Updates and Major New Developments. *Nucleic Acids Research,* 32(suppl_1), D431-D433.

[8] Di, L., Fish, P. V., & Mano, T. (2012). Bridging Solubility between Drug Discovery and Development. *Drug Discovery Today,* 17(9), 486-495.

[9] Dearden, J. C. (2016). The History and Development of Quantitative Structure-Activity Relationships (QSARs). Oncology: Breakthroughs in Research and Practice: Breakthroughs in Research and Practice, 67.

[10] Gennis, R. B. (Ed.). (2013). *Biomembranes: Molecular Structure and Function.* Springer Science & Business Media.

[11] Martin, Y. C. (2010). *Quantitative Drug Design: A Critical Introduction.* CRC Press.

[12] Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R., & Consonni, V. (2013). Quantitative Structure–Activity Relationship Models for Ready Biodegradability of Chemicals. *Journal of Chemical Information and Modeling,* 53(4), 867-878.

[13] Kubinyi, H. (2002). From Narcosis to Hyperspace: The History of QSAR. *Molecular Informatics,* 21(4), 348-356.

[14] Banfi, L. (2006). Molecular Diversity and Combinatorial Chemistry. *Molecular Diversity,* 10(1), 85-85.

[15] Feher, M., & Schmidt, J. M. (2003). Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences,* 43(1), 218-227.

[16] Salim, N., Holliday, J., & Willett, P. (2003). Combination of Fingerprint-based Similarity Coefficients Using Data Fusion. *Journal of Chemical Information and Computer Sciences,* 43(2), 435-442.

[17] Reker, D., & Schneider, G. (2015). Active-learning Strategies in Computer-Assisted Drug Discovery. *Drug Discovery Today,* 20(4), 458-465.

[18] Dahlgren, M. K., Kauppi, A. M., Olsson, I. M., Linusson, A., & Elofsson, M. (2007). Design, Synthesis, and Multivariate Quantitative Structure−Activity Relationship of Salicylanilides Potent Inhibitors of Type III Secretion in Yersinia. *Journal of Medicinal Chemistry,* 50(24), 6177-6188.

[19] Eriksson, L., Byrne, T., Johansson, E., Trygg, J., & Vikström, C. (2013). Multi-and *Megavariate Data Analysis Basic Principles and Applications.* Umetrics Academy.

[20] Roy, K. (2007). On Some Aspects of Validation of Predictive Quantitative Structure–activity Relationship Models. *Expert Opinion on Drug Discovery,* 2(12), 1567-1577.

[21] Bro, R. (2003). Multivariate Calibration: What is in Chemometrics for the Analytical Chemist? *Analytica Chimica Acta,* 500(1), 185-194.

[22] Linusson, A., Gottfries, J., Lindgren, F., & Wold, S. (2000). Statistical Molecular Design of Building Blocks for Combinatorial Chemistry. *Journal of Medicinal Chemistry,* 43(7), 1320-1328.

[23] Topliss, J. G. (1972). Utilization of Operational Schemes for Analog Synthesis in Drug Design. *Journal of Medicinal Chemistry*, 15(10), 1006-1011.

[24] Derakhshandeh, K., Erfan, M., & Dadashzadeh, S. (2007). Encapsulation of 9-nitrocamptothecin, a Novel Anticancer Drug, in Biodegradable Nanoparticles: Factorial Design, Characterization and Release Kinetics. *European Journal of Pharmaceutics and Biopharmaceutics*, 66(1), 34-41.

[25] Jaynes, J., Ding, X., Xu, H., Wong, W. K., & Ho, C. M. (2013). Application of Fractional Factorial Designs to Study Drug Combinations. *Statistics in Medicine*, 32(2), 307-318.

[26] Giraud, E., Luttmann, C., Lavelle, F., Riou, J. F., Mailliet, P., & Laoui, A. (2000). Multivariate Data Analysis using D-Optimal Designs, Partial Least Squares, and Response Surface Modeling: A Directional Approach for the Analysis of Farnesyltransferase Inhibitors. *Journal of Medicinal Chemistry*, 43(9), 1807-1816.

[27] Hasegawa, K., Miyashita, Y., & Funatsu, K. (1997). GA Strategy for Variable Selection in QSAR Sudies: GA-based PLS Analysis of Calcium Channel Antagonists⊥. *Journal of Chemical Information and Computer Sciences,* 37(2), 306-310.

[28] Smilde, A. K., Jansen, J. J., Hoefsloot, H. C., Lamers, R. J. A., Van Der Greef, J., & Timmerman, M. E. (2005). ANOVA-simultaneous Component Analysis (ASCA): A New Tool for Analyzing Designed Metabolomics Data. *Bioinformatics*, 21(13), 3043-3048.

[29] Novellino, E., Fattorusso, C., & Greco, G. (1995). Use of Comparative Molecular Field Analysis and Cluster Analysis In Series Design. *Pharmaceutica Acta Helvetiae,* 70(2), 149-154.

[30] Cronin, M. T., & Schultz, T. W. (2003). Pitfalls in QSAR. *Journal of Molecular Structure: THEOCHEM,* 622(1), 39-51.

[31] Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martínez-Mayorga, K., Langer, T., ... & Agrafiotis, D. K. (2012). Recognizing Pitfalls in Virtual Screening: A Critical Review. *Journal of Chemical Information and Modeling,* 52(4), 867-881.

[32] Doweyko, A. M. (2008). QSAR: Dead Or Alive? *Journal of Computer-Aided Molecular Design*, 22(2), 81-89.