Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

REBUILDING QUANTITATIVE STRUCTURE–ACTIVITY RELATIONSHIP MODELING AROUND CHEMICAL NEIGHBORHOODS, DOMAIN BOUNDARIES, AND DECISION-RELEVANT RELIABILITY

Lucas Pereira[1]*, Carolina Mendez[2], Felipe Rios[3], David Taylor[4]

1. ***Department of QSAR and Chemical Neighborhood Modeling, Faculty of Pharmacy, University of São Paulo, São Paulo, Brazil.***
2. ***Department of Domain Boundary Definition and Applicability, Faculty of Pharmacy, National University of Colombia, Bogota, Colombia.***
3. ***Department of Decision-Relevant Reliability in QSAR, Faculty of Sciences, University of Concepción, Concepción, Chile.***
4. ***Department of Structure–Activity Modeling for Drug Discovery, Faculty of Pharmacy, University of Melbourne, Melbourne, Australia.***

ARTICLE INFO

Received:
11 September 2024
Received in revised form:
02 November 2024
Accepted:
05 November 2024
Available online:
28 December 2024

Keywords: Quantitative structure–activity relationships, Chemical neighborhoods, Applicability domain, Activity landscapes, Local predictability, Uncertainty calibration

ABSTRACT

Quantitative structure–activity relationship modeling remains central to computational pharmaceutical science, yet its reliability is commonly summarized through aggregate performance measures that conceal where, why, and for which decisions predictions are scientifically supportable. A model may perform favorably across a retrospective test set while remaining unreliable for a sparsely represented chemical series, a locally discontinuous activity landscape, an unfamiliar molecular modality, or a decision carrying asymmetric experimental consequences. This article develops a methodological reconstruction of QSAR evaluation around chemical neighborhoods, explicit domain boundaries, calibrated uncertainty, and decision-relevant reliability. The approach synthesizes evidence concerning benchmark design, molecular representation, applicability-domain assessment, activity cliffs, local predictability, uncertainty estimation, external validation, and reproducible reporting. Its central conceptual contribution is a proposed Neighborhood–Domain–Decision Reliability framework in which reliability is assigned not to a model globally, but conditionally to a prediction–decision pair. The framework distinguishes representation-specific neighborhood support, local activity-landscape behavior, multidimensional applicability, uncertainty calibration, and intended decision consequences. It further proposes reliability tiers separating supported interpolation, boundary-adjacent or structured extrapolation, and out-of-domain use requiring abstention or hypothesis-generating interpretation. The reconstruction does not claim validated universal thresholds, prospective pharmaceutical utility, mechanistic explanation, regulatory acceptance, or deployment readiness. Instead, it offers a testable structure for designing evaluations, reporting limitations, and directing future validation. Rebuilding QSAR around these elements may improve the scientific interpretability of predictions, reduce false precision, and clarify when computational outputs can cautiously support compound prioritization, additional evidence generation, or explicit non-use.

To Cite This Article: Pereira L, Mendez C, Rios F, Taylor D. Rebuilding Quantitative Structure–Activity Relationship Modeling around Chemical Neighborhoods, Domain Boundaries, and Decision-Relevant Reliability. Pharmacophore. 2024;15(6):57-65. https://doi.org/10.51847/xPQxd9V8T7

Introduction

Quantitative structure–activity relationship modeling is frequently evaluated as though predictive reliability were a stable property of a trained model. In practice, reported performance is conditional on endpoint construction, dataset composition, molecular representation, train–test partitioning, comparator selection, and the type of generalization embedded in the evaluation. Benchmark outcomes in molecular machine learning therefore depend on task and split design, relative model behavior varies across bioactivity problems, and chemically credible evaluation requires transparent alignment among data, validation, and interpretation rather than dependence on a single headline score [1–3]. This conditionality is especially

Corresponding Author: Lucas Pereira; Department of QSAR and Chemical Neighborhood Modeling, Faculty of Pharmacy, University of São Paulo, São Paulo, Brazil. E-mail: lucas.pereira@usp.br.

consequential in pharmaceutical research, where a prediction is rarely an endpoint in itself. It may influence which compounds are synthesized, which assays are prioritized, which chemical series are expanded, or which candidates are withheld from further investigation.

The difficulty is not simply that conventional metrics are imperfect. Measures such as classification accuracy, area under a receiver-operating-characteristic curve, mean absolute error, or root-mean-square error summarize performance over a chosen evaluation population. They do not independently reveal whether a particular query compound resembles well-supported training chemistry, lies near an activity cliff, belongs to an unfamiliar modality, depends on uncertain assay labels, or falls beyond the region in which uncertainty estimates remain calibrated. A chemically meaningful evaluation must instead align the dataset, split strategy, comparator, metric, uncertainty statement, and intended use, because each element changes what the resulting performance claim can legitimately support [4]. Aggregate metrics remain useful, but they are insufficient as stand-alone evidence of prediction-specific pharmaceutical reliability.

This gap is reinforced by the character of the underlying evidence. Pharmaceutical datasets may contain heterogeneous assay protocols, imbalanced endpoints, repeated analogues, uncertain measurements, sparse chemical series, and historical selection biases. Chemical and biological data availability consequently does not establish their suitability for a particular inference or decision. The realism of computational drug-discovery claims remains bounded by data quality, assay context, chemical coverage, and biological relevance [5]. A model trained on abundant records may therefore remain weakly supported for the compound, endpoint, laboratory context, or decision under consideration. Conversely, a model with moderate global performance may provide useful evidence within a coherent local series when its boundaries and uncertainty are properly characterized.

This article proposes a methodological reconstruction termed the Neighborhood–Domain–Decision Reliability framework. The framework does not treat model reliability as a single score or permanent model attribute. It defines reliability conditionally for a specific prediction–decision pair through five connected evidentiary layers: representation-specific chemical neighborhood, local activity-landscape behavior, multidimensional applicability-domain status, calibrated uncertainty, and intended decision consequence. The contribution is conceptual and requires future empirical evaluation. It does not establish universal similarity thresholds, validated reliability tiers, mechanistic causality, therapeutic value, regulatory acceptability, or operational readiness. Its purpose is to reorganize QSAR assessment so that predictive performance, chemical support, uncertainty, and permissible use are reported as related but non-equivalent scientific claims.

Why Conventional QSAR Evaluation Obscures Domain Boundaries

Conventional evaluation obscures domain boundaries when the test design makes unsupported generalization appear equivalent to ordinary interpolation. Different applicability-domain measures can admit or reject different regions of chemical space; close analogue redundancy between training and test sets can reward recognition of familiar series rather than transfer to new chemistry; and dataset biases can provide shortcuts unrelated to the intended structure–activity problem [6–8]. These mechanisms are distinct, but they converge on the same interpretive failure: a favorable aggregate result may not disclose what the model has actually learned or where its evidence becomes insufficient. The domain boundary remains hidden because the evaluation population is treated as homogeneous even when its compounds occupy very different relationships to the training evidence.

A second source of obscurity is the molecular representation used to define similarity and distance. Chemical neighborhoods are not natural objects that exist independently of modeling choices. Fingerprints, physicochemical descriptors, graph representations, and learned embeddings emphasize different structural regularities and therefore produce different geometries of proximity. Learned and engineered molecular representations organize transferable information differently across prediction tasks, making neighborhood membership and distance-to-training-data claims representation-specific [9]. A compound can appear well supported under one representation and peripheral under another. Reporting only a model architecture or average test performance leaves this dependency invisible and risks presenting a contingent geometric choice as an objective chemical boundary.

Domain interpretation is also distorted when model, representation, dataset, and validation design are analyzed separately. Molecular property prediction is produced by interactions among data scale, chemical coverage, representation, architecture, pretraining, endpoint definition, and split strategy rather than by model class alone [10]. A high-capacity method cannot recover information absent from the training data, correct an inconsistently defined endpoint, or guarantee transport across a new assay protocol. Likewise, a sophisticated representation cannot by itself establish that the local structure–activity relationship is smooth, experimentally reliable, or relevant to the intended decision. Domain assessment must therefore evaluate the complete prediction system rather than attribute reliability to the algorithm in isolation.

The proposed reconstruction treats an applicability boundary as a declared relationship among a query compound, a standardized chemical representation, an endpoint, a trained model, a body of local evidence, and an intended use. This relationship may contain several partially independent boundaries: chemical coverage, representation support, assay consistency, model behavior, modality compatibility, temporal stability, and decision acceptability. A query may be supported in one dimension and weakly supported in another. The resulting domain status should therefore be multidimensional and open to abstention rather than compressed into an unqualified binary label. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why conventional qsar evaluation obscures domain boundaries within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Why conventional QSAR evaluation obscures domain boundaries

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Evaluation-population design	What form of generalization does the test set actually represent?	Train–test separation must reflect the chemical, temporal, or project shift relevant to the intended use.	Distinguishes familiar-analogue interpolation from transfer to less represented chemistry.	Treating random or convenient partitions as evidence of prospective generalization.	No split design universally reproduces all pharmaceutical-use conditions.
Applicability-domain measure	How is support or distance from the training evidence operationalized?	Distance-, density-, leverage-, ensemble-, and model-based approaches represent support differently.	Makes domain membership an explicit methodological choice.	Presenting a molecule as intrinsically inside or outside a domain.	Domain status is conditional on the selected representation, diagnostic, endpoint, and threshold.
Chemical-data bias	Could a model exploit dataset artifacts rather than the intended structure–activity relation?	Structural redundancy, source artifacts, class imbalance, and assay-specific shortcuts can influence apparent performance.	Separates predictive convenience from scientifically relevant generalization.	Interpreting shortcut learning as robust molecular recognition.	Bias analysis cannot prove that a model learned mechanism or causality.
Molecular representation	Which chemical features determine proximity and transferable information?	Fingerprints, descriptors, graphs, and learned embeddings induce different chemical geometries.	Defines the basis on which neighborhoods and distances are constructed.	Treating one representation as a universal account of chemical similarity.	Representation adequacy is endpoint- and decision-dependent.
Endpoint and assay consistency	Are labels scientifically comparable across compounds, sources, and time?	Assay protocols, thresholds, measurement uncertainty, and biological context influence the meaning of observed activity.	Prevents chemically similar but experimentally incomparable records from being treated as coherent support.	Conflating data availability with endpoint suitability.	A larger dataset cannot compensate automatically for inconsistent or poorly aligned measurements.
Model and workflow dependence	Which preprocessing, architecture, fitting, and validation choices shape the result?	Predictive behavior emerges from interacting data and modeling decisions.	Reframes reliability as a property of the complete prediction system.	Attributing performance or failure solely to the algorithm.	No model class is globally reliable independently of its data and validation context.
Intended decision	What action could follow, and what errors are tolerable?	Screening, analogue prioritization, experimental escalation, and abstention involve different costs and evidence requirements.	Connects applicability to pharmaceutical consequence rather than geometry alone.	Using the same domain threshold for every decision.	Domain acceptability must be defined relative to a stated use and human oversight process.

Chemical Neighborhoods, Activity Landscapes, and Local Predictability

A chemical neighborhood is the set of compounds considered locally relevant to a query under a declared representation, distance or similarity measure, and inclusion rule. Its scientific importance arises from the expectation that nearby compounds may provide more informative evidence than distant compounds. That expectation, however, is conditional rather than automatic. Nonlinear chemical distances can reorganize structure–activity space, activity cliffs can produce major activity differences among close structural analogues, and activity-landscape topology can reveal local continuity or discontinuity that global performance measures conceal [11–13]. Local predictability therefore depends not only on whether neighbors exist, but on how chemical proximity relates to endpoint behavior within the neighborhood.

Neighborhood analysis should distinguish support quantity from support quality. A densely populated region may still contain heterogeneous assays, conflicting labels, narrow analogue redundancy, or abrupt activity transitions. Conversely, a smaller but experimentally coherent series may provide stronger evidence for a bounded analogue comparison. Visual and computational analysis of molecular activity spaces can reveal clusters, transitions, and heterogeneous local regions that are obscured when the dataset is summarized globally [14]. Such analysis does not itself provide calibrated uncertainty, but it helps identify whether the neighborhood resembles a smooth local response surface, a fragmented collection of subseries, or a boundary region in which small structural changes are associated with unstable activity.

Experimental uncertainty further complicates the interpretation of local landscapes. Near a classification threshold, modest measurement variation can alter the assigned activity label and create apparent discontinuities that are partly methodological rather than chemical. Bioactivity prediction can be improved by explicitly accounting for experimental uncertainty around such thresholds, demonstrating that observed class assignments should not always be treated as error-free facts [15]. The proposed framework therefore separates at least three sources of local instability: genuine structure–activity discontinuity, measurement or assay uncertainty, and model error. These sources may coexist, but they have different consequences. A

genuine cliff may require additional structural or mechanistic investigation, uncertain labels may require repeat measurement, and systematic model error may require a different representation or modeling strategy.

Within the proposed Neighborhood–Domain–Decision Reliability framework, neighborhood evidence is summarized through four linked questions: whether the query has relevant local support, whether that support is chemically and experimentally diverse enough to avoid simple redundancy, whether the local activity landscape is sufficiently coherent for the intended inference, and whether the result remains stable across reasonable neighborhood definitions. Activity cliffs are treated as local warnings rather than isolated anomalies because they can expose prediction failures hidden by favorable aggregate metrics [12]. The framework does not assume that every cliff makes prediction impossible, nor that every smooth neighborhood is reliable. Instead, neighborhood density, landscape ruggedness, assay consistency, and prediction-specific uncertainty are carried forward as separate evidence dimensions when applicability and reliability tiers are assigned.

Reconstructing Applicability Domains around Decision Context

Applicability should be reconstructed as a property of a prediction made for a particular decision rather than as a permanent attribute of a model. A compound can be adequately supported for coarse screening prioritization yet insufficiently supported for selecting among close analogues, interpreting a threshold-sensitive endpoint, or withholding experimental testing. Prediction-specific error estimates provide a more informative representation of this heterogeneity than a model-wide average because they communicate that expected error can differ among queries [16]. The proposed framework therefore defines applicability through the relationship among the query compound, endpoint, neighborhood, model, uncertainty evidence, intended action, and consequences of error.

Decision context determines which forms of uncertainty and boundary evidence are material. In high-throughput screening, an uncertainty threshold may control the balance between retained compounds and confidence in the selected subset. Conformal prediction studies illustrate that screening gain depends on the trade-off among confidence, efficiency, and the number of actionable predictions rather than on discrimination performance alone [17]. A domain rule that is acceptable when the objective is broad enrichment may be inappropriate when false negatives could eliminate a rare chemical series. The framework consequently requires the intended decision, loss asymmetry, available experimental capacity, and abstention option to be declared before a reliability judgment is made.

Applicability is also conditional on molecular modality. A property model developed primarily from conventional small molecules may encounter a substantially different structural and physicochemical regime when applied to targeted protein degraders, macrocycles, or other atypical compounds. Evaluation of machine-learning property prediction for targeted protein degraders demonstrates why nominal endpoint continuity does not guarantee continuity of model support when size, flexibility, composition, and available training evidence shift [18]. Such predictions should not automatically be classified as ordinary interpolation. They may instead represent structured extrapolation when partial support remains interpretable or out-of-domain use when representation and validation evidence are inadequate.

The reconstructed domain is therefore multidimensional. Relevant dimensions include chemical coverage, representation support, neighborhood density, local landscape behavior, assay alignment, modality, temporal stability, model diagnostics, uncertainty calibration, and intended use. Comparative analysis of chemical-space coverage shows that alternative applicability-domain methods can encompass materially different regions, reinforcing that coverage must be declared rather than presumed [19]. The framework does not require every dimension to produce an identical status. It requires disagreements to remain visible. A query supported chemically but weakly calibrated temporally, for example, should not receive an unqualified in-domain label merely because one distance criterion is satisfied.

Reliability Tiers for Interpolation, Extrapolation, and Out-of-Domain Use

The proposed reliability tiers translate multidimensional domain evidence into bounded interpretations without converting it into a universal numerical score. Tier I, supported interpolation, applies when the query has relevant local support, the endpoint and assay context are aligned, the activity landscape does not show unresolved instability, and uncertainty behavior has been evaluated under a split resembling the intended use. Application-oriented benchmarking demonstrates that compound-activity models can be judged differently under realistic drug-discovery separations than under conventional retrospective partitions [20]. Tier I therefore requires evidence that the validation problem resembles the anticipated prediction problem; proximity under a convenient test design is not sufficient.

Tier II, boundary-adjacent or structured extrapolation, applies when one or more domain dimensions are weak but the remaining evidence allows the limitation to be described and managed. Examples include a query near the edge of a represented series, a moderate modality shift, reduced neighborhood density, uncertain local landscape continuity, or calibration supported only under related conditions. Limitations of representation learning show that learned molecular features remain constrained by their tasks and training distributions, especially when queried chemistry departs from learned structural regularities [21]. Tier II is not a declaration that extrapolation is reliable. It is a conditional status requiring sensitivity analysis, explicit assumptions, stronger human review, and often additional experimental evidence.

Tier III, out-of-domain or abstain, applies when support is insufficient for the intended decision or when conflicts among neighborhood, assay, modality, and uncertainty evidence cannot be resolved. External and industrial validation of Caco-2 permeability models illustrates how practical evaluation can expose reliability limitations that remain hidden during internal model development [22]. A Tier III designation does not imply that the numerical prediction is necessarily wrong. It means

that available evidence does not justify treating it as decision-ready. Appropriate outputs may include abstention, qualitative hypothesis generation, selection of a different model, acquisition of local measurements, or explicit escalation to expert review. Tier assignment must remain endpoint-specific and decision-specific. A model may warrant Tier I interpretation for one physicochemical property, Tier II for another property, and Tier III for a new chemical modality. Comprehensive benchmarking across toxicokinetic and physicochemical prediction tools demonstrates that predictive behavior differs across endpoints and computational methods [23]. The framework therefore prohibits assigning a single permanent reliability grade to an entire model. Tier status belongs to a versioned prediction context and may change when new compounds, assay data, validation results, representations, or decision requirements become available.

Uncertainty Calibration and Neighborhood-Aware Validation

Uncertainty calibration is necessary because confidence and observed error are not interchangeable. Conformal prediction can complement conventional QSAR evaluation by providing error-control and efficiency information under stated assumptions, but those guarantees do not eliminate the need to examine domain support [24]. A prediction may satisfy nominal calibration across an aggregate population while remaining poorly supported in a sparse or cliff-rich neighborhood. The proposed framework therefore interprets uncertainty jointly with local chemical evidence. Calibration answers whether stated uncertainty corresponds to observed error frequencies; neighborhood analysis asks whether the calibration population is scientifically relevant to the query.

Uncertainty methods must themselves be evaluated rather than assumed to be reliable because they generate intervals, probabilities, variances, or ensemble disagreement values. Comparative evaluation of scalable uncertainty-estimation methods for molecular property prediction shows that approaches differ in calibration and in their ability to identify error-prone predictions [25]. The reporting of an uncertainty value is therefore insufficient without evidence concerning its empirical calibration, ranking behavior, coverage, sharpness or efficiency, and sensitivity to distribution shift. The framework further distinguishes uncertainty arising from model parameters, limited data, measurement error, assay inconsistency, local landscape ruggedness, and unrepresented chemistry.

Neighborhood-aware validation combines global evaluation with local and boundary-sensitive analyses. It examines error as a function of neighborhood support, representation-specific distance, scaffold or series membership, activity-cliff status, assay provenance, and temporal or external shift. Neural uncertainty estimates should be assessed both for calibration and for their relationship to actual predictive error because apparent model confidence may correlate imperfectly with failure [26]. Validation should also examine selective prediction: whether abstaining from uncertain or unsupported queries improves the reliability of retained predictions without reducing coverage to an operationally unhelpful level. **Figure 1** presents the qsar applicability landscape, showing how the article's key components and boundaries are connected within uncertainty calibration and neighborhood-aware validation.

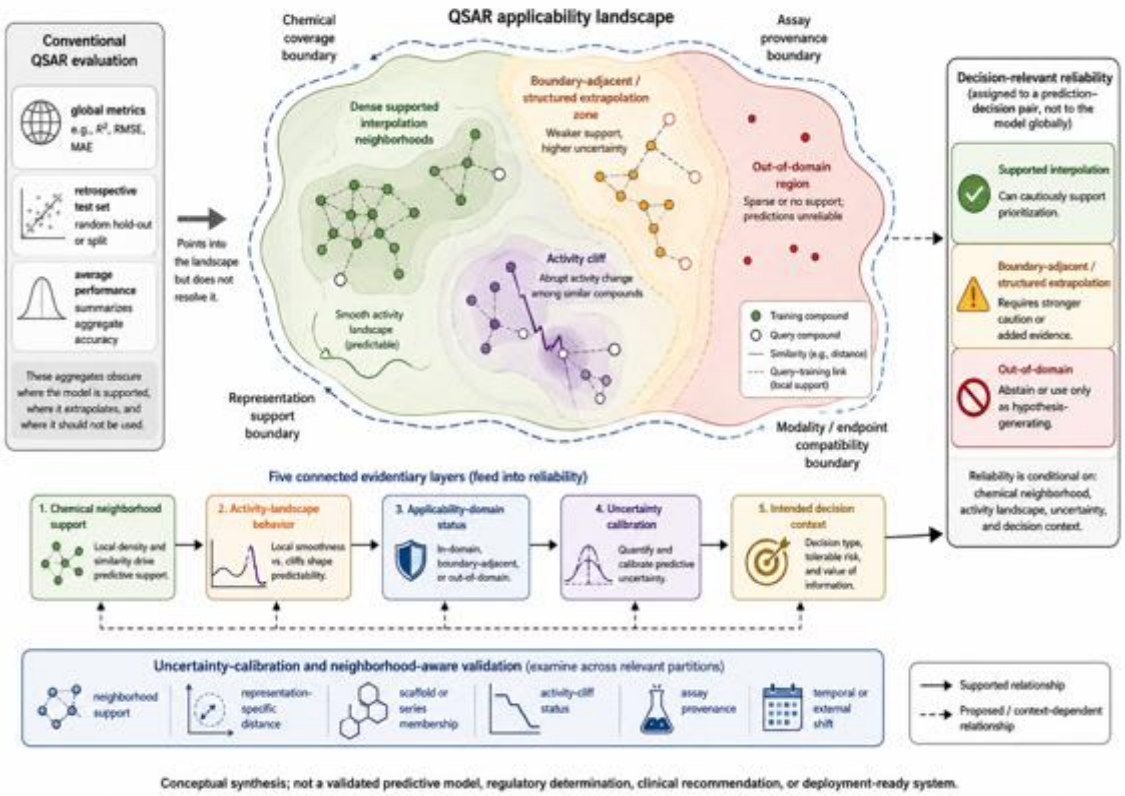

Figure 1. QSAR Applicability Landscape.

The figure is an original conceptual synthesis that organizes the article's central contribution across conventional QSAR evaluation obscures domain boundaries, chemical neighborhoods, activity landscapes, and local predictability, chemical neighborhoods, activity landscapes, local predictability, reconstructing applicability domains around decision context. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Calibration must also be tested under time and distribution change. Assessment of conformal calibration in toxicological models shows that calibration can deteriorate across chronologically distinct data and that nonexchangeability can become visible through calibration diagnostics [27]. Neighborhood-aware validation should therefore include temporal partitions, independent external data where available, drift assessment, and recalibration triggers. A model that was calibrated when developed should not be presumed calibrated indefinitely. Reliability must be maintained as a versioned evidentiary claim that can be revised when the chemical series, assay protocol, data source, or decision environment changes.

A Practical Reporting Standard for Decision-Relevant QSAR

A practical reporting standard should permit an independent reader to reconstruct not only the model but also the basis of its reliability claim. Minimum reporting should include endpoint definition, assay context, data inclusion and exclusion rules, chemical standardization, representation, model configuration, comparator methods, splitting strategy, global and local performance, neighborhood construction, applicability diagnostics, uncertainty method, calibration evidence, intended decision, tier rationale, and abstention conditions. Practical guidance for gradient boosting in molecular property prediction demonstrates that preprocessing, hyperparameter choices, and validation practices materially influence results and reproducibility [28]. Such choices must therefore be reported as scientific determinants rather than treated as incidental implementation details.

Chemical identity and preprocessing require particular attention because standardization affects the structures that enter the model and the neighborhood relations derived from them. An open QSAR-ready structure-standardization workflow illustrates the importance of recording transformations involving salts, mixtures, charges, tautomers, stereochemistry, duplicates, and invalid structures [29]. The proposed standard requires each reported prediction to be traceable to a defined input structure and a versioned preprocessing procedure. Data provenance should further identify assay sources, aggregation rules, replicate handling, thresholding decisions, and temporal cutoffs. Without this information, local support and activity-landscape claims cannot be independently evaluated.

Workflow reproducibility should connect data, split assignments, descriptors or embeddings, fitted models, domain diagnostics, uncertainty calculations, and reporting outputs. Open-source QSPR tooling demonstrates how these components can be preserved within a traceable computational workflow [30]. The proposed standard does not mandate a particular software package. It requires enough machine-readable and human-readable information to reconstruct the analysis, identify leakage, reproduce prediction contexts, and challenge domain assignments. Software availability improves transparency, but the use of an open tool does not establish that the resulting model is chemically valid, externally calibrated, or appropriate for a pharmaceutical decision.

Independent assessability is the final reporting criterion. Research on QSAR reproducibility by nonexperts shows that incomplete descriptions of data, methods, applicability, and performance can prevent independent evaluation [31]. A report should therefore state what the model supports, what remains uncertain, what evidence would change the assigned tier, and which interpretations are prohibited. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop a practical reporting standard for decision-relevant qsar within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Rebuilding Quantitative Structure–Activity Relationship Modeling around Chemical Neighborhoods, Domain Boundaries, and Decision-Relevant

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Decision-context declaration	Endpoint, intended action, decision owner, asymmetric loss, experimental capacity, abstention option	Define what reliability must mean for the intended use	Versioned intended-use and prohibited-use statement	Consequences and acceptable losses may be incompletely specified	Prospective assessment of whether the declaration matches actual decisions
Chemical identity and provenance	Standardized structures, source records, assay metadata, replicate rules, temporal cutoffs	Preserve the lineage of every model input and query	Traceable QSAR-ready dataset and query structure	Standardization choices can alter identity and neighborhood membership	Independent reconstruction of data preparation and identity rules
Representation-specific neighborhood	Descriptor, fingerprint, graph, or embedding; distance metric; inclusion rule	Identify locally relevant training evidence	Neighborhood membership, density, diversity, and sensitivity report	Results depend on representation, metric, dimensionality, and threshold	Comparison across reasonable neighborhood definitions and external compounds
Activity-landscape assessment	Local similarities, continuous activities, thresholds, assay	Characterize local smoothness, cliffs, and conflicting SAR	Local continuity or instability statement	Apparent cliffs can reflect measurement error or assay mismatch	Replicate-aware and threshold-sensitive evaluation

	uncertainty, matched analogues				
Multidimensional domain assessment	Chemical coverage, model diagnostics, assay alignment, modality, temporal stability, decision context	Combine distinct boundary dimensions without collapsing them prematurely	Domain profile with supporting and conflicting evidence	Dimensions may disagree and universal thresholds are unavailable	Predefined domain rules tested on untouched and shifted data
Uncertainty calibration	Predictive intervals, probabilities, ensembles, conformal outputs, calibration data	Relate stated uncertainty to observed error	Calibration, coverage, efficiency, and error-ranking evidence	Calibration can fail under drift or nonexchangeability	External, temporal, and neighborhood-stratified calibration analysis
Reliability-tier assignment	Neighborhood, landscape, domain, calibration, validation, and decision evidence	Assign supported interpolation, structured extrapolation, or abstention status	Tier rationale with explicit conditions and escalation triggers	Proposed tiers and thresholds are not yet empirically validated	Prospective, preregistered testing of tier discrimination and stability
Selective prediction and abstention	Decision costs, uncertainty thresholds, coverage requirements, alternative evidence pathways	Prevent unsupported predictions from being treated as decision-ready	Predict, conditionally use, abstain, or request-data recommendation	Excessive abstention may make a model operationally unhelpful	Risk–coverage and decision-utility evaluation
External and temporal validation	Independent laboratories, future project data, revised assays, new modalities	Test transportability and detect drift	Updated calibration, domain profile, and tier assignment	One external dataset cannot establish universal transfer	Repeated multi-source and longitudinal validation
Reporting and audit layer	Versioned data, code, model settings, splits, diagnostics, limitations, intended use	Make the reliability claim reproducible and challengeable	Human-readable report and machine-readable provenance record	Documentation completeness does not prove scientific correctness	Independent audit, reconstruction, and reporting-usability studies

Limitations and Boundary Conditions

The proposed Neighborhood–Domain–Decision Reliability framework is a conceptual reconstruction, not a validated prediction system. Its components are supported by methodological evidence concerning benchmark design, applicability domains, local structure–activity behavior, uncertainty, external validation, and reproducibility, but their integration into a tiered reliability structure has not been prospectively tested. The framework does not establish optimal neighborhood metrics, universal similarity cutoffs, accepted domain thresholds, or standardized tier boundaries. These elements must remain endpoint-, representation-, modality-, and decision-specific until comparative validation demonstrates otherwise.

Evidence limitations also constrain the framework. Retrospective datasets may underrepresent failed compounds, inactive series, assay changes, proprietary chemical regions, and project-specific decision processes. Activity cliffs may reflect biological discontinuity, experimental variation, or inconsistent endpoint construction, and these explanations cannot always be separated computationally. Calibration evidence may be valid only for a particular dataset or period and can deteriorate under drift [27]. Even comprehensive reporting cannot recover evidence that was never collected or correct systematic selection bias. The proposed reconstruction can expose uncertainty and unsupported use, but it cannot eliminate those limitations.

The framework must not be used to equate prediction with explanation, mechanism, causality, experimental confirmation, safety, developability, therapeutic value, or regulatory acceptability. A Tier I prediction remains a bounded computational inference requiring appropriate experimental and human evaluation. A Tier II designation should not be reframed as weak approval, and Tier III should not be interpreted as proof that a compound lacks value. The tiers describe the strength and relevance of available predictive support for a specified decision. They do not rank the intrinsic quality of molecules, replace domain expertise, or authorize autonomous pharmaceutical decisions.

Research Agenda and Implementation Priorities

The first research priority is to determine whether neighborhood-aware evidence provides reproducible information beyond conventional global metrics. Studies should compare fingerprints, descriptors, graph representations, and learned embeddings using predeclared neighborhood rules, local-error estimators, activity-cliff diagnostics, and realistic data partitions. The principal question is whether neighborhood density, diversity, and landscape stability predict error consistently across endpoints and project stages. Application-oriented benchmark designs provide an initial basis for examining realistic generalization [20], but prospective and temporal studies are needed to determine whether local diagnostics remain informative when new chemistry is generated.

The second priority is empirical evaluation of the proposed reliability tiers. Tier rules should be specified before examining outcomes and then tested on untouched retrospective sets, temporally later compounds, external laboratory data, and modality shifts. Evaluations should assess whether the tiers produce ordered differences in error, calibration, coverage, abstention, and downstream decision utility. Industrial external validation demonstrates the importance of testing models under practical

conditions [22], while conformal approaches provide tools for evaluating error control and selective prediction [24]. These methods should support tier testing without being treated as substitutes for decision-specific validation.

The third priority is implementation infrastructure. A minimum reporting schema should use controlled terminology for chemical identity, endpoint provenance, neighborhood definition, domain dimensions, uncertainty evidence, tier rationale, and prohibited use. Versioned workflows should make each prediction traceable to data, preprocessing, representation, model, and validation artifacts. Independent groups should test whether reports permit reconstruction and whether reliability information improves expert decisions rather than merely increasing documentation. Reproducibility studies indicate that assessability depends on complete reporting [31]. Staged implementation should therefore begin with retrospective audit and reporting pilots, proceed to silent prospective evaluation, and advance to bounded decision support only after context-specific evidence accumulates.

Conclusion

Rebuilding QSAR around chemical neighborhoods, domain boundaries, and decision-relevant reliability changes the central evaluative question from "How well does the model perform?" to "What evidence supports this prediction for this chemistry, endpoint, and decision?" The proposed Neighborhood–Domain–Decision Reliability framework organizes representation-specific neighborhood support, activity-landscape behavior, multidimensional applicability, calibrated uncertainty, realistic validation, reliability tiers, and reproducible reporting into one bounded methodological structure. Its purpose is not to replace predictive metrics but to prevent them from carrying claims they cannot support. Supported interpolation, structured extrapolation, and out-of-domain abstention are presented as proposed interpretive categories requiring empirical validation rather than established universal standards. By making local support, uncertainty, decision consequence, and prohibited inference visible, the reconstruction may help computational pharmaceutical science distinguish useful prediction from false precision while preserving the essential roles of experimental confirmation, human oversight, and context-specific judgment.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A
2. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/C8SC00148K
3. Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8. doi:10.1038/s41557-021-00716-z
4. Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. Nat Rev Chem. 2022;6(6):428-42. doi:10.1038/s41570-022-00391-9
5. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
6. Klingspohn W, Mathea M, ter Laak A, Heinrich N, Baumann K. Efficiency of different measures for defining the applicability domain of classification models. J Cheminform. 2017;9(1):44. doi:10.1186/s13321-017-0230-2
7. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403
8. Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712
9. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237
10. Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nat Commun. 2023;14(1):6395. doi:10.1038/s41467-023-41948-6
11. Balabin IA, Judson RS. Exploring non-linear distance metrics in the structure-activity space: QSAR models for human estrogen receptor. J Cheminform. 2018;10(1):47. doi:10.1186/s13321-018-0300-0
12. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
13. Iqbal J, Vogt M, Bajorath J. Activity landscape image analysis using convolutional neural networks. J Cheminform. 2020;12(1):34. doi:10.1186/s13321-020-00436-5

14. Kausar S, Falcao AO. A visual approach for analysis and inference of molecular activity spaces. J Cheminform. 2019;11(1):63. doi:10.1186/s13321-019-0386-z
15. Mervin LH, Trapotsi MA, Afzal AM, Barrett IP, Bender A, Engkvist O. Probabilistic Random Forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. J Cheminform. 2021;13(1):62. doi:10.1186/s13321-021-00539-7
16. Liu R, Glover KP, Feasel MG, Wallqvist A. General approach to estimate error bars for quantitative structure–activity relationship predictions of molecular activity. J Chem Inf Model. 2018;58(8):1561-75. doi:10.1021/acs.jcim.8b00114
17. Svensson F, Afzal AM, Norinder U, Bender A. Maximizing gain in high-throughput screening using conformal prediction. J Cheminform. 2018;10(1):7. doi:10.1186/s13321-018-0260-4
18. Peteani G, Huynh MTD, Gerebtzoff G, Rodríguez-Pérez R. Application of machine learning models for property prediction to targeted protein degraders. Nat Commun. 2024;15(1):5764. doi:10.1038/s41467-024-49979-3
19. Zhang Z, Sangion A, Wang S, Gouin T, Brown TN, Arnot JA, et al. Chemical space covered by applicability domains of quantitative structure–property relationships and semiempirical relationships in chemical assessments. Environ Sci Technol. 2024;58(7):3386-98. doi:10.1021/acs.est.3c05643
20. Tian T, Li S, Zhang Z, Chen L, Zou Z, Zhao D, et al. Benchmarking compound activity prediction for real-world drug discovery applications. Commun Chem. 2024;7(1):127. doi:10.1038/s42004-024-01204-4
21. Dias AL, Bustillo L, Rodrigues T. Limitations of representation learning in small molecule property prediction. Nat Commun. 2023;14(1):6394. doi:10.1038/s41467-023-41967-3
22. Wang D, Jin J, Shi G, Bao J, Wang Z, Li S, et al. ADMET evaluation in drug discovery: 21. Application and industrial validation of machine learning algorithms for Caco-2 permeability prediction. J Cheminform. 2025;17(1):3. doi:10.1186/s13321-025-00947-z
23. Gadaleta D, Serrano-Candelas E, Ortega-Vallbona R, Colombo E, Garcia de Lomana M, Biava G, et al. Comprehensive benchmarking of computational tools for predicting toxicokinetic and physicochemical properties of chemicals. J Cheminform. 2024;16(1):145. doi:10.1186/s13321-024-00931-z
24. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform. 2019;11(1):4. doi:10.1186/s13321-018-0325-4
25. 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
26. 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
27. Morger A, Svensson F, Arvidsson McShane S, Gauraha N, Norinder U, Spjuth O, et al. Assessing the calibration in toxicological in vitro models with conformal prediction. J Cheminform. 2021;13(1):35. doi:10.1186/s13321-021-00511-5
28. Boldini D, Grisoni F, Kuhn D, Friedrich L, Sieber SA. Practical guidelines for the use of gradient boosting for molecular property prediction. J Cheminform. 2023;15(1):73. doi:10.1186/s13321-023-00743-7
29. Mansouri K, Moreira-Filho JT, Lowe CN, Charest N, Martin T, Tkachenko V, et al. Free and open-source QSAR-ready workflow for automated standardization of chemical structures in support of QSAR modeling. J Cheminform. 2024;16(1):19. doi:10.1186/s13321-024-00814-3
30. van den Maagdenberg HW, Šícho M, Araripe DA, Luukkonen S, Schoenmaker L, Jespers M, et al. QSPRpred: A flexible open-source quantitative structure–property relationship modelling tool. J Cheminform. 2024;16(1):128. doi:10.1186/s13321-024-00908-y
31. Patel M, Chilton ML, Sartini A, Gibson L, Barber C, Covey-Crump L, et al. Assessment and reproducibility of quantitative structure–activity relationship models by the nonexpert. J Chem Inf Model. 2018;58(3):673-82. doi:10.1021/acs.jcim.7b00523