Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2025, 14(1):58-66
https://doi.org/10.51847/lCrWzgNVQy

Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

What Counts as External Validation in QSAR Modelling across Scaffolds, Assays, Laboratories, Chemical Series, and Historical Time?

Mark Anderson[1]*, Lisa Wong[2], Sarah Lee[1]

[1]Department of QSAR Validation and External Applicability, College of Pharmacy, University of Florida, Gainesville, United States.
[2]Department of Cross-Assay and Cross-Laboratory Validation, Faculty of Pharmacy, National University of Singapore, Singapore.

***Email:** mark.anderson@ufl.edu*

ABSTRACT

External validation is frequently treated as a binary property of quantitative structure–activity relationship models, although evaluation sets can differ from development data along chemically and experimentally distinct dimensions. A holdout may be structurally distant yet generated in the same assay and laboratory, chronologically later yet dominated by familiar chemical series, or externally sourced while retaining substantial analogue overlap. This methodological standards article proposes a multiaxial framework that distinguishes structural-distance, scaffold, chemical-series, assay, protocol, laboratory, temporal, and population or use-context externality. These dimensions are combined with cross-cutting requirements for provenance integrity, leakage control, calibration, predictive uncertainty, representation stability, and claim traceability. The framework introduces claim–validation alignment, non-compensation, weakest-relevant-dimension, provenance-break, prospective-lock, recalibration, and failure-localization rules. Minimum external-validation standards are defined around transparent split construction, independent data generation where relevant, comparison with less demanding partitions, external-set calibration and uncertainty assessment, subgroup analysis, and explicit restriction of the resulting claim. Stress tests and negative controls are positioned as mechanisms for revealing hidden dependence, leakage, activity-cliff sensitivity, and instability under reasonable analytical alternatives. A reporting checklist is proposed to make each claimed form of externality auditable without converting reporting completeness into a guarantee of validity. The contribution is an original conceptual synthesis rather than an empirically validated scoring system. It requires prospective and interlaboratory evaluation, comparison across endpoints and chemical programs, and investigation of how validation dimensions interact. It does not establish causal explanation, clinical benefit, regulatory acceptance, universal applicability, or deployment readiness.

Key words: *QSAR, Pharmacoinformatics, Chemical space, External validation, Applicability domain, Calibration*
Received: 04 November 2024; **Revised:** 12 February 2025; **Accepted:** 19 February 2025

INTRODUCTION

Quantitative structure–activity relationship modelling converts molecular information into predictions intended to support prioritization, interpretation, or experimental planning. Its credibility therefore depends not only on model construction but also on data curation, applicability-domain definition, validation design, interpretability, and the intended scientific claim [1]. Yet the expression “external validation” often compresses these distinct requirements into a single label, leaving unclear which dependence between development and evaluation data has actually been broken.

This ambiguity matters because molecular prediction is embedded in a staged drug-discovery process. A model used to rank analogues within a familiar series faces a different evidentiary problem from one used to prioritize compounds for a new target, assay platform, laboratory, or historical period. Evaluation should therefore be interpreted relative to the decision and development context rather than as an isolated algorithmic property [2]. The same performance statistic can support materially different conclusions depending on how compounds, measurements, and experimental provenance were separated.

Interpretability does not resolve this problem. Explanations may expose influential descriptors, substructures, or learned features, but an interpretable prediction can still be poorly calibrated or non-transportable. Conversely, a predictive model can generalize within a declared domain without supplying a causal or mechanistic explanation. Explainability and external validity should therefore be reported as related but non-substitutable properties [3].

This article develops an original methodological standards framework for describing what counts as external validation across structural, experimental, temporal, and use-context dimensions. It is a non-empirical conceptual contribution. The objective is not to identify a universally superior split or metric, but to align validation evidence with the claim being made, specify minimum reporting and testing requirements, and define boundaries beyond which additional experimental or prospective evidence is necessary.

The ambiguity of external validation in QSAR

A dataset can be "external" in location while remaining internally familiar in chemical content. Ligand-based benchmarks may preserve close analogues or recurring series across partitions, allowing models to exploit local similarity rather than demonstrate broader generalization [4]. Record-level separation is therefore insufficient when the intended claim concerns new scaffolds, new medicinal-chemistry series, or more distant regions of chemical space.

Split terminology also conceals differences in task difficulty and meaning. Random, scaffold, stratified, temporal, and task-specific partitions do not test the same hypothesis, even when the same endpoint and metric are used. Benchmark design accordingly determines whether performance primarily represents interpolation, structural extrapolation, chronological transfer, or another form of dependence [5]. Calling all resulting test sets "external" removes information needed to interpret the evidence.

The ambiguity is compounded because performance emerges from interacting factors. Dataset quality, molecular representation, model family, split construction, and local discontinuities such as activity cliffs can alter apparent robustness [6]. Poor performance on a difficult split may reflect genuine distribution shift, limited sample support, label noise, or representation inadequacy. Strong performance may likewise reflect transferable structure–property information or residual overlap. Externality cannot be inferred from the metric alone.

The term should therefore be reserved for a documented claim about independence. At minimum, authors should state what was external, what remained shared, how independence was operationalized, and which intended use the evaluation approximates. This interpretation does not imply that every dimension must always be crossed. It requires that the dimensions relevant to the stated claim be identified rather than hidden behind a binary label.

Validation dimensions and proposed terminology

Predictive uncertainty is a cross-cutting validity condition because external prediction is not adequately described by average discrimination or error alone. Molecular uncertainty estimators differ in their ability to identify unreliable predictions, and their behaviour must be assessed empirically in the relevant task and data regime [7]. An internal uncertainty estimate should not be assumed to remain meaningful after a substantial structural or experimental shift.

Comparative evaluations further show that uncertainty methods can differ in calibration, sharpness, and error-ranking behaviour, without one estimator dominating across all settings [8]. Uncertainty should therefore be treated as an evaluated property of a model–dataset–shift combination. It is not a certificate attached permanently to an algorithm.

Its practical meaning also depends on the decision. An uncertainty estimate used to defer compounds, trigger measurement, support active learning, or communicate risk requires different evidence from one used only for descriptive analysis [9]. Calibration, predictive uncertainty, and applicability domain are therefore related but distinct: calibration compares probabilities or intervals with observed frequencies, uncertainty expresses prediction-specific doubt, and applicability domain defines where model use is scientifically defensible.

Across molecular-property tasks, uncertainty quality is method-, dataset-, and shift-dependent rather than guaranteed by the estimator class [7–9]. Building on this evidence, the proposed framework defines eight

externality questions: structural distance; scaffold independence; chemical-series independence; assay independence; protocol independence; laboratory independence; temporal independence; and population or use-context transportability. These dimensions form an externality profile, not an additive score. Strength on one relevant dimension cannot compensate for absence on another.

Figure 1 presents the original conceptual synthesis for multiaxial external-validation framework for QSAR models, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

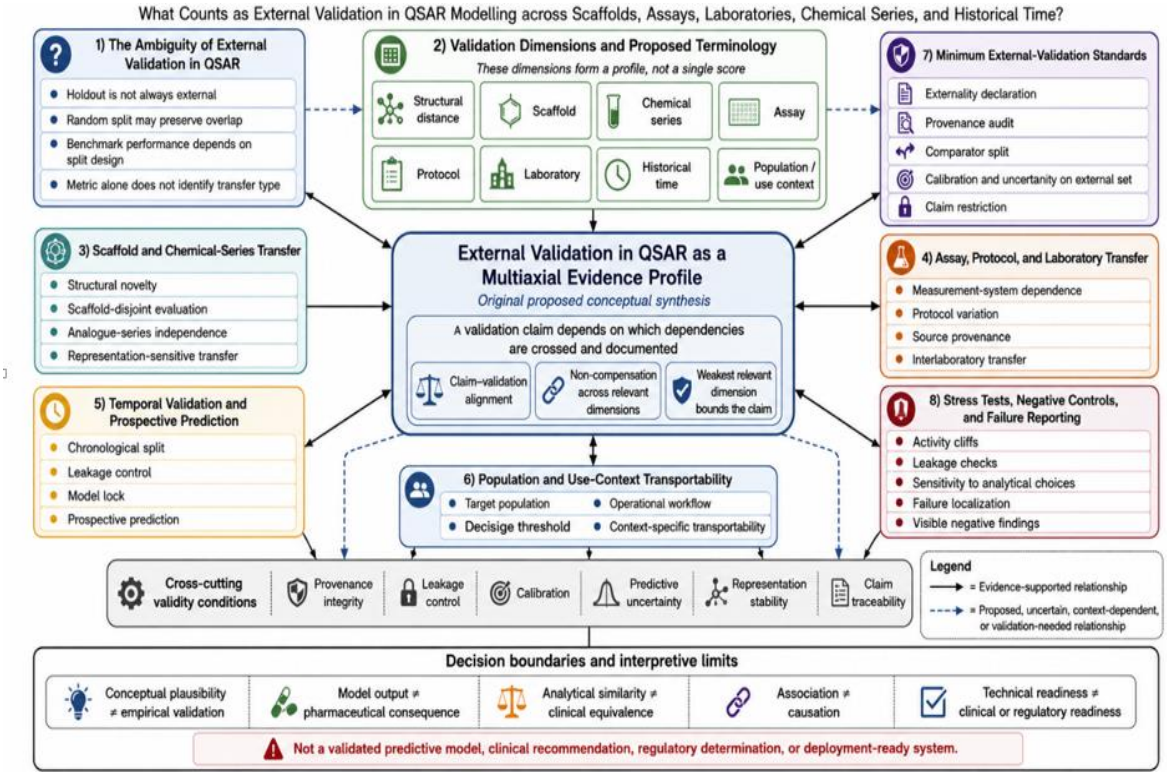

Figure 1. Multiaxial External-Validation Framework for QSAR Models

The framework separates externality dimensions from validity conditions. Provenance integrity, leakage control, external-set calibration, uncertainty assessment, representation sensitivity, and claim traceability apply across the profile but do not themselves establish that a particular structural or experimental dependency was crossed. The claim–validation alignment rule permits only claims corresponding to documented dimensions. The non-compensation and weakest-relevant-dimension rules prevent aggregate strength from concealing a missing decision-relevant test. These rules are proposed for empirical evaluation and do not constitute validated thresholds.

Scaffold and chemical-series transfer

Structural transfer begins with representation but cannot be reduced to it. Learned and engineered representations may perform differently across endpoints, dataset sizes, and chemical spaces [10]. A representation that is effective for interpolation within densely sampled analogues may not preserve the relationships needed for a novel scaffold or sparsely represented chemotype. Representation quality must therefore be assessed together with the transfer claim and split design.

Cluster-disjoint evaluation can impose stronger separation than random partitioning by reducing direct neighbourhood overlap [11]. However, cluster membership depends on fingerprints, similarity measures, dimensionality reduction, clustering choices, and thresholds. A cluster split is consequently evidence of a declared operational separation, not proof that compounds are mechanistically novel or outside every relevant analogue series.

Comparisons between descriptor-based and graph-based models likewise show that architecture rankings depend on dataset and evaluation context [12]. Graph representation does not automatically confer scaffold transfer, while conventional descriptors do not necessarily restrict a model to memorization. The relevant question is whether the

selected representation and model remain reliable under the specific structural dependency that the validation design claims to break.

The proposed terminology distinguishes structural-distance, scaffold, and chemical-series externality. Structural-distance externality concerns neighbourhood separation under declared representations and similarity measures. Scaffold externality concerns separation under a stated scaffold definition. Chemical-series externality concerns independence from analogue programs sharing design history, optimization logic, or recurring substitutions, even when scaffold labels differ. These forms are related but non-equivalent. None alone establishes assay, laboratory, temporal, causal, clinical, or regulatory transportability.

Assay, protocol, and laboratory transfer

Performance on curated bioactivity data may remain dependent on the development assay. A benchmark can separate records while retaining compound overlap, endpoint conventions, or shared measurement provenance [13]. Assay externality therefore concerns independence of the measurement system, not merely the storage location or ownership of the evaluation records.

Transitions between biological assays can change model behaviour even when the target and chemistry appear related [14]. Assay principle, biological material, readout, concentration range, censoring, and endpoint transformation may alter the response being predicted. Validation across assays should document comparability and unresolved measurement differences.

Curated resources improve consistency but do not erase source heterogeneity [15]. Protocol details, laboratory practices, aggregation rules, and publication histories may remain incomplete. Laboratory externality should be claimed only when outcome generation is organizationally independent and sufficiently documented to distinguish transfer from reused measurements.

Curated benchmarks, assay-transition studies, and harmonized resources collectively show that assay provenance must remain visible when transfer is claimed [13–15]. The proposed provenance-break rule records assay, protocol, and laboratory independence separately. Crossing one level does not establish the others, interlaboratory reproducibility, or endpoint interchangeability.

Temporal validation and prospective prediction

Temporal validation asks whether a model predicts observations generated after development. Its validity depends on the timestamp and on controls preventing compounds, tasks, or related measurements from crossing the boundary. Studies should identify whether dates represent registration, assay completion, publication, or deposition and should control molecule- and task-level leakage [16].

A chronological split can estimate a different problem from a random split because chemistry programs, assays, and target knowledge change over time [17]. Chronology alone, however, does not guarantee structural novelty or laboratory independence. Later records may contain familiar series, retrospective corrections, or earlier-generated measurements.

Prospective prediction requires the model, preprocessing, parameters, and relevant decision rules to be fixed before outcomes become available. Industrial prospective evaluation demonstrates how subsequent experimental outcomes can reveal performance change and the need for monitoring and retraining policies [18]. Retrospectively reconstructed time splits should not be labelled prospective.

The proposed prospective-lock rule distinguishes retrospective temporal validation, pseudo-prospective reconstruction, and locked prospective prediction. Reports should state the cutoff logic, frozen components, outcome-availability date, retraining policy, and post-prediction intervention. A time split is not prospective merely because its test observations are later.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, components, relationships, and intended uses in what counts as external validation in qsar modelling across scaffolds, assays.

Table 1. Definitions, components, relationships, and intended uses in What Counts as External Validation in QSAR Modelling across Scaffolds, Assays.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement

Structural-distance externality	Hidden neighbourhood overlap	Representation and similarity measure	Declared distance limits direct analogue dependence	Separates interpolation from distance transfer	Distance definition and overlap audit [10]	Representation-dependent result	Does not establish scaffold or assay transfer
Scaffold externality	Shared cores across partitions	Scaffold definition	Core-disjoint testing limits declared core reuse	Makes scaffold independence auditable	Assignment method and exceptions [11]	Definitions alter membership	Does not guarantee series independence
Chemical-series externality	Analogue programs cross partitions	Series labels and design history	Series-disjoint testing limits medicinal-chemistry continuity	Approximates unfamiliar-program transfer	Documented rules and blinded assignment	Boundaries may be subjective	Does not prove mechanistic novelty
Assay externality	Measurement-system dependence	Assay principle and endpoint	Independent assay generation tests measurement transfer	Separates chemical from assay transfer	Comparability analysis [14]	Endpoint mismatch may dominate error	Does not establish protocol independence
Protocol externality	Procedural variation	Reagents, instruments, timing, censoring	Protocol change may shift labels	Identifies operational shift	Protocol metadata [15]	Metadata may be incomplete	Does not establish laboratory transfer
Laboratory externality	Shared organizational provenance	Site, personnel, instrumentation	Independent generation reduces common-process dependence	Tests interorganizational transfer	Verifiable provenance [15]	Shared pipelines blur independence	Does not imply clinical transportability
Temporal externality	Historical change	Cutoff event and model state	Later observations test historical transfer	Approximates future-data prediction	Timestamp and leakage audit [17]	Later chemistry may remain familiar	Does not equal prospective validation
Locked prospective prediction	Retrospective reconstruction bias	Pre-outcome model lock	Prediction precedes outcome availability	Provides future-facing evidence	Immutable procedure and later outcomes [18]	Post hoc tuning invalidates claim	Does not establish decision benefit
Use-context transportability	Evaluation–use mismatch	Population, endpoint, threshold, workflow	Validation approximates intended operation	Aligns evidence with use	Context-specific assessment [19]	Context may change	Does not establish implementation readiness
Externality profile	Binary external label	Declared dimensions and controls	Proposed non-compensatory profile records crossed dependencies	Makes claims auditable	Dimension-specific evidence	Aggregate language may conceal gaps	Proposed synthesis, not a score

Population and use-context transportability

External validation should address the compounds, targets, assays, and decisions to which a model will be applied. Few-shot learning can assist low-data tasks, but success remains conditional on task construction and shared information between source and target settings [20]. Limited target data do not themselves make an evaluation external.

Transfer learning depends on source–target relatedness, available target observations, and label alignment [19]. Successful fine-tuning may show computational transfer while leaving assay, laboratory, temporal, or decision-context independence untested. Reports should identify what knowledge was transferred and which dependencies were newly evaluated.

Pretrained chemical-language models can provide reusable molecular representations [21]. Pretraining also broadens the provenance problem because large corpora may contain evaluation compounds or related chemistry. Downstream studies should disclose pretraining sources where possible, assess contamination risk, and validate the final task-specific model in its intended context.

The proposed use-context dimension asks whether the validation population, decision threshold, workflow, and consequences of error approximate intended use. Representation transfer or low-data adaptation does not independently establish clinical, organizational, or operational transportability.

Minimum external-validation standards

Minimum standards should coordinate data quality, split logic, metrics, reproducibility, and scientific interpretation [22]. The first requirement is an externality declaration stating which dimensions matter to the claim, which were crossed, and which dependencies remain shared.

The second requirement is a provenance and leakage audit covering duplicates, variants, close analogues, series overlap, shared measurements, timestamp integrity, and preprocessing fitted across partitions. Because robustness varies across out-of-distribution designs, studies should justify why the external set approximates the intended shift [23].

The third requirement is decision-relevant evaluation on the external set, including appropriate error or discrimination measures, calibration, predictive uncertainty, and meaningful strata. A less demanding comparator split should show how conclusions change with the externality profile, while the final external set remains isolated from model selection.

The fourth requirement is claim restriction: authors should state the strongest supported claim, untested dimensions, alternative explanations, and required next evidence. This is a proposed scholarly standard, not a regulator-approved protocol or guarantee of successful use.

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for testable propositions, evidence requirements, and validation criteria for what counts as external validation in qsar modelling across scaffolds, assays.

Table 2. Testable propositions, evidence requirements, and validation criteria for What Counts as External Validation in QSAR Modelling across Scaffolds, Assays.

Architecture layer or process stage	Core function	Information or material flow	Interaction with other components	Validation criterion	Potential failure mode	Human or experimental responsibility	Readiness boundary
Claim specification	Define intended prediction and use	Question to explicit claim	Selects relevant dimensions	Population, endpoint, decision, and error consequences stated	Inflated intended use	Domain scientist and modeller	No claim before use is specified
Provenance mapping	Trace compounds and outcomes	Records to auditable lineage	Enables leakage and independence checks	Series, assay, laboratory, and time provenance documented	Hidden common-origin data	Data steward and experimental owner	Missing provenance limits claims
Split construction	Break declared dependencies	Curated data to partitions	Implements externality design	Split matches the stated dimension	Convenience split tests another hypothesis	Modeller with domain review	Holdout alone is insufficient
Model lock	Prevent post-outcome adaptation	Final pipeline to immutable procedure	Required for prospective claims	Pipeline frozen before outcomes	Post hoc tuning	Model owner and custodian	Otherwise claim is retrospective
External evaluation	Measure performance under shift	External outcomes to metrics	Includes calibration and uncertainty	Metrics and relevant strata reported	Aggregate results conceal failure	Independent analyst where feasible	Does not establish causality or utility
Comparator analysis	Characterize shift sensitivity	Internal and external estimates	Interprets split severity	Same pipeline compared across partitions	Unequal tuning	Modeller and statistician	Not a universal baseline
Stress and negative controls	Challenge robustness	Cliffs, perturbations,	Tests leakage and instability	Expected response to controls	Misleadingly strong performance	Independent reviewer or validation team	Passing does not prove reliability

		alternative splits					
Failure localization	Identify transfer breakdown	Errors to dimension-specific strata	Restricts claims and guides experiments	Failures reported by series, assay, time, and uncertainty	Selective omission	Modeller and domain expert	Analysis does not validate remedies
Claim–validation alignment	Bound interpretation	Evidence profile to statement	Applies proposed non-compensation rule	Claim covers documented dimensions only	Strong axis masks missing axis	Authors, reviewers, decision owners	Proposed rule, not a score
Evidence escalation	Define next validation step	Residual uncertainty to further testing	Connects computational and experimental evidence	Next study addresses unresolved dependency	Premature implementation	Experimental or organizational owner	No deployment readiness inferred

Stress tests, negative controls, and failure reporting

External validation should include tests designed to reveal weakness. Activity cliffs are informative because small structural changes can accompany large response changes, challenging smooth similarity assumptions [24]. Performance in cliff-containing regions can expose local failure hidden by global averages.

Negative controls should target leakage and inadvertent dependence. Duplicate removal, reconstructed random splits, label permutation where appropriate, timestamp checks, and near-neighbour exclusion can reveal whether transfer depends on unavailable information. Leakage can invalidate estimates despite nominal train–test separation [25].

Robustness analysis should vary reasonable choices, including representation, preprocessing, scaffold definition, series assignment, and uncertainty estimator. Sensitivity does not automatically invalidate a model, but it restricts generalization claims. Controls should be planned before interpreting the final external result whenever feasible.

Failures should be localized by externality dimension, chemical series, assay source, period, uncertainty level, and decision-relevant subgroup. Negative findings, calibration failure, unstable rankings, and unsuccessful transfer should remain visible. Passing a stress test does not establish safety, causality, or prospective benefit.

Reporting checklist for Q1 pharmacoinformatics studies

A metric table cannot make externality auditable. Reports should identify data sources, outcome construction, split logic, optimization, model specification, and evaluation design, consistent with structured recommendations for supervised machine-learning validation [26].

Reproducibility also requires transparent data processing, software and model versions, randomization, computational environments, and evaluation code [27]. Where data cannot be released, authors should provide enough provenance and procedural detail for independent reconstruction or bounded assessment.

The proposed checklist reports intended use, externality profile, split algorithm, compound and series overlap, assay and laboratory provenance, timestamp definition, model-selection isolation, calibration, uncertainty, comparator partitions, stress tests, subgroup failures, and untested dependencies. Leakage prevention, structured reporting, and reproducibility documentation are complementary rather than substitutable safeguards [25–27].

Checklist completion should not become a quality score. Transparency makes a claim auditable but cannot establish that the external set is representative, the model is scientifically correct, or downstream decisions will improve.

Boundary conditions and unresolved problems

The framework applies primarily to predictive QSAR and related pharmacoinformatics models. Drug-design claims ultimately require evidence appropriate to experimental and translational readiness, not computational performance alone [28]. The framework cannot replace synthesis, assay confirmation, pharmacology, safety assessment, or decision-impact evidence.

Predictive transportability is also distinct from causal explanation. Reliable association-based prediction can occur without identifying an intervention mechanism, whereas causal inference requires additional assumptions and

designs [29]. Feature attribution, counterfactual explanation, or scaffold transfer should not be called causal validation without such evidence.

Unresolved problems include correlated dimensions, uncertain chemical-series boundaries, incomplete assay metadata, and undocumented pretraining exposure to evaluation chemistry. Continual retraining adds further ambiguity because an updated model has no single fixed validation state.

The proposed rules should be tested across endpoints, laboratories, periods, representations, and decision settings. Until such evidence exists, the framework remains a claim-organizing and reporting proposal, not a validated maturity model, universal standard, or regulatory pathway.

CONCLUSION

External validation in QSAR should identify which structural, experimental, temporal, and use-context dependencies were crossed rather than label any holdout as external. The proposed framework separates these dimensions while treating provenance, leakage, calibration, uncertainty, and claim traceability as cross-cutting conditions. Its decision rules are intended to make validation claims precise and auditable, but require empirical testing and do not establish causality, experimental success, clinical value, regulatory acceptance, or deployment readiness.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. Chem Soc Rev. 2020;49(11):3525-64. doi:10.1039/D0CS00098A
2. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
3. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84. doi:10.1038/s42256-020-00236-4
4. 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
5. 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
6. 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
7. 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
8. 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
9. Mervin LH, Johansson S, Semenova E, Giblin KA, Engkvist O. Uncertainty quantification in drug design. Drug Discov Today. 2021;26(2):474-89. doi:10.1016/j.drudis.2020.11.027
10. 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
11. Guo Q, Hernandez-Hernandez S, Ballester PJ. UMAP-based clustering split for rigorous evaluation of AI models for virtual screening on cancer cell lines. J Cheminform. 2025;17(1):94. doi:10.1186/s13321-025-01039-8

12. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform. 2021;13(1):12. doi:10.1186/s13321-020-00479-8
13. 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
14. Arvidsson McShane S, Ahlberg E, Noeske T, Spjuth O. Machine learning strategies when transitioning between biological assays. J Chem Inf Model. 2021;61(7):3722-33. doi:10.1021/acs.jcim.1c00293
15. Béquignon OJM, Bongers BJ, Jespers W, IJzerman AP, van de 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
16. 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
17. Morita K, Mizuno T, Kusuhara H. Investigation of a data split strategy involving the time axis in adverse event prediction using machine learning. J Chem Inf Model. 2022;62(17):3982-92. doi:10.1021/acs.jcim.2c00765
18. Fang C, Wang Y, Grater R, Kapadnis S, Black C, Trapa P, et al. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. J Chem Inf Model. 2023;63(11):3263-74. doi:10.1021/acs.jcim.3c00160
19. Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, et al. Transfer learning for drug discovery. J Med Chem. 2020;63(16):8683-94. doi:10.1021/acs.jmedchem.9b02147
20. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. ACS Cent Sci. 2017;3(4):283-93. doi:10.1021/acscentsci.6b00367
21. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell. 2022;4(12):1256-64. doi:10.1038/s42256-022-00580-7
22. 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
23. 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
24. 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
25. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
26. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, et al. DOME: Recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-7. doi:10.1038/s41592-021-01205-4
27. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-5. doi:10.1038/s41592-021-01256-7
28. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
29. Michoel T, Zhang JD. Causal inference in drug discovery and development. Drug Discov Today. 2023;28(10):103737. doi:10.1016/j.drudis.2023.103737