Available online www.ijpras.com

International Journal of Pharmaceutical Research & Allied Sciences, 2026, 15(1):54-63
https://doi.org/10.51847/ajdBgZrfdC

Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

The Continually Learning QSAR System under Chemical-Space Drift, Endpoint Redefinition, Model Change, and Requalification throughout Deployment

Daniel Fischer[1]*, Laura Meier[1], Stefan Koch[2], Thomas Braun[3]

[1]Department of Continually Learning QSAR and Chemical-Space Drift, Faculty of Pharmacy, University of Freiburg, Freiburg, Germany.
[2]Department of Endpoint Redefinition and Model Change, Faculty of Pharmacy, Technical University of Munich, Munich, Germany.
[3]Department of Requalification and Deployment, Faculty of Pharmacy, University of Kiel, Kiel, Germany.

***Email:** daniel.fischer@pharmazie.uni-freiburg.de*

ABSTRACT

Quantitative structure–activity relationship models are commonly developed as bounded predictive artefacts, yet their evidential conditions may continue to change after deployment. Incoming compounds can occupy unfamiliar chemical space, assay procedures and biological contexts can alter endpoint meaning, labels can be corrected or reinterpreted, molecular representations and software environments can change, and prediction consequences can expand beyond the original context of use. Treating the deployed model as static therefore risks preserving an unchanged model identifier while the scientific object being predicted, the data-generating process, or the supported decision has changed. This Original Lifecycle-Governance Architecture Article proposes a continually learning QSAR lifecycle in which chemical-space state, endpoint contract, data and label state, model and representation version, and qualification state are governed as interacting controlled objects. The architecture separates surveillance from adaptation: drift indicators initiate investigation and change-impact assessment rather than automatic retraining. Updating options range from data correction and recalibration to retraining, representation replacement, endpoint remodelling, quarantine, or withdrawal. Revalidation is proposed as the generation of context-specific evidence regarding predictive behavior, applicability, calibration, and uncertainty, whereas requalification determines whether the complete changed system remains supportable for its intended use. Performance monitoring incorporates chemical-space coverage, predictive uncertainty, calibration, and abstention, while auditability requires traceable change records, human authorization, reproducible version bundles, and rollback capability. The contribution is an original conceptual integration rather than a validated operating standard. Its propositions require retrospective stress testing, prospective evaluation, reproducibility assessment, and implementation studies. No universal drift threshold, retraining interval, uncertainty method, requalification category, regulatory pathway, or deployment-readiness claim is established.

Key words: *QSAR, Pharmacoinformatics, Chemical space, External validation, Applicability domain, Calibration*

Received: 21 October 2025; **Revised:** 30 January 2026; **Accepted:** 06 February 2026

INTRODUCTION

Quantitative structure–activity relationship modelling extends beyond fitting statistical associations between molecular descriptors and measured responses. Contemporary QSAR practice also encompasses data curation, molecular representation, model validation, applicability assessment, interpretation, and specification of the

domain within which predictions may be scientifically defensible [1]. These functions become more difficult to preserve when a model moves from a fixed development dataset into an environment that continuously generates new compounds, assays, labels, and decision demands.

The predictive meaning of a model cannot be inferred from architecture or internal performance alone. Molecular-learning benchmarks differ in endpoint type, sample composition, class balance, splitting strategy, and evaluation metric, and these choices can materially affect apparent performance [2]. A model that performs well under one benchmark construction may therefore support only a bounded claim about that task rather than a general claim of prospective pharmaceutical validity.

Molecular representation is likewise not a neutral technical layer. Learned and engineered representations encode different structural information, inductive assumptions, and transfer behavior, meaning that a change in representation can alter the identity and operating characteristics of the resulting model [3]. Nevertheless, deployed QSAR systems are often discussed as though the prediction function were the only object requiring control, while data provenance, endpoint meaning, domain boundaries, calibration, software dependencies, and authorization status remain external documentation.

This article develops an original, non-empirical lifecycle-governance architecture for continually learning QSAR systems. The proposed contribution treats chemical-space state, endpoint definition, label quality, molecular representation, model version, calibration, applicability, and qualification as interacting controlled objects throughout deployment. It does not present a validated software system, exhaustive review, clinical recommendation, regulatory standard, or universally applicable update procedure. Its purpose is to define an architecture, derive testable relationships, identify revalidation and requalification triggers, and establish boundaries for scientifically responsible adaptation.

Why deployed QSAR systems cannot be treated as static

A static representation assumes that the evidential meaning of a validated model persists unless its code or parameters are deliberately changed. That assumption is fragile because apparently strong validation may partly reflect similarity or redundancy between development and test data. Ligand-based classification benchmarks can reward memorization of related chemical series rather than generalization to genuinely unfamiliar compounds [4]. A frozen model may therefore remain computationally unchanged while its prospective users encounter chemistry for which the original evidence was never informative.

Model choice also does not yield a context-independent hierarchy of predictive quality. Large-scale comparisons across bioactivity tasks show that relative algorithm performance depends on the target, dataset, endpoint, representation, and evaluation design [5]. Consequently, selecting a model during development does not establish enduring superiority after the composition of incoming compounds, assays, or labels changes. The relevant question is not simply whether the model remains operational, but whether the relationship between its training evidence and its current use remains sufficiently preserved.

Direct chemical-domain evidence further shows that data distributions and predictive validity can change across time, data source, assay conditions, and chemical composition [6]. Such changes may arise from genuine expansion into new chemistry, altered experimental practice, selective testing, revised curation, or technical ingestion faults. These mechanisms are not equivalent: recalibration may address some probability distortions, whereas endpoint redefinition, local activity discontinuities, or major coverage loss may require different responses. Observed degradation should therefore not be attributed automatically to one generic form of drift.

The article proposes that a deployed QSAR system should be understood as a changing evidence relationship rather than a static model file. Its state includes the compounds being encountered, the endpoint being claimed, the labels available for learning and evaluation, the representation through which chemistry is encoded, the calibration and applicability rules governing interpretation, and the decisions for which outputs are authorized. The existence of these interacting sources of change supports lifecycle governance, but it does not establish a universal monitoring frequency, retraining threshold, or requalification procedure.

Proposed continually learning QSAR lifecycle

Concept-drift scholarship commonly separates three activities: detecting that a relevant change may have occurred, understanding its source and consequences, and adapting the system where justified [7]. The proposed QSAR lifecycle preserves this separation. A surveillance signal is not treated as proof of model failure, and adaptation is not treated as synonymous with retraining. Detection opens a controlled investigation in which chemical, experimental, computational, and decision-context explanations can be compared.

Deployment stability also depends on whether the associations learned during development remain transportable when the data-generating environment changes [8]. For QSAR, this requires more than comparing incoming descriptors with historical distributions. The lifecycle must consider whether the biological response retains the same meaning, whether assay and curation practices remain compatible, whether the representation preserves relevant distinctions, and whether the model is still being used within its authorized decision context.

Figure 1 presents the original conceptual synthesis for continually learning qsar lifecycle, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

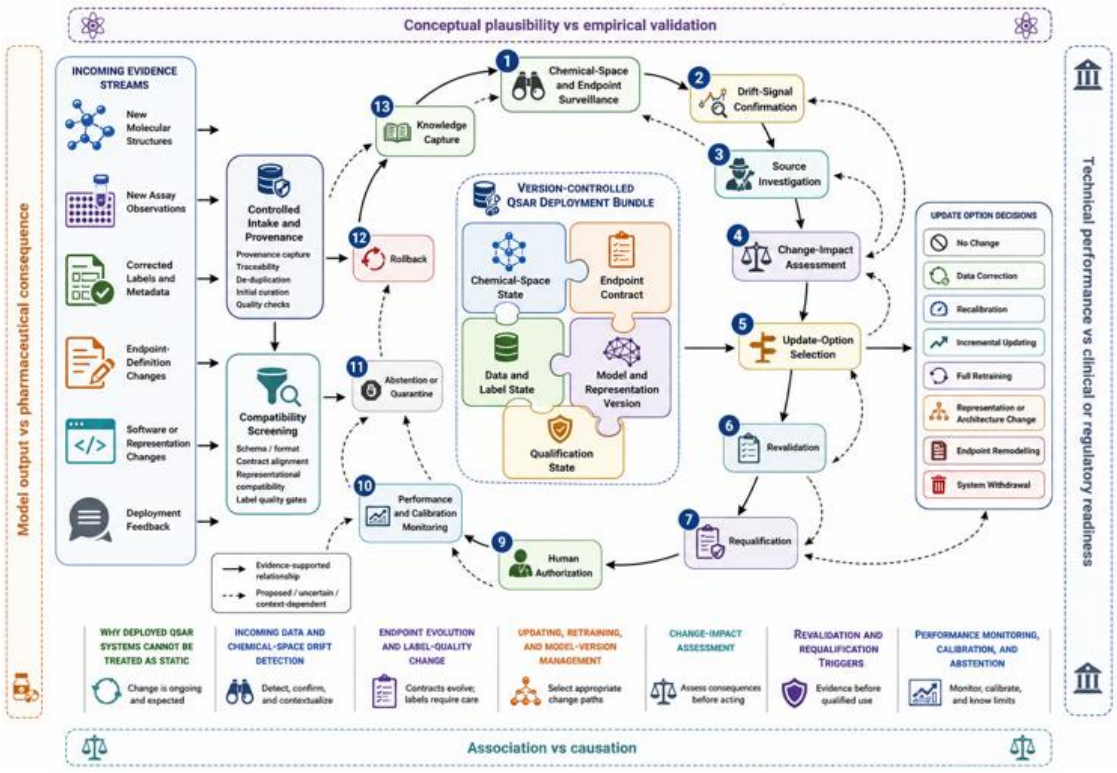

Figure 1. Continually Learning QSAR Lifecycle.

The figure is an original conceptual synthesis developed for "The Continually Learning QSAR System under Chemical-Space Drift, Endpoint Redefinition, Model Change, and Requalification throughout Deployment." Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. Incoming molecular structures, assay observations, corrected labels, endpoint changes, representation changes, software changes, and deployment feedback enter controlled intake and provenance assessment. Five interacting controlled objects—chemical-space state, endpoint contract, data and label state, model and representation version, and qualification state—are evaluated through compatibility screening, drift surveillance, investigation, change-impact assessment, update-option selection, revalidation, requalification, human authorization, controlled release, monitoring, abstention, quarantine, and rollback. Solid relationships represent concepts supported by the selected evidence, whereas dashed relationships represent the proposed lifecycle integration or relationships requiring empirical validation. The figure does not represent a validated predictive model, clinical recommendation, regulatory determination, or deployment-ready system.

Material post-deployment changes also require accountable intervention rather than unrestricted autonomous adaptation. Experience from deployed predictive systems indicates that dataset shift may require human review, restriction of use, temporary suspension, or withdrawal when the conditions supporting the original model no longer hold [9]. Transferred to QSAR governance, this principle supports an authorization boundary between detecting change and releasing a changed model, although it does not itself validate a specific pharmaceutical workflow.

The proposed lifecycle therefore governs five controlled objects. The chemical-space state represents the structural and property distribution for which evidence exists. The endpoint contract specifies the biological quantity, assay context, units, censoring rules, aggregation logic, and intended interpretation. The data and label state records provenance, curation, exclusions, and uncertainty. The model and representation version includes

the complete computational configuration rather than weights alone. The **qualification state** identifies the context in which the assembled system is presently supportable.

Around these objects, eight proposed functions form a closed lifecycle: controlled intake; compatibility and quality screening; drift and endpoint surveillance; investigation and impact assessment; update-option selection; revalidation and requalification; human authorization and controlled release; and monitoring with abstention, rollback, and knowledge capture. Continual learning is thus defined as a governed capacity to incorporate new evidence. It may lead to recalibration, retraining, endpoint remodelling, or no model change, and it remains conditional on empirical validation.

Incoming data and chemical-space drift detection

Incoming-data governance begins before model scoring. Each compound batch should be linked to source, time, intended use, structure-processing rules, assay provenance where available, and the deployed version expected to evaluate it. Applicability-domain methods based on molecular similarity can help identify compounds whose structural neighborhoods are poorly represented in the training data [10]. Such methods provide evidence of unfamiliarity, but their outputs are conditional on the representation, distance function, reference population, and endpoint.

Chemical-space distance alone is insufficient because response surfaces may be locally discontinuous. Activity cliffs demonstrate that structurally similar compounds can possess sharply different activities, allowing a model to appear globally well covered while remaining unreliable within specific local neighborhoods [11]. Drift surveillance should therefore distinguish broad distribution movement from local response-topology risk and should avoid interpreting high overall similarity as proof of supported prediction.

Predictive uncertainty adds another signal, but it is not a universal substitute for applicability analysis. Molecular uncertainty methods can behave differently across the form and severity of dataset shift, and some estimators may require complementary error models or diagnostics to remain informative [12]. A change in uncertainty distribution may indicate unfamiliar chemistry, model instability, altered noise, or estimator failure; it does not independently identify which explanation is correct.

The proposed surveillance layer combines chemical-space coverage, applicability-domain distance, local activity-cliff exposure, predictive uncertainty, calibration behavior, and observed error where labels are available [10–12]. Alerts are recorded against the affected endpoint, chemical region, data source, model version, and decision context. A confirmed signal then enters source investigation and change-impact assessment. No individual detector, fixed threshold, or alert count is proposed as sufficient to authorize retraining, requalification, or continued use.

Endpoint evolution and label-quality change

A QSAR endpoint is not merely a response column. Its meaning depends on assay type, target identity, biological system, units, qualifiers, censoring, aggregation, and provenance. Curated bioactivity resources preserve many of these attributes because identical numerical values may represent scientifically different observations [13].

Data-acquisition practices also evolve. Direct deposition, revised mappings, corrected structures, and enriched assay metadata can alter the evidence available for modelling without any change in molecular inputs [14]. Such changes may improve data quality, but they may also disrupt comparability with the endpoint originally used for qualification.

Assay context further determines whether observations sharing an endpoint label can be pooled legitimately. Differences in protocol, biological setting, detection technology, or experimental conditions may alter the interpretation of measured activity [15]. Endpoint drift must therefore be distinguished from chemical-space drift and ordinary random label noise.

The architecture proposes a version-controlled endpoint contract specifying assay principle, biological context, units, transformation rules, censoring, aggregation, activity thresholds, exclusions, and intended interpretation. Corrected labels may permit bounded maintenance, whereas substantive endpoint redefinition may require bridge analysis, parallel legacy and revised models, or construction of a new prediction target. These responses remain context-dependent and require empirical evaluation.

Updating, retraining, and model-version management

Updating should begin only after the source and consequence of change have been investigated. Possible actions include no change, metadata correction, recalibration, incremental updating, full retraining, representation

replacement, architecture change, endpoint remodelling, quarantine, or withdrawal. More data do not automatically justify a new deployed version.

Reproducible chemical-property modelling depends on explicit software, configuration, features, architecture, preprocessing, and training settings [16]. The proposed version object therefore extends beyond saved parameters to include the data snapshot, endpoint contract, representation, calibration method, applicability rule, software environment, evaluation evidence, and authorized context of use.

Representation change may be material even when the endpoint and training compounds remain nominally unchanged. Large-scale pretrained chemical representations can capture different structural and property information from task-specific encodings [17]. Their adoption should therefore be evaluated as a scientific model change rather than treated as an interchangeable software upgrade.

Architectures designed to recognize unfamiliar compounds or behave differently near chemical-space boundaries may alter both prediction and abstention behavior [18]. Such changes may be beneficial, neutral, or harmful depending on task and evaluation design. Aggregate performance alone cannot establish equivalence between versions.

Version management should preserve lineage among data, labels, endpoints, models, calibration objects, applicability rules, approvals, and rollback targets. **Table 1** organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, components, relationships, and intended uses in the continually learning qsar system under chemical-space drift, endpoint redefinition.

Table 1. Definitions, components, relationships, and intended uses in The Continually Learning QSAR System under Chemical-Space Drift, Endpoint Redefinition.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement
Chemical-space state	Unrecognized extrapolation	Incoming structures, representation, training coverage	Compare current chemistry with the qualified reference domain	Earlier identification of unsupported regions	External and temporally structured testing [10]	Similarity may miss local discontinuities	No universal distance threshold is established
Endpoint contract	Loss of target meaning	Assay context, units, qualifiers, aggregation	Version endpoint meaning independently from model code	Traceable distinction between new data and a new target	Assay-context comparability analysis [15]	Nominal labels may conceal biological differences	Shared terminology does not establish equivalence
Data and label state	Hidden provenance or curation change	Source, timestamps, corrections, exclusions	Link every model version to an immutable data snapshot	Reproducible learning evidence	Provenance and deposition records [14]	Corrected data may remain biased or incomplete	Curation does not guarantee validity
Model and representation version	Incomplete computational identity	Code, features, architecture, parameters, environment	Treat the full computational configuration as one controlled bundle	Reproducibility and impact tracing	Independent reconstruction of predictions [16]	Archived weights may not reproduce the system	Versioning is not validation
Applicability and calibration objects	Unsupported confidence	Domain rules, uncertainty outputs, calibration data	Govern these objects separately but link them to each model release	Bounded interpretation and abstention	Shift-sensitive evaluation [12]	Confidence may be miscalibrated	Eligibility does not prove correctness
Qualification state	Continued use after material change	Evidence package, intended use,	Connect supported use to a specific version bundle	Explicit authorization boundary	Context-aligned evaluation [22]	Documentation may outlive evidence	Proposed scientific state, not

		affected components					regulatory approval
Rollback point	Irreversible or partial recovery	Previous bundle, data lineage, deployment record	Restore the last supportable complete system state	Controlled recovery after failure	Reproducibility and rollback testing	Dependencies may prevent restoration	Rollback must be empirically demonstrated
Continual-learning controller	Fragmented response to change	All controlled objects and surveillance signals	**Proposed synthesis:** separate detection, investigation, change, authorization, and release	Coherent lifecycle governance	Comparative implementation studies	Excessive control or unsafe automation	The controller is conceptual and unvalidated

Change-impact assessment

Change-impact assessment asks which scientific claims may no longer hold after a detected event or proposed update. Dataset size is inadequate as a proxy because chemical collections can remain unevenly distributed across structural space despite containing many observations [19].

Out-of-distribution performance also depends on how unfamiliarity is constructed. Scaffold, cluster, property, temporal, or source-based separations can produce different robustness conclusions [20]. Assessment should therefore reflect the expected deployment challenge rather than rely on one convenient split.

Deep models do not remove applicability-domain limitations. Prediction error may remain related to structural coverage and similarity to training compounds even when flexible architectures are used [21]. Model replacement must consequently revisit domain behavior instead of inheriting the previous version's applicability claim.

The proposed assessment jointly examines chemical-space coverage, robustness under relevant OOD construction, and applicability-domain behavior [19–21]. It then traces direct and downstream effects on endpoint interpretation, calibration, uncertainty, abstention, supported decisions, monitoring, and rollback. Proposed minor, major, and critical change classes organize review scope but do not constitute validated thresholds.

Revalidation and requalification triggers

Revalidation should reproduce the evidence needed for the intended scientific claim rather than repeat a generic metric panel. Chemical-machine-learning evaluation requires alignment among datasets, splitting strategy, comparators, metrics, reproducibility, and intended use [22].

External validation is informative only when the relationship between development and validation contexts is examined. A separately collected dataset may still be too similar, while a highly different dataset may test a use that was never claimed [23]. Externality is therefore contextual rather than merely geographic or institutional.

The architecture distinguishes revalidation from requalification. Revalidation tests predictive behavior, coverage, calibration, uncertainty, reproducibility, and abstention after change. Requalification determines whether the complete changed bundle remains scientifically supportable for its stated use.

Broad review is proposed when endpoint meaning changes, chemical-space coverage expands substantially, model class or representation is replaced, calibration deteriorates without explanation, reproducibility fails, or rollback cannot be demonstrated. **Table 2** organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for testable propositions, evidence requirements, and validation criteria for the continually learning qsar system under chemical-space drift, endpoint redefinition.

Table 2. Testable propositions, evidence requirements, and validation criteria for The Continually Learning QSAR System under Chemical-Space Drift, Endpoint Redefinition.

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
Drift surveillance	Detect consequential change	Incoming chemistry and monitoring	Supplies evidence to impact assessment	Complementary detectors outperform a single global	False alerts or undetected local failure	Define deployment-relevant shift scenarios	Detection alone does not authorize updating

		signals enter investigation		measure in prespecified tests [11]			
Endpoint governance	Preserve target meaning	Assay and label changes update the endpoint record	Can trigger remodelling independently of chemistry	Comparable performance under stable versus revised endpoint definitions [13]	Incompatible labels treated as one target	Establish assay comparability experimentally	Numerical convertibility is not biological equivalence
Version management	Reproduce the changed system	Data, representation, code, calibration, and domain rules move as a bundle	Enables impact tracing and rollback	Independent reconstruction produces the intended model behaviour [16]	Partial archives or hidden dependencies	Maintain immutable artefacts and environments	Reproducibility does not prove validity
Change-impact assessment	Identify affected claims	Confirmed events are mapped to controlled objects	Determines revalidation scope	Tests cover relevant chemical-space and OOD conditions [20]	Convenient splits conceal deployment failure	Prespecify realistic challenge sets	One test partition cannot establish universal robustness
Revalidation	Test post-change behaviour	Candidate version enters comparative evaluation	Feeds requalification decision	Metrics, splits, uncertainty, and claims align with context of use [22]	Aggregate improvement conceals local deterioration	Conduct independent technical review	Better average performance does not prove equivalence
Requalification	Determine supportable use	Evidence package enters human authorization	Produces release, restriction, quarantine, or rejection	All materially affected claims receive adequate evidence [23]	Formal approval despite unresolved uncertainty	Approve, restrict, or reject with documented rationale	Proposed governance decision, not regulatory certification
Monitoring and abstention	Bound unsupported use	Predictions, uncertainty, coverage, and outcomes feed surveillance	Can trigger rollback or new investigation	Abstention improves retained-set reliability without unacceptable blind spots [24]	Overconfidence or excessive rejection	Evaluate utility and consequences prospectively	Abstention does not validate retained predictions
Rollback and closure	Restore prior state and capture learning	Deployment evidence returns to version history	Closes or reopens the change record	Full prior bundle can be restored and reproduced	Rollback restores only weights or fails operationally	Test recovery before release	Operational recovery must be demonstrated

Performance monitoring, calibration, and abstention

Monitoring should evaluate the deployed evidence relationship rather than track one performance statistic. Molecular uncertainty methods differ in calibration, error association, computational requirements, and behavior across tasks [24]. A confidence output should not therefore be interpreted as a universal probability of correctness. Comparative work likewise indicates that no tested neural uncertainty estimator is uniformly superior across molecular-property settings [25]. Method selection should be endpoint-specific and evaluated under the shifts, noise patterns, and decision consequences expected during use.

No single uncertainty output should be treated as sufficient because its quality varies across molecular tasks and shift conditions [12, 24, 25]. The proposed monitoring layer combines chemical coverage, uncertainty, calibration, error where observable, assay and label changes, and the frequency and location of abstention.

Abstention may operate at compound, chemical-region, endpoint, model-version, or system level. It can restrict extrapolation while additional evidence is generated, but it also reduces coverage and may concentrate errors among retained predictions. Its value therefore requires prospective assessment rather than assumption.

Auditability, human approval, and rollback

Auditability requires reconstruction of what changed, why it changed, which evidence was considered, who authorized release, and which use remained permitted. Explanations generated by predictive models may support investigation, but they should not automatically be treated as causal or mechanistic evidence [26].

Higher-consequence uses justify stronger interpretability and human scrutiny [27]. In the proposed architecture, approval is not ceremonial: reviewers may retain the existing version, restrict use, request additional evidence, quarantine a candidate, authorize controlled release, or withdraw the system.

Each change record should identify the trigger, affected controlled objects, competing explanations, selected response, rejected alternatives, evaluation results, residual uncertainty, approver, release date, monitoring plan, and rollback target. These records support accountability but cannot compensate for weak evidence.

Rollback should restore the complete previously supportable bundle, including data lineage, endpoint contract, representation, software environment, calibration, applicability rule, and authorization conditions. A rollback mechanism that restores only model parameters may fail to reproduce the earlier system and must not be assumed effective without testing.

Limitations and implementation priorities

Technical improvement does not establish pharmaceutical impact, reliable decision use, or implementation readiness [28]. The proposed architecture may increase traceability yet also impose substantial computational, experimental, and organizational burden, particularly for small datasets, delayed outcomes, proprietary assays, and rapidly changing model ecosystems.

Chemical and biological data remain shaped by sampling, measurement, curation, context, and incomplete knowledge [29]. Governance cannot eliminate these limitations. It can only make assumptions, changes, evidence gaps, and decision boundaries more visible.

Figure 2 presents the original conceptual synthesis for drift-triggered change-control and requalification pathway, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

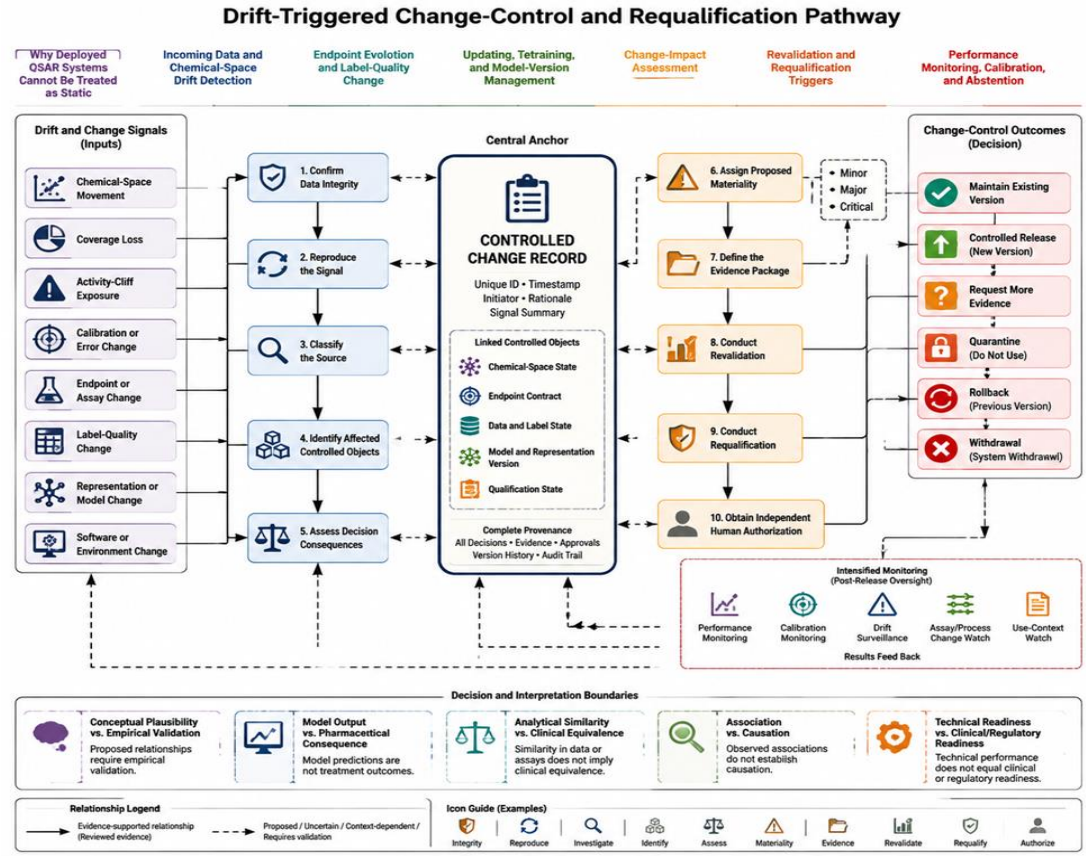

Figure 2. Drift-Triggered Change-Control and Requalification Pathway.

The figure is an original conceptual synthesis developed for "The Continually Learning QSAR System under Chemical-Space Drift, Endpoint Redefinition, Model Change, and Requalification throughout Deployment."

Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, clinical recommendation, regulatory determination, or deployment-ready system.

Priority evaluations should compare the proposed lifecycle with static and automatically retrained alternatives using retrospective temporal simulation, chemically structured external tests, endpoint-change challenges, reproducibility exercises, prospective shadow deployment, and rollback drills. Evaluation should include false-alert burden, undetected failure, abstention consequences, decision delay, and evidence quality without assuming that greater automation is preferable.

The architecture may require substantial local adaptation for multi-task models, foundation-model updates, federated learning, confidential assays, delayed or selectively observed labels, causal claims, and decisions with clinical or regulatory consequences. Its classifications and decision pathways remain proposed. They should be refined or rejected according to empirical findings rather than adopted as universal rules.

CONCLUSION

The proposed continually learning QSAR lifecycle reframes deployment as governance of a changing evidence relationship involving chemical space, endpoint meaning, labels, representation, model version, calibration, applicability, and supported use. It separates surveillance from adaptation, revalidation from requalification, and technical change from authorization. Its principal contribution is an auditable architecture linking drift detection, endpoint governance, controlled updating, impact assessment, monitoring, abstention, human approval, and rollback. The architecture remains an original conceptual synthesis requiring computational, experimental, and implementation validation; it does not establish universal thresholds, clinical recommendations, regulatory determinations, 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. 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
3. 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
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. 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. Morger A, Garcia de Lomana M, Norinder U, Svensson F, Kirchmair J, Mathea M, et al. Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data. Sci Rep. 2022;12(1):7244. doi:10.1038/s41598-022-09309-3
7. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: A review. IEEE Trans Knowl Data Eng. 2019;31(12):2346-63. doi:10.1109/TKDE.2018.2876857
8. Subbaswamy A, Saria S. From development to deployment: Dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345-52. doi:10.1093/biostatistics/kxz041
9. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283-6. doi:10.1056/NEJMc2104626

10. Liu R, Wallqvist A. Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. J Chem Inf Model. 2019;59(1):181-9. doi:10.1021/acs.jcim.8b00597
11. 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
12. Parrondo-Pizarro R, Lanini J, Rodríguez-Pérez R. Uncertainty quantification in molecular machine learning for property predictions under data shifts. J Chem Inf Model. 2026;66(2):923-35. doi:10.1021/acs.jcim.5c02381
13. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1). doi:10.1093/nar/gkw1074
14. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1). doi:10.1093/nar/gky1075
15. Schoenmaker L, Sastrokarijo EG, Heitman LH, Beltman JB, Jespers W, van Westen GJP. Toward assay-aware bioactivity model(er)s: Getting a grip on biological context. J Chem Inf Model. 2025;65(13):7013-23. doi:10.1021/acs.jcim.5c00603
16. 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
17. 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
18. van Tilborg D, Rossen L, Grisoni F. Molecular deep learning at the edge of chemical space. Nat Mach Intell. 2026;8(4):575-87. doi:10.1038/s42256-026-01216-w
19. Kretschmer F, Seipp J, Ludwig M, Klau GW, Böcker S. Coverage bias in small molecule machine learning. Nat Commun. 2025;16(1):554. doi:10.1038/s41467-024-55462-w
20. 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
21. Liu R, Wang H, Glover KP, Feasel MG, Wallqvist A. Dissecting machine-learning prediction of molecular activity: Is an applicability domain needed for quantitative structure–activity relationship models based on deep neural networks? J Chem Inf Model. 2019;59(1):117-26. doi:10.1021/acs.jcim.8b00348
22. Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. Nat Rev Chem. 2022;6(6):428-42. doi:10.1038/s41570-022-00391-9
23. Cabitza F, Campagner A, Soares F, García de Guadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external: Methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. 2021;208:106288. doi:10.1016/j.cmpb.2021.106288
24. 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
25. 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
26. 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
27. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15. doi:10.1038/s42256-019-0048-x
28. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
29. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037