STM JOURNALS

International Journal of
Cheminformatics

Volume 3, Issue 1, 2025
January–June
DOI (Journal): 10.37591/IJCI

https://journals.stmjournals.com/ijci

Review IJCI

QSAR Modeling Techniques: A Comprehensive Review of Tools and Best Practices

Sheshmani Tiwari*

Abstract

Quantitative structure–activity relationship (QSAR) modeling has become an essential tool in drug discovery, toxicity assessment, and environmental chemistry. By correlating chemical structure with biological activity or toxicity, QSAR enables the prediction of compound behavior without extensive experimental testing. This approach not only saves time and resources but also supports ethical practices by reducing reliance on animal studies. The evolution of QSAR from basic linear models to advanced machine learning and AI-based techniques has significantly improved predictive accuracy and the handling of large datasets. This review outlines the key stages of QSAR model development, including data collection, descriptor selection, algorithm choice, model validation, and result interpretation. Emphasis is placed on dataset quality, reproducibility, and clearly defining a model's applicability domain. The review also examines popular QSAR software—both commercial and open-source—that streamline model creation and evaluation. Regulatory guidelines, such as those from the Organization for Economic Co-operation and Development (OECD), are discussed to highlight best practices for ensuring model reliability in regulatory contexts. Emerging innovations like deep learning, transfer learning, and generative models are also explored. The article concludes with a discussion of current challenges and future directions, aiming to support researchers in developing robust QSAR models for chemical safety and pharmaceutical applications.

Keywords: Artificial Intelligence, chemical safety, machine learning, QSAR modeling, toxicity prediction

INTRODUCTION

QSAR modeling is a powerful and widely adopted computational approach in modern chemical, pharmaceutical, and environmental sciences. Fundamentally, QSAR modeling seeks to create a mathematical connection between the physicochemical characteristics, biological activity, and chemical structure of a compound. This technique relies on the assumption that the activity or properties of a molecule are inherently related to its structural features, which can be quantified using molecular descriptors. By analyzing these descriptors using statistical or machine learning methods, QSAR models can be developed to predict the behavior of new and untested compounds.

***Author for Correspondence**
Sheshmani Tiwari
E-mail: xtylishmanibaba@gmail.com

Student, Department of Biochemical Engineering, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India

Received Date: May 22, 2025
Accepted Date: May 29, 2025
Published Date: June 03, 2025

Citation: Sheshmani Tiwari. QSAR Modeling Techniques: A Comprehensive Review of Tools and Best Practices. International Journal of Cheminformatics. 2025; 3(1): 56–63p.

The concept of QSAR has a rich history dating back to the 1960s, with the foundational work of Hansch and Fujita demonstrating that biological activity could be quantitatively related to physicochemical parameters such as lipophilicity, electronic effects, and steric factors. Since then, the field has expanded significantly in scope and sophistication. With advances in computational power and the development of new algorithms, QSAR techniques have evolved from basic linear regression models to complex nonlinear

approaches, including support vector machines, artificial neural networks, random forests, and deep learning architectures [1–3].

In the early phases of chemical risk assessment and drug discovery, QSAR modeling is particularly helpful. It enables high-throughput virtual screening of chemical libraries, identification of lead compounds, optimization of pharmacological profiles, and the prediction of toxicological endpoints. In regulatory contexts, QSAR models are increasingly used to support chemical safety assessments and decision-making, especially when experimental data are lacking or ethical concerns limit in vivo testing. Agencies such as the Organization for Economic Co-operation and Development (OECD), European Chemicals Agency (ECHA), and U.S. Environmental Protection Agency (EPA) have established guidelines for the use of QSAR models in regulatory submissions, highlighting their growing acceptance and importance [4].

Despite its widespread application, QSAR modeling presents several challenges. Key issues include the availability of high-quality and well-curated data, selection of relevant molecular descriptors, model overfitting, interpretability of complex models, and definition of applicability domains. Furthermore, a lack of an established methodology and open reporting procedures may make QSAR investigations less reproducible. Researchers are increasingly using best practices for data preprocessing, model validation, and performance evaluation to address these concerns. Transparency and cooperation within the scientific community are fostered by the use of open-source tools and platforms.

The merging of deep learning, artificial intelligence, and data fusion techniques has led to a surge in innovation in the sector in recent years. These advances enable more accurate and generalizable QSAR models that can capture subtle nonlinear relationships between the molecular structure and activity. Furthermore, the application of QSAR methodologies is expanding beyond drug discovery to include agrochemicals, materials sciences, and nanotechnology.

The goal of this review is to present a thorough analysis of QSAR modeling methods, resources, and best practices. It is intended to serve as a guide for researchers, practitioners, and regulators seeking to understand or apply QSAR methods in their respective domains [5, 6].

HISTORICAL BACKGROUND AND EVOLUTION

The concept of QSAR modeling has its roots in the mid-20th century, with early contributions from scientists such as Corwin Hansch and Toshio Fujita. In the 1960s, Hansch introduced the idea that biological activity can be correlated with the physicochemical properties of molecules, such as lipophilicity, electronic distribution, and steric effects, through mathematical equations. This marked the birth of classical QSAR, which often employs linear regression to model the relationship between molecular descriptors and observed biological activity.

Over the next few decades, QSAR techniques have gained popularity in medicinal chemistry, where they have been used to optimize lead compounds for drug design. Initially, models were built using small datasets and limited descriptor sets, primarily focusing on two-dimensional structural information. Three-dimensional (3D) descriptors were added to QSAR as computing power and chemical and biological data increased. This has resulted in the creation of 3D-QSAR techniques, such as Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) [7, 8].

There was a change toward more sophisticated statistical and machine learning methods in the late 1990s and the early 2000s. The predictive performance was greatly enhanced by modeling nonlinear and high-dimensional data using algorithms such as support vector machines, random forests, and neural networks. The advent of cheminformatics databases, cloud computing, and open-source tools has accelerated QSAR adoption across various disciplines, including environmental chemistry and

toxicology. Today, QSAR continues to evolve with the integration of artificial intelligence, deep learning, and big data analytics, transforming it into a powerful approach for virtual screening, risk assessment, and regulatory decision-making.

KEY STEPS IN QSAR MODELING

Data Collection and Curation

Accurate QSAR models begin with high-quality datasets. This step involves sourcing, cleaning, and organizing the data to ensure consistency and reliability [9].

Data Sources

- Public databases: PubChem, ChEMBL, ToxCast, DrugBank
- In-house experimental datasets

Curation Process

- Remove duplicates and resolve conflicting entries.
- Normalize chemical structures (tautomer correction, salt stripping, stereochemistry)
- Standardize biological endpoints (e.g., converting IC_{50} to pIC_{50})

Error Handling

- Detect and handle missing values or erroneous outliers.
- Ensure units and concentration values are consistent.

Quality Assurance

- Annotate metadata and assay protocols.
- Document dataset versioning for reproducibility.

A well-curated dataset forms the cornerstone of developing accurate, interpretable, and regulatory-compliant QSAR models.

Molecular Descriptor Calculation

The descriptors numerically represent the molecular properties. These features are used by algorithms to learn structure–activity relationships.

Types of Descriptors

- 1D: Molecular weight, number of atoms
- 2D: Topological indices, functional groups
- 3D: Molecular volume, surface area
- Fingerprints: ECFP, MACCS keys (binary vector format)

Descriptor Tools

- PaDEL-Descriptor
- RDKit
- Dragon

Key Considerations

- Ensure descriptors are reproducible and well-documented.
- Avoid irrelevant or noisy descriptors.
- Select descriptors based on model purpose (e.g., ADMET prediction).

Careful descriptor calculation ensures meaningful structure-property correlations and reduces the risk of overfitting.

Feature Selection

Feature selection refines the descriptor set by retaining only the most relevant features, thereby improving model performance and interpretability.

Why It Matters

- Removes redundant, irrelevant, or noisy features.
- Reduces overfitting and improves generalization.
- Enhance computational efficiency.

Common Techniques

- *Filter methods:* Correlation thresholds, mutual information
- *Wrapper methods:* Recursive Feature Elimination (RFE)
- *Embedded methods:* Least absolute shrinkage and selection operator (LASSO), decision tree importance

Dimensionality Reduction

- Principal Component Analysis (PCA)
- t-SNE for visual exploration

Best Practices

- Use domain knowledge to guide selection.
- Evaluate performance on cross-validated data.
- Maintain transparency and reproducibility.

Well-chosen features help the model to learn relevant patterns without introducing unnecessary complexity [10].

Model Building

Model building uses selected descriptors to train a machine learning algorithm that can predict molecular activity or properties.

Classical Algorithms

- Multiple Linear Regression (MLR)
- Partial Least Squares (PLS)

Modern Machine Learning (ML) Techniques

- Support vector machines (SVM)
- Random forest (RF)
- k-nearest neighbors (k-NN)
- Artificial neural networks (ANN)

Advanced Approaches

- Ensemble models (e.g., boosting, bagging)
- Deep learning for large datasets

Toolkits and Platforms

- KNIME
- WEKA
- Scikit-learn (Python)

Key Practices

- Perform hyperparameter optimization.

- Prevent overfitting via regularization.
- Consider model transparency for regulatory use.

Choosing the right algorithm depends on the data type, size, and the trade-off between interpretability and accuracy.

Model Validation

Validation ensures that the model is robust, reliable, and generalizable to new data.

Internal Validation

- k-Fold Cross-Validation
- Leave-One-Out Cross-Validation (LOOCV)
- Bootstrapping

External Validation

- Use a hold-out or independent test set.
- Confirm generalization to unseen compounds.

Statistical Metrics

- R^2 (goodness of fit)
- Q^2 (predictive power)
- RMSE (root mean square error)
- MAE (mean absolute error)
- ROC-AUC for classification tasks

Overfitting Checks

- Y-randomization
- Learning curves

Regulatory Alignment

- Adherence to OECD validation principles

A validated model inspires confidence and is crucial for regulatory submission and scientific credibility [11].

Applicability Domain (AD)

The AD defines the chemical space where the model's predictions are considered reliable.

Importance of AD

- Prevents unreliable predictions for out-of-domain compounds.
- Ensures safe and scientifically valid application.

AD Estimation Methods

- Leverage method (Williams plot).
- Distance-based (Euclidean, Mahalanobis).
- Similarity thresholds using molecular fingerprints.

Visualization Tools

- PCA scatter plots
- Leverage vs. standardized residuals plot.

Key Tools

- OECD QSAR Toolbox
- QSARINS

Best Practices

- Report AD boundaries.
- Flag predictions made outside the AD.
- Use AD to guide experimental prioritization.

Defining the applicability domain helps ensure that QSAR models are used responsibly and accurately.

QSAR TOOLS AND SOFTWARE

Several tools facilitate QSAR modeling (Table 1).

BEST PRACTICES IN QSAR MODELING

1. *Data quality control:* Prioritize curated datasets and use standardized endpoints.
2. *Descriptor relevance:* Choose descriptors based on domain knowledge and statistical merit.
3. *Balanced datasets:* Avoid bias by addressing class imbalance.
4. *Model interpretability:* Prefer transparent models, especially for regulatory applications.
5. *Reproducibility:* Maintain version-controlled pipelines and share code and data.
6. *Regulatory compliance:* Follow OECD principles for model development and reporting.

CHALLENGES AND LIMITATIONS

Despite advances in technology and research, several challenges and limitations continue to affect the successful implementation of systems and processes. These issues must be acknowledged and addressed to ensure effective treatment outcomes.

- *Data availability and quality:* The absence of reliable, consistent, and easily accessible data is one of the main drawbacks. Analysis and decision-making can be severely hampered by inaccuracies or a lack of data.
- *Technological constraints:* Limited access to modern tools or infrastructure can restrict performance, particularly in underdeveloped regions or institutions with budget constraints.
- *Skilled workforce:* Professionals with the know-how required to handle complicated technology, run sophisticated systems, or correctly interpret data frequently in short supply.
- *High implementation costs:* Initial setup, licensing, training, and maintenance can be financially demanding, making them less feasible for small-scale organizations or startups.
- *Resistance to change:* Individuals and organizations may resist adopting new methods or technologies because of unfamiliarity or fear of failure.
- *Security and privacy issues:* To manage sensitive data, cybersecurity threats must be addressed, and privacy laws must be followed.
- *Scalability problems:* Systems designed for smaller operations may not perform well under increased loads, thereby limiting their long-term viability.

Table 1. QSAR modeling tools.

Tool	Description
KNIME	Open-source platform integrating data mining and ML for QSAR workflows
OECD QSAR Toolbox	Regulatory-focused tool for toxicity prediction
ADMET Predictor	Proprietary tool for pharmacokinetic modeling
PaDEL-Descriptor	Java-based software for descriptor calculation
QSARINS	Regression-focused tool supporting extensive validation

Addressing these challenges requires continuous investment in education, technology, and strategic planning.

RECENT TRENDS AND FUTURE DIRECTIONS

Recent advancements in deep learning and transfer learning have significantly transformed the landscape of QSAR modeling. These technologies have enabled the development of more sophisticated and accurate predictive architectures that can capture complex nonlinear relationships within chemical data. One of the most exciting developments is the application of generative models, particularly those using variational autoencoders (VAEs) for de novo molecular design. These models offer intriguing paths for drug development because they can produce new molecular structures with desired biological activity.

Moreover, there is a growing trend toward integrating QSAR approaches with multi-omics data, such as genomics, proteomics, and metabolomics, along with real-world clinical evidence. This integrative approach enhances the utility of QSAR in precision medicine, thereby allowing for more personalized and effective therapeutic strategies.

In parallel, initiatives promoting data transparency and reusability, such as the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, are becoming increasingly influential. The open science movement also encourages broader collaboration, data sharing, and reproducibility in QSAR research. These combined trends point toward a future in which QSAR models are not only more powerful and versatile but also more transparent, collaborative, and applicable to real-world healthcare challenges [12–15].

CONCLUSION

In cheminformatics, quantitative structure–activity relationship modeling is still a vital technique that provides a methodical and economical way to forecast the physicochemical characteristics and biological activity of chemical compounds. By analyzing the structural attributes of molecules and correlating them with observed activities, QSAR models significantly reduce the need for exhaustive experimental testing, thereby accelerating the drug discovery process and supporting environmental risk assessments. The growing integration of machine learning algorithms, artificial intelligence, and advanced statistical methods has further enhanced the predictive power, robustness, and interpretability of QSAR models. Moreover, the establishment of regulatory guidelines and validation protocols ensures that the developed models meet high standards of reliability and reproducibility, which is crucial for their acceptance in the scientific and regulatory domains. As research in computational chemistry evolves, QSAR modeling is expected to become even more refined by incorporating novel descriptors, larger and more diverse datasets, and multidimensional data analysis. This progress will not only improve model performance but also expand the scope of QSAR applications beyond pharmaceuticals into areas such as toxicology, cosmetics, agrochemicals, and materials science. Overall, QSAR modeling holds immense promises for advancing scientific innovation while ensuring safety, efficiency, and sustainability in chemical research and development.

REFERENCES

1. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: Where have you been? Where are you going to? J Med Chem. 2014;57(12):4977–5010. doi:10.1021/jm4004285.
2. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29(6-7):476–88. doi:10.1002/minf.201000061.
3. Organisation for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. Report No.: 69. (OECD series on testing and assessment). Paris: OECD Publishing; 2007. 162 p. doi: https://doi.org/10.1787/9789264085442-en.

4. Gramatica P. Principles of QSAR model validation: Internal and external. QSAR Comb Sci. 2007;26(5):694–701. doi:10.1002/qsar.200610151.
5. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74. doi:10.1002/jcc.21707.
6. Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. (Methods and Principles in Medicinal Chemistry). Weinheim: Wiley-VCH; 2009. doi: https://doi.org/10.1002/9783527628766.
7. Roy K, Kar S, Das RN. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Amsterdam: Academic Press; 2015. doi: https://doi.org/10.1016/C2013-0-19096-4.
8. Fourches D, Muratov E, Tropsha A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model. 2010;50(7):1189–204. doi:10.1021/ci100176x.
9. Gadaleta D, Lombardo A, Toma C, Benfenati E. A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. J Cheminform. 2018;10(1):60. doi:10.1186/s13321-018-0315-6.
10. Hanser T, Barber C, Berggren E, et al. Improving the regulatory assessment of chemicals through QSAR modeling: OECD and EU perspectives. Comput Toxicol. 2019;9:38–45.
11. Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model. 2002;20(4):269–76. doi:10.1016/S1093-3263(01)00123-1.
12. Todeschini R, Consonni V, Mauri A, et al. DRAGON: A software for molecular descriptor calculation. J Chemom. 2004;18(5):274–85.
13. Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model. 2013;53(4):783–90. doi:10.1021/ci400084k.
14. Mauri A, Consonni V, Todeschini R. The prediction of the biological activity of chemicals using QSAR: Reliability and validation. Curr Pharm Des. 2005;11(4):509–20.
15. Schüürmann G, Ebert RU, Chen J, Wang B, Kühne R. External validation and prediction employing the predictive squared correlation coefficient – test set activity mean vs training set activity mean. J Chem Inf Model. 2008;48(11):2140–5. doi:10.1021/ci800253u.

