IJTI | INTERNATIONAL JOURNAL OF THERAPEUTIC INNOVATION

Review article

QSAR and docking study: A review

Mohd Basheer, Somesh Kumar Saxena*, Shailesh Jain

SAM College of Pharmacy, SAM Global University, Raisen, Madhya Pradesh, India

Corresponding author: Somesh Kumar Saxena, ✉ somesh1207@gmail.com, **Orcid Id**: https://orcid.org/ 0000-0003-4824-2853

Received - 20-02-2025, **Revised** - 28-02-2025, **Accepted** - 12-03-2025 **(DD-MM-YYYY)**

Refer this article

Mohd Basheer, Somesh Kumar Saxena, Shailesh Jain, QSAR and docking study: A review. March-April 2025, V3 – I2, Pages - 01 – 05. Doi: https://doi.org/10.55522/ijti.v3i2.0107.

ABSTRACT

Quantitative structure–activity relationship models (QSAR models) are regression or classification models used in the chemical and biological sciences and engineering. Like other regression models, QSAR regression models relate a set of "predictor" variables (X) to the potency of the response variable(Y), while classification QSAR models relate the predictor variables to a categorical value of the response variable. In QSAR modeling, the predictors consist of physico-chemical properties or theoretical molecular descriptors of chemicals; the QSAR response-variable could be a biological activity of the chemicals. QSAR models first summarize a supposed relationship between chemical structures and biological activity in a data-set of chemicals. Second QSAR models predict the activities of new chemicals. Related terms include quantitative structure–property relationships (QSPR) when a physico-chemical property or reactivity is modeled as response variable.

Keywords: QSAR, Docking, QSPR, Drug Design, Molecular Descriptor, Biological activity.

INTRODUCTION

Quantitative Structure Activity Relationship (QSAR) is the mathematical relationship liking chemical structure and pharmacological activity in a quantitative manner for a series of compound. The aim of QSAR is to develop a correlation between forms of activity (Biological activity) and properties (physiochemical properties) for a set of molecules QSAR started with similar correlation between chemical reactivity and structure [1].

Application of QSAR to Drug Design Practice

After formation of statistically significant as well as physico-chemically significant meaningful correlation equation for a given set of compounds, the information contained in the equation can be used to design new compounds. According to the method of utilization of the information, examples could be classified into at least three categories:

Extrapolation of certain parameters towards directions enhancing the potency. If the correlation is linear in terms of certain physico-chemical parameters, structural modifications so as to extrapolate these parameters towards directions increasing the value of their terms should generate compounds of more potent activity [2].

Insights of QSAR and Docking Study

The field of drug discovery and design has significantly advanced with the application of computational techniques, particularly Quantitative Structure-Activity Relationship (QSAR) and molecular docking studies. QSAR is a mathematical approach that models the relationship between the chemical structure of a compound and its biological activity. On the other hand, molecular docking studies simulate the interaction between a ligand (usually a drug candidate) and a target protein to predict the binding affinity and stability of the complex. These methods help in understanding the molecular basis of the drug-target interaction, assisting in the optimization of drug candidates, and reducing the need for time-consuming and costly experimental procedures [3].

This article delves into the insights provided by QSAR and docking studies, discussing their principles, applications, advantages, limitations, and the synergy between them in drug discovery.

Principles of QSAR

QSAR is a computational method that correlates the physicochemical properties of molecules to their biological activity. The key idea is that the structure of a molecule determines its interaction with a biological target, which ultimately influences its pharmacological effect. The mathematical model in QSAR attempts to predict the activity of compounds based on their molecular descriptors [4].

QSAR Modelling Typically Follows Several Key Steps

Molecular Descriptor Calculation: Molecular descriptors are quantitative representations of the chemical structure of a molecule. These descriptors can be related to molecular shape, size, electrostatic properties, and hydrophobicity. Popular descriptors include topological indices, charge distribution, molecular weight, and others.

Data Set Preparation: The set of compounds and their corresponding biological activity values (e.g., IC50, EC50, Ki) are collected. The dataset is then divided into a training set and a test set for model development and validation.

Model Development: Various mathematical techniques, such as linear regression, non-linear regression, and machine learning algorithms, are applied to establish a correlation between the descriptors and biological activity.

Model Validation: The model is validated using different statistical parameters like the correlation coefficient (R^2), cross-validation, and root mean square error (RMSE) to assess its predictive power [5].

Applications of QSAR

QSAR models are extensively used in drug design, including

Prediction of Drug Activity: QSAR models help predict the biological activity of novel compounds based on their structural features. This can guide the selection of compounds for experimental testing.

Optimization of Drug Candidates: QSAR allows for the optimization of lead compounds by modifying their chemical structure to enhance desired properties (e.g., increasing potency or reducing toxicity).

Virtual Screening: QSAR can be used to screen large libraries of compounds for potential drug candidates before physical synthesis, thus saving time and resources.

Toxicological Predictions: QSAR is also used in predicting the toxicity of chemicals, thus playing an essential role in risk assessment and drug safety [6].

Challenges in QSAR

While QSAR offers powerful insights into drug design, it faces several challenges

Data Availability: QSAR models require a large and diverse dataset of compounds with known biological activities. Incomplete or biased datasets can lead to inaccurate predictions.

Model Over fitting: If the model is overly complex, it may perform well on the training set but fail to generalize to new compounds, leading to over fitting.

Descriptor Selection: The choice of molecular descriptors significantly impacts the performance of the model. Selecting the right descriptors is crucial for the success of QSAR [7].

QSAR Methods

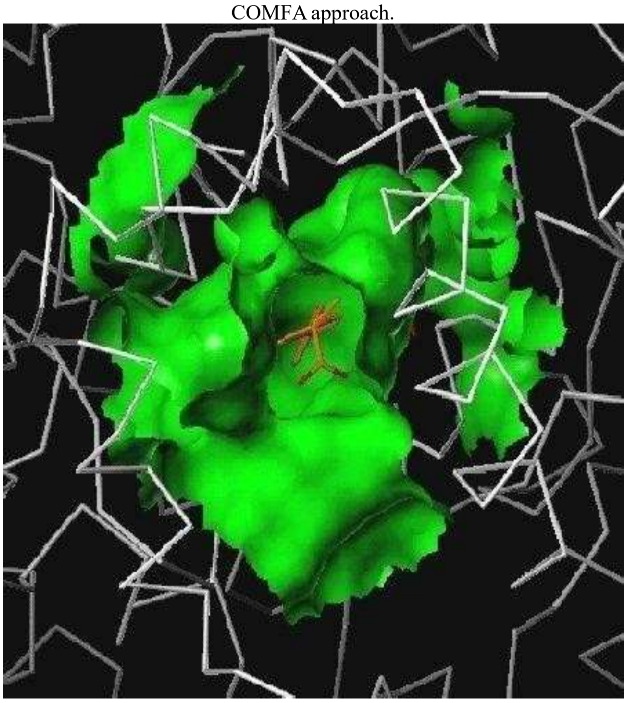
1.2-D QSAR.

It do not consider the 3D features.

Molecules are represented by descriptors, numerical values characterizing various aspect of molecular structure.

ADAPT software uses 2-D QSAR.

Figure 1: 3D structure of the molecule is considered exemplified by COMFA approach.



COMFA (Comparative molecular field analysis) uses statistical correlation techniques for the analysis of the quantitative relationship between the biological activity of a set of compound with a specific alignment and their 3Delectronic an steric properties.

The molecule is aligned on the grid and various properties are evaluated [8].

Requires accurate alignment.

Only single conformation is considered.

Fragment Based (Group Contribution

The structure (and hence the activity) of a molecule could be defined as the sum of its individual atoms, but it is better defined for QSAR purposes as the sum of its chemical fragments. Analogously, the "partition coefficient" – a measurement of differential solubility and itself a component of SAR predictions -- can be predicted either by atomic methods (known as "XLogP" or "ALogP") or by chemical fragment methods (known as "CLogP" and other variations). It has been shown that the LogP of compound can be determined by the sum of its fragments; fragment based methods are generally accepted as better predictors than atomic-based methods. Fragmentary LogP value shave been determined statistically, based on empirical data for known LogP values. This method gives mixed results and is generally not trusted to have accuracy of more than ±0.1 units [9].

Group or Fragment based QSAR is also known as GQSAR. GQSAR allows flexibility to study various molecular fragments of interest in relation to the variation in biological response. The molecular fragments could be substituents at various substitution sites in congeneric set of molecules or could be on the basis of pre-defined chemical rules in case of non-congeneric set. GQSAR also considers cross-terms fragment descriptors, which could be helpful in identification of key fragment interactions in determining variation of activity. Lead discovery using Fragnomics is an emerging paradigm. In this context FB-QSAR proves to be a promising strategy for fragment library design and in fragment-to-lead identification endeavours [10].

3D-QSAR

3D-QSAR refers to the application of force field calculations requiring three-dimensional structures, e.g. based on protein crystallography or molecule superimposition. It uses computed potentials, e.g. the Lennard-Jones potential, rather than experimental constants and is concerned with the overall molecule rather than a single substituent. It examines the steric fields (shape of the molecule), the hydrophobic regions (water-soluble surfaces), and the electrostatic fields [11].

The created data space is then usually reduced by a following feature extraction (see also dimensionality reduction). The following learning method can be any of the already mentioned machine learning methods, e.g. support vector machines. An alternative approach uses multiple-instance learning by encoding molecules as sets of data instances, each of which represents a possible molecular conformation. A label or response is assigned to each set corresponding to the activity of the molecule, which is assumed to be determined by at least one instance in the set (i.e. some conformation of the molecule).

On June 18, 2011 the CoMFA patent has dropped any restriction on the use of GRID and PLS technologies and the RCMD team has opened a 3D QSAR web server (www.3d-qsar.com) based on the 3-D QSAutogrid/R engine. 3-D QSAutogrid/R covers all the main features of CoMFA and GRID/GOLPE with implementation by multiprobe/multi-region variable selection (MPGRS) that improves the simplification of interpretation of the 3-D QSAR map. The methodology is based on the integration of the molecular interaction fields as calculated by AutoGrid and the R statistical environment that can be easily coupled with many free graphical molecular interfaces such as UCSF-Chimera, AutoDock Tools, JMol and others [12].

Evaluation of the quality of QSAR models

QSAR modeling produces predictive models derived from application of statistical tools correlating biological activity (including desirable therapeutic effect and undesirable side effects)or physico-chemical properties in QSAR models of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative of molecular structure and/or properties. QSARs are being applied in many disciplines for example risk assessment, toxicity prediction, and regulatory decisions in addition to drug discovery and lead optimization. Obtaining a good quality QSAR model depends on many factors, such as the quality of input data, the choice of descriptors and statistical methods for modeling and for validation. Any QSAR modeling should ultimately lead to statistically robust and predictive models capable of making accurate and reliable prediction of the modeled response of new compounds.

For Validation of QSAR Models Usually Various Strategies Are Adopted

Internal validation or cross-validation.

External validation by splitting the available data set into training set for model development and prediction setfor model predictivity check;

Blind external validation by application of model on new external data.

Bata randomization or Y-scrambling for verifying the absence of chance correlation between the response and the modeling descriptors.

The success of any QSAR model depends on accuracy of the input data, selection of correlation between the response and the modeling descriptors. The use of very much appropriate descriptors and statistical tools, and most importantly validation of the developed model. Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose; for QSAR models validation must be mainly for robustness, prediction performances and applicability domain of the models. Leave one-out cross-validation generally leads to an overestimation of predictive capacity, and even with external validation, no one can be sure when the selection of training and test sets was manipulated to maximize the predictive capacity of the model being published [13].

QSAR and Drug Design

Quantitative structure-activity relationships (QSAR) represent an attempt to correlate structural or property descriptors of compounds with activities. These physicochemical descriptors, which include parameters to account for hydrophobicity, topology, electronic properties, and steric effects, are determined empirically or, more recently, by computational methods. Activities used in QSAR include chemical measurements and biological assays. QSAR currently are being applied in many disciplines, with many pertaining to drug and environmental risk assessment.

Basic Requirements in QSAR Studies

All analogs belong to a congeneric series

All analogs exert the same mechanism of action

All analogs bind in a comparable manner

The effects of isosteric replacement can be predicted inding affinity is correlated to interaction energies

Current State and Perspectives of 3D - QSAR

Quantitative structure-activity relationships (QSAR) have played an important role in the design of pharmaceuticals and agrochemicals. All QSAR techniques assume that all the compounds used in analyses bind to the same site of the same biological target. However, each method differs in how itdescribes structural properties of compounds and how it finds the quantitative relationships between the properties and activities. The Hansch-Fujit approach, the so-called classical QSAR.

Despite the usefulness, classical QSAR techniques cannot be applied to all datasets due to the lack of availability of physicochemical parameters of the whole molecule or its substituents and often it is difficult to estimate those values. In addition, molecular properties based on the three dimensional (3D) structure of compounds may be useful in describing the ligand-receptor interactions. Recently, a variety of ligand-based 3D-QSAR methods such as Comparative Molecular Field Analysis (CoMFA) have been developed and widely used in medicinal chemistry.

This review describes different 3D-QSAR techniques and indicates their advantages and disadvantages. Several studies about 3D-QSAR of ADME-toxicity and perspective of 3D-QSAR are also described in this review.

Applications of QSAR in Drug Discovery

Within pharmaceutical research, it is often relevant to make the distinction between descriptive and predictive quantitative structure activity relationships (QSARs). Descriptive QSAR modeling is used extensively to understand structure activity relationships with respect to various endpoints within a chemical series and to guide structural changes driving the biological activity in a desired direction.

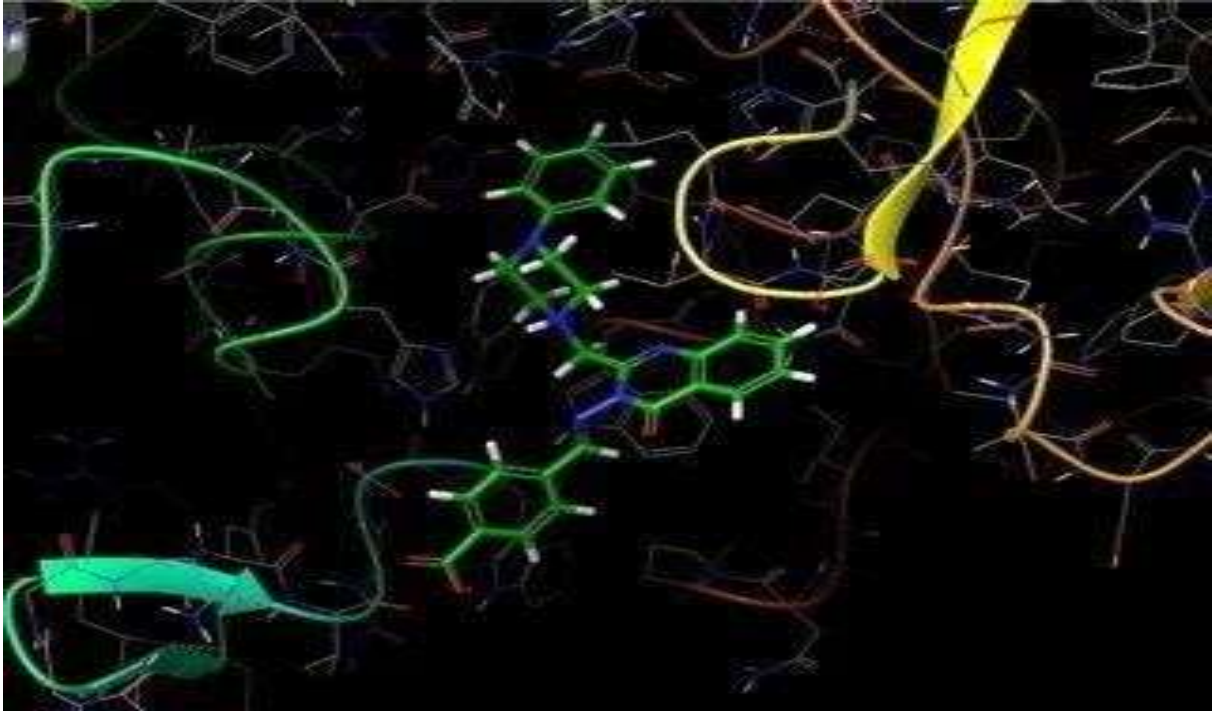
Predictive QSAR modeling is used mainly for biological responses and physical properties relevant to all pharmaceutical projects such as modeling of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET). Economical necessities and the concern for laboratory animals constantly drive the pharmaceutical industry towards replacing in vivo studies with in vitro experiments and in silico methods. Hence, predictive QSAR modeling is becoming increasingly important within drug discovery, in particular for ADMET characterization. ADMET modeling is used throughout the pre-clinical discovery and development process from hit prioritization to selection of compounds for in-vivo testing. Developing a QSAR model, intended to be applicable to all pharmaceutically relevant parts of chemical space, is exceedingly challenging. In addition, many ADMET endpoints and in particular toxicological endpoints are dependent upon a multitude of molecular mechanisms. Many of these mechanisms may remain unknown and even with a clear mechanistic understanding; the underlying physical processes are complex and structural knowledge often not available [13].

Limitations of QSAR modeling

While there are limits to the Hansch approach, it permitted complex biological systems to be modeled successfully using simple parameters. The approach has been used successfully to predict substituent effects in a wide number of biological assays. The main problem with the approach was the large number of compounds which were required to adequately explore all structural combinations. Further, the analysis methods did not lend themselves to the consideration of conformational effects. Several authors have published articles that provide additional background on the Hansch approach. The CASE program extended the techniques in ADAPT by using topological methods to define sub structural fragments which were essential for activity. CASE was able to differentiate between positional isomers. Both CASE and ADAPT are limited to analyzing structurally similar data sets.

In 1988, Richard Cramer proposed that biological activity could be analyzed by relating the shape-dependent steric and electrostatic fields for molecules to their biological activity. Additionally, rather than limiting the analysis to fitting data to a regression line, CoMFA (Comparative Molecular Field Analysis) utilized new methods of data analysis, PLS (Partial Least Squares) and cross-validation, to develop models for activity predictions [14].

Figure 2: 3-D QSAR utogrid-R MPGRS example image



CONCLUSION

QSAR and docking studies are invaluable tools in the field of drug discovery. QSAR provides a quantitative relationship between chemical structure and biological activity, while docking offers insights into the molecular interactions between a drug and its target. When used together, these techniques complement each other and enhance the efficiency and accuracy of the drug design process.

The integration of QSAR and docking can significantly reduce the time and cost of drug development by predicting potential drug candidates before experimental testing. However, both methods have their limitations, including issues with data availability, descriptor selection, receptor flexibility, and scoring function accuracy. Despite these challenges, the combination of QSAR and docking remains one of the most powerful approaches in modern drug discovery.

REFERENCES

1. Tropsha A, Gramatica P, Gombar VJ, 2003. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Comb Sci. 22, Pages 69-77. Doi: 10.1021/cr950066q (http://dx.doi.org/10.1021%2Fcr950066q).

2. Gramatica P, 2007. Principles of QSAR models validation: internal and external. QSAR & Comb Sci. 26, Pages 694-701. Doi: https://doi.org/10.1002/qsar.200610151.

3. Chirico N, Gramatica P, 2012. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J Chem Inf Model. 52(8), Pages 2044–2058. Doi: 10.1021/ci300084j (http://dx.doi.org/10.1021%2Fci300084j).

4. Thompson SJ, Hattotuwagama CK, Holliday JD, 2006. On the hydrophobicity of peptides: Comparing empirical predictions of peptide log P values. Bioinformation. 1(7), Pages 237-341. Doi: 10.6026/97320630001237.

5. Wildman SA, Crippen GM, 1999. Prediction of physicochemical parameters by atomic contributions. J Chem Inf Comput Sci. 39(5), Pages 868–873. Doi: 10.1021/ci990307l.

6. Ajmani S, Jadhav K, Kulkarni SA, 2008. Group-Based QSAR (G-QSAR): Mitigating Interpretation Challenges in QSAR. QSAR & Combinatorial Science. 28(1), Pages 36–51. Doi:10.1002/qsar.200810063.

7. Manoharan P, Vijayan RSK, Ghoshal N, 2010. Rationalizing fragment based drug discovery for BACE1: insights from FB-QSAR, FB-QSSR, multi objective (MO-QSPR) and MIF studies. Journal of Computer-Aided Molecular Design. 24(10), Pages 843–864. Doi: http://dx.doi.org/10.1007%2Fs10822-010-9378-9).

8. Tim Cheeseright, 2009. The Identification of Bioisosteres as Drug Development Candidates. Medicine.

9. Leach Andrew R, 2001. Molecular modelling: principles and applications. Englewood Cliffs. N J Prentice Hall.

10. Vert Jean-Philippe, Schl_kopf Bernhard, Schölkopf, Bernhard, 2004. Kernel methods in computational biology. Cambridge Mass. Doi: https://doi.org/10.7551/mitpress/4057.001.0001.

11. Dietterich Thomas G, Lathrop Richard H, Lozano-Pérez Tomás, 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89(1–2), Pages 31–71. Doi: 10.1016/S0004-3702(96)00034-3.

12. Ballante Flavio, Ragno Rino, 2012. 3-D QSAutogrid/R: an alternative procedure to build 3-D QSAR models. Methodologies and applications. J Chem Inf Model. 52(6), Doi: 10.1021/ci300123x.

13. Gusfield Dan, 1997. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge. UK: Cambridge University Press.

14. Helma, Christoph, 2005. Predictive toxicology. Washington. DC Taylor & Francis.