

Limitations of QSAR Modeling: Data Bias, Curation, and Predictive Reliability in Computational Drug Discovery



Amena Khatun Manica ^{1*}

Abstract

Quantitative Structure–Activity Relationship (QSAR) modeling remains a widely used approach in computational toxicology and drug discovery, enabling prediction of biological activity from molecular structure. However, despite decades of methodological development, concerns persist regarding the reliability and generalizability of QSAR models. This narrative review revisits QSAR modeling with a focus on data curation, dataset bias, and their impact on predictive reliability. Rather than viewing QSAR as a purely algorithmic process, this review emphasizes how data quality, experimental variability, and endpoint definition influence model performance. Issues such as class imbalance, sampling bias, and publication bias are shown to significantly affect predictive outcomes, often leading to overestimated model accuracy. In particular, commonly used validation metrics, including R^2 , may fail to reflect true predictive performance when external validation and applicability domain considerations are not adequately addressed. Emerging approaches, including advanced validation strategies, consensus modeling, and improved descriptor frameworks, offer partial solutions to these challenges. However, the findings suggest that QSAR reliability is fundamentally dependent on data integrity, transparency, and appropriate validation rather than computational complexity alone. Overall, this review

highlights the need for more robust data curation practices and context-aware validation frameworks to improve predictive modeling in QSAR and enhance its application in drug discovery and computational toxicology. **Keywords:** QSAR modeling; dataset curation; bias; applicability domain; model validation

Keywords

QSAR modeling; dataset curation; bias; applicability domain; model validation

1. Introduction

Quantitative Structure–Activity Relationship (QSAR) modeling—at least in principle—offers an elegant solution to a persistent problem in chemistry and toxicology: how to infer biological activity or physicochemical behavior from molecular structure alone. Since the foundational work of Corwin Hansch and Toshio Fujita in the 1960s, which formalized the relationship between chemical substituents and biological activity, QSAR has steadily evolved from relatively simple regression frameworks into a central pillar of modern computational toxicology and drug discovery (Hansch & Fujita, 1964; Dearden, 2016). Yet, despite this apparent maturity—and perhaps because of it—the field continues to wrestle with deeper methodological uncertainties, particularly those tied to how datasets are

Significance | Highlights hidden data-driven biases in QSAR, reframing reliability as a function of data integrity rather than algorithmic sophistication alone.

*Correspondence.

Amena Khatun Manica, Department of Chemistry, University of New Haven, West Haven, CT, United States. E-mail: monica.95amena@gmail.com,

Editor Vishwanadham Mandala, Ph.D. And accepted by the Editorial Board August 08, 2023 (received for review Jun 14, 2023)

Author Affiliation:

¹ Department of Chemistry, University of New Haven, West Haven, CT, United States.

Please cite this article as:

Manica, A. K. (2023). "Limitations of QSAR Modeling: Data Bias, Curation, and Predictive Reliability in Computational Drug Discovery", Bioinfo Chem, 5(1), 1-13, 10719

© 2023 Bioinfo Chem, a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)
(<https://publishing.emanresearch.org>)

defined, curated, and interpreted.

It is tempting to think of QSAR as a purely computational exercise, driven by descriptors, algorithms, and validation metrics. But this view, while convenient, is incomplete. At its core, QSAR is inseparable from the data it consumes. Each data point represents not just a number, but the outcome of a specific experimental context—conditions, protocols, measurement variability, and sometimes even unreported assumptions. As Alexander Tropsha and colleagues have repeatedly emphasized, the predictive power of QSAR models depends fundamentally on the integrity and meaning of the underlying data (Tropsha, 2010; Fourches et al., 2010). In this sense, QSAR modeling is less about discovering patterns in abstract space and more about reconciling heterogeneous fragments of experimental reality.

This becomes particularly evident when considering dataset curation, which—somewhat paradoxically—remains both the most critical and the least standardized phase of QSAR development. Public chemical and toxicological databases have grown exponentially over the past two decades, fueled by regulatory initiatives and high-throughput screening programs such as Tox21 (Tice et al., 2013). However, the aggregation of such data often introduces inconsistencies: duplicated entries, conflicting measurements, ambiguous chemical representations, and incomplete metadata. The now-familiar principle of “garbage in, garbage out” is not merely rhetorical here—it is a structural limitation. Even subtle inconsistencies, such as how salt forms or tautomers are represented, can shift descriptor calculations and distort similarity relationships, ultimately affecting model predictions (Fourches et al., 2010).

And yet, the challenge is not only technical—it is also conceptual. What, precisely, constitutes a “dataset” in QSAR modeling? When multiple assays, endpoints, or biological systems are combined under a single label, the resulting dataset may lose semantic clarity. For instance, pooling binding affinity measurements with functional assay outcomes—despite their fundamentally different biological meanings—can lead to models that appear statistically robust but are biologically incoherent. This issue of endpoint ambiguity has been noted in environmental and toxicological modeling contexts, where heterogeneous data sources are frequently merged without sufficient harmonization (Raimondo et al., 2010; Saouter et al., 2017).

Closely related to this is the problem of data sparsity. Although millions of chemical substances are registered globally, only a relatively small subset has been thoroughly characterized in terms of toxicity or biological activity (Commission of the European Communities, 2001). This imbalance creates a narrow and unevenly populated chemical space, within which QSAR models must operate. The concept of the applicability domain (AD)—the region of chemical space where predictions are considered reliable—emerges as a necessary constraint, but also as a limitation. Models trained on sparse or clustered datasets may perform well internally yet fail when extrapolated to novel compounds, particularly those with distinct scaffolds or mechanisms of action (Cherkasov et al., 2014).

If dataset definition poses one layer of difficulty, bias introduces another—arguably more insidious—layer. Bias in QSAR datasets is not always immediately visible; it often manifests through subtle distortions in class distribution, sampling strategies, or publication practices. Class imbalance, for example, is a pervasive issue in toxicology datasets, where “active” or “toxic” compounds are often overrepresented relative to inactive ones. This imbalance can inflate performance metrics such as accuracy or AUC, giving a misleading impression of model reliability while masking poor sensitivity toward minority classes (Christley, 2010).

Sampling bias further complicates the picture. Many QSAR models are developed using datasets enriched with specific chemical classes—pesticides, pharmaceuticals, or industrial chemicals—depending on the research focus. While this may improve performance within that domain, it limits generalizability. A model trained predominantly on triazole fungicides, for instance, may struggle to predict the activity of structurally unrelated compounds. In effect, the model learns not the underlying biology, but the idiosyncrasies of its training set.

Publication bias adds yet another dimension. Historically, studies reporting significant or adverse effects have been more likely to be published, leading to an overrepresentation of “positive” findings in the literature (Wandall et al., 2007). In QSAR datasets derived from published sources, this can result in an underreporting of negative or null results, skewing model training and evaluation. Moreover, smaller studies—often with limited statistical power—are more prone to false positives, which can propagate through datasets and into predictive

models (Christley, 2010).

Recognizing these challenges, regulatory bodies and international organizations have attempted to formalize best practices for QSAR development. The Organisation for Economic Co-operation and Development (OECD) established a set of validation principles in 2004, emphasizing the need for defined endpoints, transparent algorithms, clear applicability domains, and robust validation procedures (OECD, 2004). These guidelines have undoubtedly improved the transparency and reproducibility of QSAR models, particularly in regulatory contexts such as chemical risk assessment in the European Union (Kluxen et al., 2021).

However, adherence to these principles does not fully resolve the underlying issues. In practice, many QSAR models still exhibit discrepancies between internal validation metrics and external predictive performance. This gap—sometimes subtle, sometimes striking—often reflects unresolved biases or overfitting to curated datasets. As newer approaches, including machine learning and big data analytics, are integrated into QSAR workflows, the risk of amplifying these biases may increase rather than diminish (Kerner et al., 2021; Cote et al., 2016).

Emerging strategies attempt to address these limitations, though not without their own uncertainties. Read-across methods, for example, aim to infer properties of untested chemicals based on structurally similar compounds, thereby reducing reliance on experimental data (He et al., 2017). Similarly, efforts to develop nano-QSAR models extend traditional frameworks to nanomaterials, introducing new descriptors and challenges (Puzyn et al., 2009). While promising, these approaches depend heavily on the same foundational issue: the quality and representativeness of the underlying data.

In this context, it becomes increasingly clear that QSAR modeling is not merely a technical discipline, but an interpretive one. It requires not only computational rigor, but also a critical awareness of data provenance, experimental variability, and inherent bias. This narrative review, therefore, does not attempt to resolve these challenges outright. Rather, it seeks to examine them—perhaps even to sit with them for a moment—by exploring how dataset definition and bias shape the reliability of QSAR models, and what this means for their future role in drug discovery and regulatory science.

Ultimately, the question is not whether QSAR can predict chemical behavior—it clearly can—but under what conditions, and with what degree of confidence. And that, it seems, depends less on the sophistication of the algorithms and more on the often-overlooked details of the data itself.

This narrative review aims to examine the methodological challenges associated with QSAR dataset definition and to critically evaluate the sources of bias that affect model reliability, while highlighting emerging strategies to improve predictive robustness.

2. Methodology

2.1 Study Design and Conceptual Framework

This study was conducted as a narrative review, designed to critically examine methodological challenges in QSAR modeling, particularly those related to dataset definition, curation, and bias. Unlike systematic reviews that rely on predefined inclusion–exclusion criteria and quantitative synthesis, this approach allows for a more flexible and interpretive exploration of conceptual issues that underpin QSAR reliability. The review is grounded in a framework that prioritizes data integrity, applicability domain, and validation rigor as central determinants of model performance (Tropsha, 2010). Rather than focusing solely on algorithmic advancements, the methodology emphasizes how pre-modeling decisions—such as dataset construction and endpoint selection—shape downstream predictive outcomes.

2.2 Literature Identification and Selection Strategy

Relevant literature was identified through a targeted review of peer-reviewed publications in computational toxicology, cheminformatics, and QSAR modeling. Foundational studies were prioritized to capture the historical and theoretical evolution of QSAR, including early structure–activity relationship frameworks and subsequent methodological refinements (Hansch & Fujita, 1964; Dearden, 2016). Additional sources were selected to reflect contemporary challenges in dataset curation, bias, and validation, particularly those addressing issues of data heterogeneity, experimental variability, and model interpretability (Fourches et al., 2010; Cherkasov et al., 2014).

The selection process emphasized conceptual relevance rather than exhaustive coverage. Studies were included if they addressed at least one of the following domains: (i)

dataset construction and curation practices, (ii) sources of bias in QSAR modeling, or (iii) validation strategies and applicability domain considerations. Regulatory and guideline-oriented literature, such as OECD validation principles, was also incorporated to contextualize best practices in model development (OECD, 2004).

2.3 Thematic Analysis and Synthesis Approach

A qualitative thematic synthesis was employed to organize and interpret the selected literature. Key concepts were grouped into interconnected themes, including dataset definition, descriptor selection, validation strategies, and bias. This approach allowed for the identification of recurring methodological patterns, such as the influence of class imbalance, sampling bias, and publication bias on model outcomes (Christley, 2010; Raimondo et al., 2010).

Rather than extracting quantitative effect sizes, the analysis focused on conceptual relationships between data quality and predictive reliability. Comparative insights were drawn across studies to highlight how variations in dataset composition and preprocessing can lead to divergent modeling outcomes, even when similar algorithms are used. This synthesis also enabled the identification of structural limitations in commonly used validation metrics, particularly when applied without external validation.

2.4 Evaluation of Modeling and Validation Practices

The review further examined widely used QSAR modeling practices, including descriptor generation, model training, and validation procedures. Particular attention was given to the distinction between internal and external validation, as well as the limitations of traditional performance metrics such as R^2 in assessing predictive reliability (Tropsha, 2010; Roy et al., 2016). Emerging validation strategies, including double cross-validation and Y-randomization, were also evaluated for their ability to detect overfitting and chance correlations (De et al., 2022). Additionally, the concept of the applicability domain (AD) was critically assessed as a key constraint on model generalizability. Methods for defining AD boundaries were examined to understand how models determine the limits of reliable prediction.

2.5 Limitations of the Methodological Approach

As a narrative review, this study is inherently subject to interpretive bias, as the selection and synthesis of literature depend on the authors' judgment. While efforts

were made to include influential and representative studies, the absence of a systematic search strategy may limit reproducibility. Furthermore, the review emphasizes conceptual and methodological insights rather than empirical benchmarking, which may restrict its applicability for immediate technical implementation.

Despite these limitations, this methodological approach provides a comprehensive and context-sensitive perspective on QSAR modeling, highlighting the critical role of data quality and bias in shaping predictive reliability and guiding future research directions.

3. Rethinking QSAR Reliability—From Metrics to Meaning

The digital transformation of toxicology and drug discovery has, in many ways, reshaped how we think about evidence. Models now sit where experiments once dominated. Among these, Quantitative Structure–Activity Relationship (QSAR) modeling has emerged not just as a supportive tool, but as something closer to a decision-making framework—guiding regulatory judgments, prioritizing compounds, and, at times, standing in for empirical testing. And yet, there is a quiet tension here. The more we rely on QSAR models, the more we are forced to confront a deceptively simple question: What does it mean for a model to be reliable? At first glance, the answer might seem straightforward—good performance metrics, strong validation scores, and reproducibility. But the reality is less clean, perhaps even a little uncomfortable. Reliability in QSAR cannot be distilled into a single number or captured by a single validation step. It is layered, conditional, and—importantly—deeply dependent on human decisions made long before any algorithm is applied.

3.1 The Human Foundations of an Apparently Computational Field

It is easy to forget, amid the language of algorithms and descriptors, that QSAR modeling is fundamentally rooted in a human insight. The early work of Hansch and Fujita (1964) proposed that chemical structure encodes biological behavior in a predictable way—a remarkably elegant idea that still underpins the field today. Over time, this idea has been expanded, refined, and operationalized through increasingly sophisticated computational methods (Dearden, 2016).

But even now, there is something almost interpretive

about QSAR. We are not simply modeling data—we are translating chemical reality into numerical abstractions, and then asking those abstractions to speak back to us. And in that translation, choices are made. Which descriptors to include, which endpoints to trust, how to clean the data—each decision introduces subtle shifts in meaning.

As Alexander Tropsha (2010) emphasized, QSAR modeling is not just about prediction—it is about understanding the limits of prediction. A model that performs well on its training data but fails when confronted with new compounds is not merely flawed; it reveals something deeper about how knowledge has been encoded, or perhaps misrepresented.

3.2 OECD Principles and the Architecture of Trust

Recognizing the need for structure in what could otherwise become an opaque process, the Organisation for Economic Co-operation and Development (OECD) introduced a set of validation principles intended to standardize QSAR modeling practices (OECD, 2014). These principles—defined endpoints, transparent algorithms, applicability domains, robust validation, and mechanistic interpretation—are often described as the backbone of regulatory trust. And yet, even here, there is nuance. The OECD principles do not guarantee reliability; they provide a framework within which reliability can be pursued. A model may satisfy all five principles and still fail in practice if, for example, the dataset is poorly curated or the applicability domain is overly optimistic.

This is particularly evident in emerging areas such as nano-QSAR modeling, where the complexity of materials introduces additional uncertainty (Li et al., 2022). In such contexts, the OECD principles function less as strict rules and more as guiding constraints—helping to prevent the most obvious pitfalls, but not eliminating the need for critical evaluation.

3.3 The Validation Paradox: Stability Is Not Generalizability

One of the more persistent challenges in QSAR modeling lies in the distinction between internal and external validation. Internal validation—through methods such as cross-validation—assesses how well a model performs on the data it has already seen. It is, in a sense, a measure of stability. But stability is not the same as generalizability.

External validation, by contrast, asks a harder question:

can the model predict outcomes for entirely new compounds? This is where many models begin to falter. High internal performance often gives way to disappointing external results, revealing that the model has captured patterns specific to the training dataset rather than underlying chemical principles (Tropsha, 2010). This tension—the validation paradox—is not simply a technical issue. It reflects a deeper epistemological challenge. Are we building models that learn, or models that merely remember? And how do we distinguish between the two when both can produce impressive metrics? Recent methodological advances, such as double cross-validation, attempt to address this issue by separating feature selection from model evaluation (De et al., 2022). While effective, these approaches also highlight how fragile validation can be when datasets are small or heterogeneous—a common situation in toxicology.

3.4 Beyond R^2 : Rethinking Performance Metrics

For decades, the coefficient of determination (R^2) has served as the default measure of QSAR model performance. Its appeal is obvious—it is simple, interpretable, and widely understood. But simplicity, in this case, can be misleading.

R^2 is sensitive to the distribution of data and can be artificially inflated when response ranges are wide. A model with a high R^2 may still produce large prediction errors for individual compounds. In this sense, R^2 captures correlation, not necessarily accuracy. Alternative metrics, such as Mean Absolute Error (MAE), offer a more grounded perspective by quantifying the average deviation between predicted and observed values. In classification tasks, particularly those involving imbalanced datasets, metrics like the Matthews Correlation Coefficient (MCC) provide a more balanced assessment by accounting for all components of the confusion matrix. The shift toward these metrics reflects a broader realization: no single measure can fully capture model performance. Instead, reliability emerges from a constellation of indicators, each illuminating a different aspect of model behavior.

3.5 The Applicability Domain: The Boundaries of Knowledge

Perhaps one of the most conceptually important, yet often underappreciated, aspects of QSAR modeling is the Applicability Domain (AD). It is, in essence, a statement of humility—a recognition that no model can predict everything. The AD defines the region of chemical space

within which model predictions are considered reliable. Outside this domain, predictions become increasingly speculative. Methods for defining the AD vary, from distance-based approaches to leverage statistics and more recent inhomogeneity mapping techniques (De et al., 2022). What is striking, however, is how often the AD is treated as a secondary consideration, appended to models rather than integrated into their design. A model that does not clearly articulate its boundaries is, in a sense, overconfident. And in regulatory contexts, overconfidence can be more dangerous than uncertainty.

3.6 Data Curation: The Hidden Determinant of Model Quality

If there is one stage of QSAR modeling that quietly determines everything that follows, it is data curation. And yet, it is also the stage that receives the least attention in published studies.

Data curation involves more than correcting errors—it requires standardizing chemical representations, resolving inconsistencies, and ensuring that endpoints are comparable. As Fourches et al. (2010) argued, even small errors in chemical structure can propagate through descriptor calculations, ultimately distorting model predictions. The importance of curation becomes even more pronounced when dealing with large datasets. While the availability of big data has expanded the scope of QSAR modeling, it has also introduced new challenges related to data heterogeneity and noise (Ambure & Cordeiro, 2020). In this context, reliability is not something that can be added after the fact. It must be built into the dataset from the beginning.

3.7 Bias: The Invisible Distortion

Beyond technical challenges, QSAR datasets are often shaped by biases that are difficult to detect but profoundly influential. Class imbalance, for instance, can skew model performance, particularly in toxicological datasets where non-toxic compounds may dominate. Publication bias further complicates the picture. Studies reporting significant effects are more likely to be published, leading to an overrepresentation of positive results. This, in turn, affects the datasets used for QSAR modeling, potentially inflating performance metrics and reducing generalizability. Sampling bias—where certain chemical classes are overrepresented—can also limit the scope of predictions. A model trained on a narrow subset of chemical space may perform well within that domain but

fail when applied more broadly. Addressing these biases requires more than technical fixes. It requires transparency—clear documentation of dataset composition, curation steps, and validation procedures. Without this, even the most sophisticated models risk becoming black boxes.

3.8 Advanced Validation Strategies: Toward More Robust Models

In response to these challenges, a range of advanced validation techniques has been developed. Y-randomization, for example, tests whether model performance arises from genuine structure–activity relationships or from chance correlations. If randomly shuffled datasets produce similar results, the original model may be unreliable.

Consensus modeling, which combines predictions from multiple models, offers another approach to improving reliability by reducing individual model biases. Similarly, emerging methods using graph neural networks and semantic data integration aim to capture more complex relationships within chemical datasets (Romano et al., 2022). These approaches, while promising, do not eliminate the need for critical evaluation. If anything, they underscore the importance of understanding the assumptions and limitations underlying each method.

3.9 Toward Intelligent Validation

Perhaps the most important shift in recent years is the move toward what might be called “intelligent validation.” Rather than relying solely on statistical metrics, this approach emphasizes context—understanding how data were generated, how models were constructed, and where predictions are likely to succeed or fail.

This perspective aligns with broader developments in risk assessment and regulatory science, where there is increasing recognition that model outputs must be interpreted within a framework of uncertainty and evidence integration (Cote et al., 2016). In this sense, QSAR modeling becomes less about producing definitive answers and more about supporting informed decisions. Reliability, then, is not an absolute property but a conditional one—dependent on data quality, methodological rigor, and transparency.

3.10 Reliability as a Continuous Process

As QSAR modeling continues to evolve, the question of

Table 1. Key Public Databases for Nanomaterial and Chemical QSAR Modeling. This table summarizes major publicly available databases supporting QSAR and nano-QSAR modeling, highlighting their domain focus, metadata scope, and regulatory relevance. It provides insight into data curation practices and accessibility, which are critical for reproducible and regulatory-aligned modeling workflows.

Database Name	Target Domain	Metadata Provided	NP Types Covered	Curated Data	Data Sharing Policy	Regulatory Alignment	Primary Reference
NIL	Health/Safety	Physicochemical properties	Broad range	Yes	Public access	Occupational safety	Miller et al., 2007
caNanoLab	Biomedicine	Bio-characterization	Cancer-focused	Yes	Community sharing	Research/validation	Gaheen et al., 2013
DaNa	Human/Environment	Hazard information	Advanced materials	Yes	Knowledge base	Human exposure	Marquardt et al., 2013
NM Registry	Archiving	Physicochemical data	Limited	Yes	Guide-based	Systematic grouping	Mills et al., 2014
eNanoMapper	Nanosafety	Toxicity data	Engineered nanomaterials	Yes	Open framework	REACH support	Jeliazkova et al., 2015
S2NANO	Classification	Toxicity profiles	Metallic nanomaterials	Yes	Web-based	Hazard prediction	Trinh et al., 2018
NanoSolveIT	Nanoinformatics	Structural fingerprints	Diverse nanomaterials	Yes	Multi-source	In silico tools	Afantitis et al., 2020
InterNano	Manufacturing	Government reports	Industrial nanomaterials	No	Information sharing	Manufacturing	Li et al., 2022
NBI	Biological	Biocorona information	Broad nanomaterials	Yes	Unbiased interpretation	Systems biology	Li et al., 2022
NCL	Pre-clinical	Standardized assays	Cancer nanomedicines	Yes	Laboratory-generated	Standard protocols	Li et al., 2022

Table 2. Categories of Molecular Descriptors Used in QSAR Modeling. This table categorizes molecular descriptors used in QSAR modeling based on dimensionality, structural basis, and computational complexity. It highlights their functional roles in capturing chemical information relevant to biological activity and toxicity prediction.

Descriptor Class	Dimension	Structural Basis	Information Captured	Calculation Method	Complexity	Typical Usage	Primary Reference
Constitutional	0D	Molecular formula	Atom counts	Algorithmic	Low	Baseline descriptors	De et al., 2022
Topological	2D	Connectivity	Branching patterns	Matrix-based	Medium	Structure-activity	De et al., 2022
Geometrical	3D	Spatial coordinates	Molecular conformation	Optimization	High	Steric effects	De et al., 2022

Electrostatic	2D/3D	Charge distribution	Polarity/interactions	QM-based	High	Binding affinity	Li et al., 2022
Quantum-chemical	3D	Electronic states	HOMO/LUMO	Schrödinger equation	Very high	Reactivity modeling	Li et al., 2022
Fragment-based	2D	Substructures	Functional groups	Counting	Low	Toxicity alerts	De et al., 2022
Hydrophobic	1D	Partitioning	LogP/solubility	Experimental	Low	Bioavailability	Hansch & Fujita, 1964
PTDs	0D	Periodic table	Atomic features	Mapping	Very low	Nano-QSAR	Kar et al., 2014
Quasi-SMILES	1D	Encoded notation	Hybrid descriptors	Conversion	Medium	Data-gap filling	Li et al., 2022
GETAWAY	3D	Autocorrelation	3D molecular shape	Matrix-based	High	3D-QSAR	De et al., 2022

Table 3. Essential Validation Metrics for Regression QSAR Models. This table outlines key statistical metrics used to evaluate regression-based QSAR models, emphasizing model robustness, predictivity, and error estimation. It also highlights acceptable thresholds and common sources of bias affecting model reliability.

Metric Name	Mathematical Focus	Application	Interpretation	Optimal Value	Quality Threshold	Potential Bias	Primary Reference
R ²	Goodness-of-fit	Training set	Variance explained	1.00	> 0.6	Over-optimistic	Tropsha, 2010
Q ² _LOO	Robustness	Internal cross-validation	Stability measure	1.00	> 0.5	Redundancy	De et al., 2022
RMSE	Average error	External validation	Magnitude of error	0.00	Lower is better	Outlier sensitivity	De et al., 2022
MAE	Absolute error	Predictivity	Real-world error	0.00	< 10% range	Balanced	Roy et al., 2016
CCC	Accuracy & precision	External validation	Reproducibility	1.00	> 0.8	Scale invariance	De et al., 2022
Q ² _ext (F1)	Predictivity	Hold-out set	External validation	1.00	> 0.5	Training mean bias	De et al., 2022
Q ² _ext (F2)	Predictivity	Hold-out set	External validation	1.00	> 0.5	Test mean bias	De et al., 2022
r ² _m (test)	Modified R ²	Penalized fit	Reliability check	1.00	> 0.5	Overfitting detection	Ojha et al., 2011
r ² _m	Stability	Directionality	Difference measure	0.00	< 0.20	Axis dependency	Ojha et al., 2011
Y-scrambling	Chance correlation	Statistical validation	Random fit detection	0.00	R ² _p < 0.5	Random noise	Kar et al., 2014

reliability remains central—and unresolved. The field has made significant progress, from the early days of linear regression to the current landscape of machine learning and big data. Yet, the core challenge persists: how to ensure that models are not only predictive but trustworthy. The answer, it seems, lies not in any single technique or metric, but in a combination of practices—rigorous data curation, thoughtful validation, clear definition of applicability domains, and an ongoing commitment to transparency. Reliability is not a destination. It is a process—one that requires constant reassessment, critical thinking, and, perhaps most importantly, a willingness to question even our most confident models.

4. Reframing QSAR Reliability: Data Integrity, Bias, and the Limits of Predictive Modeling

4.1 Reliability, Bias, and the Quiet Complexity of QSAR Modeling

If QSAR modeling once appeared as a straightforward computational exercise—an elegant mapping of molecular structure to biological activity—it no longer feels quite so simple. The more one looks closely, the more the process reveals itself as layered, sometimes fragile, and deeply dependent on decisions that are not purely mathematical. What emerges from this synthesis is not a critique of QSAR as a discipline, but rather a more careful understanding of its limits—and, perhaps more importantly, its responsibilities.

At the center of this discussion lies a subtle but critical shift: reliability in QSAR is no longer judged solely by how well a model performs, but by how well its assumptions hold under scrutiny. And those assumptions, as it turns out, often begin long before any model is built.

4.2 The Curation Problem: Where Reliability Begins (or Fails)

It is tempting to focus on algorithms—the sophistication of machine learning methods, the elegance of descriptor selection—but the literature repeatedly points elsewhere. The most decisive step in QSAR modeling is often the least visible: data curation. Fourches et al. (2010) made this point forcefully, showing that even minor inconsistencies in chemical structures—incorrect stereochemistry, unresolved tautomers, or improper salt handling—can propagate into entirely misleading models. A summary of key public databases supporting QSAR and nano-QSAR modeling, including their metadata scope,

curation practices, and regulatory relevance, is presented in Table 1.

And yet, despite its importance, curation is often treated as a preliminary step rather than a foundational one. Large public databases, such as the Nanoparticle Information Library and eNanoMapper, have dramatically expanded data availability (Jeliazkova et al., 2015; Miller et al., 2007). But more data, as several studies caution, does not necessarily translate into better models. Without careful standardization, heterogeneity becomes noise rather than insight. Ambure and Cordeiro (2020) emphasize that rigorous curation—removing duplicates, harmonizing endpoints, verifying chemical identities—is not optional. It is, in a sense, the first act of modeling. When this step is neglected, even the most advanced algorithms cannot recover the lost integrity of the dataset.

4.3 Descriptors and Meaning: Translating Chemistry into Numbers

Once data are curated, the next challenge is representation—how to encode chemical structures in a way that preserves meaningful information. QSAR descriptors, ranging from simple atom counts to complex three-dimensional spatial parameters, serve as this bridge. But not all descriptors are equally informative. Earlier approaches relied heavily on simple constitutional descriptors, which, while useful, often lack the nuance required for complex biological interactions (Roy et al., 2015b). Over time, the field has shifted toward topological and 3D descriptors that better capture molecular geometry and steric effects (Dearden, 2016). The major classes of molecular descriptors used in QSAR modeling, categorized by dimensionality, structural basis, and computational complexity, are outlined in Table 2.

In specialized domains such as nano-QSAR, even these descriptors may fall short. Periodic Table–based descriptors (PTDs), for example, have been developed to account for the unique electronic and physicochemical properties of nanoparticles (Kar et al., 2014). This progression suggests an important insight: descriptor choice is not merely technical—it is conceptual. A model is only as interpretable as the features it uses to describe reality.

4.4 The Illusion of Performance: Why Metrics Can Mislead

Perhaps one of the more uncomfortable realizations in QSAR research is that widely used performance metrics

can, at times, be misleading. The coefficient of determination (R^2), long regarded as a hallmark of model quality, has come under increasing scrutiny. While it measures how well a model fits its training data, it says little about how well the model will perform on new, unseen compounds (Tropsha, 2010). This distinction—between fitting and predicting—is not trivial. Models with high R^2 values can still produce significant prediction errors, particularly when applied outside their training domain. As Roy et al. (2016) demonstrated, reliance on R^2 alone can create a false sense of confidence. The principal statistical metrics used to evaluate regression-based QSAR models, along with their interpretations, thresholds, and associated biases, are summarized in Table 3.

To address this, researchers have advocated for more robust metrics such as Mean Absolute Error (MAE) and Concordance Correlation Coefficient (CCC), which provide a clearer picture of predictive accuracy. In classification problems, particularly those involving imbalanced datasets, the Matthews Correlation Coefficient (MCC) has emerged as a preferred metric due to its balanced consideration of true and false predictions (Chicco, 2017). These shifts reflect a broader movement toward what might be called context-aware validation—recognizing that no single metric can capture the full complexity of model performance.

4.5 Bias and Imbalance: The Hidden Distortions

Even with well-curated data and appropriate metrics, QSAR models remain vulnerable to bias. Class imbalance is perhaps the most obvious example. In toxicological datasets, non-toxic compounds often outnumber toxic ones, creating a skew that can artificially inflate model accuracy. But bias is not limited to class distribution. Publication bias—where studies reporting significant or adverse findings are more likely to be published—can distort the underlying dataset, leading models to learn from an incomplete representation of chemical space. Similarly, sampling bias can arise when certain chemical classes are overrepresented, limiting the model's generalizability.

Christley (2010) highlights another dimension: the risks associated with small datasets. Underpowered studies are more prone to false positives, which can be incorporated into QSAR models if not carefully filtered. This “law of small numbers” is particularly relevant in niche areas of

toxicology, where data scarcity remains a persistent challenge. Addressing these biases requires more than statistical adjustments. It requires transparency—clear documentation of dataset composition, curation protocols, and validation procedures. A comprehensive overview of classification performance metrics, including their calculation logic, interpretative value, and limitations under class imbalance, is provided in Table 4.

4.6 Applicability Domain: Defining the Limits of Prediction

One of the most conceptually important aspects of QSAR modeling is the Applicability Domain (AD). It is, essentially, an acknowledgment that no model can predict everything. Each QSAR model is trained within a specific region of chemical space, and its reliability is confined to that region.

Methods such as the Williams plot provide a way to visualize this domain, identifying compounds that lie outside the model's reliable range (Sahigara et al., 2012). Yet, despite its importance, the AD is often underemphasized in practice. The OECD guidelines explicitly require a defined applicability domain (OECD, 2014), but implementation varies widely. Some models provide only vague descriptions, while others integrate AD analysis more rigorously. What becomes clear is that a model's ability to say “I don't know” may be as important as its ability to make accurate predictions.

4.7 Validation as a Process, not a Step

Another recurring theme in the literature is the idea that validation should not be treated as a single step, but as an ongoing process. Internal validation techniques, such as cross-validation, assess model stability, but they cannot guarantee predictive performance. External validation—testing the model on independent datasets—is widely regarded as the gold standard (Tropsha, 2010; Roy et al., 2016). More advanced techniques, such as Y-randomization, provide additional safeguards by testing whether observed relationships are genuine or the result of chance correlations (Kar et al., 2014). Double cross-validation further strengthens model evaluation by separating feature selection from model training.

These approaches collectively suggest that validation is not merely about confirming model performance, but about challenging it—probing its assumptions and identifying its weaknesses.

4.8 Toward a More Reflective QSAR Practice

Table 4. Essential Validation Metrics for Classification QSAR Models. This table presents commonly used metrics for evaluating classification-based QSAR models, including sensitivity, specificity, and balanced performance indicators. It emphasizes limitations such as class imbalance and the importance of using multiple complementary metrics.

Metric Name	Focus Area	Calculation Logic	Interpretation	Range	Balanced	Known Limitation	Primary Reference
Sensitivity	Potency	$TP / (TP + FN)$	True positive rate	0–1	No	Ignores negatives	De et al., 2022
Specificity	Safety	$TN / (TN + FP)$	True negative rate	0–1	No	Ignores positives	De et al., 2022
Accuracy	Overall correctness	$(TP + TN) / \text{Total}$	Global performance	0–1	No	Class imbalance	De et al., 2022
Precision	Correctness	$TP / (TP + FP)$	Prediction confidence	0–1	No	False positives	De et al., 2022
F-measure	Harmonic mean	$2 / (1/P + 1/S)$	Balance of precision/recall	0–1	No	Skewed data	De et al., 2022
G-means	Geometric mean	$(Sn \times Sp)$	Balanced performance	0–1	Yes	Geometric bias	De et al., 2022
Kappa ()	Agreement	Observed vs chance	Concordance	-1 to 1	Yes	Chance dependency	Cohen, 1960
MCC	Correlation	Confusion matrix	Overall quality	-1 to 1	Yes	None (robust metric)	De et al., 2022
False Positive Rate	Risk	$FP / (TN + FP)$	Type I error	0–1	No	Misidentification	De et al., 2022
False Negative Rate	Risk	$FN / (TP + FN)$	Type II error	0–1	No	Missed toxicity	De et al., 2022

What emerges from this discussion is a more reflective view of QSAR modeling. It is not simply a computational pipeline, but a process shaped by human judgment—by decisions about data, representation, validation, and interpretation. Reliability, in this context, becomes less about achieving perfect metrics and more about understanding limitations. A model that performs modestly but transparently may be more valuable than one that appears highly accurate but obscures its assumptions. QSAR modeling continues to play a vital role in modern toxicology and drug discovery. But as its influence grows, so too does the need for careful evaluation. The path to reliable QSAR models does not lie in increasingly complex algorithms alone, but in disciplined practices—rigorous data curation, thoughtful descriptor selection, robust validation, and clear definition of applicability domains.

In the end, reliability is not a static property. It is something that must be continually earned—through transparency, critical thinking, and a willingness to question even our most confident predictions.

5. Limitations

This review is inherently constrained by its narrative design, which, while allowing for conceptual depth, does not provide the quantitative synthesis typical of systematic reviews or meta-analyses. The interpretation of QSAR challenges is therefore shaped by selected literature and may not fully capture all emerging methodological developments. Additionally, the discussion emphasizes conceptual and methodological issues—such as dataset definition and bias—without presenting empirical benchmarking across specific datasets or algorithms. This may limit direct applicability for practitioners seeking immediate technical solutions. Furthermore, the rapidly evolving nature of machine learning and data-driven modeling means that some emerging techniques may not be comprehensively addressed. Finally, the review relies on published literature, which itself may be influenced by publication bias, potentially reinforcing some of the very distortions discussed. These limitations highlight the need for complementary empirical and systematic investigations to strengthen the conclusions presented here.

6. Conclusion

QSAR modeling remains an indispensable tool in modern chemical and toxicological research, yet its reliability is far from absolute. As this review suggests, predictive performance is deeply intertwined with how datasets are

constructed, curated, and interpreted. Metrics alone cannot guarantee validity, particularly when underlying biases and data inconsistencies persist. Moving forward, the field must adopt a more reflective approach—one that prioritizes transparency, rigorous validation, and clear boundaries of applicability. Reliability, in this sense, is not achieved through algorithmic complexity alone but through disciplined methodological practice and continuous critical evaluation of both data and models.

Author Contributions

A.K.M. conceptualized the study, designed the review framework, conducted literature analysis and synthesis, interpreted the findings, and drafted, reviewed, and finalized the manuscript.

References

- Afantitis, A., et al. (2020). NanoSolveIT project: Driving nanoinformatics research to develop innovative and integrated tools for in silico nanosafety assessment. *Computational and Structural Biotechnology Journal*, 18, 583–602. <https://doi.org/10.1016/j.csbj.2020.02.023>
- Ambure, P., & Cordeiro, M. N. D. S. (2020). Importance of data curation in QSAR studies especially while modeling large-size datasets. In K. Roy (Ed.), *Ecotoxicological QSARs* (pp. 97–109). Springer. https://doi.org/10.1007/978-1-0716-0150-1_5
- Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R. D., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., & Tropsha, A. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, 57(12), 4977–5010. <https://doi.org/10.1021/jm4004285>
- Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(35), 1–17. <https://doi.org/10.1186/s13040-017-0155-3>
- Christley, R. M. (2010). Power and error: Increased risk of false positive results in underpowered studies. *Open Epidemiology Journal*, 3, 16–19. <https://doi.org/10.2174/1874297101003010016>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Commission of the European Communities. (2001). White paper on a strategy for a future chemicals policy. European Commission.
- Cote, I., Andersen, M. E., Ankley, G. T., Barone, S., Birnbaum, L. S., Boekelheide, K., DeWoskin, R. S., Hays, S. M., Judson, R., Portier, C. J., Smith, M. T., & Yauk, C. L. (2016). The next generation of risk assessment multiyear study—Highlights of findings, applications to risk assessment, and future directions. *Environmental Health Perspectives*, 124(11), 1671–1682. <https://doi.org/10.1289/EHP233>
- De, P., Kar, S., Ambure, P., & Roy, K. (2022). Prediction reliability of QSAR models: An overview of various validation tools. *Archives of Toxicology*, 96, 1279–1295.

- <https://doi.org/10.1007/s00204-022-03252-y>
- Dearden, J. C. (2016). The history and development of quantitative structure–activity relationships (QSARs). *International Journal of Quantitative Structure-Property Relationships*, 1(1), 1–44. <https://doi.org/10.4018/IJQSPR.2016010101>
- Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, 50(7), 1189–1204. <https://doi.org/10.1021/ci100176x>
- Gaheen, S., et al. (2013). caNanoLab: Data sharing to expedite the use of nanotechnology in biomedicine. *Computational Science & Discovery*, 6, 014010. <https://doi.org/10.1088/1749-4699/6/1/014010>
- Hansch, C., & Fujita, T. (1964). -- analysis: A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86(8), 1616–1626. <https://doi.org/10.1021/ja01062a035>
- He, J., et al. (2017). The combined QSAR-ICE models: Practical application in ecological risk assessment and water quality criteria. *Environmental Science & Technology*, 51(16), 8877–8878. <https://doi.org/10.1021/acs.est.7b02736>
- Jeliazkova, N., et al. (2015). The eNanoMapper database for nanomaterial safety information. *Beilstein Journal of Nanotechnology*, 6, 1609–1634. <https://doi.org/10.3762/bjnano.6.165>
- Kar, S., et al. (2014). Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: A mechanistic QSTR approach. *Ecotoxicology and Environmental Safety*, 107, 162–169. <https://doi.org/10.1016/j.ecoenv.2014.05.026>
- Kerner, J., et al. (2021). Machine learning and big data provide crucial insight for future biomaterials discovery and research. *Acta Biomaterialia*, 130, 54–65. <https://doi.org/10.1016/j.actbio.2021.05.053>
- Kluxen, F. M., Felkers, E., Baumann, J., Morgan, N., Wiemann, C., Stauber, F., & Kuster, C. J. (2021). Compounded conservatism in European re-entry worker risk assessment of pesticides. *Regulatory Toxicology and Pharmacology*, 121, 104864. <https://doi.org/10.1016/j.yrtph.2021.104864>
- Li, J., et al. (2022). Nano-QSAR modeling for predicting the cytotoxicity of metallic and metal oxide nanoparticles: A review. *Ecotoxicology and Environmental Safety*, 243, 113955. <https://doi.org/10.1016/j.ecoenv.2022.113955>
- Marquardt, C., et al. (2013). Latest research results on the effects of nanomaterials on humans and the environment: DaNa—Knowledge Base Nanomaterials. *Journal of Physics: Conference Series*, 429, 012060. <https://doi.org/10.1088/1742-6596/429/1/012060>
- Miller, A. L., et al. (2007). The Nanoparticle Information Library (NIL): A prototype for linking and sharing emerging data. *Journal of Occupational and Environmental Hygiene*, 4, D131–D134. <https://doi.org/10.1080/15459620701683947>
- Mills, K. C., et al. (2014). Nanomaterial registry: A database that captures minimal information about nanomaterial physicochemical characteristics. *Journal of Nanoparticle Research*, 16, 2219. <https://doi.org/10.1007/s11051-013-2219-8>
- Ojha, P. K., Mitra, I., Das, R. N., & Roy, K. (2011). Further exploring rm² metrics for validation of QSPR models. *Chemometrics and Intelligent Laboratory Systems*, 107(1), 194–205. <https://doi.org/10.1016/j.chemolab.2011.03.011>
- Organisation for Economic Co-operation and Development (OECD). (2004). The report from the expert group on (quantitative) structure–activity relationships [(Q)SARs] on the principles for the validation of (Q)SARs. OECD Publishing.
- Organisation for Economic Co-operation and Development (OECD). (2014). Guidance document on the validation of (quantitative) structure–activity relationship [(Q)SAR] models. OECD Publishing.
- Puzyn, T., et al. (2009). Toward the development of nano-QSARs: Advances and challenges. *Small*, 5(22), 2494–2509. <https://doi.org/10.1002/sml.200900179>
- Raimondo, S., Jackson, C. R., & Barron, M. G. (2010). Influence of taxonomic relatedness and chemical mode of action in acute interspecies estimation models for aquatic species. *Environmental Science & Technology*, 44(19), 7711–7716. <https://doi.org/10.1021/es101630b>
- Romano, J. D., Hao, Y., & Moore, J. H. (2022). Improving QSAR modeling for predictive toxicology using publicly aggregated semantic graph data and graph neural networks. *Pacific Symposium on Biocomputing*, 27, 187–198. https://doi.org/10.1142/9789811250477_0018
- Roy, K., Das, R. N., Ambure, P., & Aher, R. B. (2016). Be aware of error measures: Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 152, 18–33. <https://doi.org/10.1016/j.chemolab.2016.01.008>
- Roy, K., Kar, S., & Das, R. N. (2015). A primer on QSAR/QSPR modeling. Springer. <https://doi.org/10.1007/978-3-319-17281-1>
- Sahigara, F., et al. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, 17, 4791–4810. <https://doi.org/10.3390/molecules17054791>
- Saouter, E., et al. (2017). Improving substance information in USEtox®, part 2: Data for estimating fate and ecosystem exposure factors. *Environmental Toxicology and Chemistry*, 36(12), 3463–3470. <https://doi.org/10.1002/etc.3903>
- Tice, R. R., Austin, C. P., Kavlock, R. J., & Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: A Tox21 update. *Environmental Health Perspectives*, 121(7), 756–765. <https://doi.org/10.1289/ehp.1205784>
- Trinh, T. X., et al. (2018). Dataset curation and nanoSAR model development for metallic nanoparticles. *Environmental Science: Nano*, 5, 1902–1910. <https://doi.org/10.1039/C8EN00061A>
- Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, 29(6–7), 476–488. <https://doi.org/10.1002/minf.201000061>
- Walter, M., Allen, L. N., de la Vega de León, A., Webb, S. J., & Gillet, V. J. (2022). Analysis of the benefits of imputation models over traditional QSAR models for toxicity prediction. *Journal of Cheminformatics*, 14, 32. <https://doi.org/10.1186/s13321-022-00611-w>
- Wandall, B., Hansson, S. O., & Rudén, C. (2007). Bias in toxicology. *Archives of Toxicology*, 81(9), 605–617. <https://doi.org/10.1007/s00204-007-0194-5>