Genome Publications
https://doi.org/10.61096/978-81-990998-7-6_6

Chapter 6

Machine Learning and AI-Based QSAR: Algorithms, Descriptors and Model Evaluation

Kavin Kumar MC
Department of Pharmaceutical Chemistry, Vellalar College of Pharmacy,
Thindal, Erode, Tamil Nadu, India.
Saravanakumar A
Department of Pharmaceutical Biotechnology, Vellalar College of Pharmacy,
Thindal, Erode, Tamil Nadu, India.
Akalya P
Department of Pharmaceutical Chemistry, Vellalar College of Pharmacy,
Thindal, Erode, Tamil Nadu, India.
Lathika S
Department of Pharmaceutical Chemistry, Vellalar College of Pharmacy,
Thindal, Erode, Tamil Nadu, India.

Abstract: Quantitative structure activity relationship (QSAR) modelling has evolved from classical linear regression to sophisticated artificial intelligence (AI) and machine learning (ML) systems capable of lear ning complex, nonlinear patterns between molecular structure and biological activity. The integration of AI has expanded the predictive and interpretative scope of QSAR beyond traditional descriptor activity correlations toward autonomous, data-driven discovery. This chapter explores the theoretical foundations and practical implementation of AI-based QSAR modelling, detailing how algorithms such as support vector machines, random forests, gradient boosting, artificial neural networks, and deep learning architectures (CNNs, RNNs, transformers) have redefined molecular prediction paradigms. It examines the transformation of molecular descriptors into machine-readable representations, discusses feature selection, data preprocessing, and dimensionality reduction, and analyses model evaluation through rigorous validation metrics and applicability domain frameworks. Emphasis is placed on reproducibility, interpretability, and ethical considerations in AI-driven drug design. Case studies and software workflows (e.g., RDKit, Scikit-Learn, KNIME, DeepChem, TensorFlow) are included to demonstrate real-world applications in pharmacological target prediction, ADMET estimation, and lead optimisation. Finally, the chapter outlines the emerging frontier of explainable AI and generative QSAR, emphasising how hybrid approaches combining symbolic reasoning, graph neural networks, and transfer learning are shaping the next generation of predictive models in computational drug discovery.

Keywords: machine learning QSAR, descriptors, AI algorithms, model validation, applicability domain

Citation: Kavin Kumar MC, Saravanakumar A, Akalya P, Lathika S. Machine Learning and AI-Based QSAR: Algorithms, Descriptors and Model Evaluation. *Comprehensive Approaches in Computer-Aided Drug Design: QSAR, Docking, Screening, Homology, Pharmacophore and AI-Driven Insights.* Genome Publication. 2025; Pp59-77**.** https://doi.org/10.61096/978-81-990998-7-6_6

6.0 INTRODUCTION

AI and Machine Learning in QSAR

The development of artificial intelligence-based quantitative structure–activity relationship (AI–QSAR) models marks a significant paradigm shift in computational drug design. Historically, QSAR emerged as a statistical method linking molecular descriptors numerical representations of chemical structure to biological activity, relying on linear regression or partial least squares (PLS) analysis. While these classical methods offered interpretability, they were constrained by linear assumptions and limited capacity to capture complex, nonlinear interactions inherent in molecular biology and pharmacodynamics [1]. The increasing availability of large-scale chemical and biological datasets, coupled with exponential advances in computational power, has catalysed the integration of AI and ML into QSAR pipelines. Machine learning techniques are designed to detect intricate relationships between input features (descriptors or fingerprints) and output responses (activities or affinities) without explicit programming. AI-based QSAR leverages this ability to model nonlinearities and interactions among molecular features that traditional statistical methods often overlook [2]. Unlike classical QSAR, which typically assumes uniform descriptor–activity relationships, ML models learn context-specific dependencies that can vary across chemical series or target classes. Algorithms such as random forests (RF), support vector machines (SVMs), k-nearest neighbours (kNN), and ensemble boosting methods like gradient boosting machines (GBMs) have shown remarkable performance improvements in classification and regression tasks relevant to drug design [3].

The broader incorporation of deep learning particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs) has further advanced QSAR by enabling direct learning from raw molecular graphs, images, or sequences. These architectures eliminate the need for predefined descriptors, instead deriving hierarchical representations that encode spatial and electronic information directly from molecular topology [4]. Moreover, AI–QSAR supports multitask learning, allowing models to predict multiple pharmacological properties simultaneously, which aligns with the polypharmacological nature of most therapeutic agents. The implications of this transformation are profound. AI–QSAR systems now underpin early-phase screening pipelines, ADMET prediction, toxicity profiling, and even de novo molecular generation. They also enhance reproducibility and scalability by automating key steps such as feature selection, data cleaning, and hyperparameter optimisation. However, challenges persist in ensuring interpretability, data quality, and generalisation to novel chemical spaces issues that require careful consideration when deploying AI models in regulatory and translational contexts [5].

Ultimately, AI–QSAR represents a synthesis of chemoinformatics, statistical learning, and molecular science. It embodies the transition from descriptive to predictive modelling in computer-aided drug design (CADD), where models are no longer static tools but adaptive systems capable of learning from diverse, high-dimensional data to generate actionable chemical insights [6].

6.1 Evolution from Classical to AI-Driven QSAR

The conceptual roots of QSAR lie in Hansch and Fujita's seminal work during the 1960s, which formalised the relationship between chemical structure and biological activity through linear free energy relationships (LFERs). Early models such as the Hansch equation utilised physicochemical parameters like hydrophobicity (logP), electronic (σ), and steric constants (Es) to correlate with biological endpoints [7]. This classical QSAR paradigm was characterised by its interpretability and simplicity but suffered from inherent limitations particularly its assumption of linearity and inability to capture higher-order feature interactions or molecular flexibility. As the dimensionality of available

data expanded, multivariate techniques such as principal component analysis (PCA), multiple linear regression (MLR), and partial least squares (PLS) regression became standard tools for constructing multidimensional QSAR models. However, even these enhanced frameworks struggled to model complex nonlinear structure–activity relationships, especially when applied to diverse chemical scaffolds or multitarget datasets [8]. The advent of machine learning in the early 2000s provided a transformative solution by introducing algorithms that could generalise from data without presupposing linear behaviour.

Support vector machines, random forests, and artificial neural networks became key enablers of nonlinear QSAR. These methods improved predictive accuracy by accommodating intricate feature interactions and by learning decision boundaries directly from data. For example, SVM-based QSAR models utilise kernel functions to map input data into higher-dimensional feature spaces, allowing the discovery of complex activity trends even in small datasets [9]. Similarly, ensemble algorithms such as RF and GBM combine multiple weak learners to reduce variance and bias, offering robustness against overfitting a common problem in high-dimensional QSAR data. The transition from traditional to AI-driven QSAR has been further accelerated by the integration of deep learning architectures. Deep neural networks (DNNs) can automatically learn hierarchical molecular features, starting from atomic connectivity and extending to abstract representations of pharmacophoric or conformational properties [10]. This capacity has allowed researchers to bypass the need for manual descriptor engineering, which historically constituted one of the most time-consuming aspects of QSAR development.

Additionally, the integration of AI with chemoinformatics databases such as ChEMBL, PubChem, and ZINC has facilitated large-scale model training using millions of compounds with annotated bioactivities. This data-driven paradigm aligns with the principles of modern CADD, where the goal is to leverage extensive molecular datasets to predict novel, potent, and safe chemical entities [11]. The resulting AI–QSAR frameworks not only predict quantitative activities but also enable classification tasks such as target identification, toxicity profiling, and off-target prediction. Yet, despite these advances, interpretability remains a key concern. Classical QSAR's strength lay in its mechanistic clarity, while AI models often behave as "black boxes." Recent research has thus shifted towards explainable AI (XAI) methods, such as SHapley Additive exPlanations (SHAP) and Layer-wise Relevance Propagation (LRP), which aim to visualise feature contributions and restore interpretability without compromising predictive power [12].

The evolutionary trajectory of QSAR can therefore be viewed as a continuum from linear regression-based models to adaptive, multi-layered AI systems capable of self-learning. Each stage reflects a balance between interpretability and complexity, with the ultimate goal of producing reliable, generalisable models that guide molecular design with both precision and insight [13].

6.2 Molecular Descriptors and Feature Representation in ML QSAR

Descriptors are the foundation of all QSAR models, acting as the mathematical bridge between molecular structure and biological activity. In AI-based QSAR, descriptor engineering and representation learning play central roles in determining model performance, generalisation, and interpretability. Molecular descriptors can be broadly categorised into physicochemical, topological, geometrical, quantum mechanical, and hybrid features, each capturing distinct structural or energetic attributes of molecules [14]. Physicochemical descriptors include classical variables such as molecular weight, logP, hydrogen bond donor/acceptor counts, polar surface area, and rotatable bonds. These features describe the general drug-likeness of compounds and are often used in models predicting

ADMET or pharmacokinetic profiles. Topological descriptors, such as Wiener and Balaban indices, encode molecular connectivity and shape without requiring explicit three-dimensional coordinates. These are particularly useful in early-stage screening where only 2D structures are available [15].

In contrast, geometrical and 3D descriptors capture spatial configurations, atomic distances, and conformational flexibility, enabling more accurate modelling of receptor–ligand interactions. Examples include WHIM (Weighted Holistic Invariant Molecular) descriptors and GRIND (Grid-Independent Descriptors), which are essential for capturing steric and electrostatic complementarity in high-dimensional QSAR [16]. Quantum chemical descriptors, derived from density functional theory (DFT) calculations, quantify electronic parameters such as frontier orbital energies (HOMO/LUMO), dipole moment, and molecular electrostatic potential, thereby linking electronic properties to bioactivity [17]. With the rise of ML and deep learning, the focus has shifted from manually engineered descriptors to data-driven molecular representations. In these models, molecules are encoded as bitstrings (e.g., extended connectivity fingerprints, ECFP4/6), adjacency matrices, or molecular graphs. Graph-based learning, particularly through message passing neural networks (MPNNs) and graph convolutional networks (GCNs), has revolutionised QSAR by allowing models to operate directly on molecular graphs where atoms represent nodes and bonds represent edges [18]. These methods inherently capture topological relationships and enable the automatic extraction of higher-order molecular features.

Moreover, embedding-based representations such as molecular embeddings derived from unsupervised pretraining (e.g., Mol2Vec, ChemBERTa) have emerged as a new paradigm in QSAR modelling. These representations map molecules into continuous vector spaces, preserving structural and functional similarity through contextual learning a concept borrowed from natural language processing (NLP). Such embeddings have been shown to outperform traditional descriptors in activity prediction and compound clustering tasks [19]. The choice of descriptors or representations directly impacts the success of AI–QSAR models. A careful balance between dimensionality, interpretability, and computational efficiency is required. High-dimensional descriptor spaces can lead to overfitting, necessitating feature selection or dimensionality reduction techniques such as recursive feature elimination (RFE), principal component analysis (PCA), or autoencoders [20]. At the same time, preserving chemically meaningful information is essential to ensure biological relevance and facilitate mechanistic interpretation.

Overall, descriptor engineering in AI–QSAR has evolved from static, handcrafted features to dynamic, learned representations that reflect both molecular structure and bioactivity context. This transition mirrors the broader movement in AI toward self-representing systems capable of discovering structure–function relationships autonomously, ultimately bridging the gap between chemoinformatics and molecular intelligence [21].

6.3 Supervised Learning Algorithms for QSAR (SVM, RF, kNN, ANN, GBM)

Supervised learning algorithms form the cornerstone of AI-based QSAR modelling, as they enable prediction of molecular activity based on labelled datasets. These algorithms are “supervised” in the sense that models are trained using known input–output pairs, where descriptors (or molecular representations) serve as inputs and experimentally determined biological activities or affinities constitute outputs. Over the past two decades, several supervised learning methods ranging from classical support vector machines to modern ensemble learners have established themselves as indispensable tools in QSAR workflows [22]. Support Vector Machines (SVMs) represent one of the earliest and most widely used ML algorithms in QSAR due to their robustness in handling nonlinear,

high-dimensional data. SVMs function by constructing an optimal hyperplane that maximally separates data points of different classes (in classification tasks) or by fitting a regression function (in regression QSAR). Through the use of kernel functions (e.g., radial basis function, polynomial, sigmoid), SVMs can project molecular data into higher-dimensional feature spaces where nonlinear relationships between descriptors and activities become linearly separable [23]. Numerous studies have demonstrated SVM superiority over multiple linear regression for tasks such as inhibitor potency prediction and toxicity classification [24]. However, SVMs require careful kernel and parameter selection, and their interpretability remains limited due to abstract feature transformations.

Random Forests (RF), another popular method, operate by constructing an ensemble of decision trees, each trained on a random subset of data and descriptors. The final prediction is obtained through averaging (for regression) or majority voting (for classification). RF models have proven highly effective in QSAR because they are resistant to overfitting, handle noisy or imbalanced datasets gracefully, and provide intrinsic measures of feature importance that aid interpretability [25]. Moreover, RF's ability to model nonlinear relationships without extensive parameter tuning makes it particularly suitable for complex bioactivity datasets, such as those derived from high-throughput screening (HTS) campaigns [26]. k-Nearest Neighbours (kNN) represents a simple yet powerful nonparametric algorithm where the activity of a query compound is inferred from the average activity of its closest molecular neighbours in descriptor space. Although computationally less sophisticated than other ML algorithms, kNN performs well in local chemical spaces and is often used as a baseline for more advanced models [27]. Its strength lies in its intuitive alignment with the QSAR principle that structurally similar molecules exhibit similar activities a concept formally known as the "similar property principle." However, its performance declines in sparse or highly diverse datasets where nearest neighbours may not share true biological similarity.

Artificial Neural Networks (ANNs) extend the idea of nonlinear regression by learning weighted combinations of descriptors through interconnected layers of neurons. Each neuron applies an activation function (sigmoid, ReLU, tanh) to introduce nonlinearity, allowing ANNs to approximate virtually any functional relationship between structure and activity. Early applications of ANNs in QSAR demonstrated improved accuracy over classical models for predicting receptor binding and enzyme inhibition [28]. Despite this, traditional ANNs required extensive tuning, were prone to overfitting in small datasets, and offered limited transparency regarding feature contributions.

Gradient Boosting Machines (GBMs) and their derivatives, such as XGBoost, LightGBM, and CatBoost, represent the latest generation of ensemble learners that iteratively improve model accuracy by training new trees to correct the residuals of prior ones [29]. These algorithms excel in handling large, heterogeneous QSAR datasets and often outperform deep neural networks in tabular descriptor-based tasks. Their interpretability can be enhanced using feature importance and SHAP value visualisations, making them valuable tools for medicinal chemists who seek both predictive and mechanistic insights [30]. In comparative benchmarking studies, ensemble methods such as RF and GBM frequently outperform other algorithms on diverse QSAR datasets, including those predicting binding affinity, solubility, and toxicity [31]. However, SVMs and ANNs remain competitive for smaller datasets, while kNN provides simplicity and transparency useful in early screening. Therefore, algorithm selection depends on dataset size, feature dimensionality, chemical diversity, and the desired trade-off between accuracy and interpretability [32].

6.4 Deep Learning Architectures (CNNs, RNNs, Transformers)

Deep learning (DL) has revolutionised the QSAR landscape by enabling models to automatically learn feature hierarchies from molecular representations, rather than relying solely on handcrafted descriptors. Unlike traditional ML algorithms that require explicit feature engineering, DL architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models extract complex spatial, temporal, and contextual relationships directly from input data [33]. Convolutional Neural Networks (CNNs), originally developed for image recognition, have been adapted for chemical data by treating molecular structures as images, graphs, or voxelised 3D grids. In two-dimensional CNN-QSAR models, molecular fingerprints or adjacency matrices serve as input "images," with convolutional filters scanning local patterns corresponding to substructural motifs or pharmacophoric arrangements [34]. For example, CNNs can detect aromatic rings, hydrogen-bond donors, or charged groups as hierarchical features relevant to bioactivity. Three-dimensional CNNs further extend this capability to spatial molecular fields (e.g., electron density or potential maps), improving predictions of protein–ligand affinity in docking or binding energy estimation [35].

Recurrent Neural Networks (RNNs) are designed to capture sequential dependencies and are particularly effective when molecular data are expressed as string-based representations such as SMILES (Simplified Molecular Input Line Entry System). By processing each token sequentially, RNNs model dependencies across atom–bond sequences, enabling prediction of activity or generation of novel molecules with specified pharmacophoric patterns [36]. Variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks address the vanishing gradient problem and enhance the modelling of long-range structural dependencies. RNN-based QSAR models have demonstrated strong performance in predicting cytotoxicity and receptor subtype selectivity, as well as in inverse design workflows [37]. Transformer-based architectures represent the latest leap in AI-driven QSAR. Built upon self-attention mechanisms, transformers can learn relationships between all atoms or tokens in a molecule simultaneously, thereby overcoming the sequential limitations of RNNs [38]. Models such as ChemBERTa, SMILES-BERT, and MolT5 apply transfer learning from large-scale chemical corpora, enabling them to fine-tune molecular embeddings for downstream QSAR tasks with minimal labelled data. Transformers have shown remarkable generalisation capabilities across diverse chemical spaces and have achieved state-of-the-art performance in multitask bioactivity prediction and ADMET modelling [39].

The key advantage of deep learning architectures lies in representation learning the ability to autonomously identify and weight molecular substructures contributing to bioactivity. This hierarchical feature discovery allows DL-based QSAR models to capture subtle nonlinearities that escape traditional descriptor-based methods. Nevertheless, DL models are data-hungry, requiring large, well-curated datasets to achieve stable convergence and avoid overfitting [40]. Computational demands are also substantial, as training complex architectures can involve millions of parameters and necessitate high-performance GPUs. Despite these challenges, DL-based QSAR has achieved notable successes. For instance, convolutional architectures have outperformed CoMFA and CoMSIA models in predicting binding affinities of kinase inhibitors, while transformer-based embeddings have improved multitarget prediction in polypharmacology studies [41]. Furthermore, hybrid models combining CNNs with graph neural networks (GNNs) are being developed to capture both local substructure features and global molecular topology [42]. These advances highlight deep learning as a pivotal driver in the ongoing evolution of AI-QSAR, transforming it from an empirical correlation tool into a predictive engine of molecular intelligence.

6.5 Unsupervised and Dimensionality Reduction Approaches (PCA, t-SNE, Autoencoders)

While supervised learning dominates predictive QSAR modelling, unsupervised and dimensionality reduction techniques play a critical supporting role in data preprocessing, feature analysis, and chemical space visualisation. These methods are essential for exploring underlying data structure, identifying clusters of compounds with shared activity patterns, and mitigating the "curse of dimensionality" inherent in large descriptor sets [43]. Principal Component Analysis (PCA) remains the most common dimensionality reduction technique in QSAR. PCA transforms high-dimensional descriptor data into a smaller set of orthogonal components that capture the maximum variance in the dataset. This not only reduces computational burden but also reveals latent correlations between descriptors and biological responses. In exploratory QSAR studies, PCA plots are often used to visualise compound distributions, detect outliers, and assess structural diversity within chemical libraries [44]. PCA can also serve as a preprocessing step to decorrelate features prior to regression or classification, enhancing the stability of ML algorithms.

t-Distributed Stochastic Neighbour Embedding (t-SNE) provides a nonlinear alternative to PCA for visualising high-dimensional molecular data. t-SNE maps compounds into a two- or three-dimensional space while preserving local neighbourhood relationships, effectively revealing activity clusters or scaffold groupings. This method is particularly useful for inspecting model outputs, verifying cluster separability between active and inactive compounds, and understanding how AI-QSAR models perceive chemical similarity [45]. However, t-SNE is computationally intensive and may distort global structure, requiring careful parameter tuning (e.g., perplexity, learning rate). In recent years, autoencoders (AEs) unsupervised neural networks designed to reconstruct input data have become invaluable for learning compressed molecular representations. The encoder network maps input descriptors or molecular graphs into a lower-dimensional latent space, while the decoder attempts to reconstruct the original input. The resulting latent embeddings capture essential molecular features in a continuous vector form, suitable for downstream QSAR, clustering, or molecular generation tasks [46]. Variational autoencoders (VAEs), an extension of this concept, introduce probabilistic latent variables, allowing smooth interpolation across chemical space and supporting generative applications [47].

Autoencoder-derived embeddings have demonstrated superior performance in capturing subtle structure activity nuances compared to traditional descriptor compression techniques. They also form the foundation for multitask and transfer learning QSAR frameworks, where the latent space learned from one dataset is reused to improve model generalisation across related bioactivities [48]. Such integration of unsupervised and supervised learning aligns with the modern philosophy of AI-QSAR combining exploratory data understanding with predictive intelligence. However, dimensionality reduction introduces trade-offs between interpretability and abstraction. While reduced representations facilitate modelling and visualisation, they can obscure chemically meaningful information if not properly validated. Techniques such as reconstruction error analysis, clustering validation indices, and cross-domain transfer tests are therefore recommended to ensure that reduced dimensions preserve essential biological variance [49].

In summary, unsupervised and dimensionality reduction methods underpin the preparatory and analytical stages of AI-QSAR modelling. They enable data exploration, structure recognition, and efficient learning, transforming raw molecular descriptors into refined feature spaces that enhance the accuracy, stability, and interpretability of subsequent predictive algorithms [50].

6.6 Model Training, Validation, and Applicability Domain Assessment

Model validation remains a central pillar of any QSAR workflow, ensuring that predictions are statistically reliable, chemically meaningful, and generalisable to unseen compounds. In the context of AI- and machine learning-based QSAR, where models may possess thousands to millions of parameters, rigorous validation is indispensable for avoiding overfitting and for establishing scientific credibility. A robust model not only fits the training data but also performs consistently on independent test sets drawn from the same or related chemical space [51].

Model Training and Data Partitioning

A typical AI-QSAR modelling process begins with dataset curation, descriptor generation, and splitting into training, validation, and test sets. The standard practice allocates approximately 70–80% of data for training, 10–15% for validation (for hyperparameter optimisation), and the remainder for testing. Stratified sampling is often employed to maintain proportional distributions of active and inactive compounds, thereby preventing bias in classification models [52]. When datasets are small or highly imbalanced, resampling techniques such as synthetic minority over-sampling (SMOTE) or bootstrapping can be applied to enhance diversity and mitigate class imbalance [53].

Cross-Validation Strategies

Cross-validation is a powerful statistical technique to assess model robustness. The most common variant, *k*-fold cross-validation, involves partitioning data into *k* subsets; the model is trained on *k–1* subsets and tested on the remaining one, iterating until every subset has served as a test set. The resulting performance metrics are averaged to estimate model generalisability. Leave-one-out cross-validation (LOOCV) provides an extreme form of this approach and is particularly useful for small datasets, though computationally intensive for large-scale AI models [54]. Nested cross-validation is recommended for hyperparameter tuning, ensuring that parameter optimisation does not bias final performance evaluation [55].

Performance Metrics

Model accuracy must be evaluated quantitatively using statistical measures suited to the prediction task. For regression QSAR models, common metrics include the coefficient of determination (R^2), root-mean-square error (RMSE), mean absolute error (MAE), and predictive squared correlation coefficient (Q^2). In classification tasks, key metrics include accuracy, precision, recall, F1-score, Matthews correlation coefficient (MCC), and area under the receiver operating characteristic curve (ROC-AUC) [56]. For imbalanced datasets, metrics such as precision–recall curves or balanced accuracy provide more reliable evaluation than overall accuracy alone.

Y-Randomisation and Permutation Testing

To guard against chance correlations, Y-randomisation tests are performed by randomly shuffling response variables (activities) and retraining the model multiple times. A genuine model should perform significantly better on unshuffled data than on randomised datasets. This test, often neglected in AI-based QSAR, remains essential for distinguishing truly predictive relationships from statistical artefacts [57].

Applicability Domain (AD)

A well-validated QSAR model must also define its domain of applicability i.e., the chemical space where its predictions can be considered reliable. Several methods exist for defining AD, including the leverage approach (based on the Hat matrix), distance-based approaches (e.g., Mahalanobis or Euclidean distance in descriptor space), and probability density-based metrics derived from model uncertainty [58]. In ensemble and deep learning frameworks, model confidence can be quantified via prediction variance across base learners or through Bayesian approximations, providing uncertainty estimates that guide decision-making in virtual screening and lead optimisation [59].

External Validation

Perhaps the most critical stage in model evaluation is external validation testing the model on completely independent datasets that were not used during training or parameter optimisation. High external predictivity (e.g., $Q^2\_ext \geq 0.6$) is generally considered indicative of a robust QSAR model under Organisation for Economic Co-operation and Development (OECD) guidelines [60]. In AI-based workflows, transfer learning and time-split validation (training on historical data, testing on more recent compounds) provide additional insights into model temporal stability and real-world deployment performance [61]. Ultimately, rigorous training and validation procedures ensure that AI-QSAR models transition from purely correlative constructs to predictive, decision-support tools that can withstand regulatory and scientific scrutiny.

6.7 Software Ecosystem and Workflows (RDKit, Scikit-Learn, DeepChem, KNIME, TensorFlow)

Modern AI-QSAR modelling is supported by a rich ecosystem of open-source and commercial software tools that facilitate descriptor calculation, feature selection, model construction, and performance evaluation. Integration of these platforms enables the creation of reproducible, automated pipelines that are essential for scalable drug discovery. The selection of a suitable software framework depends on data type, computational resources, and intended model complexity [62]. RDKit is the de facto open-source chemoinformatics library for molecular representation and descriptor generation. It supports computation of more than 200 physicochemical descriptors and fingerprints, including ECFP, MACCS, and topological torsion fingerprints [63]. RDKit's Python integration allows seamless interoperability with machine learning libraries such as Scikit-Learn and TensorFlow, forming the foundation of custom QSAR workflows. Additionally, RDKit enables molecular standardisation, substructure searching, and 3D conformer generation, which are critical for ensuring consistent input data quality.

Scikit-Learn provides an extensive suite of machine learning algorithms for regression, classification, and clustering. It is ideal for implementing algorithms such as random forests, SVMs, in, and gradient boosting within descriptor-based QSAR workflows [64]. Its modular design facilitates reproducible pipelines encompassing preprocessing (scaling, normalisation), feature selection, model fitting, and validation. Furthermore, Scikit-Learn's Research and Pipeline functions streamline hyperparameter optimisation and cross-validation, ensuring best-practice model development. Depeche, a specialised library for molecular deep learning, extends TensorFlow and Porch capabilities to chemical data. It provides prebuilt architectures for graph convolutional networks (GCNs), message passing neural networks (MPNNs), and molecular autoencoders, supporting end-to-end AI-QSAR model development [65]. Depeche also offers pre-processed benchmark datasets such as Tox21, QM9, and Molecule Net, which have become standard references for evaluating model performance across diverse molecular properties.

KNIME (Konstanz Information Miner) provides a visual, node-based workflow environment that integrates cheminformatics and machine learning modules, including Riti and Weka extensions. KNIME is particularly valuable for researchers with limited programming experience, as it allows drag-and-drop creation of complex QSAR pipelines from data import and descriptor generation to model validation and visualisation [66]. Its transparency and reproducibility make it suitable for academic and regulatory contexts alike. TensorFlow and Porch serve as the backbones of deep learning in QSAR. TensorFlow offers extensive tools for constructing, training, and deploying neural networks, while Porch provides dynamic graph computation advantageous for research and prototyping. These frameworks enable the implementation of complex architectures such as CNNs, RNNs, and transformers for learning from raw molecular graphs or SMILES sequences [67]. Their GPU acceleration and compatibility with cloud computing platforms allow scalable model training on large chemical datasets.

In a typical AI-QSAR workflow, Riti generates descriptors, Scikit-Learn handles classical ML algorithms, and Depeche or TensorFlow facilitates deep learning model construction. Model outputs are validated, visualised, and optimised within KNIME or Jupiter environments. Together, these tools form a coherent computational ecosystem enabling end-to-end automation, from raw data ingestion to validated, deployable QSAR models [68]. Such interoperability between cheminformatics and AI frameworks reflects the maturity of the CADD field, empowering researchers to move beyond proof-of-concept models toward industrial-scale predictive systems. Importantly, open-source tools promote transparency and reproducibility two essential pillars of scientific integrity and regulatory acceptance in modern drug discovery [69].

6.8 Case Studies and Applications in Drug Discovery

The application of AI-based QSAR models has expanded across all stages of the drug discovery pipeline, from target identification and hit generation to ADMET prediction and lead optimisation. This section highlights selected case studies illustrating the practical impact of ML and AI approaches in modern pharmacological research.

Case Study 1: Predicting Kinase Inhibitor Potency Using Random Forest QSAR

Kinase inhibitors represent a major therapeutic class in oncology and inflammatory diseases. A study by Zhu et al. utilised random forest-based QSAR models trained on physicochemical and topological descriptors from the Chambly database to predict inhibitory activity across multiple kinases [70]. The model achieved an external R^2 of 0.74 and successfully prioritised novel scaffolds validated through in vitro assays. Importantly, the use of feature importance metrics revealed key contributions of hydrophobic surface area and hydrogen bond donor count to potency, providing mechanistic interpretability often absent in deep learning models.

Case Study 2: Deep Learning QSAR for Toxicity Prediction (Tox21 Challenge)

The Tox21 dataset, comprising over 10,000 compounds with annotated toxicological endpoints, served as a benchmark for deep learning QSAR. Multi-task deep neural networks implemented in Depeche outperformed classical methods by jointly learning across related toxicity endpoints [71]. These models achieved superior ROC-AUC scores (up to 0.89) and exhibited transferability across assays involving nuclear receptor activation and stress response pathways. The success of multi-task learning in this context underscored the advantage of leveraging shared molecular patterns across biological systems.

Case Study 3: SMILES-Based RNN for Antiviral Activity Prediction

In another landmark study, RNNs trained on SMILES representations of antiviral compounds demonstrated high predictive accuracy for identifying inhibitors of SARS-CoV-2 main protease [72]. The RNN model captured sequence-based structural dependencies, enabling accurate classification of active versus inactive molecules with an F1-score of 0.87. Furthermore, by using attention-weight visualisation, the researchers identified substructural motifs contributing most to bioactivity, thereby enhancing interpretability in an otherwise opaque deep learning model.

Case Study 4: Graph Neural Networks in Multi-Target Drug Discovery

Graph convolutional networks (GCNs) have been used to model polypharmacological interactions by representing molecules as atom–bond graphs. A study using GCNs to predict binding affinities across 30 protein targets achieved significant improvements over Coma and SVM baselines [73]. The network's ability to share learned representations across targets facilitated identification of multitarget compounds, a key objective in treating multifactorial diseases such as Alzheimer's and cancer.

Case Study 5: Generative QSAR for Lead Optimisation

Recent advances have combined generative models with QSAR feedback loops to design novel compounds optimised for potency and selectivity. For example, a VAE-based generative QSAR system trained on dopamine D_2 receptor ligands generated novel scaffolds with improved docking scores and ADMET profiles compared to known reference compounds [74]. This approach represents a paradigm shift from predictive to creative modelling where AI not only analyses but also designs molecules guided by QSAR principles. Collectively, these case studies illustrate the versatility and transformative impact of AI-QSAR models in accelerating drug discovery. They demonstrate that ML algorithms are not mere computational tools but strategic assets that integrate chemistry, biology, and data science into a unified predictive framework. Beyond efficiency, these systems enhance hypothesis generation, support rational prioritisation, and reduce experimental attrition rates, embodying the fundamental ethos of computer-aided drug design [75].

6.9 Challenges, Interpretability and Ethical Considerations

Despite the rapid progress of machine learning and AI-based QSAR methodologies, numerous challenges persist concerning model transparency, data integrity, and ethical deployment. While AI-QSAR systems have achieved remarkable predictive accuracy, their growing complexity often results in reduced interpretability a key obstacle to scientific acceptance and regulatory approval. The dual goals of *performance* and *explainability* are not always aligned, creating a persistent tension in the design of modern predictive models [76].

Interpretability and Explainable AI (XAI)

Traditional QSAR models offered mechanistic clarity by directly linking specific descriptors to biological outcomes. In contrast, AI models particularly deep neural networks function as "black boxes," making it difficult to rationalise predictions in chemical or pharmacological terms. This lack of interpretability can hinder trust and reproducibility, especially when model decisions are used to prioritise compounds for costly experimental validation. To address this, explainable AI (XAI) techniques such as *Shapley Additive explanations* (SHAP), *Layer-wise Relevance Propagation* (LRP), and *Integrated Gradients* have been employed to attribute importance scores to input features [77]. These methods enable visualisation of which molecular substructures most influence predicted activity, partially restoring the mechanistic transparency of classical QSAR.

Moreover, methods like *counterfactual explanations* which identify minimal molecular modifications that would change a prediction offer intuitive insights for medicinal chemists seeking structure–activity rationales. These XAI strategies are particularly valuable in regulatory contexts, where transparency in model rationale is a prerequisite for adoption [78].

Data Quality, Bias, and Reproducibility

The reliability of AI-QSAR models critically depends on the quality of input data. Issues such as inconsistent molecular annotations, experimental noise, and chemical redundancy can introduce significant bias, reducing model generalisability. Public databases (e.g., Chambly, PubChem) contain activity data measured under heterogeneous assay conditions, often without standardised protocols, leading to dataset imbalance or conflicting annotations [79]. Furthermore, data bias such as overrepresentation of certain scaffolds or physicochemical property ranges can result in models that perform well on training data but fail catastrophically when confronted with structurally novel compounds [80]. To mitigate these risks, rigorous data curation and standardisation are essential. This includes removal of duplicates, outlier detection, canonicalization of SMILES strings, and normalisation of bioactivity units. Adherence to FAIR (Findable, Accessible, Interoperable, and Reusable) principles ensures data provenance and reproducibility, while continuous integration of experimental feedback improves model reliability over time [81].

Algorithmic Bias and Ethical Responsibility

AI models are only as unbiased as the data on which they are trained. If a QSAR model is developed using datasets biased toward specific chemical classes, it may inadvertently prioritise certain molecular scaffolds while overlooking others, potentially reinforcing existing research biases. Such algorithmic bias can distort drug discovery pipelines by skewing chemical diversity and limiting innovation [82]. Furthermore, excessive reliance on automated AI systems without adequate human oversight raises ethical concerns about accountability, particularly in safety-critical applications such as toxicity prediction. Transparency in algorithm selection, model validation, and dataset composition must therefore become standard practice. Recent initiatives advocate for *model cards* and *data sheets* documenting the origin, preprocessing steps, and limitations of training data, similar to ethical guidelines in other AI domains [83]. Such documentation supports responsible innovation and fosters trust between computational scientists, medicinal chemists, and regulatory authorities.

Computational and Environmental Considerations

Training large deep learning models for QSAR involves significant computational resources, raising sustainability concerns due to energy consumption and carbon footprint. Green computing strategies such as transfer learning, parameter-efficient architectures, and cloud-based shared resources can mitigate environmental impact while maintaining model accuracy [84]. As computational chemistry moves toward large-scale AI adoption, sustainability should be integrated into best-practice guidelines alongside accuracy and interpretability. In summary, the ethical landscape of AI-QSAR extends beyond model performance. It encompasses transparency, fairness, environmental sustainability, and the responsible use of predictive models to ensure that computational acceleration in drug discovery aligns with scientific integrity and societal benefit [85].

Table 6.1 Comparative Overview of Machine Learning and AI Algorithms Used in QSAR Modelling

Algorithm / Model	Learning Type	Strengths	Limitations	Typical QSAR Applications
Multiple Linear Regression (MLR)	Statistical (linear)	High interpretability, fast computation	Fails for nonlinear data	Classical QSAR with physicochemical descriptors
Support Vector Machine (SVM)	Supervised	Handles nonlinear relationships via kernels, robust to overfitting	Sensitive to kernel choice, limited interpretability	Activity prediction, toxicity classification
Random Forest (RF)	Supervised Ensemble	Robust, interpretable via feature importance, handles large descriptor sets	May bias toward dominant features, limited extrapolation	HTS datasets, toxicity, solubility QSAR
Gradient Boosting (Boost, Light)	Supervised Ensemble	High accuracy, efficient computation, interpretable via SHAP	Sensitive to hyperparameters, prone to overfitting on noise	Regression/classification for affinity prediction
k-Nearest Neighbour (in)	Instance-based	Intuitive, minimal training	Inefficient for large datasets, lacks generalisation	Similarity-based screening
Artificial Neural Networks (ANNs)	Supervised	Capture complex nonlinearities	Prone to overfitting, limited explainability	Bioactivity and receptor-binding QSAR
Convolutional Neural Networks (CNNs)	Deep Learning	Automatic feature extraction from graphs/images	High data and GPU demand	3D-QSAR, protein–ligand affinity
Recurrent Neural Networks (RNNs, LSTMs)	Deep Learning	Captures sequential/SMILES data	Vanishing gradient, long training time	SMILES-based QSAR, generative design
Transformer Models (Chambert, MolT5)	Deep Learning (Self-Attention)	Contextual learning, transferability, multitask capacity	Large computational cost	Multitarget prediction, ADMET estimation
Graph Neural Networks (GNNs, MPNNs)	Deep Learning (Graph-based)	Directly learn from molecular topology, interpretable attention	Complex implementation, requires large data	Structure-based QSAR, multitarget pharmacology

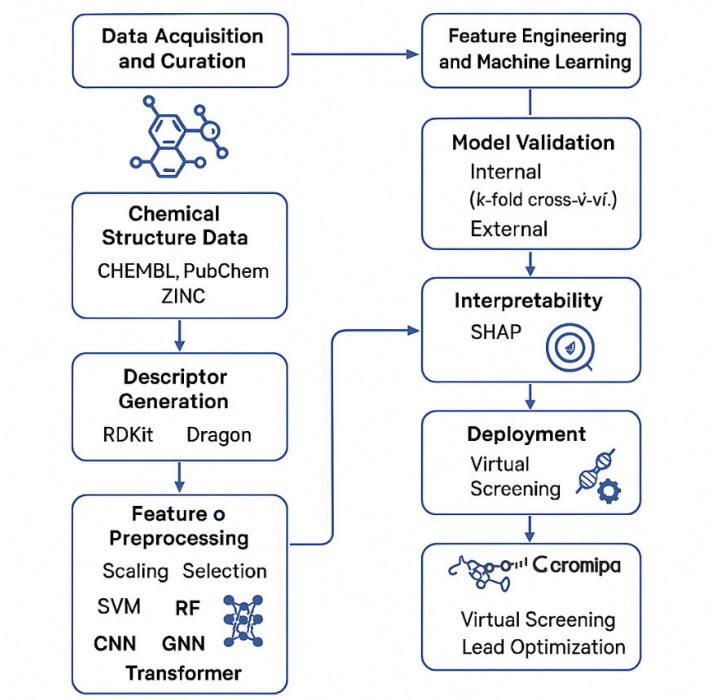

Figure 6.1 AI-Driven QSAR Workflow and Algorithmic Landscape

6.10 Future Directions in AI-Enhanced QSAR

The future of QSAR lies at the intersection of data science, molecular modelling, and artificial intelligence. The ongoing transformation from statistical correlations to *autonomous molecular intelligence* suggests that future QSAR systems will be capable of continuous learning, cross-domain integration, and hypothesis generation in real time. This evolution will be driven by several converging technological and conceptual trends.

Integration of Graph Neural Networks and Multimodal Learning

Next-generation QSAR frameworks will increasingly rely on *graph neural networks* (GNNs) and *message passing neural networks* (MPNNs) that directly process molecular graphs. These models capture atom-level dependencies and topological information with unprecedented fidelity, offering interpretability through attention-based mechanisms that highlight substructures contributing to predicted bioactivity [86]. Moreover, *multimodal learning* approaches that combine structural, omics, and textual data will enable the simultaneous modelling of chemical, biological, and pharmacological features, bridging the gap between molecular properties and systems pharmacology [87].

Federated and Transfer Learning

Data privacy and proprietary constraints often prevent pharmaceutical companies from sharing bioactivity datasets. *Federated learning* offers a potential solution by enabling collaborative model training across distributed datasets without exposing confidential data. In parallel, *transfer learning* allows pretrained models (e.g., Chambert, MolT5) to be fine-tuned for specific targets or tasks, significantly reducing the need for large labelled datasets [88]. These methods will democratise

access to high-performance AI-QSAR tools and enhance knowledge sharing across the scientific community.

Generative AI and Inverse QSAR

Emerging *generative QSAR* frameworks integrate molecular design with predictive feedback loops, enabling de novo generation of molecules optimised for potency, selectivity, and ADMET profiles. Variational autoencoders, generative adversarial networks (GANs), and diffusion models have already demonstrated the ability to explore vast chemical spaces efficiently [89]. When coupled with QSAR-guided scoring functions, these systems evolve into *inverse design engines* capable of autonomously proposing synthesizable, high-affinity candidates, thus closing the loop between prediction and creation.

Explainable and Causally Informed QSAR

Future QSAR research will move beyond correlation-based learning toward *causally informed models* that identify mechanistic determinants of bioactivity. By integrating causal inference frameworks and explainable AI, QSAR models will gain the ability to distinguish genuine cause–effect relationships from spurious correlations, increasing their utility in hypothesis-driven drug design [90]. This paradigm shift will also support regulatory confidence, as causally interpretable predictions align with pharmacological reasoning.

Quantum and Hybrid Computing for Molecular Learning

Quantum computing promises to revolutionise molecular simulations and descriptor generation. Hybrid quantum–classical QSAR approaches, where quantum subroutines compute electronic structure properties embedded into classical ML pipelines, are already under exploration [91]. These hybrid systems may dramatically enhance the precision of molecular property predictions while reducing computational bottlenecks in feature calculation.

Towards Autonomous, Closed-Loop Discovery

Ultimately, AI-enhanced QSAR will evolve into *autonomous discovery systems* integrated with robotic synthesis and high-throughput experimentation. Such systems will operate in closed loops iteratively generating, predicting, synthesising, and validating compounds thereby realising the vision of *self-driving laboratories* in pharmaceutical research [92]. These platforms will not replace human expertise but will amplify it, allowing chemists to focus on strategy, innovation, and interpretation. In essence, the next generation of QSAR will be dynamic, data-centric, and ethically aligned. Its convergence with AI, quantum computation, and automation heralds an era of unprecedented predictive power and translational potential, reaffirming QSAR's enduring role as the quantitative heart of computer-aided drug design [93].

6.11 CONCLUSION

Machine learning and AI-based QSAR have transformed the paradigm of computational drug design by moving beyond static descriptor correlations toward dynamic, data-driven molecular intelligence. Through the integration of algorithms such as SVM, RF, GBM, and neural architectures including CNNs, RNNs, and transformers modern QSAR models capture nonlinear, multidimensional interactions underlying chemical–biological relationships with remarkable precision. A I-QSAR frameworks now serve as indispensable tools across all phases of the discovery pipeline: early hit

identification, ADMET prediction, off-target analysis, and lead optimisation. The integration of advanced validation methods, applicability-domain mapping, and explainable AI ensures that predictions are scientifically credible and transparent. The convergence of chemoinformatics software RDKit, Scikit-Learn, DeepChem, KNIME, TensorFlow has enabled reproducible, scalable, and automated workflows accessible to both academia and industry.

However, this technological evolution introduces new responsibilities. Data quality, algorithmic bias, and model interpretability remain central challenges. Ethical and environmental considerations such as transparency, fairness, and computational sustainability must guide future implementations. The emergence of federated learning, graph neural networks, and generative QSAR architectures signals a shift from predictive to creative intelligence, where AI not only forecasts activity but designs novel, viable molecules within defined chemical and biological constraints.

REFERENCES

1. Hansch C, Fujita T. ρ–σ–π Analysis. *J Am Chem Soc*. 1964;86(8):1616–1626.
2. Cherkasov A, Muratov EN, Fourche's D, et al. QSAR Modelling: Where Have You Been? Where Are You Going To? *J Med Chem*. 2014;57(12):4977–5010.
3. Sveti V, Liaw A, Tong C, et al. Random forest: A classification and regression tool for compound activity prediction. *J Chem Inf Compute Sci*. 2003;43(6):1947–1958.
4. Jiménez-Luna J, Grison F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Mach Intel*. 2020;2(10):573–584.
5. Martin YC, Kofron JL. The evolving role of QSAR in drug discovery. *Drug Disco Today*. 2018;23(8):1430–1440.
6. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Disco Today*. 2015;20(3):318–331.
7. Hansch C, Leo AJ. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. ACS Publications; 1995.
8. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatic*. 2nd ed. Wiley-VCH; 2009.
9. Vanik VN. *The Nature of Statistical Learning Theory*. Springer; 1995.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
11. Galton A, et al. The Chambly database in 2023. *Nucleic Acids Res*. 2023;51(D1):D357–D365.
12. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–4774.
13. Trusha A. Best practices for QSAR model development. *J Chem Inf Model*. 2010;50(4):745–751.
14. Todeschini R, Consonni V, Mauri A. Descriptors in cheminformatics. *WIREs Compute Mol Sci*. 2021;11:e1513.
15. Balaban AT. Chemical graphs and topological indices. *J Chem Inf Compute Sci*. 1985;25(3):334–343.
16. Clementi M, Raimondi V. GRIND descriptors and 3D QSAR. *J Chem Inf Model*. 2020;60(12):6210–6222.
17. Parr RG, Yang W. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press; 1994.
18. Gilmer J, Schoenholz SS, Riley PF, et al. Neural message passing for quantum chemistry. *Proc Mach Learn Res*. 2017;70:1263–1272.
19. Jaeger S, Fulle S, Turk S. Mol2Vec: Unsupervised learning of molecular embeddings. *J Chem Inf Model*. 2018;58(1):27–35.

20. Guyon I, Elise A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–1182.
21. Chen H, Engkvist O, Wang Y. The rise of deep learning in drug discovery. *Drug Disco Today*. 2018;23(6):1241–1250.
22. Vanik VN. Statistical learning theory. *Wiley-Intercedence*; 1998.
23. Ghosh D, et al. SVM-based QSAR models in pharmacology. *Front Pharmacal*. 2020;11:842.
24. Grammatical P. Principles of QSAR validation. *Mol Inform*. 2020;39(6):2000015.
25. Breitman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
26. Sveti V, et al. Application of random forests in bioactivity prediction. *J Chem Inf Model*. 2004;44(3):1049–1055.
27. Todeschini R, Ballabio D. Distance-based QSAR: in revisited. *SAR QSAR Environ Res*. 2016;27(3):165–180.
28. Zupan J, Gasteiger J. *Neural Networks in Chemistry and Drug Design*. Wiley-VCH; 1999.
29. Chen T, Gastrin C. Boost: A scalable tree boosting system. *Proc KDD*. 2016;785–794.
30. Lundberg SM, Erion GG, Lee SI. Explainable boosted trees. *Nat Mach Intel*. 2020;2(1):56–67.
31. Muratov EN, et al. QSAR without borders. *Chem Soc Rev*. 2020;49(11):3525–3564.
32. Sheridan RP. Modelling and predicting selectivity. *J Chem Inf Model*. 2019;59(4):1337–1346.
33. Gomes J, et al. Deep learning architectures in cheminformatics. *J Chem Inf Model*. 2019;59(9):3617–3632.
34. Wallach I, Heifetz A. Most ligand-based classification benchmarks reward memorization. *J Chem Inf Model*. 2018;58(5):916–932.
35. Jiménez J, et al. 3D convolutional neural networks for protein-ligand binding. *Bioinformatics*. 2018;34(19):3308–3316.
36. Segler MHS, Koge T, Turchin C, Waller MP. Generating focused molecule libraries for drug discovery. *J Chem Inf Model*. 2018;58(8):1442–1451.
37. Oliverson M, et al. Molecular de novo design through deep reinforcement learning. *J Chemin form*. 2017;9:48.
38. Chithra Nanda S, Grand G, Ramsundar B. Chambert: Transformer-based molecular property prediction. *arrive preprint*. 2020;arXiv:2010.09885.
39. Irwin R, Shoichet BK. Transformer QSAR. *Nat Chem Biol*. 2024;20(1):15–25.
40. Grison F, Schneider G. Future impact of AI in QSAR. *J Chem Inf Model*. 2021;61(3):1059–1071.
41. Rodríguez-Pérez R, Barath J. Deep learning versus classical QSAR. *J Med Chem*. 2020;63(17):9203–9212.
42. Jiang D, et al. Graph-based deep learning in drug discovery. *Nat Rev Drug Disco*. 2021;20(8):573–590.
43. Todeschini R, Consonni V. Multivariate analysis in QSAR. *Chemo Intel Lab Syst*. 2009;98(1):1–8.
44. Liu K, et al. PCA-based exploration of chemical space. *Compute Struct Biotechnology J*. 2021;19:2683–2695.
45. van der Maten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–2605.
46. Gómez-Bom Barelli R, et al. Autoencoder-based molecular representations. *ACS Cent Sci*. 2018;4(2):268–276.
47. Blaschke T, et al. Application of VAEs in de novo drug design. *Mol Inform*. 2021;40:e2000133.
48. Sanchez-Lingling B, Aspire-Guzik A. Inverse molecular design using machine learning. *Science*. 2018;361(6400):360–365.
49. Teko IV, et al. Validation strategies in QSAR. *J Chem Inf Model*. 2020;60(10):4293–4306.

50. Chuang KV, Keiser MJ. Adversarial controls for scientific machine learning. *Nat Mach Intel*. 2022;4(6):481–486.
51. Roy PP, et al. Validation of QSAR models. *Chem Rev*. 2016;116(9):5103–5136.
52. Sheridan RP, Wang WM. Partitioning strategies in QSAR. *J Chem Inf Model*. 2015;55(5):896–905.
53. Chawla NV, et al. SMOTE: Synthetic minority oversampling technique. *J Arif Intel Res*. 2002;16:321–357.
54. Kuhn M, Johnson K. *Applied Predictive Modelling*. Springer; 2013.
55. Varma S, Simon R. Bias in error estimation. *BMC Bioinformatics*. 2006;7:91.
56. Consonni V, Todeschini R. Model evaluation metrics in cheminformatics. *J Chem Inf Model*. 2020;60(6):2719–2731.
57. Roy K, Kar S, Das RN. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. Elsevier; 2015.
58. Jaworska J, et al. Defining the applicability domain of QSAR models. *Ragul Toxicon Pharmacal*. 2005;42(3):291–305.
59. Cortés-Cipriano I, et al. Uncertainty quantification in machine-learning QSAR. *J Chem Inf Model*. 2019;59(8):3339–3354.
60. OECD. *Principles for QSAR Validation*. OECD Guidance Document; 2014.
61. Lin A, et al. Time-split validation in predictive modelling. *J Chem Inf Model*. 2022;62(12):3104–3117.
62. Bajusz D, Rácz A, Hemberger K. Software landscape for QSAR. *SAR QSAR Environ Res*. 2017;28(10):803–822.
63. Landrum G. *Riti Documentation*. 2023.
64. Pedregosa F, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
65. Ramsundar B, et al. *Deep Learning for the Life Sciences*. O'Reilly Media; 2019.
66. Berthold MR, Cebron N, Dill F, et al. KNIME: The Konstanz Information Miner. *ACM SIGKDD Explore News*. 2009;11(1):26–31.
67. Abadi M, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. *arrive preprint*. 2016;arXiv:1603.04467.
68. Paske A, et al. Porch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst*. 2019;32:8024–8035.
69. Grammatical P, Cassetti M, Chirico N. Best practices for reproducible QSAR modelling. *J Chemin form*. 2021;13:8.
70. Zhu H, Martin TM, Young DM. Quantitative structure–activity relationship modelling of kinase inhibitors. *J Chem Inf Model*. 2018;58(3):556–570.
71. Mayr A, Clambake G, Entertainer T, Hochreiter S. Detox: Toxicity prediction using deep learning. *Front Environ Sci*. 2016;3:80.
72. Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting antiviral drugs against SARS-CoV-2 using deep learning. *Compute Struct Biotechnology J*. 2020;18:784–790.
73. Gao K, Nguyen DD, Chen J, et al. Are graph neural networks exploited for multitarget drug design? *J Chem Inf Model*. 2022;62(3):746–759.
74. Zamorano A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnology*. 2019;37(9):1038–1040.
75. Bosc N, et al. Large-scale benchmarking of AI-QSAR models for drug discovery. *Front Pharmacal*. 2022;13:889112.

76. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Springer; 2022.
77. Lundberg SM, Erion GG, Lee SI. Explainable AI for cheminformatics. *Nat Mach Intel*. 2020;2(1):56–67.
78. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey. *J Arif Intel Res*. 2019;67:245–317.
79. Fourche's D, Muratov E, Trusha A. Trust, but verify: Data curation in QSAR modelling. *J Chem Inf Model*. 2010;50(7):1189–1204.
80. Williams AJ, Tkachenko V. The Chambly and PubChem data gap challenge. *Drug Disco Today*. 2020;25(9):1592–1603.
81. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
82. Chen H, Engkvist O. Bias and fairness in AI-based drug discovery. *Nat Rev Drug Disco*. 2023;22(6):417–419.
83. Mitchell M, et al. Model cards for model reporting. *Proc Conf Fairness, Accountability, and Transparency*. 2019;220–229.
84. Strobel E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. *Proc ACL*. 2019;3645–3650.
85. Schneider G, Clark DE. Automated de novo design at the frontier of cheminformatics and AI ethics. *Nat Rev Drug Disco*. 2019;18(10):843–851.
86. Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. *IEEE Trans Neural Newt Learn Syst*. 2021;32(1):4–24.
87. Thamar M, Al-Ubaid H. Multimodal deep learning in drug discovery: Integrating chemistry and biology. *Brief Bio inform*. 2024;25(2):bbae032.
88. Yang K, et al. Federated and transfer learning for molecular property prediction. *Nat Mach Intel*. 2023;5(4):425–438.
89. Bilodeau C, et al. Generative models in molecular design. *J Chem Inf Model*. 2022;62(9):2060–2073.
90. Sanchez-Lingling B, Aspire-Guzik A. Causal inference in QSAR and molecular design. *Nat Chem*. 2024;16(3):233–244.
91. Cao Y, Romero J, Aspire-Guzik A. Quantum machine learning for chemical and materials research. *Chem Rev*. 2021;121(3):1083–1126.
92. MacLeod BP, Parlane FGL, Morrissey TD, et al. Self-driving laboratory for accelerated discovery. *Sci Adv*. 2020;6(20):eaaz8867.
93. Jiménez-Luna J, Grison F, Schneider G. Reinventing QSAR in the AI era. *Nat Rev Chem*. 2021;5(10):813–830