CHEMISTRY

Applications and Significance of Quantitative Structure–Activity Relationship (QSAR) in Rational Molecular Design: A Review

Vineeta Asati[1], Varsha Nigam[1], Mehzish Ahmed[2], V. Abhishek[2], Shweta Sharma[2], Neelu Singh[3]

[1]Sanjeev Agrawal Global Educational (SAGE) University, Bhopal, Madhya Pradesh, India
[2]Career College (Autonomous), Bhopal, Madhya Pradesh, India
[3]Sagar Institute of Research and Technology – Excellence, Bhopal, Madhya Pradesh, India

Published November 2026

Abstract

Understanding how molecular structure governs biological activity is fundamental to rational drug discovery and predictive toxicology. While a qualitative evaluation of structural fragments influencing biological response is termed Structure-Activity Relationship (SAR), expressing these correlations mathematically produces a Quantitative Structure-Activity Relationship (QSAR). QSAR establishes a statistically significant mathematical link between the biological activity of a congeneric series of compounds and their physicochemical or structural descriptors. As a cornerstone of Computer-Aided Drug Design (CADD), QSAR enables the in silico prediction of pharmacological potency, pharmacokinetic profiles, and toxicological endpoints. This review outlines the historical evolution, multidimensional classifications (1D to 6D), key physicochemical parameters, statistical modeling algorithms, and validation criteria of QSAR methodologies in modern molecular design.

Keywords: QSAR; computer-aided drug design (CADD); 3D-QSAR; CoMFA; CoMSIA; physicochemical descriptors; chemometrics

1 Introduction

Explaining the relationship between chemical architecture and biological response is a primary objective in medicinal chemistry and molecular design. A qualitative assessment of how specific pharmacophoric fragments or substructures modulate biological effects is known as Structure-Activity Relationship (SAR), which forms the conceptual basis of Computer-Aided Drug Design (CADD). When these structural features are quantified using numerical descriptors and correlated mathematically with biological endpoints, the methodology advances to Quantitative Structure-Activity Relationship (QSAR).

The fundamental premise of QSAR is that the molecular structure of a compound dictates its physicochemical properties and chemical reactivity, which in turn govern its biological and toxicological behavior [1, 2]. These molecular descriptors—quantifying hydrophobicity, topology, electronic distribution, and steric bulk—are derived either empirically or computationally. Beyond accelerating lead optimization, robust QSAR models significantly reduce development costs and minimize reliance on in vivo animal testing [3, 4].

1.1 Historical evolution of QSAR

The conceptual origins of QSAR date back to the late 19th century, evolving into a rigorous computational discipline over the following decades:

- **1860s–1900s (Early Foundations):** Crum-Brown and Fraser proposed that physiological action is a function of chemical constitution, followed by Meyer and Overton, who independently correlated the narcotic potency of organic compounds with their oil–water partition coefficients.
- **1960s (Modern QSAR):** Corwin Hansch and Toshio Fujita established modern QSAR by integrating Linear Free Energy Relationships (LFER) with biological activity through multiparameter linear regression [1, 2]. Simultaneously, Free and Wilson introduced an additive mathematical model based on substituent contributions [5].
- **1980s–1990s (3D-QSAR Era):** Richard Cramer introduced Comparative Molecular Field Analysis (CoMFA), shifting QSAR from two-dimensional descriptors to three-dimensional spatial interaction fields.
- **2000s–Present (Machine Learning & Big Data):** QSAR has integrated advanced machine learning algorithms, chemoinformatics, and multidimensional modeling (4D–6D QSAR) for high-throughput virtual screening and regulatory toxicology [6, 7].

2 Classification of QSAR methodologies

Based on the dimensionality of the molecular descriptors utilized, QSAR methodologies are categorized into six distinct classes (Table 1).

Table 1. Classification of QSAR methodologies.

Dimensionality	Description and parameters used
1D-QSAR	Correlates biological activity with global, bulk molecular properties and constitutional descriptors (e.g., molecular weight, pKa, log P).
2D-QSAR	Incorporates two-dimensional structural topology, connectivity indices, and physicochemical parameters without considering spatial orientation.
3D-QSAR	Correlates biological activity with non-covalent interaction fields (steric, electrostatic, hydrophobic) surrounding the 3D conformation of aligned ligands.
4D-QSAR	Incorporates conformational flexibility and ensemble sampling of ligand–receptor binding modes alongside 3D spatial properties.
5D-QSAR	Accounts for receptor flexibility by explicitly modeling multiple induced-fit conformations within the 4D-QSAR framework.
6D-QSAR	Integrates different solvation models and desolvation energy profiles into 5D-QSAR to simulate physiological environments accurately.

2.1 Advances in 3D-QSAR

- **Comparative Molecular Field Analysis (CoMFA):** Developed by Cramer et al., CoMFA predicts biological activity by evaluating steric (Lennard-Jones) and electrostatic (Coulombic) interaction energies at grid points surrounding superimposed bioactive conformations. While highly informative, CoMFA is sensitive to molecular alignment and potential chance correlations if grid parameters are over-adjusted.

- **Comparative Molecular Similarity Indices Analysis (CoMSIA):** CoMSIA overcomes the limitations of CoMFA's steep grid cut-off potentials by employing Gaussian-type distance-dependent functions derived from a modified SEAL algorithm. It evaluates steric, electrostatic, hydrophobic, and hydrogen-bonding fields, yielding smoother, more easily interpretable contour maps with comparable statistical robustness.

3 Core physicochemical parameters in QSAR

In classical Hansch analysis, biological activity is expressed as a linear or parabolic function of three primary physicochemical parameters [1, 2], as shown in Equation (1), or, for hydrophobicity, in the parabolic form of Equation (2):

$$\log(1/C) = a\pi + b\sigma + cE_s + d \tag{1}$$

$$\log(1/C) = -k_1(\log P)^2 + k_2(\log P) + k_3 \tag{2}$$

where C is the molar concentration required to elicit a standard biological response, π denotes hydrophobicity, σ represents electronic effects, E_s signifies steric factors and a, b, c and d are regression coefficients.

3.1 Hydrophobicity (log *P* and π)

Hydrophobicity governs the ability of a drug molecule to partition from an aqueous phase across lipophilic biological membranes to reach its target receptor.

- **Partition Coefficient (P):** Defined as the ratio of unionized drug concentrations between n-octanol and an aqueous phosphate buffer (pH 7.4), commonly determined via the shake-flask method (Equation 3).

$$P = \frac{[\text{Drug}]_{\text{octanol}}}{[\text{Drug}]_{\text{aqueous}}} \tag{3}$$

- **Substituent Hydrophobicity Constant (πX):** Quantifies the hydrophobic contribution of a specific substituent (X) relative to a parent unsubstituted molecule (H) [2] (Equation 4).

$$\pi_X = \log P_X - \log P_H \tag{4}$$

3.2 Electronic effects (σ)

Electronic properties dictate the ionization state (pKa) and polarity of a molecule, critically influencing both membrane permeability and electrostatic drug–receptor interactions.

- **Hammett Substituent Constant (σ):** Measures the electron-withdrawing (+σ) or electron-donating (-σ) capacity of meta- and para-substituents on an aromatic ring based on the ionization constants of substituted benzoic acids [8] (Equation 5).

$$\sigma_X = \log K_X - \log K_H \tag{5}$$

Note: Classical Hammett constants are inapplicable to ortho-substituents due to steric interference near the reaction center, requiring specialized treatments such as Fujita–Nishioka analysis [8, 9].

3.3 Steric effects (E_s)

Steric parameters, such as Taft's steric factor (Es) or molar refractivity (MR), quantify the bulk, size, and spatial geometry of substituents. Complementary steric fit between the ligand and the receptor's active site is essential for optimal binding affinity.

3.4 Craig plot

A Craig plot is a two-dimensional graphical optimization tool plotting the Hammett electronic parameter (σ) on the x-axis against the hydrophobic parameter (π) on the y-axis. It partitions substituents into four quadrants (+π/+σ, +π/-σ, -π/+σ, -π/-σ), enabling medicinal chemists to systematically select substituents that independently modulate electronic and lipophilic properties.

4 QSAR workflow and statistical modelling algorithms

4.1 Standard QSAR workflow

Developing a predictive QSAR model follows a structured five-stage pipeline [3, 9]:

1. **Data Curation & Structure Generation:** Compilation of reliable biological assay data and generation of 1D, 2D, or 3D molecular structures.
2. **Descriptor Calculation:** Computation of constitutional, topological, electronic, geometric, and quantum-chemical descriptors.
3. **Data Pre-processing:** Normalization, standardization, removal of collinear or zero-variance variables, feature selection, and outlier detection.
4. **Multivariate Statistical Modeling:** Establishing mathematical correlations between selected descriptors and biological activity.
5. **Model Validation:** Internal and external statistical evaluation using metrics such as correlation coefficient (R), coefficient of determination (R^2), cross-validated correlation coefficient (Q^2), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) [3].

4.2 Regression and machine learning algorithms

- **Multiple Linear Regression (MLR):** Generates simple, interpretable linear equations. However, standard MLR is vulnerable to multicollinearity when the number of descriptors exceeds the number of compounds. Advanced variants include Best Multiple Linear Regression (BMLR) [10], Heuristic Method (HM) for rapid descriptor screening [10–12], and Genetic Algorithm-MLR (GA-MLR) for evolutionary feature selection [13].
- **Partial Least Squares (PLS):** Projects correlated, high-dimensional descriptor matrices onto orthogonal latent variables [14, 15]. Notable extensions include Genetic PLS (GPLS) for evolutionary variable selection, Factor Analysis PLS (FA-PLS) for dimensionality reduction, and Orthogonal Signal Correction PLS (OSC-PLS) to filter out systematic orthogonal noise [16–19].
- **Artificial Neural Networks (ANN):** Biologically inspired computational architectures capable of modeling complex, non-linear structure–activity relationships across large datasets [6, 20].
- **Support Vector Machines (SVM):** Utilizes kernel functions to map descriptors into high-dimensional feature spaces, providing robust non-linear regression and classification resistant to outliers [7, 20, 21].

- **Advanced Non-Linear Methods:** Emerging techniques include Gene Expression Programming (GEP), which evolves explicit mathematical expressions via symbolic regression [22]; Projection Pursuit Regression (PPR), which models non-linearities using smooth projection functions [10, 23]; and Local Lazy Regression (LLR), which constructs local approximations based on Manhattan or Euclidean nearest-neighbor distances [24, 25].

5 OECD principles for QSAR validation and applications

To ensure regulatory acceptance and scientific reliability (per OECD guidelines), a valid QSAR model must satisfy five core criteria [3]:

1. **A defined endpoint:** A clear, standardized biological or physicochemical measurement.
2. **An unambiguous algorithm:** A transparent and reproducible mathematical formulation.
3. **A defined domain of applicability:** The defined chemical space within which model predictions are statistically reliable.
4. **Appropriate measures of goodness-of-fit, robustness, and predictivity:** Rigorous internal cross-validation (Q^2) and external test-set validation (R^2_{pred}).
5. **A mechanistic interpretation:** Establishing a plausible biochemical or physicochemical rationale, wherever possible.

5.1 Key applications

- **Pharmaceutical Drug Discovery:** Lead identification, potency optimization, ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profiling, and elucidation of receptor binding mechanisms [6, 10].
- **Predictive Toxicology & Regulatory Science:** Screening environmental chemicals, pesticides, and engineered nanomaterials for ecotoxicity and carcinogenicity without animal testing [4, 7, 12, 26, 27].
- **Physical & Material Chemistry:** Predicting physicochemical properties (QSPR), such as boiling points, aqueous solubility, chromatographic retention indices, and polymer capacity factors [11, 19–22, 28–31].

6 Conclusion

Quantitative Structure-Activity Relationship (QSAR) modeling has matured from classical linear free energy correlations into an indispensable, multidimensional discipline within modern computational chemistry, drug discovery, and regulatory toxicology [1, 3, 6]. By translating complex molecular architectures into quantifiable physicochemical and spatial descriptors, QSAR enables researchers to predict biological potency, optimize pharmacokinetic profiles, and identify potential toxicological hazards prior to chemical synthesis [4, 7]. The progression from 1D and 2D descriptors to 3D–6D methodologies—such as CoMFA and CoMSIA—alongside advanced machine learning algorithms like SVM, ANN, and evolutionary regression techniques, has markedly improved the accuracy and applicability of predictive models across diverse chemical spaces [20, 22, 24]. Nevertheless, the reliability of any QSAR model remains fundamentally dependent on high-quality experimental data, rigorous feature selection, and strict adherence to OECD validation principles, particularly defining the domain of applicability [3]. As artificial intelligence, big data analytics, and molecular simulation

continue to converge, QSAR will play an increasingly pivotal role in accelerating rational molecular design while substantially reducing development costs and animal testing.

References

[1] Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*. 1962;194(4824):178–180. https://doi.org/10.1038/194178b0

[2] Hansch C, Fujita T. ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc*. 1964;86(8):1616–1626. https://doi.org/10.1021/ja01062a035

[3] Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*. 2003;22(1):69–77. https://doi.org/10.1002/qsar.200390007

[4] Borm PJ, Robbins D, Haubold S, Kuhlbusch T, Fissan H, Donaldson K, et al. The potential risks of nanomaterials: A review carried out for ECETOC. *Part Fibre Toxicol*. 2006;3:11. https://doi.org/10.1186/1743-8977-3-11

[5] Free SM Jr, Wilson JW. A mathematical contribution to structure–activity studies. *J Med Chem*. 1964;7(4):395–399. https://doi.org/10.1021/jm00334a001

[6] Yap CW, Li ZR, Chen YZ. Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods. *J Mol Graph Model*. 2006;24(5):383–395.

[7] Zhao CY, Zhang HX, Zhang XY, Liu MC, Hu ZD, Fan BT. Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology*. 2006;217(2–3):105–119.

[8] Tsuno Y, Ibata T, Yukawa Y. Resonance effect in Hammett relationship. I. The substituent effect in the acid catalyzed decompositions of ω-diazoacetophenones in acetic acid. *Bull Chem Soc Jpn*. 1959;32(9):960–965.

[9] Katritzky AR, Lobanov VS, Karelson M. *Comprehensive descriptors for structural and statistical analysis (CODESSA) reference manual, version 2.7.10*. Gainesville (FL): University of Florida; 2007.

[10] Yuan YN, Zhang RS, Hu RJ, Ruan XF. Prediction of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas based on the heuristic method, support vector machine and projection pursuit regression. *Eur J Med Chem*. 2009;44(1):25–34.

[11] Ma WP, Luan F, Zhao CY, Zhang XY, Liu MC, Hu ZD, et al. QSAR prediction of the penetration of drugs across a polydimethylsiloxane membrane. *QSAR Comb Sci*. 2006;25(10):895–904.

[12] Luan F, Ma WP, Zhang XY, Zhang HX, Liu MC, Hu ZD, et al. Quantitative structure–activity relationship models for prediction of sensory irritants (log RD50) of volatile organic chemicals. *Chemosphere*. 2006;63(7):1142–1153.

[13] Fisz JJ. Combined genetic algorithm and multiple linear regression (GA-MLR) optimizer: Application to multi-exponential fluorescence decay surface. *J Phys Chem A*. 2006;110(48):12977–12985.

[14] Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnaiah PR, editor. *Multivariate analysis*. New York: Academic Press; 1966. p. 391–420.

[15] Jöreskog KG, Wold H. *Systems under indirect observation: Causality, structure, prediction*. Amsterdam: North-Holland; 1982.

[16] Yin PY, Mohemaiti P, Chen J, Zhao XJ, Lu X, Yimiti A, et al. Serum metabolic profiling of abnormal savda by liquid chromatography/mass spectrometry. *J Chromatogr B*. 2008;871(2):322–327.

[17] Niazi A, Jafarian B, Ghasemi J. Kinetic spectrophotometric determination of trace amounts of palladium by whole kinetic curve and a fixed time method using resazurine sulfide reaction. *Spectrochim Acta A*. 2008;71(3):841–846.

[18] Niazi A, Amjadi E, Nori-Shargh D, Bozorghi SJ. Simultaneous voltammetric determination of lead and tin by adsorptive differential pulse stripping method and orthogonal signal correction-partial least squares in water samples. *J Chin Chem Soc*. 2008;55(2):276–285.

[19] Luan F, Liu HT, Wen YY, Zhang XY. Prediction of quantitative calibration factors of some organic compounds in gas chromatography. *Analyst*. 2008;133(7):881–887.

[20] Chen HF. Quantitative predictions of gas chromatography retention indexes with support vector machines, radial basis neural networks and multiple linear regression. *Anal Chim Acta*. 2008;609(1):24–36.

[21] Yuan YN, Zhang RS, Hu RJ, Ruan XF. Prediction of volatile components retention time in blackstrap molasses by least-squares support vector machine. *QSAR Comb Sci*. 2008;27(5):535–542.

[22] Si HZ, Zhang KJ, Hu ZD, Fan BT. QSAR model for prediction capacity factor of molecular imprinting polymer based on gene expression programming. *QSAR Comb Sci*. 2007;26(1):41–50.

[23] Friedman JH, Stuetzle W. Projection pursuit regression. *J Am Stat Assoc*. 1981;76(376):817–823. https://doi.org/10.1080/01621459.1981.10477729

[24] Gunturi SB, Archana K, Khandelwal A, Narayanan R. Prediction of hERG potassium channel blockade using kNN-QSAR and local lazy regression methods. *QSAR Comb Sci*. 2008;27(11–12):1305–1317.

[25] Guha R, Dutta D, Jurs PC, Chen T. Local lazy regression: Making use of the neighborhood to improve QSAR predictions. *J Chem Inf Model*. 2006;46(4):1836–1847. https://doi.org/10.1021/ci060064e

[26] Nel A, Xia T, Mädler L, Li N. Toxic potential of materials at the nanolevel. *Science*. 2006;311(5761):622–627. https://doi.org/10.1126/science.1114397

[27] Chatterjee R. The challenge of regulating nanomaterials. *Environ Sci Technol*. 2008;42(2):339–343.

[28] Alivisatos P. The use of nanocrystals in biological detection. *Nat Biotechnol*. 2004;22(1):47–52. https://doi.org/10.1038/nbt927

[29] Chen FQ, Gerion D. Fluorescent CdSe/ZnS nanocrystal–peptide conjugates for long-term, nontoxic imaging and nuclear targeting in living cells. *Nano Lett*. 2004;4(10):1827–1832.

[30] Zhou K, Wang R, Xu B, Li Y. Synthesis, characterization and catalytic properties of CuO nanocrystals with various shapes. *Nanotechnology*. 2006;17(15):3939–3943.

[31] Yuranova T, Laub D, Kiwi J. Synthesis, activity and characterization of textiles showing activity under daylight irradiation. *Catal Today*. 2007;122(1–2):109–117.