



Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

QSAR Models Fail Quietly When Applicability Domains Are Treated as Administrative Boundaries

Sofía Martínez^{1*}, Carlos Gómez¹, Lucia Navarro²

¹Department of QSAR Modeling and Applicability Domains, Faculty of Pharmacy, Polytechnic University of Madrid, Madrid, Spain.

²Department of Predictive Toxicology and Model Validation, Faculty of Pharmacy, University of Barcelona, Barcelona, Spain.

*Email: sofia.martinez@upm.es

ABSTRACT

Applicability domains are intended to delimit the region in which a quantitative structure–activity relationship model has defensible predictive support. In practice, however, domain assessment is often compressed into an administrative decision: a query compound is classified as inside or outside a boundary, after which the binary label is allowed to stand in for prediction reliability. This Perspective argues that this compression obscures several analytically distinct quantities. Chemical-space coverage, local training support, response-surface regularity, model uncertainty, distribution shift, calibration quality, and decision consequence need not change together. A compound may therefore occupy apparently familiar chemical space while remaining difficult to predict, whereas a geometrically unusual compound may sometimes receive a defensible prediction from a well-calibrated model. The article develops an evidence-bounded distinction between domain membership and prediction confidence, examines why static boundaries can fail under realistic molecular distribution shifts, and prepares the basis for a proposed risk-calibrated applicability framework. The central claim is not that applicability domains should be abandoned, but that they should be interpreted as one component of a broader reliability assessment whose evidential meaning is explicitly validated and reported.

Keywords: QSAR applicability domain, Chemical space, Molecular out-of-distribution prediction, Uncertainty quantification, Conformal prediction, Model validation

INTRODUCTION

A quantitative structure–activity relationship model can appear accurate without demonstrating the type of generalization required when medicinal chemistry moves beyond the compounds that made the model look successful. Ligand-based classification benchmarks, for example, can reward memorization when structural redundancy permits held-out molecules to remain closely related to training examples [1]. More broadly, molecular-property prediction depends not only on algorithm choice but also on the interaction among dataset composition, molecular representation, model architecture, and evaluation design [2]. Realistic molecular out-of-distribution problems further show that “unseen” chemistry is not a single condition: different forms of chemical and dataset shift impose different generalization demands [3].

Applicability-domain analysis is supposed to make this problem explicit. Yet its operational use often becomes narrower than its scientific purpose. A molecule is labelled inside or outside a specified region, and the label is then interpreted as though it were a reliability certificate. That interpretation is difficult to sustain when local high-error cohorts can occur among predictions that an existing applicability-domain procedure has classified as

reliable [4]. It is also inconsistent with the broader uncertainty literature, in which prediction uncertainty is decomposed, estimated, calibrated, and evaluated using methods that are not equivalent to geometric domain membership [5].

The methodological problem addressed here is therefore not whether QSAR models need boundaries. It is whether a boundary should terminate the reliability inquiry. This Perspective treats applicability-domain membership as evidence about model support rather than as an administrative permission to trust a prediction. The distinction matters because structural support, local prediction difficulty, predictive uncertainty, calibration, dataset shift, and the consequence of an erroneous decision can diverge. The article develops these distinctions sequentially and uses them to motivate a proposed risk-calibrated interpretation of applicability in which a hard boundary remains useful when warranted, but cannot substitute for evidence about what the model actually knows at the queried molecule.

What an applicability domain is supposed to represent

At its most defensible, an applicability domain represents the relationship between a query molecule and the evidence used to construct a model. Similarity-based approaches operationalize this relationship by asking whether a query occupies a sufficiently represented neighborhood of training chemistry. Molecular-similarity domain metrics can identify compounds that are structurally remote from the training set and therefore poorly supported by the learned chemical space [6]. Distance-based Boolean procedures similarly demonstrate that restricting predictions to better-supported regions can improve retrospective classification behavior in particular high-throughput screening settings [7]. These findings justify treating molecular proximity as useful evidence of support.

They do not establish that proximity is equivalent to correctness. The definition of “near” depends on the representation, metric, endpoint, and sampled chemistry. A fingerprint neighborhood can be dense while the corresponding biological response surface remains irregular; conversely, geometric remoteness does not logically guarantee that a prediction will be poor. The appropriate interpretation is narrower: distance and similarity describe support under a specified representation.

Calibration-based methods make this distinction clearer. Large-scale comparisons of conventional QSAR and conformal prediction show that conformal methods provide prediction sets or confidence-related information with properties analytically different from a point prediction plus a hard domain flag [8]. Local conformal approaches extend the idea by making applicability prediction-specific rather than imposing one global boundary on every query molecule [9]. Their assumptions and empirical performance remain model- and dataset-dependent, but their structure exposes an important conceptual error in administrative-domain thinking. Applicability is not necessarily a permanent attribute of a molecule. It can depend on the model, endpoint, local neighborhood, calibration evidence, and prediction being requested.

For this article, an applicability domain is therefore best interpreted as a model-relative statement about evidential support. It should answer, “How strongly is this prediction connected to chemistry represented during model construction or calibration?” It should not, by itself, answer the stronger question, “How confident should a medicinal chemist be that this numerical prediction is sufficiently reliable for the present decision?” The latter requires additional evidence.

Chemical-space coverage and predictive support

Chemical space is often described as though it were a fixed map on which training compounds occupy a territory and test compounds either fall inside or outside. QSAR practice is less stable. Apparent generalization changes with data partitioning: random, scaffold-sensitive, and task-specific splits can create materially different tests of molecular prediction [10]. Consequently, a domain defined around a conveniently sampled training set may describe the dataset more faithfully than the chemistry encountered after deployment.

Coverage also depends on what is being predicted and how the domain is constructed. Large-scale analysis of physicochemical-property models across more than 81,000 inventory chemicals showed that applicability-domain coverage varied with the property, chemical class, and domain definition [11]. “Coverage” is therefore not an endpoint-independent percentage of chemical universe. It is conditional on representation, sampled compounds, model formulation, target property, and the rule used to decide support.

The support region can also change when evidence changes. In a mitochondrial-toxicity setting, integrating Cell Painting phenotypic information with active learning enabled acquisition of compounds that broadened the usable model space relative to relying only on the original structural evidence [12]. This result is context-specific and

should not be generalized to every QSAR endpoint, but it demonstrates that applicability is partly a property of the training evidence rather than an immutable molecular characteristic.

These relationships become difficult to follow when coverage, support, transfer, and failure are described sequentially in prose. **Table 1** therefore separates the molecular state being queried from the type of support available, the transfer boundary that remains, and the specific way an apparently acceptable domain decision can still fail.

Table 1. When chemical-space coverage becomes predictive support rather than a geometric administrative label

scenario	Molecular or chemical state	Observable support evidence	Decision consequence	Transfer boundary that remains	Failure signature	Experimental or data counterpart	Prospective reliability test	Consequence for quiet QSAR failure
Represented local neighborhood	Query resembles compounds repeatedly sampled for the same endpoint	Dense representation under the chosen similarity scheme	Prediction can proceed as structurally supported	Response behavior may still vary within the neighborhood	Acceptable proximity accompanies unexpectedly large local error	Additional close analogues measured under the same assay conditions	Test held-out near-neighbors rather than infer reliability from similarity alone	Domain membership can overstate confidence if structural support is treated as sufficient
Sparse chemical-class edge	Query remains recognizable but approaches a poorly sampled class boundary	Declining neighborhood support or unstable coverage classification	Prediction warrants stronger qualification	Boundary location changes with representation and domain rule	Small methodological changes move the compound between inside and outside status	Targeted sampling of the underrepresented class	Compare support classifications across justified representations	A nominal cutoff can make a continuous loss of evidence look categorical
Split-sensitive apparent coverage	Query looks supported under a permissive validation partition	High apparent overlap between training and held-out chemistry	Retrospective performance appears reassuring	Deployment may impose scaffold or distribution changes absent from the split	Performance decreases when chemically separated evaluation is used	Scaffold-sensitive or deployment-matched holdout data	Re-evaluate under a split reflecting intended use	Administrative inclusion may preserve benchmark optimism rather than deployment evidence
Property-dependent support	Same chemical inventory is evaluated for different modeled properties	Coverage changes with endpoint and chemical class	Applicability must be interpreted separately for each property model	Support for one endpoint does not transfer automatically to another	A molecule is supported for one property but weakly represented for another	Endpoint-specific experimental measurements	External assessment for each modeled property	Molecule-level “in-domain” status becomes misleading when detached from endpoint

Evidence-expanded support	New orthogonal measurements guide acquisition of informative chemistry	Training evidence gains structurally or phenotypically distinct compounds	Effective supported region can broaden	Expansion depends on the added measurement platform and assay context	Original model fails where newly acquired evidence changes local support	Targeted phenotypic or experimental acquisition	Retrain and challenge the model on the newly sampled region	Applicability should be updated with evidence rather than treated as a permanent label

Table note: The final column expresses the article's proposed synthesis of how evidence-supported support scenarios can generate quiet model failure. No numerical ranking or universal threshold is implied.

Distance-, leverage-, density-, and uncertainty-based domains

Distance, leverage, neighborhood density, and uncertainty are often placed under the same applicability-domain umbrella because each can identify predictions that deserve caution. Their mathematical meanings, however, are different. Distance asks how far a query lies from training observations under a selected representation. Leverage asks how strongly the query occupies an extrapolative position relative to the fitted design space. Local density describes how much nearby evidence is available. Predictive uncertainty attempts to characterize expected error or dispersion in the prediction itself. Treating these quantities as interchangeable turns a family of diagnostics into a single administrative boundary.

Leverage illustrates the problem particularly well. In polyparameter linear free-energy relationship models, prediction error generally increased as leverage increased, but the conventional leverage cutoff could be excessively strict, excluding extrapolative predictions that remained accurate; prediction intervals represented observed error behavior more directly in that studied setting [13]. This does not make leverage uninformative. It shows that a geometric cutoff is not automatically calibrated to prediction risk, even in a well-characterized linear setting.

The same caution applies more strongly to modern molecular machine learning. Comparative studies of scalable molecular uncertainty methods, neural-network uncertainty estimation, and uncertainty-guided chemical prediction show that uncertainty estimators differ in calibration, error ranking, scalability, and their ability to recognize poorly supported predictions [14–16]. No single uncertainty family can therefore be substituted mechanically for an applicability-domain rule.

Hybrid approaches make the complementary nature of these signals explicit. Deep QSAR work combining different uncertainty sources was motivated in part by the tendency of individual methods to become overconfident outside well-supported regions [17]. Yet evaluating an uncertainty estimator is itself nontrivial. Different uncertainty-quality metrics emphasize calibration, ranking, sharpness, or association with error and can produce different judgments about the same estimator [18]. Prediction-specific QSPR uncertainty methods likewise reinforce that reliability varies across individual compounds and properties rather than belonging only to the model as a whole [19]. Explainable uncertainty methods go further by localizing contributions to uncertainty within molecular representations, although such attribution should not be confused with mechanistic chemical explanation [20].

The implication is methodological rather than terminological. Distance, leverage, density, and uncertainty should remain separate until evidence shows that a particular combination is useful for the intended endpoint and decision. A molecule can be close to training chemistry yet uncertain because the response surface is unstable. Another can be relatively distant while receiving a narrow, empirically defensible interval from a model whose extrapolative behavior has been tested. A third can receive low estimated uncertainty from a poorly calibrated model and therefore appear safer than it is. The administrative-domain error occurs when these distinct possibilities are collapsed into a single membership flag.

Silent extrapolation beyond training support

The most consequential extrapolation is not always the molecule that obviously lies beyond a structural boundary. Quiet extrapolation occurs when the representation declares a query familiar while the property being predicted behaves in a way that the training evidence does not adequately constrain.

Activity cliffs provide a direct molecular example. Highly similar compounds can exhibit large differences in measured activity, making local prediction difficult precisely where a similarity-based view might suggest strong

support. More generally, molecular-property landscapes with greater roughness tend to be less modellable out of sample, although landscape roughness is an association with difficulty rather than a deterministic cause of every error [21, 22]. These observations shift the relevant question from “Is the query close to training compounds?” to “Is the property locally learnable at the resolution required for this prediction?”

Retrospective applicability-domain analyses reinforce the same distinction from another direction. High-error subspaces can occur among predictions already designated reliable [4]. Large chemical-space analyses show that domain coverage changes with the property and domain construction [11], while leverage analyses demonstrate that geometrically extrapolative predictions are not uniformly inaccurate [13]. Together, these findings rule out a one-to-one mapping between domain membership and prediction correctness.

Silent extrapolation therefore includes at least two conceptually different failures. The first is support extrapolation, in which the query enters chemistry insufficiently represented by the training evidence. The second is response extrapolation, in which nominally represented chemistry contains a local structure–property relationship that is too discontinuous, sparse, noisy, or endpoint-specific for the model to resolve reliably. This distinction is a proposed synthesis of the evidence, not an established universal taxonomy. Its value is that it explains why moving the geometric boundary alone cannot eliminate every quiet failure.

Applicability-domain failure under dataset shift

Applicability domains are usually computed from a fixed training distribution, but deployment changes the distribution that gives the domain meaning. New scaffolds, altered chemical-class frequencies, different assay conditions, changing label distributions, or transfer to a related molecular-property task can all produce dataset shift without generating an obvious geometric “outside” event.

Uncertainty estimation is not automatically protected against this problem. Molecular active-learning studies comparing in-distribution and out-of-distribution settings show that uncertainty estimators differ in their ability to rank OOD molecules and in how effectively their uncertainty signals support acquisition decisions [23]. High uncertainty may be useful evidence of weak support, but low uncertainty cannot be assumed safe unless the estimator has itself been challenged under the relevant shift.

Task transfer creates another boundary. A transferability-map approach across molecular-property datasets showed that the relationship between source and target tasks affects whether transfer is likely to be useful [24]. Task relatedness is not identical to molecular similarity. A query may occupy apparently familiar structural space while the predictive task, endpoint definition, or learned representation has changed enough to undermine the meaning of the original domain.

These risks make reproducibility part of applicability analysis. Modern molecular-prediction software can expose data splitting, molecular featurization, architecture, training, and evaluation choices in a reproducible workflow [25]. Such transparency does not validate external generalization, but without it an apparent domain failure cannot be cleanly separated from a change in preprocessing, split logic, software configuration, or model construction. Realistic OOD specifications similarly show that different molecular shifts answer different questions [3], while scaffold-sensitive benchmark design demonstrates that the apparent difficulty of generalization depends on how chemical separation is imposed [10].

A static boundary can therefore fail through a sequence rather than a single event: the deployment distribution changes, the original support relation weakens or changes meaning, the uncertainty estimator may or may not respond, and observed prediction error becomes difficult to interpret unless the shift and model configuration are recoverable. **Table 2** records that sequence because the precursor state, shift trigger, competing explanation, validation discriminator, and downstream reliability consequence cannot be represented adequately by a simple inside/outside column.

Table 2. Sequence of quiet applicability-domain failure when deployment no longer matches the evidence that defined the boundary

Shift state	Precursor condition	Transition trigger	Observable readout	Competing explanation	Dominant uncertainty source	Validation discriminator	Downstream consequence	Condition that interrupts or reverses the sequence	QSAR-specific synthesis consequence
Nominally supported	Query resemb	Deployment	Domain flag	Ordinary model	Unrecognized	Compare deployed	Apparent	Demonstrate matched	Inside-domain

deployment	les development chemistry	begins outside the original sampling process	remains acceptable	noise rather than shift	distribution mismatch	nt descriptors and endpoint context with development evidence	continuity of trust	distribution and stable error behavior	status should be treated as provisional evidence, not a terminal decision
Chemical-distribution drift	Scaffold or class frequencies move away from training experience	New chemistry accumulates	Similarity or density support becomes uneven	Descriptor choice creates apparent drift	Epistemic uncertainty from sparse support	Shift-aware external or scaffold-separated evaluation	Reliability becomes locally heterogeneous	Acquire representative compounds and re-evaluate support	A fixed boundary can become stale even when its mathematical definition is unchanged
Task or endpoint transfer	Structural chemistry remains partly familiar	Source model is reused for a changed target property or task	Predictions remain numerically available	Poor transfer may reflect endpoint mismatch rather than chemical novelty	Task-transfer uncertainty	Evaluate source–target relatedness and target-task performance	Original AD loses part of its evidential meaning	Recalibrate or retrain using target-task evidence	Applicability belongs to a model–endpoint pair, not to the molecule alone
Uncertainty-response mismatch	Model continues to produce confidence estimates	OOD chemistry or task shift alters error structure	Low or poorly ranked uncertainty accompanies weak predictions	Estimator miscalibration rather than domain geometry	UQ model failure	Test calibration and error ranking under the intended shift	Quiet overconfidence becomes possible	Requalify the estimator on shifted data or adopt an alternative uncertainty method	Confidence requires its own domain of demonstrated validity
Reproducibility ambiguity	Performance deterioration is observed	Data, split, software, or preprocessing differs from development	Error changes cannot be assigned to one cause	Implementation drift mimics chemical shift	Procedural uncertainty	Reconstruct the complete modeling workflow and repeat the comparison	Domain failure cannot be diagnosed cleanly	Recover versioned data, configuration, split, and model specification	Applicability reporting must preserve the provenance of the boundary itself

Table note: The ordered states constitute a proposed sequence-aware synthesis. They do not imply that every deployment progresses through all states, that the transitions are irreversible, or that a universal numerical shift threshold exists.

Separating domain membership from prediction confidence

A domain decision and a confidence statement answer different questions. Domain membership asks whether the query is sufficiently connected to the evidence used to train or calibrate the model. Prediction confidence asks

how uncertain the particular estimate remains after the model, data, and local chemical environment are considered. Evidential deep learning demonstrates that prediction-specific uncertainty can be modeled and used to guide molecular selection without reducing the result to a geometric inside/outside flag [26]. Such uncertainty remains an estimate generated by a model; it is not an observed probability that a prediction is chemically or pharmacologically correct.

Confidence therefore requires its own qualification. UNIQUE formalizes benchmarking across complementary uncertainty criteria rather than treating any reported uncertainty value as self-validating [27]. Conformalized graph learning similarly shows that calibrated prediction intervals can be constructed for molecular ADMET tasks under the evaluated calibration and split conditions [28]. Marginal interval coverage, however, does not guarantee correctness for an individual compound, nor does it establish that a model has captured the biological mechanism governing the endpoint.

The practical distinction is consequential. A query can be structurally well supported but receive unstable or poorly calibrated predictions. Another may be geometrically unusual yet receive a defensible interval because the model's extrapolative behavior has been explicitly tested. Domain membership should therefore be interpreted as one evidential coordinate, not as prediction confidence itself.

This separation is easiest to recover visually because support, uncertainty, local error, and shift can diverge while still feeding the same medicinal-chemistry decision. **Figure 1** therefore keeps those trajectories analytically separate until they converge at an explicitly proposed decision layer.

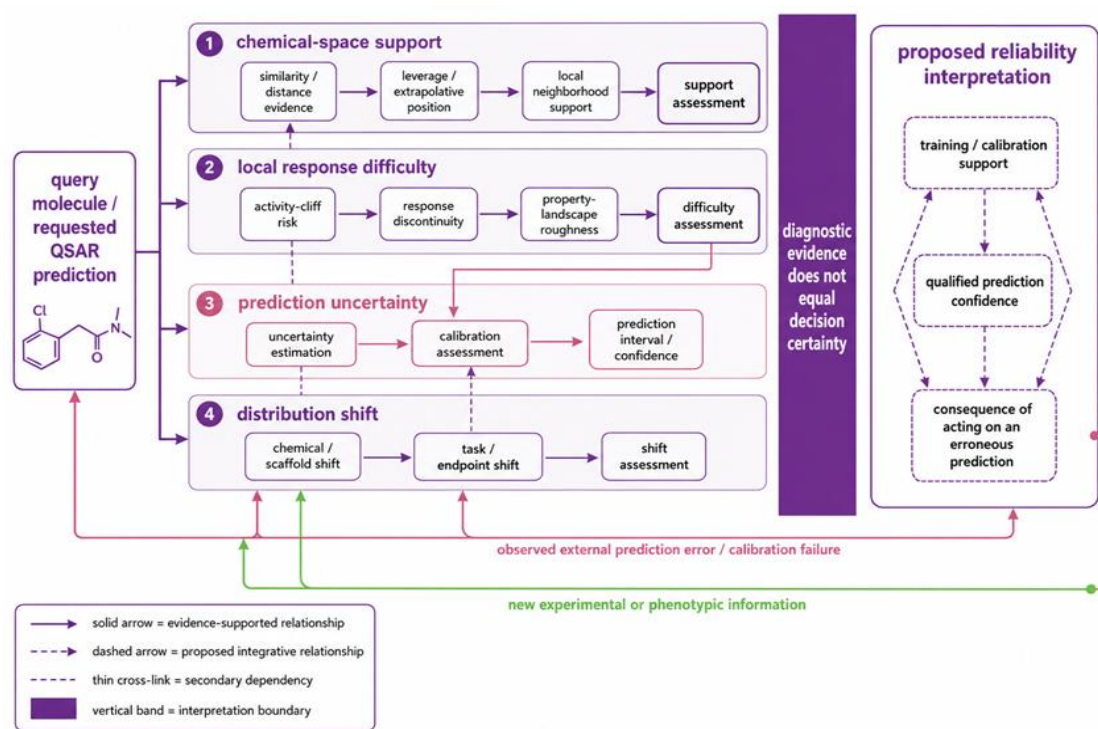


Figure 1. Where structural support stops being a sufficient proxy for QSAR prediction confidence

A risk-calibrated applicability framework

The preceding evidence suggests that the failure is not the existence of an applicability boundary but the decision to stop reasoning once that boundary has been crossed. This Perspective therefore proposes a risk-calibrated applicability framework in which reliability is assessed through three distinguishable questions: whether the query is supported by relevant training or calibration evidence, whether prediction-specific confidence has been independently qualified, and what consequence follows if the prediction is wrong.

The first question remains recognizably an applicability-domain problem. Local domain analysis, dynamic applicability and leverage behavior all show why the answer may need to be graded rather than treated as permanently binary [4, 9, 13]. The second question concerns confidence rather than membership. Hybrid uncertainty estimation, complementary UQ metrics and systematic UQ benchmarking collectively show why uncertainty should enter a decision only after its calibration and error-discrimination behavior have been examined [17, 18, 27]. The third question is explicitly proposed here: identical statistical uncertainty need not justify

identical action when one prediction is used for exploratory compound prioritization and another is used to terminate or advance a costly development decision.

Risk calibration in this sense is not a numerical scoring system. No universal probability, threshold or weighting is implied. Instead, the framework changes the order of inference. A domain flag initiates a reliability assessment; it does not conclude one. Evidence of weak support increases the burden placed on uncertainty qualification, external testing or new sampling. Strong structural support does not erase activity-cliff or local-error concerns. Low predicted uncertainty is insufficient when the uncertainty estimator has not been tested under the relevant shift.

Table 3 compresses these conditional states into a sequence-aware ledger so that the proposed framework cannot be misread as another renamed inside/outside classification.

Table 3. Decision states in the proposed risk-calibrated applicability framework

Proposed assessment state	Classification basis	Sampling or data support	Confidence evidence	Mechanistic interpretation permitted	Competing explanation that must remain open	Decision consequence	Condition that changes the state	Unresolved methodological question
Supported and qualified	Query is represented under the endpoint-specific domain and confidence procedure performs acceptably under relevant validation	Local training or calibration evidence is adequate for the intended use	Prediction-specific uncertainty has been independently evaluated	Model-supported prediction only; no mechanistic proof	Hidden local response discontinuity or untested deployment shift	Prediction may be used with ordinary model-specific qualification	New external error, shift or contradictory measurement	How much external evidence is sufficient before confidence is considered portable?
Supported but uncertain	Structural or leverage evidence is acceptable but confidence remains broad, unstable or unqualified	Chemical support exists without equivalent certainty about prediction error	UQ is weak, inconsistent or not demonstrated under the intended conditions	Structural familiarity may be stated	Model misspecification, activity cliff or calibration failure	Require stronger validation or additional measurements before consequential use	Successful recalibration or new local data	Which uncertainty diagnostics should be mandatory for high-consequence predictions?
Weakly supported but informative	Query lies near or beyond a support boundary	Training evidence is sparse or extrapolative	Confidence may still be informative if extrapolative calibration has been	Extrapolative prediction may be reported as such	Apparent confidence may reflect estimator overconfidence	Treat prediction as hypothesis-generating unless stronger evidence exists	Targeted sampling moves the query into better-supported evidence space	When can tested extrapolation be preferable to an excessively strict

		demonstrated					domain cutoff?
Shift-compromised	Original domain no longer matches deployment chemistry, endpoint or task	Historical support is not demonstrably representative of current use	Existing confidence model may also be out of domain	No transfer claim without direct testing	Distribution shift rather than intrinsic model weakness	Reassess support, calibration and workflow before relying on the prediction	How should distinct chemical, scaffold and task shifts be reported comparably?
Unresolved reliability	Support, confidence or provenance cannot be reconstructed sufficiently	Evidence base is incomplete or workflow provenance is ambiguous	Confidence cannot be qualified	No mechanistic or reliability interpretation beyond the observed output	Data leakage, implementation drift, hidden shift or model error	Abstention or new evidence is methodologically preferable	Recoverable provenance, external testing or new measurements
							When should formal abstention replace a nominal applicability-domain decision?

Table note: All assessment states and transitions are the article's proposed synthesis. They are qualitative methodological states, not validated risk categories, regulatory classifications or numerical thresholds.

Validation and reporting requirements

A domain claim is interpretable only if another investigator can reconstruct how the boundary was produced. QSPRpred illustrates that data preparation, splitting, descriptors, modeling and evaluation can be organized in an explicit reproducible QSAR/QSPR workflow [29]. Chemprop similarly demonstrates the practical feasibility of versioned, configurable molecular-prediction workflows [25]. Reproducibility is not external validity, but it is a prerequisite for diagnosing whether a changed result originates from chemistry, model configuration or implementation.

Validation should also challenge the intended mode of generalization. Benchmark structures can reward memorization [1], molecular OOD regimes require explicit definition [3], and scaffold-sensitive evaluation tests a different question from a random holdout [10]. Uncertainty estimators should likewise be examined under the shifts for which they will be interpreted [23]. No single split should therefore be mandated universally; the validation design should reproduce the chemical or task separation relevant to the intended deployment.

Reporting should keep the principal reliability objects separate. The applicability-domain method and representation should be specified independently of any prediction-interval or UQ procedure. Leverage, distance or similarity thresholds should not be described as calibrated risk limits unless evidence establishes that interpretation [13]. Uncertainty metrics should identify whether they assess calibration, ranking, sharpness or error association rather than presenting a single undifferentiated confidence statistic [18]. Frameworks for UQ benchmarking can make those distinctions reproducible [27], while conformalized molecular prediction illustrates why interval coverage and point accuracy should remain analytically separate [28].

For this Perspective, the minimum defensible reporting logic is consequently straightforward: state what evidence defines support, identify the representation and endpoint to which that support applies, document how the model was challenged beyond ordinary random resampling, report prediction confidence with its calibration evidence, disclose known local or distributional failure modes, and make the rule for abstention or further testing explicit. These are proposed reporting requirements, not an assertion of current regulatory consensus.

CONCLUSION

Applicability domains remain essential because QSAR predictions should not be interpreted without reference to the evidence from which they were learned. Their weakness appears when a scientifically conditional support statement is converted into an administrative boundary and allowed to substitute for reliability.

Chemical-space proximity, leverage, local density, response-surface regularity, uncertainty, calibration and dataset shift describe different aspects of prediction support. None should be assumed to determine the others. A structurally familiar molecule can remain difficult to predict, a nominal extrapolation can sometimes remain informative, and an uncertainty estimate can itself fail when deployment departs from the conditions under which it was qualified.

The proposed risk-calibrated interpretation therefore makes domain membership the beginning of a decision rather than its end. It separates support from confidence, requires both to be interpreted under the intended distribution, and increases evidential demands as the consequence of an erroneous prediction rises. This architecture remains a methodological proposal requiring prospective testing across endpoints, representations, model classes and realistic medicinal-chemistry shifts. Its immediate value is narrower but important: it prevents a binary applicability label from silently carrying more scientific meaning than the underlying evidence can support.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. 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
2. 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
3. Tossou P, Wognum C, Craig M, Mary H, Noutahi E. Real-world molecular out-of-distribution: Specification and investigation. *J Chem Inf Model*. 2024;64(3):697-711. doi:10.1021/acs.jcim.3c01774
4. Mora JR, Marquez EA, Pérez-Pérez N, Contreras-Torres E, Perez-Castillo Y, Agüero-Chapin G, et al. Rethinking the applicability domain analysis in QSAR models. *J Comput Aided Mol Des*. 2024;38(1):9. doi:10.1007/s10822-024-00550-8
5. Heid E, McGill CJ, Vermeire FH, Green WH. Characterizing uncertainty in machine learning for chemistry. *J Chem Inf Model*. 2023;63(13):4012-29. doi:10.1021/acs.jcim.3c00373
6. Liu R, Wallqvist A. Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. *J Chem Inf Model*. 2019;59(1):181-9. doi:10.1021/acs.jcim.8b00597
7. Berenger F, Yamanishi Y. A distance-based Boolean applicability domain for classification of high throughput screening data. *J Chem Inf Model*. 2019;59(1):463-76. doi:10.1021/acs.jcim.8b00499
8. 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
9. Oršolić D, Šmuc T. Dynamic applicability domain (dAD): Compound-target binding affinity estimates with local conformal prediction. *Bioinformatics*. 2023;39(8). doi:10.1093/bioinformatics/btad465
10. 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
11. Zhang Z, Sangion A, Wang S, Gouin T, Brown T, 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
12. Herman D, Kańduła MM, Freitas LGA, van Dongen C, Le Van T, Mesens N, et al. Leveraging Cell Painting images to expand the applicability domain and actively improve deep learning quantitative structure-activity relationship models. *Chem Res Toxicol*. 2023;36(7):1028-36. doi:10.1021/acs.chemrestox.2c00404
13. Endo S. Applicability domain of polyparameter linear free energy relationship models evaluated by leverage and prediction interval calculation. *Environ Sci Technol*. 2022;56(9):5572-9. doi:10.1021/acs.est.2c00865
14. 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

15. 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
16. Janet JP, Duan C, Yang T, Nandy A, Kulik HJ. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem Sci*. 2019;10(34):7913-22. doi:10.1039/C9SC02298H
17. Wang D, Yu J, Chen L, Li X, Jiang H, Chen K, et al. A hybrid framework for improving uncertainty quantification in deep learning-based QSAR regression modeling. *J Cheminform*. 2021;13(1):69. doi:10.1186/s13321-021-00551-x
18. Rasmussen MH, Duan C, Kulik HJ, Jensen JH. Uncertain of uncertainties? A comparison of uncertainty quantification metrics for chemical data sets. *J Cheminform*. 2023;15(1):121. doi:10.1186/s13321-023-00790-0
19. Brown TN, Sangion A, Arnot JA. Identifying uncertainty in physical-chemical property estimation with IFSQSAR. *J Cheminform*. 2024;16(1):65. doi:10.1186/s13321-024-00853-w
20. 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
21. 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
22. Aldeghi M, Graff DE, Frey N, Morrone JA, Pyzer-Knapp EO, Jordan KE, et al. Roughness of molecular property landscapes and its impact on modellability. *J Chem Inf Model*. 2022;62(19):4660-71. doi:10.1021/acs.jcim.2c00903
23. Yin T, Panapitiya G, Coda ED, Saldanha EG. Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. *J Cheminform*. 2023;15(1):105. doi:10.1186/s13321-023-00753-5
24. Yao S, Song J, Jia L, Cheng L, Zhong Z, Song M, et al. Fast and effective molecular property prediction with transferability map. *Commun Chem*. 2024;7(1):85. doi:10.1038/s42004-024-01169-4
25. Heid E, Greenman KP, Chung Y, Li SC, Graff DE, Vermeire FH, et al. Chemprop: A machine learning package for chemical property prediction. *J Chem Inf Model*. 2024;64(1):9-17. doi:10.1021/acs.jcim.3c01250
26. Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent Sci*. 2021;7(8):1356-67. doi:10.1021/acscentsci.1c00546
27. Lanini J, Huynh MTD, Scebbba G, Schneider N, Rodríguez-Pérez R. UNIQUE: A framework for uncertainty quantification benchmarking. *J Chem Inf Model*. 2024;64(22):8379-86. doi:10.1021/acs.jcim.4c01578
28. 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
29. van den Maagdenberg HW, Šicho 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