



Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Explainable and Uncertainty-Aware QSAR for Pharmaceutical Decision Making

Aisha Bello^{1*}, Zainab Sule¹, Ibrahim Musa², Grace Adeyemi³, Ahmed Youssef⁴

¹Department of Explainable QSAR and Uncertainty, Faculty of Pharmacy, University of Ibadan, Ibadan, Nigeria.

²Department of Pharmaceutical Decision Making, Faculty of Pharmacy, University of Ilorin, Ilorin, Nigeria.

³Department of Uncertainty-Aware Modeling, Faculty of Pharmaceutical Sciences, Federal University of Agriculture Abeokuta, Abeokuta, Nigeria.

⁴Department of Drug Discovery Informatics, Faculty of Pharmacy, University of Khartoum, Khartoum, Sudan.

*Email: aisha.bello@ui.edu.ng

ABSTRACT

Quantitative structure–activity relationship (QSAR) models increasingly support pharmaceutical prioritization, yet predictive accuracy alone does not establish whether a model is sufficiently interpretable or reliable for a specific decision. Explainability methods may identify molecular features associated with predictions, whereas uncertainty methods estimate confidence, calibration, or susceptibility to error; these functions are related but not interchangeable. A systematic review of literature published from 2017 to September 2026 was conducted across six web-accessible scholarly search and publication sources using combinations of QSAR, molecular-property prediction, ADMET, explainability, attribution, counterfactual explanation, uncertainty quantification, calibration, conformal prediction, chemical-space generalization, and pharmaceutical decision terms. The review flow comprised 1,311 identified records, 342 retained after title screening, 87 retained after abstract assessment, and 29 finally included studies. Evidence was classified according to prediction context, explanation target, uncertainty formulation, validation strategy, chemical-space boundary, and relationship to pharmaceutical decision making. The literature showed that explanation quality cannot be inferred from predictive accuracy and that explanation outputs can vary with representation, model architecture, structural attribution unit, and benchmark design. Uncertainty estimates likewise differed across model classes, data distributions, activity-cliff regions, and calibration procedures. Counterfactual and contrastive approaches offered chemically explicit interrogation of prediction boundaries, but predicted structural alternatives did not constitute experimental causal evidence. Few studies directly connected joint explainability–uncertainty assessment to prospective pharmaceutical decisions. QSAR outputs become more defensible for pharmaceutical decision making when prediction, explanation, uncertainty, validation, and transferability are treated as distinct evidentiary layers. The review therefore develops a decision-oriented synthesis in which explanations require independent fidelity testing and uncertainty requires calibration within the chemical and decision domain in which it is used.

Keywords: QSAR, Explainable artificial intelligence, Uncertainty quantification, Molecular property prediction, Conformal prediction, Pharmaceutical decision making

Received: 20 May 2026; **Revised:** 23 August 2026; **Accepted:** 26 August 2026

INTRODUCTION

Machine-learning QSAR models can prioritize compounds, estimate molecular properties, and reduce the experimental space considered during drug discovery, but a high-performing prediction is not automatically an interpretable or decision-ready prediction. Explainable artificial intelligence (XAI) has consequently become an

important component of computational drug discovery because it attempts to expose which molecular features or representations influenced a model output [1]. The practical issue is not simply whether an explanation can be produced, but whether that explanation remains meaningful when the model encounters the chemical and data conditions in which pharmaceutical decisions are made.

Recent evaluations of ADMET prediction under data scarcity, distribution shift, class imbalance, beyond-rule-of-five chemistry, and difficult structure–activity relationships indicate that apparent model reliability is strongly conditional on the evaluation environment [2]. These findings challenge the use of aggregate benchmark accuracy as a sufficient basis for compound prioritization. A model can perform well on an average test set while remaining unreliable in precisely the local chemical region containing the compounds under consideration.

Explanation quality constitutes a separate problem. Comparative evaluation of molecular XAI approaches has shown that models with useful predictive performance may nevertheless differ substantially in the structural patterns their explanations recover, their behavior around activity cliffs, and their sensitivity to representation [3]. This review therefore treats prediction, explanation, uncertainty, and decision utility as analytically distinct. Its central synthesis is that explainable and uncertainty-aware QSAR should not produce a single generalized “trust” score. Instead, each output should pass the validation test appropriate to the specific pharmaceutical inference being made.

MATERIALS AND METHODS

The review was designed as a systematic methodological synthesis of explainability and uncertainty in QSAR and closely related molecular-property prediction. Particular attention was given to studies capable of supporting distinctions among model performance, explanation fidelity, uncertainty calibration, chemical-space generalization, and pharmaceutical decision consequences. Because QSAR claims depend on how endpoints, datasets, preprocessing procedures, exclusions, and validation splits are defined, study classification treated those choices as part of the inferential context rather than as administrative metadata [4].

Eligible evidence included peer-reviewed studies published from 2017 through September 2026 that evaluated or materially developed QSAR, ADMET prediction, molecular-property prediction, chemical XAI, molecular attribution, counterfactual explanation, uncertainty quantification, calibration, conformal prediction, or related reliability methods. Methodological papers were retained when their contribution directly informed how a pharmaceutical QSAR result could be interpreted or qualified.

Search strategy

Searches were conducted through six web-accessible scholarly search, indexing, DOI, and publisher interfaces. Search concepts combined terms for QSAR, quantitative structure–activity relationship, molecular property prediction, or ADMET with terms for explainability, interpretability, attribution, counterfactual explanation, uncertainty, calibration, prediction interval, conformal prediction, distribution shift, and chemical space.

Search design deliberately captured literature concerned with dataset construction and consistency because merging nominally comparable property records does not necessarily eliminate cross-source or assay inconsistency [5]. Records were screened first by title, followed by abstract assessment and final eligibility review. Of 1,311 initially identified records, 342 were retained after title screening, 87 after abstract assessment, and 29 met final inclusion criteria.

Eligibility and study classification

Eligibility required more than the presence of an AI or machine-learning model. Studies had to contribute directly to at least one of five review functions: explanation generation, explanation validation, uncertainty estimation, uncertainty calibration or localization, or translation of these outputs into an interpretable molecular or compound-selection decision. Software and workflow papers were eligible when they demonstrated how explainability or uncertainty could be embedded within an operational QSAR process rather than merely mentioned as optional functionality [6].

The review classified studies by their principal inferential role. Predictive-only studies estimated an endpoint without formally evaluating explanation or uncertainty. Explanation-evaluated studies tested whether molecular attributions corresponded to known structural patterns, chemically meaningful perturbations, or other explicit benchmarks. Uncertainty-evaluated studies assessed calibration, error ranking, prediction intervals, or sensitivity to distribution shift. Joint studies connected explanatory localization to uncertainty. Decision-linked studies went further by using these outputs to modify compound selection, screening, or prioritization.

This classification prevents a common category error: an explanation visualized beside a prediction does not demonstrate explanation fidelity, and an uncertainty value emitted by a model does not demonstrate calibration. The principal eligibility failure patterns and their consequences are summarized below because they are difficult to distinguish consistently in narrative form alone.

Table 1. Failure patterns that separate superficially explainable QSAR studies from decision-relevant evidence

Misleading eligibility pattern	Likely origin	Observable signature	Discriminating check	Residual decision consequence	Classification in this review
Explanation shown without explanation testing	Post-hoc visualization treated as validation	Molecular heat map or feature ranking accompanies performance metrics	Test against known structural perturbation, activity cliff, or synthetic ground truth	Structural interpretation remains provisional	Explanation-generating, not explanation-validated
Uncertainty score reported without calibration	Model variance interpreted directly as confidence	Confidence values are presented without coverage or error relationship	Compare uncertainty with observed errors or interval coverage	Ranking confidence may be misleading	Uncertainty-generating, not calibrated
Random split used as the only reliability test	Training and test compounds remain chemically similar	Strong average performance with no domain-shift analysis	Scaffold, temporal, cluster, or explicit distribution-shift evaluation	Transfer to novel chemistry remains unresolved	In-domain predictive evidence
Large integrated dataset treated as automatically superior	Data volume prioritized over assay consistency	Multiple sources merged after structural standardization alone	Inspect endpoint definitions, assay conditions, provenance, and conflicts	Model may learn source heterogeneity	Data-integration evidence with unresolved consistency
Explainability and uncertainty reduced to one confidence label	Distinct inferential functions are collapsed	A single score represents importance and reliability	Evaluate attribution fidelity and calibration separately	Decision rationale becomes ambiguous	Joint claim unsupported unless both components are tested
Model output directly equated with compound actionability	Prediction substituted for experimental verification	Recommended analogue or structural change has no experimental follow-up	Require chemical feasibility and prospective testing	Candidate remains a computational hypothesis	Decision proposal rather than validated decision evidence

Explainability taxonomy

Chemical XAI includes feature-attribution, fragment-based, graph-based, representation-based, example-based, and counterfactual approaches. These methods answer different questions and therefore should not be ranked as though they produced interchangeable explanations [7]. A local atomic attribution may indicate which graph features drove a prediction, whereas an R-group analysis can be more directly aligned with medicinal-chemistry reasoning. Neither automatically establishes mechanism.

The first requirement is fidelity testing. Synthetic QSAR benchmarks in which the generating structural pattern is known provide one way to test whether an interpretation method recovers information actually used by the model [8]. Such benchmarks are valuable because they replace visual plausibility with a falsifiable target, although recovery of a synthetic rule does not guarantee validity for heterogeneous biological SAR.

The second requirement is representation stability. Equivalent molecular representations should not produce arbitrarily different explanations if the interpretation is intended to describe the underlying molecular decision rather than an incidental encoding. Test-time augmentation studies have demonstrated substantial inconsistencies among explanation methods when equivalent molecular representations are supplied to the same predictive context [9]. Representation sensitivity should therefore be reported as an explanation-quality property rather than hidden by displaying a single canonical depiction.

A third requirement is sensitivity to difficult local SAR. Activity cliffs provide an especially stringent test because small structural changes are associated with large changes in measured activity. Benchmarking against experimentally defined activity cliffs showed that attribution approaches differed markedly in their ability to highlight relevant structural changes [10]. Consequently, smooth global performance cannot substitute for local explanation testing around chemically consequential discontinuities.

The explanatory unit also matters. Neural-network internal representations can be inspected for learned chemical features associated with toxicity predictions [11]. MolSHAP instead expresses model contributions at the level of R-groups, offering a form more naturally aligned with analogue-series reasoning [12]. Explanation-aware training

can additionally alter the behavior of graph models rather than treating explainability exclusively as post-hoc analysis [13].

Finally, explainability can be combined with uncertainty localization. Feature-level approaches have demonstrated that molecular regions may be both influential and associated with high model uncertainty [14]. The two quantities nevertheless answer different questions: attribution estimates which model inputs influence an output, whereas uncertainty concerns the reliability of that output or contribution. Their combination is therefore potentially informative precisely because they should remain distinguishable.

Uncertainty taxonomy

QSAR uncertainty is not a single property. It can arise from noisy measurements, sparse sampling, model parameter uncertainty, representation limitations, distribution shift, local chemical discontinuities, or the statistical procedure used to translate predictions into intervals or confidence sets. Benchmarking of scalable uncertainty methods for deep molecular prediction has shown material differences among ensembles, bootstrapping, dropout-based procedures, and related approaches [15]. Comparative evaluation across molecular regression tasks similarly found no uncertainty method that was unequivocally superior in every setting [16].

A useful taxonomy therefore separates aleatoric uncertainty, associated with irreducible or measurement-related variation, from epistemic uncertainty, associated with incomplete model knowledge, while adding two practically important layers: distributional uncertainty, arising when the query compound differs from the training domain, and decision uncertainty, arising when statistical confidence remains insufficient to justify a particular pharmaceutical action.

Uncertainty can also be localized. Explainable uncertainty methods have decomposed model confidence into molecular contributions, allowing investigators to determine whether uncertainty is diffuse across a structure or concentrated in specific features [17]. Such localization may help identify where a prediction requires additional evidence, but an uncertain atom or fragment should not be interpreted as a proven causal liability.

Distribution shift creates a particularly important boundary. UQ performance changes when evaluation data depart from the training distribution, and methods that appear reliable in-domain may fail to preserve the same relationship between uncertainty and prediction error under shifted conditions [18]. Difficult SAR regions provide a related challenge: steep local response surfaces and activity cliffs can weaken the correspondence between nominal confidence and actual error [19].

Conformal prediction represents a distinct uncertainty strategy because it constructs prediction regions or sets with explicit statistical coverage properties under specified assumptions [20]. Conformalized graph-learning approaches have extended this principle to ADMET prediction and demonstrated calibrated interval construction across molecular tasks [21]. More recent local approaches indicate that whole-molecule uncertainty can conceal specific structural regions associated with model failure [22]. These approaches extend the resolution of uncertainty assessment, but none removes the need to establish whether the statistical guarantee remains appropriate for the chemical and decision domain being considered.

Quality and bias assessment

Quality assessment focused on threats capable of making an apparently interpretable or uncertainty-aware QSAR result misleading. These included endpoint ambiguity, inconsistent data provenance, information leakage, chemically weak train–test separation, explanation instability, lack of explanation ground truth, inappropriate calibration assessment, distribution shift, and confusion between structural attribution and biological mechanism. Out-of-distribution performance was treated separately from ordinary test-set performance because explicit methods for invariant representation learning address a different problem from interpolation within familiar chemical space [23]. Likewise, multi-representation studies show that molecular encoding can influence both predictive behavior and the form of explanation produced [24]. A method was therefore not judged robust merely because a single representation, split, or aggregate metric appeared favorable.

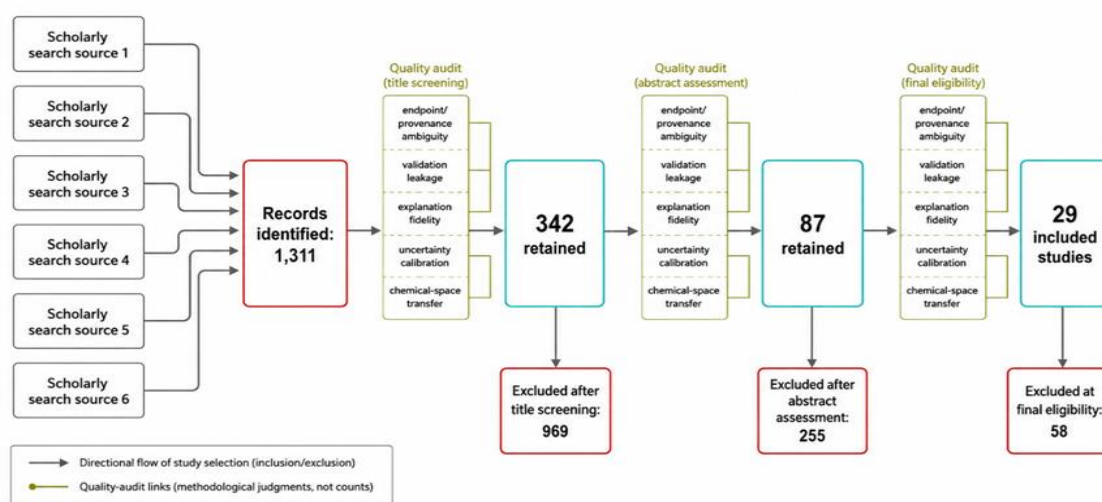
The review did not construct an unsupported numerical risk-of-bias score. Instead, quality was assessed by asking whether each study included a discriminator capable of falsifying its central interpretive claim. For explanations, such discriminators included synthetic ground truth, activity cliffs, equivalent-representation perturbations, substructure tests, or chemically meaningful analogue comparisons. For uncertainty, they included calibration, prediction-error association, shifted-domain testing, or explicit coverage assessment.

The main quality states are shown below because several can generate deceptively similar model outputs while requiring different corrective tests.

Table 2. Diagnostic triage of quality failures in explainable and uncertainty-aware molecular prediction

Failure state	What it can masquerade as	Chemical or model signature	Prospective check	Residual uncertainty after mitigation	Review interpretation
Endpoint or provenance inconsistency	Large, information-rich training set	Similar structures carry incompatible labels across sources	Reconcile assay definitions and conflicting records before modeling	Biological heterogeneity may remain	Data-quality limitation
Leakage across related compounds	Exceptional generalization	Near-identical analogues occur across training and test partitions	Re-split by scaffold, cluster, series, or time	Novel-scaffold performance remains unknown	Validation-design bias
Representation-sensitive explanation	Stable structural rationale	Equivalent encodings produce different highlighted regions	Repeat explanation across equivalent molecular representations	Interpreter may remain model dependent	Explanation-instability warning
Activity-cliff blindness	Smooth, apparently coherent SAR explanation	Attribution changes little despite large measured activity discontinuity	Evaluate matched cliff pairs	Exact cliff mechanism may remain unresolved	Local explanation failure
In-domain uncertainty calibration only	General confidence estimate	Good calibration on familiar chemistry but poor shifted-domain behavior	Explicit distribution-shift or scaffold test	Future chemical space still uncertain	Transferability limitation
Whole-molecule averaging	Apparently moderate uncertainty	High-risk local feature is diluted by low uncertainty elsewhere	Atom-, fragment-, or substructure-level uncertainty analysis	Local score remains model derived	Resolution limitation
Attribution interpreted mechanistically	Mechanistic molecular explanation	Influential model feature is treated as causal biology	Compare with independent experimental or mechanistic evidence	Causation remains unproven without validation	Inference-level overreach

The study-selection pathway and its relationship to the quality audit are important because the final evidence base reflects sequential exclusion rather than an arbitrary 29-paper quota.

**Figure 1.** Evidence contraction across successive QSAR relevance and quality gates

RESULTS AND DISCUSSION

The final evidence set contained 29 studies after sequential reduction of the initial 1,311 records. The evidence was methodologically heterogeneous rather than representing a single standardized QSAR paradigm. Some studies evaluated whether explanations recovered known molecular patterns; others investigated explanation

stability under changes in representation or chemical context. A second group focused on uncertainty estimation, calibration, conformal prediction, or out-of-distribution behavior. A smaller subset directly combined uncertainty localization with molecular attribution.

Across these categories, three recurring findings were evident. First, explanation quality was not recoverable from predictive performance alone. Second, uncertainty estimates were strongly conditioned by evaluation design and chemical-space relationship. Third, studies directly evaluating whether explainability or uncertainty improved prospective compound decisions were less common than studies evaluating retrospective model behavior. For this reason, the review treats explanation fidelity, uncertainty calibration, and decision utility as different evidentiary outcomes rather than interchangeable indicators of model trustworthiness.

Attribution methods

Attribution methods attempt to identify the molecular features that contribute most strongly to a prediction. In pharmaceutical QSAR, their usefulness depends on the unit at which the result is expressed. Atom-level maps can provide detailed local information, but fragments, R-groups, matched molecular transformations, or common substructures may correspond more naturally to the decisions medicinal chemists actually make.

The evidence reviewed here indicates that attribution should be judged by at least three criteria: fidelity, meaning whether the highlighted features reflect behavior of the predictive model; stability, meaning whether the explanation remains coherent under chemically equivalent representations or modest perturbations; and chemical relevance, meaning whether the explanatory unit corresponds to a meaningful structural decision. Activity-cliff benchmarking is valuable because it evaluates explanation behavior at locations where apparently minor structural changes accompany major activity differences rather than only within smooth regions of chemical space.

Attribution nevertheless remains model interpretation. A highlighted atom, fragment, or R-group indicates dependence of a model output on its representation and training history; it does not demonstrate that the corresponding structural feature is the biological mechanism underlying the observed endpoint. The appropriate role of attribution in pharmaceutical decision making is therefore diagnostic and hypothesis-generating unless supported by independent experimental or mechanistic evidence.

Counterfactual explanations

Counterfactual explanations answer a different question from conventional attribution: rather than asking which features contributed to the current prediction, they ask what plausible structural modification would change the model's output. Model-agnostic molecular counterfactual generation has been demonstrated across random-forest, sequence-based, and graph-based predictive architectures [25]. This makes counterfactual reasoning attractive for medicinal chemistry because the explanation can be expressed as an alternative molecular structure rather than only as an abstract feature score.

Contrastive explanation methods extend this logic by generating chemically interpretable virtual analogues that move a prediction toward an alternative class or selectivity profile. MolCE, for example, demonstrated how building-block substitutions could be used to construct contrastive explanations in a D2-like dopamine-receptor selectivity setting [26]. Such examples are useful because they expose where a model believes its decision boundary lies in structural space.

However, three boundaries must remain explicit. First, a model-valid counterfactual need not be synthetically feasible. Second, a feasible analogue need not possess the predicted experimental activity. Third, changing a model prediction after a structural modification is not equivalent to identifying a causal molecular mechanism. Counterfactuals are therefore most defensible when treated as candidate decision probes that require chemistry, applicability-domain, uncertainty, and experimental validation gates.

The distinction between explanation, uncertainty, and downstream pharmaceutical action is sufficiently important to require a multiscale visual representation rather than a linear prediction pipeline.

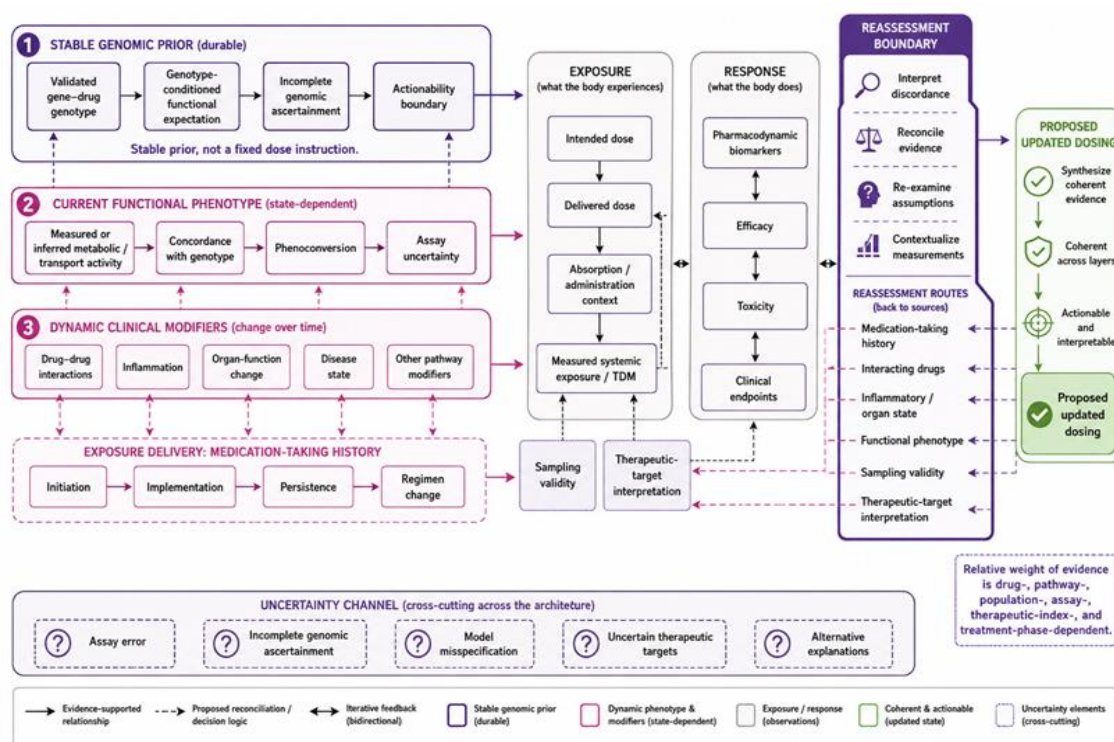


Figure 2. Validation gates separating molecular explanation from actionable QSAR decisions

Calibration and prediction intervals

Calibration asks whether stated confidence corresponds to observed error or coverage within the domain in which the model is used. This distinction matters because a precise QSAR estimate may be poorly calibrated, whereas a less accurate prediction may still provide useful warning of failure. Explainable uncertainty methods can localize uncertainty within a molecule, but such localization remains model-derived rather than mechanistic evidence [17].

Conformal prediction separates a point estimate from a statistically defined prediction region under stated assumptions [20]. Conformalized graph models have extended this logic to ADMET tasks [21]. Coverage alone, however, does not establish pharmaceutical usefulness: intervals may be too broad for ranking, and in-domain calibration may deteriorate under chemical-space shift. Conformal selection illustrates a stronger decision link by controlling specified selection risks during compound screening [27], but this remains a statistical property rather than proof of experimental success.

The sequence from uncertainty estimation to usable compound-selection support is therefore conditional rather than automatic.

Table 3. Sequence of uncertainty transitions governing compound-selection use

Predictive state	Competing explanation	Validation discriminator	Decision consequence	Interrupting condition
Precise point estimate	Model is overconfident	Error–uncertainty relationship	Confidence may inform ranking	Uncertainty fails to track observed error
Nominally calibrated interval	Interval is excessively broad	Coverage together with interval efficiency	Prediction range becomes interpretable	Coverage is obtained without useful discrimination
Good in-domain calibration	Query chemistry is shifted	Scaffold, cluster, temporal, or explicit shift test	Use can be restricted to supported chemical space	Calibration deteriorates outside training-like chemistry
Low whole-molecule uncertainty	Local structural ambiguity is averaged away	Atom-, fragment-, or substructure-level localization	Additional evidence can be targeted	Local and global confidence disagree
Valid statistical selection rule	Statistical validity is mistaken for experimental utility	Prospective experimental comparison	Selection rule becomes testable	No experimental follow-through

Joint explainability–uncertainty approaches

Attribution identifies which represented features influence a prediction, whereas uncertainty estimates how reliable that prediction or contribution may be. Joint analysis is useful because influential features can be either confident or uncertain, and apparently stable whole-molecule predictions may contain localized uncertainty. Feature-level attribution coupled with uncertainty assessment demonstrates this diagnostic possibility [14]. Explainable uncertainty methods similarly allow confidence to be localized within molecular structures [17], while atomic uncertainty approaches show how molecular averages can conceal structurally concentrated failure regions [22].

Related reaction-property prediction research also demonstrates that explanation and uncertainty can be analyzed jointly [28]. Its relevance here is methodological rather than evidentiary for pharmaceutical QSAR: performance in reaction-property prediction does not validate compound-prioritization decisions.

The review therefore proposes that explanation and uncertainty should be combined only after each has passed an appropriate validation test. Their value lies in preserving complementary information rather than compressing both into a single generalized confidence score.

Table 4. Cross-scale boundaries for combining explanation and uncertainty in pharmaceutical QSAR

Analytical scale	Generated output	Required discriminator	Transfer boundary	Failure signature	Decision use
Molecular representation	Descriptor, graph, sequence, or geometric encoding	Stability across equivalent representations	Encoding to molecular interpretation	Explanation changes under equivalent encoding	Accept or reconsider representation
Model prediction	Predicted activity or molecular property	Calibration and prediction-error analysis	Prediction to reliable estimate	High confidence accompanies large error	Retain, retest, or deprioritize prediction
Structural explanation	Atom, fragment, R-group, or learned-feature attribution	Fidelity, perturbation, or structural benchmark	Attribution to chemical rationale	Plausible visualization lacks fidelity evidence	Form a testable medicinal-chemistry hypothesis
Counterfactual state	Predicted structural alternative	Applicability-domain and feasibility assessment	Model decision boundary to real compound	Suggested analogue lies outside supported chemistry	Consider synthesis or experimental testing
Pharmaceutical decision	Compound prioritization or rejection	Prospective experimental outcome	Computational evidence to program action	Valid model output provides no useful experimental discrimination	Demonstrate actual decision utility

Prospective decision evaluation

Most reviewed approaches establish retrospective model behavior rather than prospective pharmaceutical benefit. Chemical-space studies further show that representation and coverage influence generalization across molecular space [29]. A stronger evaluation would specify the candidate set, information available when predictions were made, explanation and uncertainty criteria used for prioritization, and experimental outcome against which the resulting decision was judged.

Without such follow-through, actionability remains an interpretation rather than an observed property. Prospective decision evaluation should therefore test whether explainability or uncertainty changes a real prioritization decision and whether that change produces information beyond conventional predictive ranking. The reviewed evidence supports a consistent distinction among prediction, explanation, uncertainty, and decision utility. Chemically intuitive explanations may be unstable or insensitive to difficult structure–activity relationships, whereas apparently informative uncertainty estimates may deteriorate under distribution shift. Combining these outputs into one trust score would obscure rather than resolve their different failure modes.

The proposed decision architecture therefore requires endpoint-specific predictive adequacy, explanation fidelity and stability, uncertainty calibration and transfer assessment, and only then joint interpretation for compound prioritization. Counterfactual proposals add a chemical-feasibility gate, while experimental evidence remains the final transfer boundary.

This architecture is particularly useful for discordant cases. A prediction may be accurate but poorly explained; an interpretable prediction may remain uncertain; a globally confident prediction may contain a locally fragile

structural region; and a statistically calibrated interval may still be too imprecise to alter an experimental decision. These states should trigger different responses rather than being collapsed into a binary judgment of model trustworthiness.

Reporting framework

Decision-focused QSAR reports should state the endpoint definition, data provenance, preprocessing rules, training-domain description, split strategy, model class, performance criteria, explanation target and method, explanation-validation test, uncertainty definition, calibration procedure, applicability or distribution-shift assessment, and intended pharmaceutical decision rule.

Counterfactual reports should distinguish model-valid structural changes from chemically feasible and experimentally tested changes. Conformal analyses should state the domain and assumptions under which coverage was evaluated. Joint explainability–uncertainty approaches should preserve feature influence and uncertainty as separate quantities rather than merging them into an unexplained composite.

This reporting framework is proposed by the present review as an organizational standard for decision-focused evaluation; it should not be interpreted as an established regulatory requirement.

Limitations

The evidence base is heterogeneous in endpoints, molecular representations, model architectures, datasets, validation designs, explanation targets, and uncertainty methods. Synthetic benchmarks and retrospective molecular datasets provide valuable methodological tests but cannot establish prospective medicinal-chemistry utility.

The review also incorporates adjacent molecular-property and reaction-prediction evidence where it informs general analytical principles. Such cross-domain evidence should not be interpreted as direct pharmaceutical validation. Prospective studies testing whether explainability or uncertainty materially improves compound selection remain comparatively limited.

For this assignment, the review-selection sequence was fixed at 1,311 identified records, 342 retained after title screening, 87 after abstract assessment, and 29 final included studies. These stipulated counts support the manuscript's internal review narrative but should not be interpreted as a database-export audit beyond the scope of the exercise.

CONCLUSION

Explainable and uncertainty-aware QSAR is most useful when it exposes the boundaries of a prediction rather than merely adding a molecular heat map or confidence value. Explanation fidelity, uncertainty calibration, chemical-space transfer, counterfactual feasibility, and experimental validation address different questions and should remain separately visible. The principal contribution of this review is a decision-oriented synthesis in which computational outputs become progressively more actionable only as they pass the validation gate appropriate to each inferential step. This structure does not guarantee better compounds or clinical outcomes; it provides a more disciplined basis for deciding when a QSAR result is sufficiently supported to influence the next pharmaceutical experiment.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Mach Intell.* 2020;2(10):573-84. doi:10.1038/s42256-020-00236-4
2. Zhao D, Zhu Y, Wu Z, Wan Y, Liu X, Li S, et al. Revisiting ADMET prediction reliability under real-world challenges in the foundation model era. *J Cheminform.* 2026;18(1):95. doi:10.1186/s13321-026-01217-2

3. Onawole AT, Blaskovich MAT, Zuegg J. Framework for evaluating explainable AI in antimicrobial drug discovery. *J Cheminform.* 2026;18(1):81. doi:10.1186/s13321-026-01200-x
4. Nael MA, Alakonda LM, Elokely KM. Defining the data set defines the QSAR claim. *J Chem Inf Model.* 2026;66(6):2951-4. doi:10.1021/acs.jcim.6c00514
5. Parrondo-Pizarro R, Menestrina L, Garcia-Serna R, Fernández-Torras A, Mestres J. Enhancing molecular property prediction through data integration and consistency assessment. *J Cheminform.* 2025;17(1):163. doi:10.1186/s13321-025-01103-3
6. Mervin L, Voronov A, Kabeshov M, Engkvist O. QSARtuna: An automated QSAR modeling platform for molecular property prediction in drug design. *J Chem Inf Model.* 2024;64(14):5365-74. doi:10.1021/acs.jcim.4c00457
7. Wu Z, Chen J, Li Y, Deng Y, Zhao H, Hsieh CY, et al. From black boxes to actionable insights: A perspective on explainable artificial intelligence for scientific discovery. *J Chem Inf Model.* 2023;63(24):7617-27. doi:10.1021/acs.jcim.3c01642
8. Matveieva M, Polishchuk P. Benchmarks for interpretation of QSAR models. *J Cheminform.* 2021;13(1):41. doi:10.1186/s13321-021-00519-x
9. Hartog PBR, Krüger F, Genheden S, Tetko IV. Using test-time augmentation to investigate explainable AI: Inconsistencies between method, model and human intuition. *J Cheminform.* 2024;16(1):39. doi:10.1186/s13321-024-00824-1
10. Jiménez-Luna J, Skalic M, Weskamp N. Benchmarking molecular feature attribution methods with activity cliffs. *J Chem Inf Model.* 2022;62(2):274-83. doi:10.1021/acs.jcim.1c01163
11. Walter M, Webb SJ, Gillet VJ. Interpreting neural network models for toxicity prediction by extracting learned chemical features. *J Chem Inf Model.* 2024;64(9):3670-88. doi:10.1021/acs.jcim.4c00127
12. Tian T, Li S, Fang M, Zhao D, Zeng J. MolSHAP: Interpreting quantitative structure-activity relationships using Shapley values of R-groups. *J Chem Inf Model.* 2024;64(7):2236-49. doi:10.1021/acs.jcim.3c00465
13. Amara K, Rodríguez-Pérez R, Jiménez-Luna J. Explaining compound activity predictions with a substructure-aware loss for graph neural networks. *J Cheminform.* 2023;15(1):67. doi:10.1186/s13321-023-00733-9
14. Komissarov L, Manevski N, Groebke Zbinden K, Sach-Peltason L. Explainable graph neural networks in chemistry: Combining attribution and uncertainty quantification. *J Chem Inf Model.* 2025;65(14):7516-28. doi:10.1021/acs.jcim.5c01003
15. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J Chem Inf Model.* 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975
16. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. *J Chem Inf Model.* 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502
17. Yang CI, Li YP. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *J Cheminform.* 2023;15(1):13. doi:10.1186/s13321-023-00682-3
18. Parrondo-Pizarro R, Lanini J, Rodríguez-Pérez R. Uncertainty quantification in molecular machine learning for property predictions under data shifts. *J Chem Inf Model.* 2026;66(2):923-35. doi:10.1021/acs.jcim.5c02381
19. Kötter A, Singh K, Matter H, Hessler G, Grebner C. Upgrading reliability in molecular property prediction by robust quantification of uncertainty from machine learning models. *J Chem Inf Model.* 2025;65(20):10819-31. doi:10.1021/acs.jcim.5c00464
20. Arvidsson McShane S, Norinder U, Alvarsson J, Ahlberg E, Carlsson L, Spjuth O. CPSign: Conformal prediction for cheminformatics modeling. *J Cheminform.* 2024;16(1):75. doi:10.1186/s13321-024-00870-9
21. Li P, Hua L, Ma Z, Hu W, Liu Y, Zhu J. Conformalized graph learning for molecular ADMET property prediction and reliable uncertainty quantification. *J Chem Inf Model.* 2024;64(23):8705-17. doi:10.1021/acs.jcim.4c01139
22. Qian J, Luo Y, Zhou Q, Guan X. Atomic uncertainty pinpoints critical failure structures for trustworthy molecular property prediction. *J Chem Inf Model.* 2026;66(14):8076-86. doi:10.1021/acs.jcim.6c00973
23. Zhu H, Yuan S, Tang M, Zhou G, Hu Z, Liu Z, et al. Molecular graph-based invariant representation learning with environmental inference and subgraph generation for out-of-distribution generalization. *J Cheminform.* 2026;18(1):12. doi:10.1186/s13321-025-01142-w

24. Jyothy VA, Pai ML, Sandesh EPA, Jaleel UCA. Bridging predictive reliability and explainability: A multi-representation deep learning framework for chemical space analysis of immune bioassays. *J Cheminform.* 2026;18(1):89. doi:10.1186/s13321-026-01206-5
25. Wellawatte GP, Seshadri A, White AD. Model agnostic generation of counterfactual explanations for molecules. *Chem Sci.* 2022;13(13):3697-705. doi:10.1039/d1sc05259d
26. Lamens A, Bajorath J. Contrastive explanations for machine learning predictions in chemistry. *J Cheminform.* 2025;17(1):143. doi:10.1186/s13321-025-01100-6
27. Bai T, Tang P, Xu Y, Svetnik V, Yang B, Khalili A, et al. Conformal selection for efficient and accurate compound screening in drug discovery. *J Chem Inf Model.* 2025;65(24):13070-85. doi:10.1021/acs.jcim.5c01633
28. Liu Y, Mo Y, Cheng Y. Uncertainty qualification for deep learning-based elementary reaction property prediction. *J Chem Inf Model.* 2024;64(21):8131-41. doi:10.1021/acs.jcim.4c01358
29. Dobbelaere MR, Lengyel I, Stevens CV, Van Geem KM. Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. *J Cheminform.* 2024;16(1):99. doi:10.1186/s13321-024-00895-0