ISSN: 0976-7126
Chaturvedi & Gupta, 16(11):11-16, 2025

International Journal of Pharmacy & Life Sciences
Open Access to Researcher

Innovations in QSAR for Accelerated Development of Selective Anti-Breast Cancer Molecules

Prerna Chaturvedi* and Arun Kumar Gupta

Chameli Devi Institute of Pharmacy, Indore, (M.P)-India

Article info

Received: 01/09/2025

Revised:25/10/2025

Accepted: 30/10/2025

© IJPLS

www.ijplsjournal.com

Abstract

Quantitative structure–activity relationship (QSAR) modeling remains a cornerstone of ligand-based drug design. Recent methodological innovations — including deep learning, graph neural networks, multi-task models, transfer learning, and integration with structure-based methods — have increased QSAR's power to predict and prioritize anti-breast cancer compounds. This review synthesizes modern QSAR developments relevant to breast cancer drug discovery, examines datasets and validation best practices, highlights representative case studies where QSAR accelerated identification of selective inhibitors, and discusses persisting challenges and future directions. Key themes include descriptor design and representation learning, multi-target and context-aware models for tumor subtypes, interpretability and model uncertainty, and practical integration with ADMET prediction and experimental pipelines. (Keywords: QSAR, deep learning, breast cancer, 3D-QSAR, GNN, ADMET, ChEMBL).

Keywords: QSAR, Breast Cancer, Molecules

Introduction

Breast cancer remains a leading cause of cancer mortality among women globally and exhibits substantial molecular heterogeneity (You et al., 2022). Differences across subtypes — estrogen receptor (ER)-positive, HER2-positive, and triple-negative breast cancer (TNBC) — demand selective therapeutic strategies tailored to distinct molecular drivers. Ligand-based in silico methods such as QSAR can accelerate early-stage discovery by predicting biological activity from chemical structure, prioritizing molecules for synthesis and testing, and reducing cost and time compared with purely experimental screening (Soares et al., 2022).

QSAR traditionally relied on handcrafted molecular descriptors and classical statistical models (e.g., multiple linear regression, PLS). Over the last decade, machine learning (ML) and, more recently, deep learning (DL) methods — including graph neural networks (GNNs) and transformer-based molecular encoders — have transformed QSAR capabilities, enabling more accurate and generalizable predictions across larger and noisier datasets (Soares et al., 2022; Li, 2025). At the same time, integration with structure-based methods (docking, molecular dynamics), ADMET modelling, and multi-objective optimization has made virtual screening pipelines more realistic and directly useful for anti-breast cancer drug development (El Rhabori et al., 2025; Zarougui et al., 2024).

This review focuses on these innovations and how they can be applied to accelerate development of selective anti-breast cancer molecules.

***Corresponding Author**

E.mail: prernachaturvedi12@gmail.com

Targets and biological context in anti-breast cancer QSAR

Effective QSAR for breast cancer requires clear definition of targets and biological endpoints. QSAR models have been applied at different levels:

Single-target biochemical assays: e.g., aromatase, estrogen receptor α (ERα), HER2 kinase domain, CDKs, tubulin. These models predict binding or inhibition against purified protein/assay endpoints and are useful for lead optimization (Bhatia et al., 2025).

Cellular endpoints: e.g., cytotoxicity (MCF-7, MDA-MB-231), proliferation IC50, apoptosis markers. These incorporate cellular context but are influenced by permeability, efflux, and metabolism (Karampuri et al., 2024).

Phenotypic endpoints: invasion, migration, colony formation — more distal but biologically relevant.

Multi-assay/multi-target models: combine heterogeneous endpoints (binding across several proteins or cell-line panels) to predict polypharmacology or tumor-subtype specificity (Karampuri et al., 2024; PTML approaches).

Choosing the correct endpoint and dataset is crucial because models trained on biochemical binding do not always translate to cellular efficacy (El Rhabori et al., 2025).

Datasets and curation

Large, curated datasets fuel modern QSAR. Public resources include **ChEMBL** (bioactivity and assay metadata), PubChem BioAssay, GDSC/CCLE for cell-line sensitivity data, and specialized literature datasets focused on breast cancer targets (ChEMBL; Karampuri et al., 2024).

Recent reviews emphasize the importance of:

Assay harmonization (standardizing units, assay formats).

Activity threshold selection (e.g., pIC50 cutoffs) or regression targets.

Removing duplicates and inconsistent records.

Annotating assay context (assay type, cell line, species).

Large aggregated datasets enable ML/DL training but introduce noise; careful curation and metadata-aware modeling (e.g., multi-task or PTML approaches) help mitigate this (Soares et al., 2022; PTML literature).

Molecular representations and descriptors

Representation is central to QSAR performance.

Classical descriptors

1D/2D descriptors: molecular weight, logP, topological polar surface area (TPSA), counts of specific atom types, rotatable bonds.

Fingerprinting: ECFP (circular/Morgan fingerprints), MACCS keys — widely used for similarity, clustering, and as inputs for ML (Soares et al., 2022).

3D descriptors and 3D-QSAR

3D-QSAR (CoMFA, CoMSIA) uses aligned 3D conformations to model steric/electrostatic fields affecting binding. For breast cancer targets where ligand conformations and binding modes are well defined (e.g., ER), 3D-QSAR can yield interpretable SAR maps guiding optimization (Kim et al., 2022; Zarougui et al., 2024).

Learned representations

Deep learning enables **end-to-end** learning from raw representations:

Graph neural networks (GNNs) encode molecular graphs (atoms/nodes; bonds/edges) and learn features via message passing.

Sequence/SMILES models (RNNs, Transformers) learn patterns in SMILES strings.

Pretrained molecular encoders and **transfer learning** (pretraining on large chemical corpora and fine-tuning for breast cancer endpoints) improve performance in low-data regimes (Soares et al., 2022).

These learned representations often outperform handcrafted descriptors for complex, non-linear activity relationships.

Machine learning and deep learning architectures

Classical ML methods

Random forests, support vector machines, gradient boosting (XGBoost/LightGBM) have strong performance, particularly with engineered features and moderate-sized datasets.

Deep learning

Advances include:

GNNs (GCN, GAT, MPNN) for structure aware learning.

Graph Transformers and attention-based architectures.

Multi-task neural networks predicting multiple endpoints simultaneously (beneficial when

endpoints are related, e.g., activity across cell lines).

Hybrid models combining fingerprints with GNN embeddings or docking scores.

Soares et al. (2022) and other recent reviews document superior performance of deep learning when properly regularized and when sufficient data or transfer learning is available.

Multi-task, multi-target, and context-aware QSAR

Cancer is complex; compounds frequently act on multiple targets and their efficacy varies across cell lines and contexts. Innovations include:

Multi-task learning (MTL): trains a shared model to predict several related endpoints; helps transfer information between assays and reduces overfitting (Karampuri et al., 2024).

Perturbation-aware models (PTML): incorporate assay metadata as features to model cross-assay heterogeneity (PTML literature).

Contextual modeling: cell-line or omics features are combined with compound features to predict cell sensitivity (integrating genomic features of tumor models) (Karampuri et al., 2024; You et al., 2022).

These strategies are especially useful for predicting subtype-selective activity (e.g., compounds preferentially active in ER+ vs TNBC cell lines).

Integrating QSAR with structure-based methods, docking and MD

Combining ligand-based QSAR with docking, molecular dynamics (MD), and physics-based scoring improves confidence and can suggest binding modes for molecules prioritized by QSAR (El Rhabori et al., 2025; Zarougui et al., 2024). Typical integrated workflow:

Large chemical library screened by fast QSAR/GNN models.

Top candidates docked to target structures; docking scores and interaction fingerprints appended as additional descriptors.

MD used to validate docking poses and compute binding free energy estimates for top candidates.

ADMET filters applied before synthesis.

This multi-layered approach leverages strengths of each method: QSAR's speed and patterns across known actives, and docking/MD's mechanistic insight.

ADMET and toxicity prediction in anti-breast cancer pipelines

Predicting ADME/Tox early reduces attrition. Modern QSAR integrates ADMET predictions (logP, solubility, CYP inhibition, hERG liability, DILI) as part of multi-objective optimization (Mostafa et al., 2024). Deep learning models trained on large toxicology datasets provide a second line of filters to remove compounds with high predicted liabilities.

Validation, applicability domain, and uncertainty quantification

Robust validation is essential to avoid misleading claims:

Cross-validation strategies: scaffold split (more realistic) vs random split; scaffold split approximates prospective generalization (Soares et al., 2022).

External test sets: gold standard.

Applicability domain (AD): define chemical space where predictions are reliable (distance-based, similarity thresholds).

Uncertainty estimation: ensembles, Bayesian neural networks, Monte Carlo dropout help quantify model confidence — invaluable for prioritizing compounds for experimental follow-up.

Recent reviews stress scaffold-aware splits and explicit AD reporting as necessary for credible QSAR (Soares et al., 2022; Li, 2025).

Interpretability and explainability

Adoption in drug discovery benefits from interpretable models:

Feature importance (SHAP, permutation) for fingerprint/descriptor models.

Attention maps and substructure attribution for GNNs/transformers.

3D-QSAR contour maps showing steric/electrostatic regions to modify.

Explainability helps chemists design structural edits to optimize potency/selectivity.

Representative case studies (selected)

Aromatase and ERα inhibitors

Aromatase and ERα are classic targets for ER+ breast cancer. QSAR and 3D-QSAR studies (CoMFA/CoMSIA) plus docking have guided optimization of flavonoid derivatives and steroidal scaffolds (Awasthi et al., 2015; Bhatia et al., 2025).

Tubulin inhibitors and mitotic targets

Recent QSAR + docking studies identified novel tubulin inhibitors with activity in breast cancer models (Moussaoui et al., 2024). Combined 3D-QSAR and MD elucidated binding features.

Cell-line specific QSAR

Karampuri et al. (2024) developed combinational QSAR models that integrate multiple assays and cell-line metadata to predict activity across breast cancer cell lines. Multi-task approaches improved generalization.

Deep learning in virtual screening

Several recent studies show GNNs and transformer models trained or pretrained on large chemical libraries can prioritize actives against breast cancer targets and, when combined with docking/ADMET filters, accelerate hit discovery (Soares et al., 2022; practical cheminformatics analyses).

Practical workflows and software/tools

Open-source and commercial tools support modern QSAR pipelines:

Descriptor calculators: RDKit, PaDEL.

Fingerprints: RDKit (Morgan), CDK.

ML frameworks: scikit-learn, XGBoost, PyTorch, TensorFlow, DGL/PyG for GNNs.

Databases: ChEMBL, PubChem, GDSC/CCLE.

Docking: AutoDock Vina, Glide (commercial).

3D-QSAR: SYBYL/Forge (commercial), open approaches with RDKit + custom grids.

ChEMBL and other curated repositories are primary sources for activity data (ChEMBL).

Challenges and limitations

While innovations improved QSAR, challenges remain:

Data quality and assay heterogeneity — noisy labels, inconsistent endpoints hamper learning.

Generalization — even advanced models can fail on novel scaffolds without transfer learning or domain adaptation (Variational blog discussion).

Explainability vs performance tradeoffs.

Integration with biological heterogeneity — tumor microenvironment, heterogeneity across patients not captured by cell lines.

Regulatory acceptance and reproducibility — need for standardized reporting, code/data sharing, and external prospective validations.

Future directions and opportunities

Key future angles to accelerate selective anti-breast cancer discovery:

Pretrained chemical language models & transfer learning: pretrain on billions of molecules to improve low-data predictions (Soares et al., 2022).

Multi-omics and systems-level integration: include tumor genomics/epigenomics as context features to predict subtype-specific activity (You et al., 2022).

Active learning and closed-loop experimentation: iterative QSAR-guided design + rapid synthesis/testing to converge on optimized leads.

Generative models constrained by QSAR & ADMET: use generative DL to propose molecules optimized for potency, selectivity, and ADMET profiles simultaneously.

Benchmarking & standardized datasets: curated, cross-validated benchmark sets for breast cancer endpoints to compare models fairly (Li, 2025).

Uncertainty-aware prioritization for better resource allocation to experimental validation.

Conclusions

QSAR remains highly relevant for anti-breast cancer drug discovery. Innovations in molecular representation (GNNs, transformers), multi-task/contextual models, integration with structure-based methods, and ADMET prediction have improved predictive power and practical utility. Careful dataset curation, robust validation (scaffold splits and applicability domains), uncertainty quantification, and interpretability remain essential. Integration of QSAR with experimental and systems biology data, together with closed-loop active learning, promises to

further accelerate discovery of selective and safe anti-breast cancer agents.

References

1. Abdullahi, S. H., Uzairu, A., Shallangwa, G. A., Uba, S., & Umar, A. B. (2023). Ligand-based drug design of quinazolin-4(3H)-ones as breast cancer inhibitors using QSAR modeling, molecular docking, and pharmacological profiling. *Journal of the Egyptian National Cancer Institute, 35*, 24.
2. Angelova, V. T., et al. (2025). Novel indole-based sulfonylhydrazones as potential anti-breast cancer agents: Synthesis, cytotoxicity and QSAR analysis. *Pharmaceuticals, 18*(8), 1231.
3. El Rhabori, S., Alaqarbeh, M., Naanaai, L., El Allouche, Y., El Aissouq, A., Bouachrine, M., Zaitan, H., Chtita, S., & Khalil, F. (2024). Design, 3D-QSAR, molecular docking, ADMET studies and molecular dynamics simulations of new 1,4-quinone and quinoline derivatives as potential anti-breast cancer agents. *Journal of Molecular Graphics and Modelling, 125*, Article 108465.
4. John, A., et al. (2025). QSAR modeling, DFT, molecular docking, molecular dynamics simulation and ADMET analysis of 1,3-diphenyl-1H-pyrazole derivatives as novel ER-positive breast cancer inhibitors. *Beni-Suef University Journal of Basic and Applied Sciences*. Advance online publication.
5. Karampuri, A., et al. (2024). A breast cancer–specific combinational QSAR model integrating cell-line metadata and deep learning for prediction of drug-drug combinations. *Frontiers in Bioinformatics, 4*, Article 1328262.
6. Masand, V. H., et al. (2024). Estrogen receptor-alpha binders for hormone-dependent breast cancer: e-QSAR, molecular docking and molecular dynamics analyses. *ACS Omega*. Advance online publication.
7. Rajagopal, K., et al. (2024). Identifying potent breast cancer inhibitors against ER-α target using 3D-QSAR modeling and in-vitro validation. *Chemistry – A European Journal*. Advance online publication.
8. Saghiri, K., Daoud, I., Melkemi, N., & Mesli, F. (2022). QSAR study, molecular docking/dynamics simulations and ADME prediction of 2-phenyl-1H-indole derivatives as potential breast cancer inhibitors. *Biointerface Research in Applied Chemistry, 13*(2), 154.
9. Shoombuatong, W., et al. (2018). Towards understanding aromatase inhibitory activity via quantitative structure–activity relationships: A review of steroidal and non-steroidal inhibitors. *Current Drug Discovery Technologies, 15*(4), 265–277.
10. Subramani, A. K., Sivaperuman, A., Natarajan, R., Bhandare, R. R., & Shaik, A. B. (2022). QSAR and molecular docking studies of pyrimidine-coumarin-triazole conjugates as prospective anti-breast cancer agents. *Molecules, 27*(6), 1845.
11. Vasilev, B., & Atanasova, M. (2025). A (Comprehensive) review of the application of quantitative structure–activity relationship (QSAR) in the prediction of new compounds with anti-breast cancer activity. *Applied Sciences, 15*(3), 1206.
12. Xie, H., et al. (2014). 3D-QSAR, pharmacophore modeling and virtual screening of steroidal aromatase inhibitors. *International Journal of Molecular Sciences, 15*(6), 10226–10239.
13. El Rhabori, S., et al. (2022). 3D-QSAR, molecular docking and ADMET studies of new thioquinazolinone derivatives as potential breast cancer agents. *European Journal of Medicinal Chemistry, 221*, 113534.
14. Baammi, S., et al. (2023). Potent VEGFR-2 inhibitors for resistant breast cancer: Synthesis, biological evaluation, and 3D-QSAR studies of triazolopyrazine analogs. *Chemistry & Biodiversity, 20*(10), e20230034.
15. Soares, T. A., Nunes-Alves, A. F., Mazzolari, A., Ruggiu, F., Wei, G.-W., & Merz, K. (2022). The (Re)-evolution of

Quantitative Structure–Activity Relationship (QSAR) studies propelled by the surge of machine learning methods. *Journal of Chemical Information and Modeling, 62*(22), 5317–5320.

16. El Rhabori, S., Alaqarbeh, M., Naanaai, L., EL Allouche, Y., El Aissouq, A., Bouachrine, M., Zaitan, H., Chtita, S., & Khalil, F. (2025). Integrative computational strategy for anticancer drug discovery: QSAR-ANN modeling, molecular docking, ADMET prediction, molecular dynamics and MM-PBSA simulations, and retrosynthetic analysis. *New Journal of Chemistry, 49*, 14748–14768.
17. Zarougui, S., et al. (2024). 3D computer modeling of inhibitors targeting the MCF-7 breast cancer cell line. *Frontiers in Chemistry, Article 1384832.*
18. Wang, L., et al. (2021). Quantum chemical descriptors in quantitative structure–activity relationship (QSAR) modeling: principles, advantages, and recent applications. *Environmental Toxicology and Pharmacology, 87*, Article 103720.

Cite this article as:
Chaturvedi P. and Gupta A.K. (2025). Innovations in QSAR for Accelerated Development of Selective Anti-Breast Cancer Molecules. *Int. J. of Pharm. & Life Sci.,* 16(11):11-16.

Source of Support: Nil
Conflict of Interest: Not declared
For reprints contact: ijplsjournal@gmail.com