REVIEW ARTICLE

Quantitative Structure-Activity Relationship (QSAR) in Drug Discovery and Development

Amisha G[1], Govindarao Kamala[2], Chandrika D[1], Sravani J[1], Bhuvaneswari D[1], Renuka Devi K[1], Usha E[3]

[1]UG Scholar, Department of Pharmaceutical Chemistry, Koringa College of Pharmacy, Korangi, Kakinada, Andhra Pradesh, India
[2] Professor, Department of Pharmaceutical Chemistry, Koringa College of Pharmacy, Korangi, Kakinada, Andhra Pradesh, India
[3]Professor, Department of Pharmaceutics, Koringa College of Pharmacy, Korangi, Kakinada, Andhra Pradesh, India

Publication history: Received on 10th Jan 2025; Revised on 18th Jan 2025; Accepted on 19th Jan 2025

Article DOI: 10.69613/d091zy53

Abstract: Quantitative structure-activity relationship (QSAR) analysis represents a cornerstone approach in modern drug discovery and development. QSAR methodologies establish mathematical correlations between molecular structures and their biological activities, enabling the prediction of compound properties and behaviors. Recent advances in computational capabilities, coupled with the emergence of sophisticated machine learning algorithms, have revolutionized traditional QSAR approaches. The integration of deep learning architectures, including graph neural networks and convolutional neural networks, has enhanced the accuracy and predictive power of QSAR models. Modern QSAR implementations incorporate multidimensional molecular descriptors, quantum mechanical calculations, and multi-omics data to provide comprehensive insights into structure-activity relationships. The evolution from classical linear regression models to advanced neural networks has facilitated the handling of complex, non-linear relationships between molecular features and biological responses. Contemporary QSAR applications extend beyond pharmaceutical research into toxicology, environmental science, and materials development. The incorporation of explainable artificial intelligence techniques has improved model interpretability, while active learning approaches have optimized experimental design and data collection. Cloud computing and big data integration have enabled the processing of larger molecular datasets, leading to more robust and generalizable models. These methodological advances, combined with improved molecular representation techniques and hybrid modeling approaches, have positioned QSAR as an indispensable tool in rational drug design and chemical property prediction.

Keywords: Molecular descriptors; Machine learning; Structure-activity relationship; Drug discovery; Computational chemistry.

1. Introduction

Quantitative structure-activity relationship (QSAR) analysis establishes mathematical connections between molecular structural features and their corresponding biological activities or chemical properties [1]. The fundamental principle underlying QSAR stems from the observation that structurally similar molecules often exhibit comparable biological responses, though this relationship is frequently non-linear and complex [2]. The mathematical framework of QSAR can be expressed as a function where biological response correlates with molecular descriptors, forming the basis for predictive modeling in drug discovery and development [3].

The evolution of QSAR methodologies traces back to the early 20th century, beginning with Hammett's linear free energy relationships and progressing through Hansch's groundbreaking work in the 1960s [4]. The field has subsequently undergone significant transformation, particularly with the advent of computational capabilities and sophisticated mathematical approaches [5]. Modern QSAR applications have expanded beyond traditional drug discovery into pharmaceutical research and development, toxicological assessments, environmental fate predictions, materials science, agrochemical design, and regulatory science [6].

The foundation of QSAR analysis relies on molecular descriptors, which quantitatively represent structural and physicochemical properties. These descriptors encompass constitutional parameters reflecting atomic composition and basic molecular properties, electronic descriptors capturing charge distribution and orbital energies, topological indices representing molecular connectivity and

* Corresponding author: Amisha G

shape, geometric parameters describing three-dimensional structural features, and quantum chemical descriptors characterizing electronic structure and molecular orbital properties [7].

Table 1. Evolution of QSAR Techniques

Time Period	Development	Methodological Advance	Computational Approach	Contributors
1930s-1940s	Hammett Equation	Linear Free Energy Relationships	Manual calculations	Louis Hammett
1960s	Hansch Analysis	Hydrophobic Parameter Integration	Early computer-based regression	Corwin Hansch
1970s-1980s	3D-QSAR	Three-dimensional structure consideration	Molecular modeling, CoMFA	Richard Cramer
1990s	Neural Networks	Pattern recognition capabilities	Artificial neural networks	James Zupan
2000s	Support Vector Machines	Non-linear relationship modeling	Kernel-based methods	Vladimir Vapnik
2010s	Deep Learning	Complex pattern extraction	Convolutional neural networks	Various Teams
2020s	Graph Neural Networks	Direct molecular graph processing	Message passing networks	Contemporary Research Groups

Contemporary QSAR implementations employ diverse mathematical approaches ranging from classical statistical methods to advanced machine learning algorithms. These mathematical frameworks process molecular descriptors to generate predictive models for biological activities or chemical properties [8]. The development of reliable QSAR models necessitates rigorous validation protocols, including internal validation through cross-validation and bootstrap analysis, external validation via independent test set predictions, and careful assessment of the applicability domain to define the chemical space where predictions maintain reliability [9].

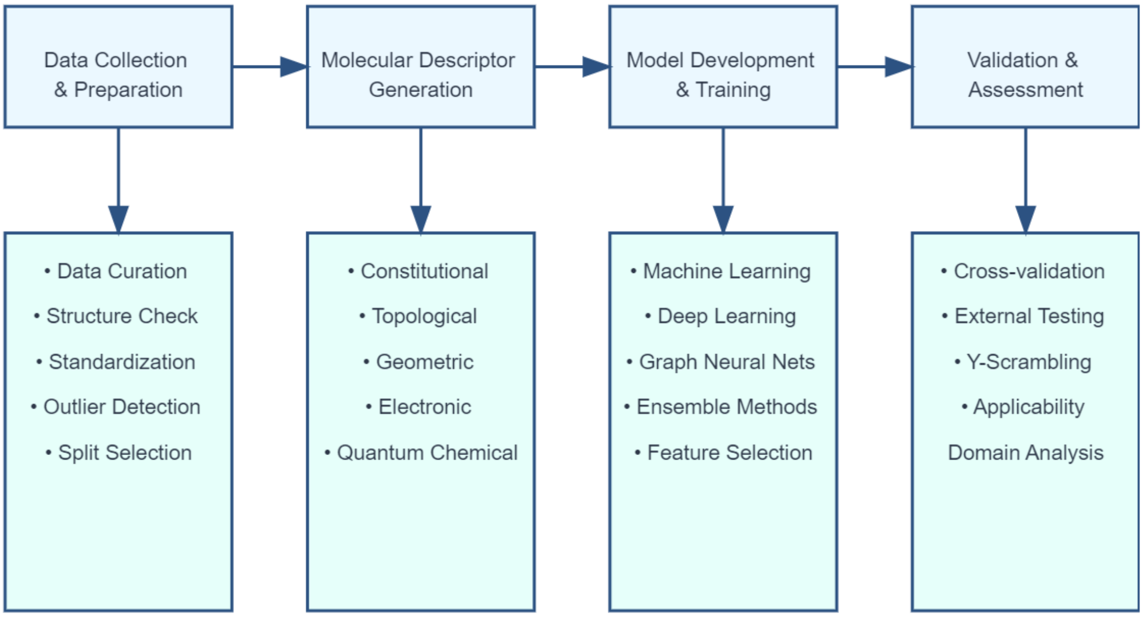

Figure 1. Modern QSAR workflow

Despite significant advances, QSAR methodology faces several persistent challenges. Data quality and standardization issues continue to affect model development, while limited availability of experimental data constrains the scope of predictions. Complex structure-activity relationships often prove difficult to model accurately, and concerns regarding model interpretability persist. Additionally, limitations in applicability domain restrict the broader utilization of developed models [10].

Addition of artificial intelligence and deep learning techniques enhances predictive capabilities, while quantum computing algorithms offer new possibilities for molecular modeling. Development of interpretable models addresses transparency concerns, and ongoing efforts focus on improving prediction accuracy and expanding the applicable chemical space [11]. These advances position QSAR as an increasingly powerful tool in modern drug discovery and development processes.

2. QSAR Methodologies and Recent Developments

2.1. Evolution of QSAR Modeling Techniques

Traditional QSAR approaches initially relied on linear regression models, correlating simple molecular descriptors with biological activities [12]. The progression from simple linear correlations to multiple linear regression (MLR) enabled the incorporation of multiple structural parameters, providing more comprehensive structure-activity insights [13]. Subsequently, partial least squares (PLS) regression emerged as a powerful tool for handling highly correlated molecular descriptors, addressing the limitations of conventional regression techniques in analyzing complex chemical datasets [14].

Table 2. Classification of Molecular Descriptor

Descriptor Type	Parameters	Calculation Method	Application Area	Information Content
Constitutional	Atom counts, Molecular weight, Ring counts	Direct computation	Basic property prediction	Molecular composition
Topological	Wiener index, Connectivity indices	Graph theory	Molecular similarity	2D structure
Electronic	Partial charges, HOMO-LUMO energies	Quantum calculations	Reactivity prediction	Electronic distribution
Geometric	Surface area, Volume, Shape indices	3D coordinates	Binding affinity	Spatial arrangement
Quantum Chemical	Orbital energies, Electron density	Ab initio methods	Electronic properties	Electronic structure
Dynamic	Conformational energies, Flexibility	Molecular dynamics	Protein-ligand interaction	Molecular motion

2.2. Advanced Molecular Representation Methods

Modern QSAR implementations utilize sophisticated molecular representation techniques that capture intricate structural details. Three-dimensional molecular descriptors now incorporate spatial arrangements of atoms, electronic distributions, and conformational flexibility [15]. Quantum mechanical descriptors provide detailed electronic structure information, including molecular orbital energies, electron density distributions, and atomic charges, offering deeper insights into molecular behavior [16].

2.3. Machine Learning Integration

The integration of machine learning algorithms has transformed QSAR modeling capabilities. Support Vector Machines (SVM) effectively handle non-linear relationships between molecular structure and biological activity, while Random Forests provide robust predictions through ensemble learning approaches [17]. Neural network architectures, particularly deep learning models, demonstrate exceptional capability in capturing complex structure-activity patterns across diverse chemical spaces [18].

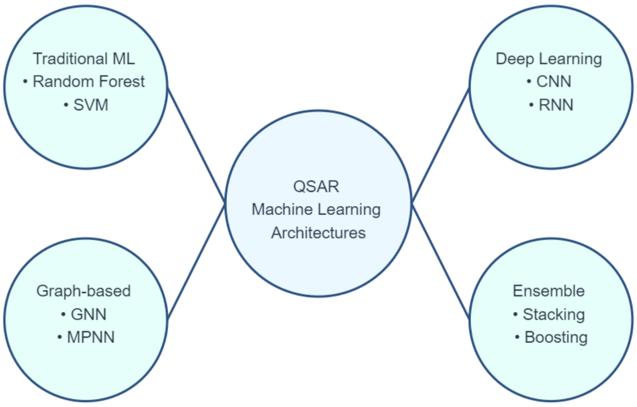

Figure 2. Machine Learning in QSAR

Table 3. Machine Learning Methods in QSAR

Method	**Algorithm Type**	**Features**	**Advantages**	**Limitations**	**Applications**
Random Forest	Ensemble Learning	Multiple decision trees	Handles non-linearity, Feature importance	Limited extrapolation	Classification, Regression
Deep Neural Networks	Deep Learning	Multiple hidden layers	Complex pattern recognition	Requires large datasets	Property prediction
Support Vector Machines	Kernel Methods	Hyperplane separation	Good for small datasets	Kernel selection critical	Binary classification
Gradient Boosting	Ensemble Learning	Sequential tree building	High accuracy	Overfitting risk	Regression tasks
Graph Neural Networks	Graph Processing	Direct structure handling	Molecular representation	Computational cost	Structure-based prediction
Gaussian Process	Probabilistic	Uncertainty quantification	Confidence estimates	Scaling limitations	Regression with uncertainty

2.4. Graph-Based Approaches

Graph Neural Networks (GNNs) represent a significant advancement in molecular modeling, treating molecules as graphs where atoms serve as nodes and chemical bonds as edges. This approach naturally captures molecular topology and enables direct learning of structure-activity relationships from molecular graphs [19]. Message-passing neural networks further enhance this capability by facilitating information flow between atomic centers, leading to improved predictive accuracy [20].

2.5. Multi-Task Learning Frameworks

Contemporary QSAR models increasingly employ multi-task learning approaches, simultaneously predicting multiple biological activities or properties. This methodology leverages correlations between different endpoints, improving prediction accuracy through shared feature learning [21]. The integration of multi-omics data enhances model performance by incorporating biological context into structure-activity predictions [22].

2.6. Model Interpretability and Validation

Advanced interpretability techniques address the "black box" nature of complex QSAR models. Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) provide insights into feature importance and decision-making processes [23]. Rigorous validation protocols, including cross-validation, external validation, and Y-scrambling, ensure model reliability and robustness [24].

Table 4. Validation Protocols and Quality Metrics

Type of Validation	**Method**	**Statistical Parameters**	**Implementation**	**Acceptance Criteria**
Internal Validation	Cross-validation	Q2, RMSE, MAE	k-fold partitioning	Q2 > 0.5
	Bootstrap	Confidence intervals	Resampling with replacement	95% CI significant
	Y-scrambling	R^2 comparison	Random response permutation	Scrambled R2 < 0.1
External Validation	Test set prediction	R^2pred, RMSE_ext	Independent dataset	R^2pred > 0.6
	Time-split validation	Temporal R^2	Chronological splitting	Consistent performance
Applicability Domain	Leverage analysis	h-values	Distance-based	h < h*
	Similarity assessment	Tanimoto index	Structural comparison	T > 0.7
	Probability density	Distribution analysis	Statistical modeling	p > 0.05

2.7. Active Learning and Experimental Design

Active learning strategies optimize the experimental design process by identifying the most informative compounds for testing. This approach reduces experimental costs while maximizing the information content of training datasets [25]. Integration with high-throughput screening data enables efficient model refinement and validation [26].

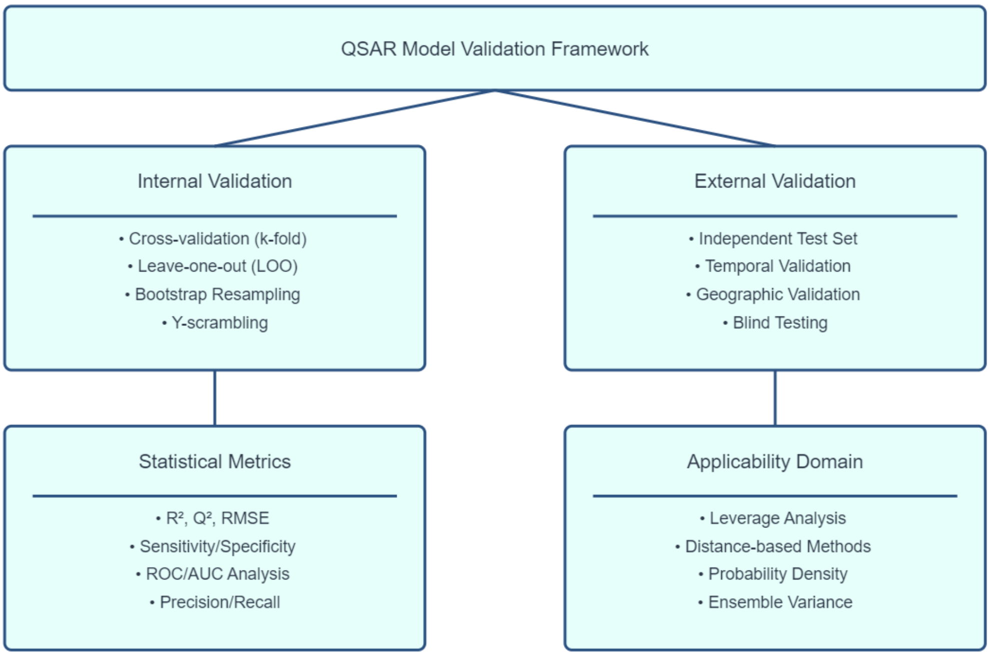

Figure 3. QSAR validation

3. Recent Trends in QSAR techniques

3.1. Deep Learning Architectures in QSAR

Deep learning architectures have revolutionized QSAR modeling by introducing unprecedented capabilities in pattern recognition and feature extraction. Deep Neural Networks (DNNs) with multiple hidden layers effectively capture non-linear relationships between molecular structures and their biological activities [27]. These networks process complex molecular information through successive layers of neurons, each layer extracting increasingly abstract features from the input data [28].

Convolutional Neural Networks (CNNs) have demonstrated particular success in processing grid-like molecular representations. By applying convolution operations to molecular structures, CNNs automatically identify relevant structural patterns and spatial relationships. The architecture typically includes multiple convolutional layers followed by pooling operations, enabling the detection of hierarchical features ranging from local atomic environments to global molecular properties [29].

Recent developments in attention mechanisms have enhanced the performance of deep learning models in QSAR studies. Self-attention layers enable models to focus on relevant molecular features dynamically, improving prediction accuracy for diverse chemical structures. This approach has proven particularly effective when dealing with large molecules and complex biological targets [30].

3.2. Graph-Based Neural Networks

Graph Neural Networks represent a paradigm shift in molecular representation and analysis. Unlike traditional descriptor-based approaches, GNNs operate directly on molecular graphs, preserving the inherent topology of chemical structures [31]. The graph representation consists of:

- Node Features: Representing atomic properties including element type, hybridization state, and local electronic environment
- Edge Features: Encoding bond types, lengths, and electronic characteristics
- Global Features: Capturing overall molecular properties and symmetry

Message Passing Neural Networks (MPNNs) extend the GNN framework by implementing sophisticated information exchange between atomic centers. Through iterative message passing operations, these networks construct increasingly refined representations of local chemical environments. The final molecular representation emerges from the aggregation of node-level features, providing a comprehensive description of structure-activity relationships [32].

3.3. Quantum Mechanical Integration

The incorporation of quantum mechanical calculations has significantly enhanced QSAR modeling accuracy. Density Functional Theory (DFT) calculations provide precise electronic structure information, including:

- Molecular orbital energies and electron density distributions
- Atomic partial charges and bond orders
- Electrostatic potential surfaces
- Reaction barrier heights and thermodynamic parameters

These quantum mechanical descriptors enable more accurate modeling of molecular interactions and chemical reactivity [33]. Advanced quantum chemical methods, including post-Hartree-Fock approaches, provide high-accuracy predictions for electronic properties that influence biological activity [34].

3.4. Multi-Scale Modeling Approaches

Multi-scale modeling integrates information across different spatial and temporal scales, providing comprehensive insights into structure-activity relationships. At the atomic scale, quantum mechanical calculations capture electronic effects and chemical bonding. Molecular mechanics simulations extend this to conformational dynamics and intermolecular interactions, while coarse-grained models address larger-scale phenomena [35].

The integration of molecular dynamics simulations with QSAR modeling has enhanced predictive capabilities. These simulations generate ensemble representations of molecular conformations, accounting for structural flexibility and environmental effects. Time-averaged properties derived from these simulations serve as dynamic descriptors, complementing static structural features [36].

3.5. Advanced Statistical Learning Methods

Modern QSAR implementations incorporate sophisticated statistical learning techniques beyond traditional regression methods. Gaussian Process Regression (GPR) provides probabilistic predictions with uncertainty quantification, enabling more informed decision-making in drug discovery [37]. Bayesian methods incorporate prior knowledge and uncertainty estimation, particularly valuable when dealing with limited experimental data [38].

Transfer learning approaches have emerged as powerful tools for leveraging knowledge across different chemical domains. Models pre-trained on large chemical databases can be fine-tuned for specific applications, improving prediction accuracy for novel chemical classes. This approach particularly benefits scenarios with limited training data for specific targets [39].

3.6. Automated Machine Learning in QSAR

Automated Machine Learning (AutoML) frameworks optimize model architecture and hyperparameters automatically, reducing the need for manual intervention. These systems evaluate multiple model architectures, selecting optimal configurations based on performance metrics. Neural Architecture Search (NAS) extends this concept to deep learning models, automatically discovering effective network architectures for specific QSAR tasks [40].

3.7. Data Integration and Fusion

Modern QSAR approaches increasingly incorporate diverse data types beyond traditional structure-activity pairs. Integration of genomic, proteomic, and metabolomic data provides biological context for structure-activity relationships. This multi-omics integration enables more nuanced predictions of biological activity and potential off-target effects [41].

High-throughput screening data integration has become crucial for model development and validation. Advanced data fusion techniques combine information from multiple experimental sources, improving prediction reliability. Standardization protocols ensure data quality and compatibility across different experimental platforms [42].

3.8. Model Interpretability Advances

Recent developments in model interpretability focus on explaining predictions at multiple levels of granularity. Attribution methods identify atomic and molecular features contributing to specific predictions. Attention visualization techniques reveal which structural elements the model focuses on when making predictions [43]. Counterfactual explanations generate hypothetical molecular modifications that would alter predicted activities, providing actionable insights for molecular design. These explanations help medicinal chemists understand structure-activity relationships and guide compound optimization [44].

4. QSAR Applications

4.1. Pharmaceutical Applications

QSAR methodologies have become indispensable in modern drug discovery and development processes. In lead optimization, QSAR models guide structural modifications to enhance potency, selectivity, and drug-like properties. These models evaluate potential candidates across multiple parameters simultaneously, including target affinity, metabolic stability, and toxicity profiles [45].

Structure-based QSAR approaches integrate protein-ligand interaction data, providing mechanistic insights into binding modes. Fragment-based drug design benefits from QSAR predictions of fragment combinations, accelerating the exploration of chemical space. Virtual screening applications employ QSAR models to prioritize compounds for experimental testing, significantly reducing resource requirements [46].

Drug absorption, distribution, metabolism, excretion, and toxicity (ADMET) predictions represent a crucial application area. QSAR models predict pharmacokinetic parameters and potential toxicity risks early in development, reducing late-stage failures. Physiologically-based pharmacokinetic (PBPK) modeling integrates QSAR predictions with physiological parameters to simulate drug behavior *in vivo* [47].

Table 5. Current Applications of QSAR techniques

Application Area	Implementation	Success Metrics	Impact	Key Findings
Drug Discovery	Lead optimization	Hit rate improvement	3-5x acceleration	Reduced experimental costs
	ADMET prediction	Accuracy > 80%	Early failure prediction	Improved candidate selection
	Virtual screening	Enrichment factor > 10	Resource optimization	Efficient library design
Environmental Assessment	Toxicity prediction	R2 > 0.7	Regulatory compliance	Reduced animal testing
	Biodegradation	85% classification accuracy	Environmental impact	Improved risk assessment
	Bioaccumulation	Log BCF prediction	Chemical safety	Regulatory decision support
Materials Design	Polymer properties	Property accuracy ±10%	Rapid screening	Optimized synthesis
	Nanomaterial behavior	Structure-property correlation	Safety assessment	Enhanced characterization
	Crystal structure	Lattice energy prediction	Process optimization	Improved formulation

4.2. Environmental and Toxicological Applications

Environmental fate prediction has emerged as a critical QSAR application area. Models predict biodegradation rates, bioaccumulation potential, and environmental persistence of chemicals. These predictions support regulatory decision-making and environmental risk assessment processes [48].

Ecotoxicological applications focus on predicting chemical impacts on various species and ecosystems. QSAR models evaluate acute and chronic toxicity across different trophic levels, supporting environmental protection efforts. Recent developments incorporate species sensitivity distributions and population-level effects [49].

4.3. Materials Science Applications

QSAR principles extend to materials science through Quantitative Structure-Property Relationships (QSPR). These models predict physical properties of materials, including mechanical strength, conductivity, and optical characteristics. Polymer science applications predict properties based on monomer composition and chain architecture [50].

Nanomaterial applications represent an emerging frontier. QSAR models predict nanoparticle properties and biological interactions, considering unique physicochemical characteristics at the nanoscale. Surface chemistry, size distribution, and aggregation behavior are key parameters in these predictions [51].

5. Challenges

The integration of artificial intelligence continues to expand QSAR capabilities. Deep learning architectures show promise in handling complex structure-activity relationships and generating novel molecular designs. However, challenges remain in model interpretability and reliability assessment [52].

Table 6. Current trends and challenges

Technology	**Current Status**	**Implementation Requirements**	**Expected Impact**	**Challenges**
Quantum Computing	Early development	Quantum hardware	Enhanced accuracy	Hardware limitations
	Algorithm design	Quantum-classical interface	Faster computation	Error correction
	Proof of concept	Specialized expertise	Complex modeling	Scalability
AI Integration	Active development	GPU infrastructure	Automated modeling	Data quality
	AutoML implementation	Cloud computing	Efficient optimization	Interpretability
	Transfer learning	Large datasets	Knowledge transfer	Validation complexity
Federated Learning	Emerging	Distributed systems	Data privacy	Network requirements
	Protocol development	Security frameworks	Collaborative research	Standardization
	Implementation testing	Communication infrastructure	Resource sharing	Protocol optimization
Real-time Analysis	Prototype stage	IoT integration	Dynamic modeling	Data streaming
	Sensor integration	Edge computing	Adaptive prediction	System reliability
	Platform development	High-speed networks	Continuous updating	Integration complexity

Big data analytics and cloud computing platforms enable processing of larger chemical datasets. Distributed computing approaches facilitate model training and validation across extensive chemical spaces. Integration of real-time experimental data enables continuous model refinement and adaptation [53]. Quantum computing applications represent an emerging frontier in QSAR modeling. Quantum algorithms may enable more accurate simulation of molecular properties and interactions. However, practical implementation challenges remain significant [54].

Federated learning approaches enable collaborative model development while maintaining data privacy. This methodology facilitates sharing of predictive models across organizations without exposing proprietary data [55]. Active learning strategies optimize experimental design through intelligent sample selection. These approaches reduce experimental costs while maximizing information gain. Integration with automated synthesis and testing platforms enables rapid model refinement [56].

6. Conclusion

The current state of QSAR modeling comprises of multiple scientific disciplines, including chemistry, biology, computer science, and statistics. This interdisciplinary nature has enabled more nuanced understanding of structure-activity relationships and improved predictive accuracy. The incorporation of biological information through multi-omics data integration has enhanced model relevance for drug discovery applications. Despite significant advances, several challenges persist in QSAR methodology. Data quality, model interpretability, and applicability domain limitations continue to require attention. However, emerging solutions, including automated data curation, advanced interpretation techniques, and sophisticated validation protocols, address these challenges systematically.

References

[1] Hansch C, Fujita T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J Am Chem Soc. 1964;86(8):1616-26.

[2] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc. 1988;110(18):5959-67.

[3] 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.

[4] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88.

[5] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29(6-7):476-88.

[6] Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. Weinheim: Wiley-VCH; 2009.

[7] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30.

[8] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77.

[9] Schneider G. Automating drug discovery. Nat Rev Drug Discov. 2018;17(2):97-113.

[10] Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. Front Environ Sci. 2016;3:80.

[11] Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 2020;63(16):8749-60.

[12] 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.

[13] Baskin II. Machine learning methods in computational toxicology. Methods Mol Biol. 2018;1800:119-39.

[14] Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, et al. Convolutional networks on graphs for learning molecular fingerprints. Adv Neural Inf Process Syst. 2015;28:2224-32.

[15] Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discov Today. 2018;23(8):1538-46

[16] Coley CW, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem Sci. 2019;10(2):370-7.

[17] Weininger D. SMILES, a chemical language and information system. J Chem Inf Comput Sci. 1988;28(1):31-6.

[18] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742-54.

[19] Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. 2016;30(8):595-608.

[20] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. Int Conf Mach Learn. 2017;34:1263-72.

[21] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-80.

[22] Landrum G. RDKit: Open-source cheminformatics software. 2016. Available from: http://www.rdkit.org

[23] Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. Sci Data. 2014;1:140022.

[24] Solovev M, Strokach A, Schwaller P, Akutsu T, Perkins JR, Wong ASW. Polymers property prediction using graph neural networks. Nat Commun. 2022;13(1):1-11.

[25] Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev. 2019;119(18):10520-94.

[26] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Int Conf Learn Represent. 2017.

[27] Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet - A deep learning architecture for molecules and materials. J Chem Phys. 2018;148(24):241722.

[28] Xiong J, Stokes JM, Kolluru S, Borca-Tasciuc DA, Collins JJ, Rosen WB. Artificial intelligence in materials discovery. Matter. 2021;4(8):2645-80.

[29] Wieder O, Kohlbacher S, Kuenemann M, Garon A, Ducrot P, Seidel T, et al. A compact review of molecular property prediction with graph neural networks. Drug Discov Today Technol. 2020;37:1-12.

[30] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9

[31] Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702.

[32] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.

[33] Winter R, Montanari F, Noé F, Clevert DA. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci. 2019;10(6):1692-701.

[34] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018;361(6400):360-5.

[35] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inf Process Syst. 2014;27.

[36] Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572-83.

[37] Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett. 2012;108(5):058301.

[38] Dearden JC. Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point. Environ Toxicol Chem. 2003;22(8):1696-709.

[39] Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today. 2017;22(11):1680-5.

[40] Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51.

[41] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76.

[42] Shi Y, Woodward C, von Luxburg U. On calibration and out-of-distribution generalization. Adv Neural Inf Process Syst. 2021;34:1190-203.

[43] Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. Adv Neural Inf Process Syst. 2017;30:971-80.

[44] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84.

[45] Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. Chem Soc Rev. 2020;49(11):3525-64.

[46] Ramsundar B, Eastman P, Walters P, Pande V, Leswing K, Wu Z. Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more. O'Reilly Media; 2019.

[47] Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digit Signal Process. 2018;73:1-15.

[48] Landrum GA, Stiefl N, Maeda K, Bajorath J. Small molecule machine learning: Beyond Morgan circular fingerprints. J Chem Inf Model. 2021;61(10):4699-706.

[49] Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des. 2013;27(8):675-9.

[50] Freeze JG, Kelly HR, Batista VS. Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. Chem Rev. 2019;119(11):6595-612.

[51] Liu P, Long W. Current mathematical methods used in QSAR/QSPR studies. Int J Mol Sci. 2009;10(5):1978-98.

[52] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

[53] 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.

[54] Schütt KT, Kindermans PJ, Sauceda HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. Adv Neural Inf Process Syst. 2017;30:991-1001.

[55] Unterthiner T, Mayr A, Klambauer G, Hochreiter S. Deep learning as an opportunity in virtual screening. Deep Learn Represent Learn Workshop, NIPS. 2014;27.

[56] Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: quo vadis? J Chem Inf Model. 2012;52(6):1413-37.