



Review Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Machine Learning QSAR Under Dataset Shift and Chemical Novelty

Patrick O'Connor^{1*}, Grace Murphy¹, Niamh Walsh²

¹Department of ML QSAR and Dataset Shift, School of Pharmacy, University College Cork, Cork, Ireland.

²Department of Chemical Novelty and Model Generalization, Faculty of Pharmacy, University of Galway, Galway, Ireland.

*Email: patrick.oconnor@ucc.ie

ABSTRACT

Machine-learning quantitative structure–activity relationship models are commonly compared under validation schemes that only partially reproduce the chemical, assay, temporal, and local structure–activity changes encountered during prospective drug discovery. Apparent performance can therefore reflect the difficulty of the selected test distribution as strongly as the intrinsic capability of the model. Peer-reviewed literature published from 2017 through 2025 was systematically examined for evidence concerning molecular-property and bioactivity prediction under scaffold displacement, assay or property change, chronological separation, external chemical novelty, activity cliffs, applicability-domain analysis, and predictive uncertainty. Model representation, validation design, distribution-shift mechanism, applicability assessment, and uncertainty treatment were coded separately. Source-level review-flow counts were not reconstructed from the final bibliography and are consequently not reported as numerical selection statistics. The evidence does not support a single hierarchy of QSAR model classes or a universal definition of out-of-distribution performance. Random splitting can overstate transferable performance, but scaffold splitting is not uniformly equivalent to chronological or prospective validation. Local activity cliffs constitute a separate challenge from broad scaffold novelty. Assay context, property range, representation, transfer strategy, and chemical-space support further modify apparent generalization. Applicability-domain membership and predictive uncertainty are informative but cannot independently establish reliability. QSAR evaluation under chemical novelty should begin by defining the mechanism producing distribution shift rather than by selecting a conventional splitter by default. Model comparison becomes more decision-relevant when global chemical support, local structure–activity discontinuity, assay context, temporal separation, model representation, and uncertainty are evaluated as distinct but interacting dimensions.

Keywords: QSAR, Dataset shift, Chemical novelty, Scaffold split, Applicability domain, Molecular machine learning

Received: 19 March 2025; **Revised:** 13 June 2025; **Accepted:** 17 June 2025

INTRODUCTION

Machine learning has expanded the representational and predictive repertoire of quantitative structure–activity relationship modelling, particularly through graph-based learning, learned molecular representations, multitask architectures, and increasingly large pretrained systems [1]. These developments have altered how molecular descriptors are constructed and how nonlinear structure–property relationships can be learned, but they have not removed the classical QSAR requirement that predictions be interpreted relative to the data and chemical domain from which the model was derived [2].

This distinction becomes critical when benchmark performance is used as evidence of prospective generalization. Ligand-based benchmarks can reward memorization when compounds in the training and test sets remain strongly

related, while standardized benchmark collections improve reproducibility without necessarily reproducing the distribution encountered during medicinal-chemistry progression [3, 4]. Consequently, a high test metric may describe interpolation within a well-supported chemical neighborhood rather than reliable prediction of genuinely novel chemistry.

Large-scale ChEMBL studies illustrate the resulting tension. Deep neural networks have achieved strong comparative performance across many bioactivity tasks [5], and independent analyses have likewise shown substantial advantages over established approaches within specified ChEMBL benchmark conditions [6]. Those findings establish that deep models can be highly effective; they do not establish that the same model ordering persists when the validation distribution changes.

The validation scheme is therefore part of the scientific question rather than a neutral technical detail. Chronological testing, when reliable timestamps are available, has been proposed as a closer approximation to prospective use than random splitting, and scaffold splitting may sometimes approximate the resulting separation [7]. Yet model rankings and conclusions can change when benchmark construction, splitting, hyperparameter selection, or validation assumptions are reanalysed [8]. Performance is thus inseparable from the process that generated the training–test boundary.

The central problem considered here is not simply whether one QSAR algorithm performs better than another. It is whether an evaluation design exposes the type of novelty for which a pharmaceutical prediction will actually be used. Structural novelty, local structure–activity discontinuity, movement into poorly represented property regions, assay-context change, chronological drift, and external chemistry can overlap, but they are not interchangeable.

This review therefore treats dataset shift as a family of analytically separable conditions. It examines how those conditions alter the interpretation of model performance and develops a proposed shift-aligned evaluation logic in which validation design, chemical-space support, local structure–activity difficulty, representation, applicability assessment, and uncertainty are considered jointly without being collapsed into a single out-of-distribution label.

MATERIALS AND METHODS

The review was designed and reported with the transparency principles of PRISMA 2020 as the methodological reference for search documentation, eligibility reasoning, evidence handling, and review-flow reporting [9]. The evidence window was restricted to peer-reviewed journal literature published from 2017 through 2025 and relevant to machine-learning QSAR, molecular-property prediction, ligand-based activity modelling, validation under chemical novelty, applicability assessment, or predictive uncertainty.

Eligible methodological and computational studies had to contribute information that could be assigned to at least one predefined analytical dimension: the origin of distribution shift; molecular representation; model family; training-domain composition; validation design; assay or endpoint context; chemical-space support; local structure–activity behavior; applicability analysis; uncertainty estimation; or external transfer. Studies were not treated as equivalent simply because they used the term “out-of-distribution.”

The review deliberately separated evidence describing model performance from evidence describing why a test set was difficult. It also separated global distributional displacement from local activity cliffs, representation choice from validation design, and confidence estimation from applicability. The integrated decision architecture developed later in the review is therefore a proposed synthesis of the evidence rather than an established consensus framework.

A numerical PRISMA flow is not reported because the source-specific export, deduplication, title/abstract-screening, retrieval, and full-text exclusion counts could not be reconstructed independently without inference. The final bibliography was not used as a substitute for the number of records identified, screened, or included.

Search and eligibility

Search logic combined terms for QSAR, molecular machine learning, molecular-property prediction, compound activity prediction, chemical space, scaffold splitting, chronological or temporal splitting, distribution shift, out-of-distribution prediction, applicability domain, uncertainty quantification, activity cliffs, and external validation. Methodological sources were also sought when they clarified reproducible model construction or the variables that should be extracted from QSAR workflows.

Reproducibility-oriented software environments illustrate why eligibility cannot be reduced to a model name. Chemprop provides a standardized implementation environment for molecular-property prediction and supports explicit control over representations, data partitions, and model training [10]. QSPRpred similarly exposes configurable elements of the QSPR workflow, reinforcing the need to distinguish data preparation, molecular representation, model specification, validation, and applicability analysis during evidence extraction [11].

Eligibility was therefore based on whether a report made its inferential setting sufficiently recoverable. A study could contribute evidence on model architecture without providing evidence for external chemical validity, while a validation study could be highly informative about distribution shift without establishing a general mechanistic explanation. Studies were coded according to the claim they could support rather than being promoted to stronger evidence classes because they used a complex algorithm.

This distinction is particularly important for the present review because “external validation,” “scaffold split,” and “OOD test” do not automatically identify equivalent inferential conditions. A test set can be structurally separated yet remain close in local property space; conversely, compounds that retain recognizable scaffolds can occupy a difficult region because of activity cliffs, assay change, or temporal changes in medicinal-chemistry strategy.

Table 1 separates the evidence that defines an eligible shift analysis from interpretations that would exceed what the underlying evaluation actually demonstrates.

Table 1. Eligibility distinctions that prevent validation labels from being mistaken for equivalent evidence of chemical generalization

What must be recoverable from the study	What may be concluded	Competing interpretation that must remain possible	Unresolved question for QSAR decision use
Manuscript-derived eligibility scenario Randomly partitioned molecular dataset			
Dataset origin, split procedure, endpoint, representation, model and metric	Performance under interpolation-dominant sampling	Structural overlap may inflate apparent generalization	Would performance persist after meaningful chemical separation?
Manuscript-derived eligibility scenario Scaffold-separated test chemistry			
Scaffold definition, splitter implementation and residual train–test similarity	Performance under the specified scaffold partition	The splitter may generate only modest practical novelty	Does the partition reproduce prospective medicinal-chemistry change?
Manuscript-derived eligibility scenario Chronologically separated evaluation			
Reliable timestamps, temporal cut point, endpoint continuity and assay context	Performance across the observed historical boundary	Time may introduce assay or project-policy changes in addition to chemistry	Which component of degradation is chemical rather than procedural?
Manuscript-derived eligibility scenario Assay- or property-shifted evaluation			
Endpoint definition, measurement context and relationship between source and target tasks	Transfer under the stated assay or property change	Apparent failure may reflect measurement or label shift rather than structural novelty	Can structural and measurement contributions be distinguished?
Manuscript-derived eligibility scenario Explicitly external chemical series			
Provenance of external chemistry and similarity/support relative to training data	Generalization to the stated external series	External origin alone does not quantify novelty or local SAR difficulty	How far outside training support is the new chemistry?

Manuscript-derived eligibility scenario Activity-cliff-rich test cases			
Matched molecular relationships and activity discontinuity	Ability to resolve local SAR discontinuity	Poor performance may occur despite strong global similarity	Does the model detect the chemically decisive local modification?
Manuscript-derived eligibility scenario Applicability- or uncertainty-filtered predictions			
Domain or uncertainty definition and validation of the reliability signal	Association between support/confidence and error in the tested setting	Confidence may remain miscalibrated under a new shift	Does the reliability measure remain valid under deployment-like novelty?

A standards-compliant numerical review-selection figure cannot be produced without verified source-level search arithmetic. Rather than manufacturing those values, the selection architecture is retained as an explicit audit boundary.

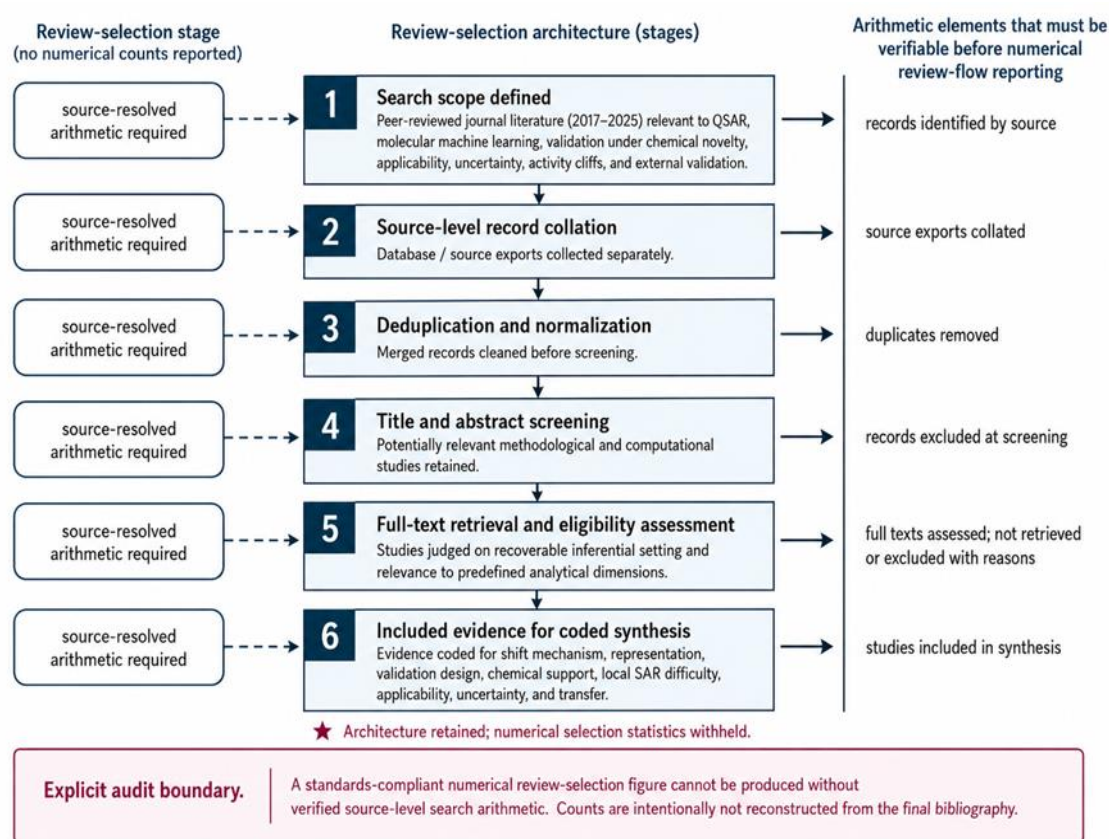


Figure 1. Review-selection stages that require source-resolved arithmetic before QSAR shift evidence can be treated as systematically enumerated

Dataset-shift taxonomy

The reviewed literature supports treating dataset shift as a property of how the training and evaluation distributions become different, not simply as a binary designation attached to a test set. In a large 2025 benchmark spanning multiple datasets, model families and splitting strategies, OOD performance varied substantially with the mechanism used to construct the test distribution; standard scaffold-based splitters were not invariably the most difficult [12]. Real-world-oriented compound-activity benchmarking similarly shows that assay and evaluation context influence the apparent transferability of predictive models [13]. These findings justify a proposed taxonomy based on the shift generator rather than the nominal name of the validation split.

A first axis is structural displacement at the level of scaffolds or molecular neighborhoods. A second is local response discontinuity. Activity cliffs expose cases in which highly similar molecules differ sharply in measured activity, and such cases can degrade performance across otherwise strong molecular-learning methods [14]. They therefore should not be absorbed into a broad global measure of train–test similarity.

The distinction becomes stronger when both members of an activity-cliff relationship are unseen during model training. QSAR analyses across dopamine D2 receptor, factor Xa, and SARS-CoV-2 main-protease examples showed that prediction of activity-cliff behavior is especially difficult when the relevant local relationship is absent from the training information [15]. This is a different inferential problem from asking whether a model has seen the same Bemis–Murcko scaffold.

A third axis is measurement or endpoint shift. Assays can differ in target state, protocol, readout, dynamic range, biological context, and label noise. A fourth is temporal shift, in which chemistry, assay practice, project priorities, or sampling policies may change together. A fifth is external chemical novelty defined by the relationship of candidate compounds to the actual support of the training set rather than by external provenance alone.

The article therefore uses “dataset shift” as an umbrella term while preserving the mechanism that produces the shift. The proposed taxonomy does not assert that the categories are independent. A chronologically later series may also contain new scaffolds, new local SAR, and different assay characteristics. Its purpose is to prevent those components from becoming analytically invisible.

Model and validation coding

Model comparison under distribution shift requires molecular representation to be coded independently of algorithm family. Context-dependent hierarchical graph representations can improve performance on standard molecular-property benchmarks and on activity-cliff-oriented tasks, indicating that the structural scale represented by the model can affect robustness to difficult local chemistry [16]. Such gains, however, remain architecture- and task-dependent.

Interpretability adds another distinction. Substructure-masking approaches can produce graph-neural-network explanations that correspond more closely to recognizable chemical fragments [17]. This can support medicinal-chemistry interrogation of a prediction, but an explanatory attribution is not equivalent to evidence that the highlighted substructure is the biological cause of the measured activity.

Model coding must also resist the assumption that graph learning intrinsically dominates engineered chemical descriptors. Comparative analyses across multiple molecular-property datasets have shown that descriptor-based approaches can equal or outperform graph neural networks on a substantial fraction of tasks, while graph models perform better in others [18]. Representation is consequently a conditional design variable rather than a universal ranking criterion.

For each study, validation coding should therefore record the representation, model family, optimization procedure, split mechanism, test-set relationship to training chemistry, endpoint, metric, and whether hyperparameter selection uses information that could reduce effective separation. A performance estimate is interpretable only together with that evaluation context.

Quality assessment

Quality assessment in shift-sensitive QSAR cannot be reduced to reporting whether cross-validation was performed. It must ask whether the chosen evaluation design interrogates the intended deployment claim. Transfer-learning studies reinforce this principle. Task-relatedness measures can identify molecular-property settings in which transferred information is useful, but transferability varies across source–target relationships [19].

The same conditionality appears in multi-fidelity learning. Large lower-fidelity datasets can improve prediction for smaller, higher-fidelity tasks when the relationship between fidelities contains transferable signal [20]. More training observations therefore do not automatically increase reliability if the added information changes the learned problem without supporting the target decision.

Foundation-model pretraining raises a related issue. Large-scale molecular pretraining can yield useful transferable representations across downstream tasks [21], but downstream benchmark improvement should not be recoded as proof of external chemical validity. Pretraining evidence and OOD evidence answer different questions.

The review consequently treats four features as quality-critical: whether the validation split represents the claimed use; whether model selection is isolated from the final evaluation set; whether the chemical and assay support of test compounds is described; and whether uncertainty or applicability claims are themselves evaluated rather than assumed from method choice.

RESULTS AND DISCUSSION

Across the reviewed evidence, the most consistent result is heterogeneity rather than a universal ordering of methods. Algorithms explicitly designed to learn representations robust to OOD change can improve performance across several molecular-property datasets [22], demonstrating that distribution shift can be targeted during model development. These results remain benchmark-specific and do not establish that a single OOD-learning architecture will dominate across medicinal-chemistry programs.

A second result is that OOD itself has multiple geometries. Work on extreme property-value prediction emphasizes that compounds or materials can be out of distribution because the target property lies beyond the range represented during training even when structural novelty is not maximal [23]. Structural similarity, scaffold identity, target-value support, and local SAR therefore provide overlapping but non-equivalent information.

The reviewed studies also reveal recurrent disagreement between apparently similar validation labels. Scaffold splits can be moderately demanding in one dataset and comparatively mild in another. Pretrained representations can improve some shifted tasks while providing limited or negative transfer elsewhere. Applicability-domain methods can identify unsupported chemistry while still leaving heterogeneous error within the nominally supported region. These discrepancies are not treated as noise to be averaged away; they are evidence that transferability depends on how novelty is constructed and measured.

Scaffold shift

Scaffold splitting is attractive because it imposes a chemically interpretable boundary between training and test compounds, but its meaning is dataset-dependent. Earlier molecular-property benchmarking found that scaffold-based evaluation could approximate chronological testing when timestamps were unavailable [7]. The later multi-dataset OOD benchmark, however, showed that conventional scaffold and generic-scaffold splits could be among the less challenging splitters relative to alternatives deliberately designed to generate stronger distributional displacement [12].

These findings are compatible if scaffold identity is treated as one component of novelty rather than a universal measure of it. A test set can contain formally unseen scaffolds while retaining strong functional-group, physicochemical, or local neighborhood support. Conversely, a model can encounter severe predictive difficulty inside a familiar scaffold when small substitutions generate large response changes.

Activity-cliff evidence makes this local limitation explicit [14]. Cliff-focused QSAR analyses further show that the relevant decision state changes according to whether both members, one member, or neither member of the local SAR relationship has appeared during training [15]. Hierarchical and context-dependent representation may improve some such cases [16], but does not remove the underlying dependence on training support.

Table 2 is therefore used to distinguish scaffold novelty from the local evidence states that determine what a medicinal-chemistry prediction can reasonably support.

Table 2. Decision states for interpreting scaffold-separated predictions when local structure–activity support is unequal

New scaffold with dense neighboring SAR New scaffold with sparse neighborhood support Familiar scaffold with unseen substitution pattern			
Decision state	New scaffold with dense neighboring SAR	New scaffold with sparse neighborhood support	Familiar scaffold with unseen substitution pattern
Information available at prediction time	Scaffold identity is new but related transformations and property ranges are represented	Global descriptors remain partly represented but few close activity analogues exist	Core scaffold is represented but local substitution space is not
Unresolved uncertainty	Whether learned local relationships transfer onto the new core	Whether apparent similarity reflects the activity-determining features	Potential activity-cliff behavior
Permissible interpretation	Prediction may be treated as chemically supported but externally shifted	Prediction should be treated as extrapolative	Scaffold familiarity alone is insufficient for confidence
Failure signature requiring reassessment	Error rises despite close analogical support	Confidence remains high while local errors increase	Small structural change produces large residual

Unseen activity-cliff pair Partially observed cliff neighborhood Scaffold shift combined with assay change			
Decision state	Unseen activity-cliff pair	Partially observed cliff neighborhood	Scaffold shift combined with assay change
Information available at prediction time	Neither side of the local discontinuity is represented	One or related cliff members are represented	Chemical core and measurement context both change
Unresolved uncertainty	Direction and magnitude of local SAR change	Whether the decisive transformation has been learned	Relative contribution of chemistry and assay shift
Permissible interpretation	Prediction is hypothesis-generating rather than confirmatory	Prediction can use local evidence but requires explicit qualification	Chemical and assay transfer cannot be inferred separately
Failure signature requiring reassessment	Model smooths the discontinuity toward neighboring activity values	Error concentrates on the unseen member or transformation	Performance drop cannot be localized to one source

Property and assay shift

Chemical novelty is only one reason that the relationship between molecular structure and a prediction target may change. Applicability-domain analyses across large chemical collections demonstrate that chemical-space coverage varies substantially by property and chemical class [24]. A region adequately represented for one endpoint may therefore be poorly supported for another.

The implication is stronger than the statement that every model needs an applicability domain. Conventional domain membership can conceal heterogeneous prediction error. Reanalysis of QSAR applicability procedures has shown that compounds classified as being within a nominal domain can still occupy regions with substantially different reliability [25]. Applicability should therefore be treated as graded evidence about support rather than an automatic certificate of correctness.

Assay shift adds a measurement-layer boundary. A model trained on one assay does not necessarily face the same prediction problem after changes in protocol, target construct, biological context, or response definition. Real-world-oriented activity benchmarks show that assay context affects apparent model transfer [13]. Consequently, structural similarity between training and test compounds cannot establish that labels remain exchangeable across assay conditions.

The proposed taxonomy therefore keeps property shift and assay shift separate. Property shift concerns movement into a response regime poorly represented by training observations; assay shift concerns alteration of the measurement process or biological context producing those observations. They can occur simultaneously, but they generate different explanations for prediction failure and different requirements for validation.

Temporal shift

Temporal validation addresses a question that random splitting cannot: whether a model trained on information available before a specified time retains value for chemistry encountered later. Where reliable timestamps exist, chronological splitting provides a direct way to reproduce this ordering and has been used as a practical reference when evaluating alternative molecular split strategies [7].

Temporal shift should nevertheless not be interpreted as a purely chronological version of scaffold shift. The later period can differ because new scaffolds were synthesized, because optimization moved into a new physicochemical region, because assay procedures changed, because project objectives changed, or because medicinal chemists deliberately sampled chemistry that earlier models considered uncertain. The distribution shift is generated by the evolving discovery process rather than by time itself.

The 2025 OOD benchmark strengthens this distinction by showing that the difficulty of a molecular test set depends strongly on how the split is generated [12]. Real-world activity benchmarking further indicates that assay and discovery context can modify apparent model performance [13]. A chronological holdout may therefore combine several shift mechanisms at once.

This creates a multiscale inference problem. At the molecular scale, the later compounds may be structurally or physicochemically novel. At the local SAR scale, new substitutions may generate previously unseen response discontinuities. At the assay scale, measurement conditions may change. At the model scale, representation and hyperparameter choices determine which relationships can be transferred. At the decision scale, the same

prediction error can have different implications depending on whether the model is used for ranking analogues, triaging distant chemistry, or prioritizing experiments.

Figure 2 is used to keep these temporal components visually separate and to show where validation must occur before a model-performance result can be translated into a medicinal-chemistry decision.

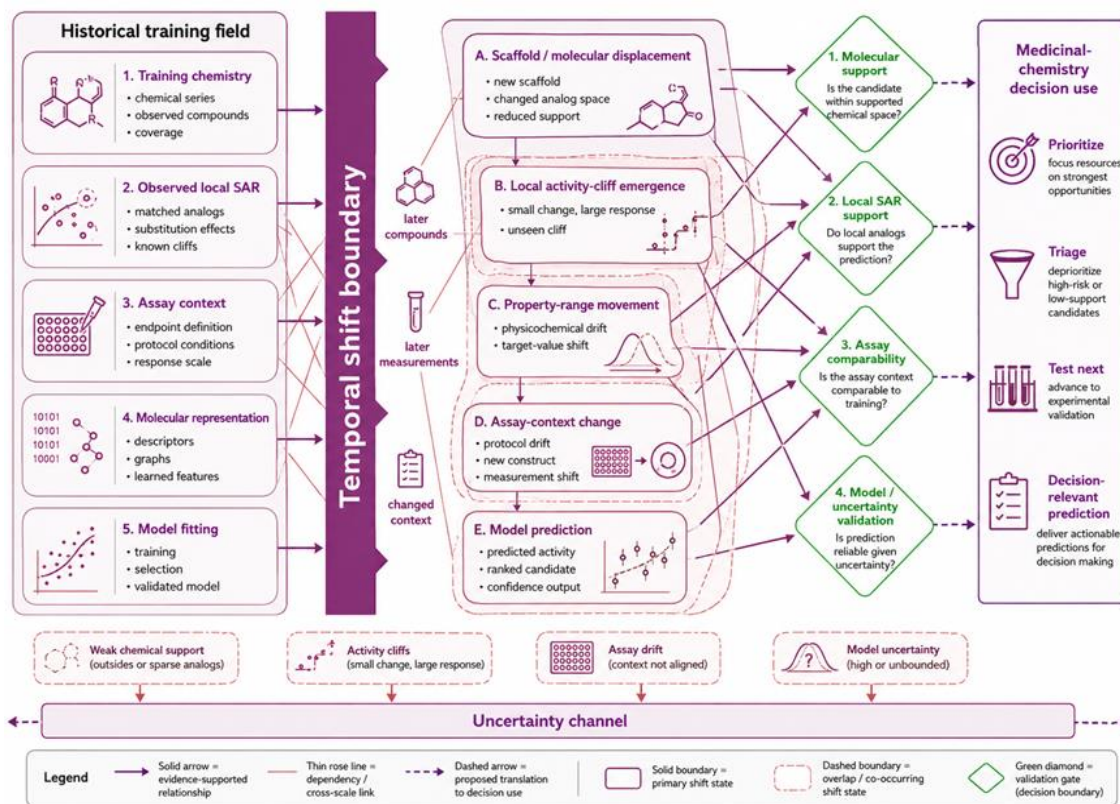


Figure 2. Temporal validation separates chemical evolution, local SAR change, assay drift, and decision transfer in prospective QSAR use

External chemical novelty

External chemical novelty is often treated as self-evident when a test compound originates from a different database, project, vendor collection, or medicinal-chemistry series. Provenance, however, does not quantify whether the model has meaningful support for the new molecule. A compound can be external yet remain close to the training distribution in descriptor, scaffold, physicochemical, or learned-representation space. Conversely, an internally generated compound may occupy a poorly represented region despite sharing familiar structural motifs.

Distance-based applicability analysis provides one route to describing this distinction. Feature-space dissimilarity can identify regions in which model error tends to increase, supporting the interpretation of chemical distance as evidence about training support rather than as an intrinsic property of the molecule [26]. Such relationships remain model- and representation-dependent. A distance calculated from fingerprints, engineered descriptors, latent embeddings, or learned graph features does not necessarily encode the same type of novelty.

External novelty must also remain distinct from property-value extrapolation. A structurally plausible molecule may require prediction in a target-value regime not represented adequately during training [23]. The reverse is also possible: a structurally distant compound may occupy a familiar property range. These conditions impose different inferential burdens and should not be represented by a single scalar novelty score.

Prospective domain expansion provides an especially important test of this assumption. In the CYP2B6 study, extending the chemical search domain increased structural diversity and enabled identification of additional experimentally active compounds, yet the anticipated improvement in the predictive model or its applicability domain was limited [27]. This negative result shows why additional diversity cannot automatically be interpreted as additional predictive support. New molecules may broaden observed chemistry without resolving the specific representation, sampling, or response discontinuities responsible for model error.

The proposed interpretation is therefore sequence-aware. External chemical novelty begins with a candidate's relationship to the existing training support, proceeds through examination of local SAR and property coverage, and only then supports a judgment about whether the prediction is interpolative, weakly supported, or meaningfully extrapolative. A subsequent experimental result can either strengthen the supported region or expose a new failure boundary.

The distinction matters operationally because medicinal-chemistry selection is not simply a question of whether a candidate is “inside” or “outside” an abstract domain. The more relevant question is what evidence connects that candidate to the model's learned response surface and what observation would demonstrate that the connection was misleading.

The sequence is summarized in **Table 3** so that external provenance, chemical support, local response behavior, and experimental reassessment are not collapsed into a binary domain label.

Table 3. Sequence states linking external chemical novelty to prediction support and reassessment

External-novelty state	Evidence structure	Boundary and consequence	External-novelty state	Evidence structure	Boundary and consequence
Supported external candidate	Precursor molecular or chemical state Molecule originates outside the training source but lies within well-populated chemical and property neighborhoods	Condition interrupting or reversing the sequence Experimental result conflicts with apparently strong support	Peripheral chemical candidate	Precursor molecular or chemical state Candidate approaches the edge of represented descriptor or latent space	Condition interrupting or reversing the sequence New measurement reveals systematic error at the boundary
	Transition trigger External compound enters prediction workflow Observable or model output Stable prediction with agreement among support diagnostics Available sampling or data support Multiple related structures and response patterns are represented			Transition trigger Declining neighborhood density or rising distance Observable or model output Prediction remains available but support weakens Available sampling or data support Partial structural or physicochemical overlap	
Structurally novel but property-supported candidate	Precursor molecular or chemical state New scaffold or topology enters a familiar response regime	Condition interrupting or reversing the sequence Local SAR behaves differently on the new core	Property-extreme candidate	Precursor molecular or chemical state Chemistry may remain recognizable while the expected response lies outside training range	Condition interrupting or reversing the sequence Experimental value extends beyond learned response range
	Transition trigger Scaffold or structural displacement Observable or model output Model predicts within historically observed property range Available sampling or data support Property support exceeds structural support			Transition trigger Target-value extrapolation Observable or model output Numerically plausible but weakly anchored prediction Available sampling or data support Structural support may remain moderate	
		Unresolved QSAR question Which training relationships actually justified transfer?			Unresolved QSAR question Is the boundary representation-specific or biologically meaningful?

Local discontinuity state	Precursor molecular or chemical state Candidate resembles known compounds globally but occupies uncertain local SAR Transition trigger Small chemical transformation with potential large response consequence Observable or model output Smooth model prediction despite ambiguous local evidence Available sampling or data support Global support strong; local response support sparse	Condition interrupting or reversing the sequence Activity-cliff behavior is experimentally observed Unresolved QSAR question Which local chemical change invalidated the smooth response assumption?	Domain-expansion state	Precursor molecular or chemical state New chemistry is deliberately sampled to broaden training support Transition trigger Experimental acquisition of novel compounds Observable or model output Updated model or applicability assessment Available sampling or data support Newly measured chemistry becomes available	Condition interrupting or reversing the sequence Added diversity fails to improve predictive reliability Unresolved QSAR question Did the added chemistry address the original failure mode?
	Precursor molecular or chemical state New experimental data reveal previously missing relationships Transition trigger Model and support diagnostics are updated Observable or model output Revised prediction and uncertainty structure Available sampling or data support Expanded measured neighborhood	Condition interrupting or reversing the sequence Further external chemistry produces a new unsupported regime Unresolved QSAR question Has support genuinely expanded or merely shifted location?			

Uncertainty and applicability methods

Applicability and uncertainty answer related but different questions. Applicability methods ask whether a new observation resembles the circumstances under which the model was constructed and validated. Uncertainty methods estimate how unreliable a specific prediction may be under a chosen probabilistic or empirical formulation. Neither automatically substitutes for the other.

Comparative studies of molecular-property prediction show substantial variation among scalable uncertainty approaches [28]. Neural-network uncertainty methods likewise do not exhibit a single ranking that persists across datasets and evaluation criteria [29]. Evidential deep learning offers another strategy by learning parameters intended to represent predictive evidence directly and has demonstrated utility for molecular prediction and acquisition under evaluated conditions [30]. These studies collectively argue against selecting an uncertainty method solely because it has performed well on an unrelated benchmark.

Interpretation also depends on the source of uncertainty. Aleatoric components can reflect irreducible response variability or measurement noise, whereas epistemic uncertainty is intended to reflect limitations in what the model has learned [31]. In practice, these components may not be cleanly identifiable, particularly under distribution shift. A model can be confidently wrong because its uncertainty estimator has not encountered the type of novelty presented by the new chemistry.

Explainable uncertainty methods attempt to localize which molecular regions contribute to low confidence and can therefore provide information that a global uncertainty value hides [32]. This is potentially useful when uncertainty arises from unfamiliar substitutions or motifs. As with predictive explanations, however, localization should not be interpreted as proof of biological mechanism.

Uncertainty can also become an acquisition variable. Active-learning studies show that uncertainty-guided compound selection can accelerate improvement of molecular-property models when the acquisition strategy is aligned with the target distribution [33]. This does not imply that selecting the most uncertain molecules is universally optimal; some highly uncertain samples may be uninformative, erroneous, or irrelevant to the desired chemical domain.

The need for task-specific evaluation is formalized by uncertainty-benchmarking frameworks showing that no single metric or UQ technique is consistently optimal across use cases [34]. Conformalized graph prediction provides a complementary route by constructing prediction intervals with empirical coverage properties under specified assumptions and has shown useful behavior on ADMET tasks [35]. Coverage guarantees nevertheless depend on the relationship between calibration and future data and therefore should not be interpreted as immunity to unrestricted dataset shift.

The interaction with local SAR is particularly important. Recent activity-prediction analyses show that steep SAR, sparse chemical support, and validation-split choice can materially change the apparent quality of uncertainty estimates [36]. An uncertainty method should therefore be evaluated under the same shift mechanism for which the final QSAR model will be used.

Table 4 treats applicability and uncertainty methods as coupled design choices rather than as competing labels for the same concept.

Table 4. Design-space constraints for combining applicability evidence with predictive uncertainty under chemical shift

Reliability design choice	Enabling condition	Coupled constraint Trade-off Non-negotiable boundary	Performance Prospective test implied by the framework
Scientific anchor	Enabling condition	Coupled constraint Trade-off Non-negotiable boundary	Performance Prospective test implied by the framework
Chemical-distance applicability assessment	Representation captures relationships relevant to the endpoint	Coupled constraint Distance meaning changes with representation and density Trade-off that must remain visible Simple interpretation versus incomplete response information Non-negotiable boundary Distance alone cannot establish correctness	Performance or diagnostic readout Neighborhood support or distance-to-training distribution Prospective test implied by the framework Test whether error increases in the intended external chemical region
Ensemble-derived predictive uncertainty	Multiple independently trained models provide meaningful variation	Coupled constraint Shared training bias can make ensembles collectively overconfident Trade-off that must remain visible Computational cost versus robustness Non-negotiable boundary Agreement is not evidence of external validity	Performance or diagnostic readout Prediction variance, calibration and error ranking Prospective test implied by the framework Evaluate calibration after deployment-relevant splitting
Evidential uncertainty	Training objective supports direct uncertainty estimation	Coupled constraint Learned evidence can inherit dataset-specific biases Trade-off that must remain visible Efficient single-model inference versus model-assumption dependence Non-negotiable boundary Low uncertainty cannot override missing chemical support	Performance or diagnostic readout Predictive evidence and calibration behavior Prospective test implied by the framework Challenge with chemistry beyond the training support

Localized or explainable uncertainty	Molecular contribution to confidence is scientifically interpretable	Coupled constraint Attribution remains model-dependent	Performance or diagnostic readout Atom-, bond- or substructure-level uncertainty patterns
		Trade-off that must remain visible Diagnostic detail versus risk of mechanistic overinterpretation	Prospective test implied by the framework
		Non-negotiable boundary Attribution does not establish causal activity determinants	Compare highlighted regions with experimentally observed SAR failures
Uncertainty-guided acquisition	New measurements can be selected iteratively	Coupled constraint High uncertainty may identify irrelevant or noisy samples	Performance or diagnostic readout Improvement after successive acquisition rounds
		Trade-off that must remain visible Exploration versus task-specific exploitation	Prospective test implied by the framework
		Non-negotiable boundary Acquisition value must be measured by subsequent model improvement	Compare uncertainty-driven selection with chemically informed alternatives
Conformal prediction interval	Calibration and prediction distributions satisfy the method's assumptions sufficiently	Coupled constraint Coverage can deteriorate when exchangeability is violated	Performance or diagnostic readout Empirical interval coverage and width
		Trade-off that must remain visible Reliability coverage versus informativeness	Prospective test implied by the framework
		Non-negotiable boundary Nominal coverage under one distribution is not universal OOD coverage	Reassess coverage after scaffold, temporal, or assay displacement
Joint applicability-uncertainty assessment	Chemical support and calibrated confidence are both available	Coupled constraint Diagnostics can disagree	Performance or diagnostic readout Combined support, residual and uncertainty pattern
		Trade-off that must remain visible More reliable interpretation versus greater workflow complexity	Prospective test implied by the framework
		Non-negotiable boundary Neither diagnostic may be converted into a universal acceptance threshold	Prospectively test whether joint diagnostics identify failures better than either alone

Comparative model performance

The reviewed evidence does not justify a universal ranking of QSAR model classes. Descriptor-based methods remain competitive with graph neural networks across numerous molecular-property tasks [18], while alternative graph representations can alter both predictive performance and interpretation across pharmaceutical and general molecular datasets [37]. Hierarchical self-supervised graph learning can improve downstream classification and regression under specified benchmark conditions [38], but such improvements do not establish prospective superiority under an unseen shift.

The same qualification applies to earlier demonstrations of deep-learning performance in ChEMBL-scale bioactivity prediction [5]. Strong benchmark performance establishes capability within the tested evaluation regime. It does not establish that the ordering will survive more severe structural, temporal, assay, or local-SAR displacement. The substantial dependence of OOD performance on split construction reinforces this point [12]. Comparative performance should consequently be interpreted conditionally: model A outperforms model B under a specified representation, endpoint, training distribution, model-selection procedure, and shift generator. Removing those qualifiers transforms an empirical comparison into a stronger claim than the evidence supports. Uncertainty performance follows the same rule. A model with marginally better average predictive accuracy may be less useful if its failures cannot be recognized under the intended deployment shift, whereas a slightly less accurate model may provide better-calibrated decision support. Recent evidence that uncertainty behavior

deteriorates under difficult SAR and weak chemical support illustrates why predictive accuracy and reliability diagnostics should be examined together [36].

The main finding of this review is that machine-learning QSAR performance under chemical novelty cannot be interpreted independently of the process generating novelty. Dataset shift is not one homogeneous test condition. Scaffold displacement, local activity cliffs, property-range extrapolation, assay change, temporal evolution, and externally sourced chemistry impose different informational deficits on the model.

This explains an apparent contradiction in the scaffold-splitting literature. Scaffold separation can be a useful approximation to chronological validation in some datasets [7], but conventional scaffold splitters need not create particularly severe OOD conditions in broader comparisons [12]. The appropriate conclusion is therefore not that scaffold splitting is either valid or invalid. Its relevance depends on whether the resulting test chemistry reproduces the novelty expected at use time.

A second recurring theme is that model complexity does not remove the need for domain reasoning. Graph architectures, hierarchical representation, transfer learning, and molecular foundation models can improve prediction in selected settings, but those improvements occur within specified datasets and evaluation designs. Chemical support and validation structure remain logically prior to claims of deployment reliability.

Applicability and uncertainty likewise represent evidence, not guarantees. Applicability analysis describes how a candidate relates to training support; uncertainty describes the model's confidence or predictive dispersion under a particular method. Either can fail independently. A compound can lie close to training data while participating in an activity cliff, and a model can remain overconfident when a new shift violates the conditions under which uncertainty was calibrated.

The review therefore proposes that prediction failure should itself be classified. A failure under global chemical displacement is not equivalent to failure inside a familiar scaffold caused by local SAR discontinuity. A failure after assay change cannot automatically be attributed to molecular representation. A failure under chronological testing may combine chemical, measurement, and project-selection changes. Making those alternatives explicit transforms validation from a single performance estimate into a diagnosis of what information the model lacks.

Recommended evaluation framework

The proposed shift-aligned evaluation framework begins by defining the intended prediction use and the mechanism expected to separate future molecules from the training distribution. Validation should then reproduce that mechanism as directly as available data permit rather than defaulting automatically to random or scaffold splitting.

Chemical support should next be evaluated at more than one scale. Global scaffold or representation-space distance should be complemented by local neighborhood analysis capable of exposing sparse SAR and possible activity cliffs. Property-range support and assay comparability should be considered separately so that structural novelty is not used as a proxy for measurement or response shift.

Representation and model families should then be compared under the same shift design. This prevents apparent architecture advantages from being confounded with easier test chemistry or different model-selection procedures. Where transfer learning or pretraining is used, source–target relatedness should be treated as an empirical variable rather than assumed from dataset size or model scale.

Applicability and uncertainty should subsequently be evaluated against observed prediction error within the intended shift condition. The central question is not whether a method produces an uncertainty number or an applicability label, but whether that signal identifies unreliable predictions when the model encounters the kind of novelty relevant to its intended use.

Finally, prospective or temporally later measurements should be used to reassess both the predictive model and the reliability diagnostics. A newly measured compound can extend the supported region, reveal an activity cliff, expose assay dependence, or demonstrate that a previously trusted uncertainty measure was miscalibrated. Evaluation therefore becomes recursive rather than a one-time gate.

This framework is a proposed synthesis of the reviewed evidence. It has not itself been prospectively validated as a unified QSAR standard and should not be presented as a regulatorily accepted or universally optimal workflow.

Limitations

The evidence base is heterogeneous in endpoints, molecular representations, dataset sizes, model families, split construction, performance metrics, and definitions of out-of-distribution data. Direct quantitative ranking across studies would therefore require assumptions that are not justified by the reviewed designs.

Much of the available literature remains benchmark-centered. Even increasingly realistic scaffold, temporal, assay-aware, and OOD evaluations cannot fully reproduce prospective medicinal-chemistry decisions, where compounds are chosen adaptively and new experimental results influence subsequent sampling.

Evidence for uncertainty and applicability is similarly context-dependent. Calibration demonstrated on one benchmark or split cannot be assumed to persist after a different form of chemical or assay displacement. No universal chemical-distance threshold, uncertainty cutoff, or model-performance value can be derived from the reviewed literature.

A methodological limitation is that the source-specific retrieval, deduplication, screening, report-retrieval, and full-text exclusion arithmetic was not independently recoverable from the final evidence record. The bibliography was therefore not reverse-engineered into a numerical PRISMA flow. This prevents unsupported reporting of review-selection frequencies but limits complete reproduction of the search-selection chain.

Finally, the proposed shift-aligned evaluation framework integrates relationships demonstrated separately across multiple studies. Its value is conceptual and organizational at this stage. Prospective testing is required to determine whether the combined architecture improves compound-selection decisions, failure detection, or experimental efficiency relative to simpler validation strategies.

CONCLUSION

Machine-learning QSAR under chemical novelty is fundamentally an evaluation problem as well as a modelling problem. Predictive performance becomes interpretable only after the mechanism separating future chemistry from the training data has been identified.

Scaffold novelty, local activity discontinuity, property-range extrapolation, assay change, temporal evolution, and external chemical provenance should therefore remain distinct analytical states. Model architecture, molecular representation, applicability assessment, and uncertainty can modify performance within those states, but none independently establishes transferability.

The proposed shift-aligned framework organizes these dimensions around a simple principle: predictions should be judged against the specific form of novelty they are expected to encounter. This reframes external validation from a generic test-set exercise into an explicit assessment of chemical support, local response difficulty, measurement continuity, uncertainty, and decision context.

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. 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
3. 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
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. 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
6. 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

7. 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
8. Robinson MC, Glen RC, Lee AA. Validating the validation: Reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *J Comput Aided Mol Des*. 2020;34(7):717-30. doi:10.1007/s10822-019-00274-0
9. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
10. 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
11. 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
12. 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
13. Tian T, Li S, Zhang Z, Chen L, Zou Z, Zhao D, et al. Benchmarking compound activity prediction for real-world drug discovery applications. *Commun Chem*. 2024;7(1):127. doi:10.1038/s42004-024-01204-4
14. 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
15. 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
16. Wan Y, Wu J, Hou T, Hsieh CY, Jia X. Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation. *Nat Commun*. 2025;16(1):413. doi:10.1038/s41467-024-55082-4
17. Wu Z, Wang J, Du H, Jiang D, Kang Y, Li D, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat Commun*. 2023;14(1):2585. doi:10.1038/s41467-023-38192-3
18. 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
19. 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
20. Buterez D, Janet JP, Kiddle SJ, Oglic D, Lió P. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nat Commun*. 2024;15(1):1517. doi:10.1038/s41467-024-45566-8
21. Méndez-Lucio O, Nicolaou CA, Earnshaw B. MolE: A foundation model for molecular graphs using disentangled attention. *Nat Commun*. 2024;15(1):9431. doi:10.1038/s41467-024-53751-y
22. Wen X, Liu H, Long W, Wei S, Zhu R. Consistent semantic representation learning for out-of-distribution molecular property prediction. *Brief Bioinform*. 2025;26(2):bbaf147. doi:10.1093/bib/bbaf147
23. Segal N, Netanyahu A, Greenman KP, Agrawal P, Gómez-Bombarelli R. Known unknowns: Out-of-distribution property prediction in materials and molecules. *NPJ Comput Mater*. 2025;11(1):345. doi:10.1038/s41524-025-01808-x
24. 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
25. 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
26. Schultz LE, Wang Y, Jacobs R, Morgan D. A general approach for determining applicability domain of machine learning models. *NPJ Comput Mater*. 2025;11(1):95. doi:10.1038/s41524-025-01573-x
27. Vignaux PA, Harris JS, Urbina F, Ekins S. Applicability domain-expansion studies for machine learning models reveal new inhibitors of CYP2B6. *Drug Metab Dispos*. 2025;53(10):100160. doi:10.1016/j.dmd.2025.100160

28. 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
29. 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
30. 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
31. 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
32. 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
33. 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
34. 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
35. 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
36. 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
37. Kengkanna A, Ohue M. Enhancing property and activity prediction and interpretation using multiple molecular graph representations with MMGX. *Commun Chem.* 2024;7(1):74. doi:10.1038/s42004-024-01155-w
38. Zang X, Zhao X, Tang B. Hierarchical molecular graph self-supervised learning for property prediction. *Commun Chem.* 2023;6(1):34. doi:10.1038/s42004-023-00825-5