



Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Structure–Activity Models Should Predict Failure Modes as Explicitly as Potency

Maria Gonzales^{1*}, Jose Santos¹, Anna Cruz², Carlos Reyes³

¹Department of SAR Modeling and Failure Prediction, Faculty of Pharmacy, University of the Philippines Manila, Manila, Philippines.

²Department of Potency and ADME Optimization, Faculty of Pharmacy, University of Santo Tomas, Manila, Philippines.

³Department of Drug Discovery Informatics, Faculty of Pharmacy, Ateneo de Manila University, Quezon City, Philippines.

*Email: maria.gonzales@upm.edu.ph

ABSTRACT

Quantitative structure–activity relationship models are commonly judged by how accurately they predict potency, yet medicinal-chemistry decisions also depend on whether a prediction is being made under conditions in which the model is likely to fail. Aggregate predictive performance can conceal chemically local errors, assay-dependent labels, representation-specific blind spots, and losses of validity under prospective distribution shift. This Methodological Perspective argues that these failure modes should become explicit prediction targets rather than remain post hoc explanations for inaccurate potency estimates. Evidence from molecular machine learning, activity-cliff analysis, assay-aware modeling, multimodal phenotypic prediction, applicability-domain studies, and uncertainty quantification indicates that model reliability is conditional on several separable factors: chemical support, local structure–activity continuity, measurement context, representational adequacy, and deployment regime. On that basis, we develop a proposed distinction between potency prediction and failure-risk prediction. The objective is not to replace conventional QSAR metrics, nor to claim that a universal failure classifier already exists, but to reorganize validation around decision-relevant questions: what could fail, why the current prediction is exposed to that failure, and which additional evidence would discriminate between competing explanations. Such an architecture would make uncertainty more actionable and prospective model evaluation more chemically interpretable.

Keywords: QSAR, Applicability domain, Activity cliffs, Assay context, Uncertainty quantification, Out-of-distribution prediction

Received: 14 March 2025; **Revised:** 06 June 2025; **Accepted:** 08 June 2025

INTRODUCTION

Modern quantitative structure–activity relationship modeling increasingly incorporates deep neural networks, learned molecular representations, transfer learning, and richer chemical or biological data while retaining the core objective of estimating molecular properties or bioactivities from structure [1]. This expansion has produced impressive benchmark results. Deep neural networks, for example, have outperformed established methods on selected ChEMBL bioactivity tasks [2]. Yet strong benchmark potency prediction is not equivalent to knowing whether an individual medicinal-chemistry decision is well supported.

That distinction matters because apparent model superiority depends on evaluation design. Large-scale comparisons on ChEMBL demonstrate that algorithm ranking can change with data composition, compound-

series structure, and modeling protocol [3]. MoleculeNet helped standardize molecular-machine-learning evaluation and encouraged chemically informed splitting strategies [4], but no generic benchmark can reproduce every future synthesis campaign, assay environment, or distributional shift. Representation also matters: learned graph representations can be highly effective, but their relative advantage varies across tasks, data regimes, and train–test partitioning [5].

The methodological problem considered here is therefore narrower than the general question of whether machine learning can predict potency. It is whether structure–activity models should explicitly estimate how and where their own predictions may fail. We argue that potency error, chemical-space extrapolation, activity-cliff exposure, assay-label instability, representational insufficiency, and deployment shift should not automatically be collapsed into one residual error term. The proposed architecture instead treats them as partially separable failure processes that may require different evidence, different validation tests, and potentially different predictive outputs.

The potency-centric QSAR paradigm

The conventional QSAR objective is naturally potency-centric because potency or another molecular endpoint is the supervised target. Contemporary deep QSAR substantially expands how molecular information can be represented and learned [1], but the optimization target usually remains a property estimate, class probability, or ranking score. A low aggregate error therefore answers a specific question—whether predicted values resemble held-out labels under the selected validation regime—not the broader question of whether a prediction is trustworthy for a particular new molecule.

This difference becomes important when structurally similar molecules occupy different local response regimes. A representation may capture the dominant structure–property trend while remaining insensitive to rare transformations, context-specific assay effects, or unfamiliar chemical neighborhoods. Likewise, a sophisticated learned representation can perform strongly without demonstrating that its latent variables correspond to pharmacologically causal mechanisms [5]. Predictive representation and mechanistic explanation are related possibilities, not interchangeable evidential claims.

We therefore use potency-centric as an analytical description rather than a criticism of all existing QSAR practice. Applicability-domain analysis, conformal prediction, uncertainty estimation, activity-cliff modeling, and assay-aware approaches already address aspects of reliability. The methodological gap is that these outputs are rarely organized as a coordinated failure model. The perspective advanced here proposes that potency and failure risk should be treated as coupled but distinct prediction problems.

What constitutes a model failure

A model failure cannot be defined solely as a large numerical residual. Benchmark design itself can create apparent success. Ligand-based classification benchmarks may contain train–validation similarity structures that reward memorization and produce performance estimates substantially more optimistic than chemically independent evaluation would suggest [6]. In such cases, the failure is partly epistemic: the validation experiment has overstated what the model has learned.

Model complexity does not remove this problem. A large systematic evaluation of molecular-property prediction found that dataset size, molecular representation, and activity-cliff structure materially influenced performance across tens of thousands of trained models [7]. Local prediction difficulty therefore deserves separate attention from global error.

Activity cliffs provide a particularly clear example. Across curated bioactivity datasets, models that perform adequately on the overall test distribution can produce substantially larger errors for structurally similar molecules separated by large potency differences. Increasing the potency discontinuity between analogues likewise increases the difficulty of accurate local prediction, while experimental error and dataset characteristics impose additional limits on attainable predictivity [8–10]. A global RMSE can consequently coexist with a medically important local failure.

These observations motivate a proposed failure taxonomy containing at least five analytically distinct states: incorrect potency within otherwise supported chemical space; unsupported chemical extrapolation; unstable or context-incompatible labels; representation-specific insufficiency; and confidence that is poorly aligned with observed error. The categories can overlap. Their value is not that they establish a universal ontology, but that they prevent different reasons for being wrong from being hidden inside one aggregate performance statistic.

Chemical-space failure

Chemical-space failure is often discussed through applicability domains, yet local structure–activity discontinuity and global domain displacement are not identical. Activity-cliff prediction experiments show that explicitly identifying cliff-forming relationships is a different problem from ordinary molecular potency prediction [11]. Large-scale cliff analyses further indicate that increasing algorithmic complexity does not eliminate this difficulty [12]. A molecule can therefore be close to the training distribution by conventional similarity measures while remaining locally difficult because a small structural change lies on a steep response surface.

Conversely, a compound may be chemically displaced from the training distribution without necessarily belonging to an activity cliff. Deployment-oriented molecular out-of-distribution studies show that performance and calibration can deteriorate as test compounds become less representative of the training domain. Broad comparisons of OOD splitting strategies likewise demonstrate that the label “scaffold split” does not itself guarantee a severe or realistic distributional shift, while distance-aware approaches show that prediction difficulty can increase with appropriately defined displacement in structure–property space [13–15].

The implication is that chemical-space failure is better treated as multidimensional. At minimum, a model should distinguish support distance from local response roughness. The former asks whether the query resembles what the model has seen; the latter asks whether nearby analogues obey a sufficiently continuous structure–activity relationship for interpolation to be dependable. A binary inside/outside-domain flag can obscure this distinction. For medicinal-chemistry use, the proposed failure output would therefore indicate why a molecule is considered difficult rather than merely stating that it is outside an applicability domain. No universal distance threshold is justified by the present evidence. The relevant boundary remains dependent on representation, endpoint, chemical series, training distribution, and intended decision.

Assay and label failure

A structurally reasonable prediction can still fail because the supervised label is not a stable description of one biological quantity. Combining nominally corresponding IC₅₀ or K_i measurements from different experimental sources can introduce substantial noise even after apparently careful curation [16]. Bioactivity labels therefore carry experimental provenance as well as molecular information.

Assay-aware modeling strengthens this argument. Incorporating assay descriptions and biological context into proteochemometric models can improve predictive fit and reveal systematic differences among assay categories [17]. The result should not be interpreted as proof that text embeddings recover biological mechanism. It demonstrates the narrower point that assay context contains predictive information that a structure-only label may omit.

Curated resources such as Papyrus reduce avoidable heterogeneity through standardization and provenance-aware processing [18], but curation does not make biologically distinct measurements interchangeable. Likewise, probabilistic treatment of experimental uncertainty can improve classification near potency thresholds rather than forcing uncertain observations into deterministic labels [19].

The analytical need is therefore to distinguish a chemically wrong prediction from a prediction trained against an unstable or context-dependent target. The boundary conditions are summarized below.

The comparison is most efficiently represented as a register because the same numerical label can imply different failure mechanisms depending on how that label was generated and transferred.

Table 1. When a Correct Structure Meets an Unstable Label: Boundary Conditions for Attributing QSAR Failure

Failure signature	Experimental counterpart
	Prospective test
	Molecular
	Unresolved article question
Assay or label case	
Measurements pooled across nominally equivalent but operationally distinct assays	

Model residuals persist despite plausible structural similarity	Experimental counterpart Re-measurement under harmonized conditions
	Prospective test Compare predictions before and after assay-specific harmonization
	Molecular or chemical state Chemically supported query with uncertain label equivalence
	Unresolved article question Is the apparent structure–activity error primarily chemical or measurement-derived?
	Assay or label case Same endpoint name embedded in different biological assay contexts
Systematic prediction differences track assay context	Experimental counterpart Context-matched assay replication
	Prospective test Fit or evaluate an assay-conditioned model
	Molecular or chemical state Same molecule under different measurement environments
	Unresolved article question When should assay identity become part of the prediction target?
	Assay or label case Activity converted to a hard class near an uncertain threshold
Label instability concentrated close to the decision boundary	Experimental counterpart Replicate measurements or probabilistic labels
	Prospective test Compare deterministic and uncertainty-aware classification
	Molecular or chemical state Molecule near an operational potency cutoff
	Unresolved article question Should failure probability include uncertainty in the observed class itself?
	Assay or label case Curated multi-source bioactivity retained after standardization
Reduced avoidable inconsistency but residual context differences remain possible	Experimental counterpart Metadata-sensitive reanalysis
	Prospective test Stratify evaluation by assay provenance and context
	Molecular or chemical state Structurally supported compound in a heterogeneous evidence pool
	Unresolved article question How much apparent model error remains attributable to unresolved experimental heterogeneity?

Note: The table distinguishes evidence-supported sources of label instability from article-level questions proposed for future validation. It does not assign numerical failure probabilities or imply that any single assay category is intrinsically unreliable.

Mechanistic heterogeneity

Structure–activity models can also fail because the representation supplied to the model does not expose the biological or molecular variation controlling the endpoint. This is different from ordinary chemical-space extrapolation. A molecule may be well represented structurally yet remain difficult because the chosen representation omits an informative interaction geometry, cellular phenotype, transcriptional response, or platform-specific chemical behavior.

Kinase–drug modeling illustrates the first possibility. Calibrated geometric deep learning can improve binding prediction when target–ligand structural information is represented explicitly [20]. The result supports the usefulness of geometry in the evaluated kinome setting, but it does not establish that the learned representation is a complete mechanistic explanation.

At another scale, Cell Painting phenotypes and conventional chemical fingerprints provide overlapping but nonidentical predictive information. Similarity-aware merging of these modalities improved prediction across more bioactivity assays than either modality alone in the evaluated study [21]. A larger comparison of chemical structures, Cell Painting features, and gene-expression profiles similarly showed that each modality predicted a different subset of assay outcomes and that their combination expanded predictive coverage [22]. The important

methodological observation is not that one modality is universally better, but that failure can be representation-specific.

Chemical modality itself adds another boundary. Targeted protein degraders occupy a design space in which molecular size, architecture, and property relationships differ from those of many conventional small molecules; dedicated machine-learning studies show that useful prediction is possible within that space while simultaneously emphasizing the need for modality-specific modeling assumptions [23]. Transfer from conventional SAR cannot simply be presumed.

Table 2 separates these mechanistic and representational states because they can produce similar potency-level outcomes while demanding different validation experiments.

Table 2. Representation Agreement Is Not Mechanistic Agreement: Design Constraints for Interpreting Heterogeneous Structure–Activity Signals

Mechanistic-heterogeneity state	Boundary	Boundary	Validation	Decision	Interpretation
Scientific anchor	Competing explanation	Principal uncertainty source	Validation discriminator	Medicinal-chemistry consequence	Unresolved article question
Structure-only model succeeds without explicit target geometry	Correlation may reflect recurring chemical motifs rather than interaction geometry	Missing structural state or target context	Compare against geometry-aware prediction under matched data support	Potency may be useful while mechanistic interpretation remains weak	When does geometry change failure risk rather than only average accuracy?
Chemical fingerprints and cellular morphology agree	Both may encode the same underlying activity through different proxies	Redundant versus complementary information	Test modality-specific errors and concordant successes	Agreement may increase confidence but does not prove mechanism	How should agreement between modalities modify a failure estimate?
Chemical structure succeeds while morphology or transcriptomic representation fails, or vice versa	Information relevant to the endpoint may be present only in one observation channel	Representation-specific coverage	Stratified evaluation by modality and assay	A single-model rejection could discard a rescuable compound	Should failure be attached to the molecule or to the representation–task pairing?
Multiple biological representations disagree	Assay context, cellular state, or feature coverage may differ	Cross-modal discordance	Orthogonal experimental readout	Discordance can identify cases requiring mechanistic follow-up	Which disagreement patterns warrant abstention rather than model averaging?
Chemical modality shifts toward targeted degraders or other atypical architectures	Conventional descriptors may incompletely describe modality-specific property relationships	Transfer across chemical platforms	Modality-specific prospective validation	Apparent extrapolation may require a different model family	At what point does domain shift become a representation-design problem?

Note: Rows define manuscript-derived design states, not empirically validated universal categories. “Mechanistic” refers to the level of information represented or hypothesized to matter; predictive success alone is not treated as causal proof.

The same distinctions are easier to recover visually when chemical support, assay context, representation, and mechanistic interpretation remain separate rather than being collapsed into a single reliability score.

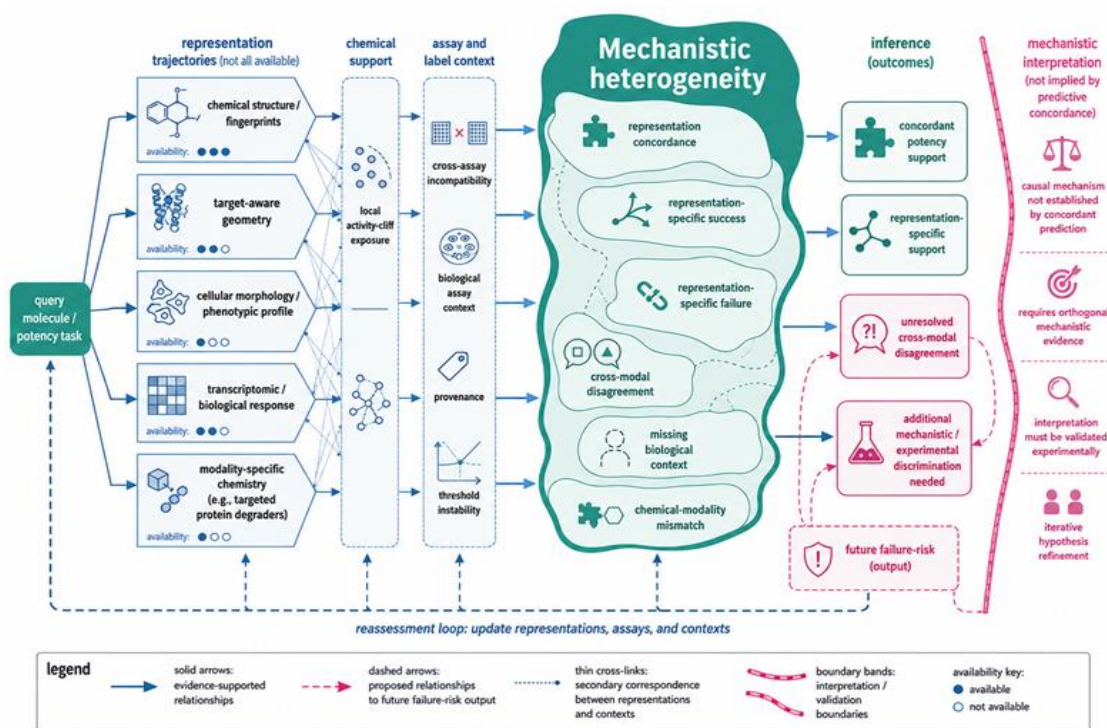


Figure 1. Mechanistic Heterogeneity Splits a Single Potency Task into Representation-Specific Failure Routes

Dataset shift and prospective failure

Retrospective model validation becomes increasingly fragile when the molecules encountered during deployment differ systematically from those used during training. The problem is not restricted to mean predictive error. Uncertainty estimates can also become poorly aligned with error as the evaluation distribution moves away from the data on which both the predictor and its uncertainty model were constructed.

Systematic evaluation of uncertainty-based active learning illustrates this dependence. No uncertainty method performed best across every evaluation dimension, and performance under out-of-distribution conditions remained particularly challenging; uncertainty-guided acquisition improved some learning tasks, but the gains were generally modest and method dependent [24]. The result cautions against treating an uncertainty score as a context-independent indicator of failure.

Transfer learning addresses a related but different problem. Molecular pretraining can improve activity prediction when labeled endpoint-specific data are limited [25]. Such gains demonstrate useful information transfer under the evaluated conditions, not automatic robustness to a future chemical series. Source–target relatedness, representation, endpoint definition, and the nature of prospective shift remain material.

For this reason, the proposed failure architecture distinguishes three validation regimes. Retrospective validation estimates performance within a pre-existing data universe. Deployment-like validation deliberately constructs chemical or temporal displacement intended to resemble future use. Prospective validation evaluates compounds, assays, or decision contexts that genuinely arise after model development. These states should not be treated as interchangeable. Scaffold or temporal splitting may approximate aspects of prospective use, but neither guarantees that the resulting test distribution reproduces the chemistry, assay conditions, or decision pressures of an actual medicinal-chemistry campaign.

A failure-aware model should therefore carry its validation context with its prediction. A potency estimate supported only by retrospective similarity should not be presented as equivalent to one whose chemical support, assay context, representation, and uncertainty behavior have survived a realistic prospective shift. The methodological task developed in the remaining sections is to convert these distinct exposures into explicit failure-related outputs without pretending that uncertainty, distance, or activity-cliff status alone already constitutes a calibrated probability of failure.

Failure-probability modelling

Predictive uncertainty is an obvious candidate for failure-aware QSAR, but uncertainty and failure probability are not interchangeable. Neural-network uncertainty methods can differ substantially in their ability to rank prediction errors, remain calibrated, or identify unfamiliar molecular regions [26]. A high uncertainty value may indicate epistemic unfamiliarity, model disagreement, noisy labels, or simply a property of the chosen estimator. None of these interpretations alone establishes the probability that a medicinal-chemistry decision will fail.

This distinction becomes clearer when uncertainty methods are evaluated against different operational objectives. Benchmarking frameworks for molecular uncertainty quantification show that methods can rank differently depending on whether the objective is calibration, error discrimination, or applicability assessment. Steep structure–activity relationships create an additional difficulty because uncertainty estimators may themselves perform poorly around activity cliffs. Conformal prediction provides a complementary output by targeting prediction-set or interval coverage under its stated assumptions rather than estimating a universal biochemical confidence value [27–29]. These outputs answer related but different questions.

A failure-aware model must therefore define its failure event before assigning a probability. Possible events include a potency error exceeding a decision-relevant tolerance, a wrong active/inactive classification, an unsupported extrapolation, failure to recognize an activity cliff, assay-context incompatibility, or unacceptable miscalibration. The appropriate event depends on how the prediction will be used. A medicinal chemist deciding which analogue to synthesize faces a different loss function from a virtual-screening system eliminating millions of compounds.

We propose that failure probability should be conditioned on several evidence channels rather than inferred from potency variance alone. The relevant inputs may include chemical-domain support, local landscape roughness, assay and label integrity, representation agreement, and the distributional relationship between development and deployment data. The proposal does not require all signals to be combined in one neural architecture. Separate models, calibrated post hoc estimators, conformal outputs, or rule-based abstention layers may sometimes provide cleaner validation.

Critically, no numerical weighting among these components is justified by the present evidence. Their relative importance is endpoint-, dataset-, representation-, and decision-dependent. **Table 3** therefore describes a sequence of failure assessment rather than assigning a universal reliability score.

Table 3. Converting Predictive Difficulty into a Decision-Relevant Failure Assessment

Training-domain support Local structure–activity continuity Assay and label integrity Representation sufficiency Predictive uncertainty and calibration					
Classification basis	Training-domain support	Local structure–activity continuity	Assay and label integrity	Representation sufficiency	Predictive uncertainty and calibration
Transfer boundary	Query becomes progressively less represented by the development chemistry	Query lies in a steep or discontinuous response neighborhood	Training and deployment measurements differ in experimental context or uncertainty	Relevant molecular, target, phenotypic, or modality-specific information is absent from the chosen input	Deployment violates assumptions under which uncertainty was estimated
Observable failure signature	Error or calibration deteriorates with chemical displacement	Similar molecules receive unexpectedly divergent observed or predicted activities	Residuals cluster by assay provenance, readout, or threshold instability	Alternative representations disagree systematically	Confidence no longer tracks error, coverage, or abstention utility
Experimental or analytical counterpart	Prospective or deployment-like evaluation across increasing domain distance	Matched-analogue or activity-cliff challenge set	Harmonized remeasurement or assay-conditioned evaluation	Orthogonal representation or multimodal comparison	Calibration, coverage, selective-risk, or error-ranking evaluation

Article-level test	Does reported confidence remain meaningful as chemical support weakens?	Can the model identify locally dangerous interpolation despite acceptable global accuracy?	Does apparent chemical failure persist after experimental context is controlled?	Is failure attributable to the molecule or to the representation–task pairing?	Is the uncertainty output valid for the decision it is being asked to support?
--------------------	---	--	--	--	--

Joint potency–failure prediction

Once failure is made an explicit target, potency prediction no longer has to carry every aspect of model reliability implicitly. Activity-landscape methods demonstrate that local discontinuity can be characterized separately from the predicted activity value itself [30]. Such measures are not failure probabilities, but they illustrate that the geometry of the response landscape contains information that a potency estimate alone does not expose.

Recent activity-cliff modeling makes the separation more concrete. Task-specific models can be trained to predict cliff-forming relationships rather than treating cliffs only as unusually large potency residuals [31]. This reframes an important failure mode as an explicit learning objective. Likewise, transferring representations learned from activity-cliff prediction into drug–target interaction modeling has shown that a failure-focused auxiliary task can contribute information to a broader activity-prediction problem [32]. These results do not establish a universally superior multitask architecture, but they provide a methodological basis for asking whether failure-relevant supervision can improve primary molecular prediction.

We therefore propose a dual-output structure–activity model. One output estimates the molecular endpoint of interest. A second output, or set of outputs, estimates defined failure states such as unsupported extrapolation, local cliff exposure, assay-context mismatch, representational disagreement, or calibrated probability of exceeding a prespecified prediction-error event. The outputs may share molecular representations while retaining different objective functions and validation criteria.

Separating the targets has several advantages at the level of interpretation. A compound could receive a favorable potency prediction but high failure risk because it occupies a steep local activity landscape. Another could receive moderate potency with low failure risk because it lies within a densely sampled, internally consistent analogue series. A third could be chemically well supported but flagged because its activity label derives from an assay context that differs materially from the training evidence. Such distinctions are more informative than a single confidence value whose biological or statistical meaning is unspecified.

Joint prediction also creates new failure possibilities. Multitask learning can produce negative transfer when the failure task and potency task emphasize incompatible patterns. A failure classifier trained on historical errors may reproduce the biases of historical medicinal-chemistry selection. Rare but consequential failure states may be inadequately represented. The failure output itself can also become miscalibrated under distribution shift. Explicit failure prediction therefore requires validation at least as rigorous as the potency model it accompanies.

The architecture should accordingly remain modular. In some settings, an integrated multitask model may be appropriate; in others, an independent applicability-domain model, activity-cliff detector, assay-context layer, and calibrated uncertainty estimator may be preferable. The central methodological claim is not that one architecture should dominate, but that failure should become an explicit object of prediction and validation rather than an explanation supplied only after a potency model has already been wrong.

Validation metrics for decision use

A failure-aware structure–activity model cannot be evaluated adequately by adding another aggregate metric to RMSE, MAE, correlation, or AUC. Conventional metrics remain necessary because the primary potency task must still be accurate. They are insufficient because chemically important errors can be concentrated in specific local or distributional regimes. Activity-cliff analyses demonstrate that acceptable global prediction can coexist with severe analogue-level errors [9]. Performance should therefore be stratified where scientifically justified.

Chemical-space evaluation should similarly describe the actual distributional relationship between training and test chemistry. Deployment-oriented OOD evaluation shows that performance loss depends on how the shift is constructed [13]. Broad OOD benchmarking further demonstrates that a named splitting method does not automatically create the same degree of extrapolation across datasets [14]. Validation reports should therefore characterize support, distance, scaffold novelty, or other relevant shift properties rather than treating “scaffold split” as a universal proxy for prospective difficulty.

Failure outputs require their own validation targets. An uncertainty score intended to rank risky compounds should be tested for error discrimination. A probability intended to represent a defined failure event should be tested for calibration. A conformal prediction system should be evaluated for coverage and efficiency under conditions compatible with its assumptions. An abstention mechanism should be evaluated by selective risk or another decision-relevant measure. A model used to prioritize new experiments may additionally require prospective assessment of acquisition utility. These objectives should not be collapsed into one generic uncertainty score.

Assay and biological context also require stratification where sufficient metadata exist. Cross-assay variation can contribute directly to apparent prediction error [16], meaning that overall error may partly measure uncontrolled measurement heterogeneity. Reporting performance separately by assay context, provenance, or readout can help distinguish chemical-model failure from label-system failure.

Validation should further include deliberately difficult molecular subsets rather than only average-case test sets. Activity cliffs, low-support regions, modality shifts, and representation disagreements are useful because they challenge specific claims the model makes about reliability. The purpose is not to construct an impossibly adversarial benchmark, but to test whether the stated failure output actually responds to the condition it claims to detect.

For decision use, calibration must also be linked to a declared consequence of error. A probability has no universal medicinal-chemistry meaning without specifying the event whose probability is estimated. The acceptable error for deciding whether to synthesize an analogue may differ from the acceptable error for removing a compound from a development program. Accordingly, no universal confidence threshold or failure-probability cutoff is supported by the present evidence.

A practical validation hierarchy would therefore report: primary endpoint accuracy; chemically stratified performance; local activity-cliff behavior; assay- or context-stratified error where possible; robustness under deployment-relevant distribution shift; calibration or coverage of failure-related outputs; and, ultimately, prospective decision performance. The hierarchy is proposed rather than empirically established. Its purpose is to make the evidential burden explicit. A model should not claim to predict failure merely because it produces an uncertainty number; it should demonstrate that the output identifies a predeclared failure event under the conditions in which the model will be used.

CONCLUSION

Structure–activity modeling has become increasingly capable of predicting molecular endpoints, but better potency prediction does not eliminate the need to know when the prediction is unreliable. The evidence considered here shows that failure can arise from chemically local discontinuity, weak domain support, incompatible or uncertain assay labels, incomplete representations, and changes between development and deployment distributions. These processes are related but not interchangeable.

The central proposal of this Perspective is therefore to elevate failure from a retrospective explanation to an explicit prediction target. Potency and failure risk should be modeled as coupled outputs with distinct meanings, evidence requirements, and validation criteria. Failure estimates should specify what event is being predicted, what evidence makes that event plausible, and under what chemical, experimental, representational, and deployment conditions the estimate has been calibrated.

This proposal does not establish a universal failure classifier, probability threshold, or preferred machine-learning architecture. Nor does it imply that uncertainty quantification, activity-cliff analysis, applicability domains, assay-aware modeling, or multimodal prediction are individually sufficient. Their value is complementary: together they reveal why a numerically plausible potency estimate can still be decisionally fragile.

A more useful structure–activity model would therefore answer two questions at once: what activity should be expected, and what could make that expectation fail? Prospective testing of that distinction—especially under new chemical series, changed assay conditions, steep local structure–activity relationships, and representation shifts—should determine whether explicit failure prediction improves medicinal-chemistry decisions beyond potency prediction accompanied by generic confidence measures.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: The emergence of deep QSAR. *Nat Rev Drug Discov.* 2024;23(2):141-55. doi:10.1038/s41573-023-00832-0
2. Lenselink EB, ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, et al. Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform.* 2017;9(1):45. doi:10.1186/s13321-017-0232-0
3. 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
4. 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
5. 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
6. 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
7. 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
8. 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
9. Janela T, Bajorath J. Anatomy of potency predictions focusing on structural analogues with increasing potency differences including activity cliffs. *J Chem Inf Model.* 2023;63(22):7032-44. doi:10.1021/acs.jcim.3c01530
10. Sheridan RP, Karnachi P, Tudor M, Xu Y, Liaw A, Shah F, et al. Experimental error, kurtosis, activity cliffs, and methodology: What limits the predictivity of quantitative structure-activity relationship models? *J Chem Inf Model.* 2020;60(4):1969-82. doi:10.1021/acs.jcim.9b01067
11. Dablander M, Hanser T, Lambiotte R, Morris GM. Exploring QSAR models for activity-cliff prediction. *J Cheminform.* 2023;15(1):47. doi:10.1186/s13321-023-00708-w
12. Tamura S, Miyao T, Bajorath J. Large-scale prediction of activity cliffs using machine and deep learning methods of increasing complexity. *J Cheminform.* 2023;15(1):4. doi:10.1186/s13321-022-00676-7
13. 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
14. Fooladi H, Vu TNL, Mathea M, Kirchmair J. Evaluating machine learning models for molecular property prediction: Performance and robustness on out-of-distribution data. *J Chem Inf Model.* 2025;65(19):9871-91. doi:10.1021/acs.jcim.5c00475
15. Kim JY, Vlachos DG. Distance-aware molecular property prediction in nonlinear structure-property space. *J Chem Inf Model.* 2025;65(13):6744-56. doi:10.1021/acs.jcim.5c01037
16. Landrum GA, Riniker S. Combining IC50 or Ki values from different sources is a source of significant noise. *J Chem Inf Model.* 2024;64(5):1560-67. doi:10.1021/acs.jcim.4c00049
17. Schoenmaker L, Sastrokarijo EG, Heitman LH, Beltman JB, Jespers W, van Westen GJP. Toward assay-aware bioactivity model(er)s: Getting a grip on biological context. *J Chem Inf Model.* 2025;65(13):7013-23. doi:10.1021/acs.jcim.5c00603
18. Béquignon OJM, Bongers BJ, Jespers W, IJzerman AP, van der Water B, van Westen GJP. Papyrus: A large-scale curated dataset aimed at bioactivity predictions. *J Cheminform.* 2023;15(1):3. doi:10.1186/s13321-022-00672-x
19. 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
20. Luo Y, Liu Y, Peng J. Calibrated geometric deep learning improves kinase-drug binding predictions. *Nat Mach Intell.* 2023;5(12):1390-401. doi:10.1038/s42256-023-00751-0

21. Seal S, Yang H, Trapotsi MA, Singh S, Carreras-Puigvert J, Spjuth O, et al. Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data. *J Cheminform.* 2023;15(1):56. doi:10.1186/s13321-023-00723-x
22. Moshkov N, Becker T, Yang K, Horvath P, Dancik V, Wagner BK, et al. Predicting compound activity from phenotypic profiles and chemical structures. *Nat Commun.* 2023;14(1):1967. doi:10.1038/s41467-023-37570-1
23. 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
24. 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
25. Li X, Fourches D. Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. *J Cheminform.* 2020;12(1):27. doi:10.1186/s13321-020-00430-x
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. 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. 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
29. 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
30. López-Pérez K, Miranda-Quintana RA. iCliff Taylor's version: Robust and efficient activity cliff determination. *J Chem Inf Model.* 2025;65(11):5801-10. doi:10.1021/acs.jcim.5c00506
31. Niu D, Deng Z, Wang X, Li Z. Task-specific activity cliff prediction method based on transfer learning and a hyper connection graph model. *J Chem Inf Model.* 2025;65(15):8345-59. doi:10.1021/acs.jcim.5c00949
32. Ibragimova R, Iliadis D, Waegeman W. Enhancing drug-target interaction prediction through transfer learning from activity cliff prediction tasks. *J Chem Inf Model.* 2025;65(13):6558-67. doi:10.1021/acs.jcim.5c00484