



Molecular Models Should Adapt at Test Time When Drug-Like Chemical Space Shifts Beyond the Distributions Seen During Training

Mateo Alvarez^{1*}, Sofia Herrera², Diego Cruz¹, Andres Castro³

¹Department of Test-Time Adaptation for Molecular Models, Faculty of Pharmacy, National University of Colombia, Bogota, Colombia.

²Department of Chemical Space Distribution Shifts, Faculty of Pharmacy, University of Chile, Santiago, Chile.

³Department of Adaptable Drug-Like Space Modeling, Faculty of Pharmacy, University of Antioquia, Medellin, Colombia.

ABSTRACT

Molecular machine-learning models are generally optimized before deployment and subsequently treated as fixed predictors, although prospective drug discovery continuously presents compounds, property regimes, assay contexts, and target domains that may differ from those represented during training. This article develops an original computational framework for molecular test-time adaptation under such distribution shift. The analysis distinguishes structural or scaffold displacement, property-support shift, target- or assay-domain change, and mixtures of these conditions from ordinary in-distribution prediction error. It then separates information legitimately available during inference—molecular structure, alternative encodings, neighborhood support, predictive uncertainty, test-batch composition, and retained source information—from unavailable experimental bioactivity labels. The proposed framework treats adaptation as a conditional decision rather than a default response to novelty. Unlabeled self-supervised, consistency-based, entropy-related, transductive, or support-assisted objectives may be activated only when shift evidence, information sufficiency, and reliability criteria justify updating. A parallel risk channel addresses pseudo-label reinforcement, unstable sequential updates, and deterioration of previously reliable predictions, while uncertainty and applicability information support retention, restricted adaptation, rollback, or abstention. The framework further separates label-free test-time adaptation from later active recovery once experimental measurements become available. Its central implication is that prospective molecular prediction should incorporate controlled model plasticity without treating every unfamiliar molecule as an invitation to retrain. The framework remains methodological: molecular test-time adaptation requires prospective validation across chemically, biologically, and operationally realistic shifts before performance, discovery utility, or deployment readiness can be inferred.

Key Words: Molecular machine learning, Distribution shift, Test-time adaptation, Uncertainty quantification, Applicability domain, Drug discovery

eIJPPR 2026; 16(1):117-125

HOW TO CITE THIS ARTICLE: Alvarez M, Herrera S, Cruz D, Castro A. Molecular Models Should Adapt at Test Time When Drug-Like Chemical Space Shifts Beyond the Distributions Seen During Training. Int J Pharm Phytopharmacol Res. 2026;16(1):117-125. <https://doi.org/10.51847/tK3vQmeAj5>

INTRODUCTION

A molecular predictor is usually evaluated as though training defines the model and testing merely interrogates it. Prospective discovery reverses that convenience. Newly synthesized compounds, changing medicinal-chemistry programs, emerging target classes, and assay-specific data streams can progressively move inference away from the

chemical and statistical support that shaped the fitted representation. Recent molecular evidence shows that predictive reliability can deteriorate near the edge of familiar chemical space, that molecular out-of-distribution status depends on how the shift itself is specified, and that apparent robustness changes substantially across alternative OOD evaluation regimes [1–3]. These findings

Corresponding author: Mateo Alvarez

Address: Department of Test-Time Adaptation for Molecular Models, Faculty of Pharmacy, National University of Colombia, Bogota, Colombia.

E-mail: ✉ mateo.alvarez@unal.edu.co

Received: 28 October 2025; **Revised:** 26 January 2026; **Accepted:** 27 January 2026

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.



establish a deployment problem, but they do not establish that changing the model during inference will solve it.

The central premise developed here is therefore narrower than “models should continuously learn.” A discovery model should first determine whether its present world differs meaningfully from the one encoded during training, then ask what information is actually available and whether any update can be justified without access to new experimental labels. In this view, test-time adaptation is a bounded computational intervention between two safer alternatives: retaining a sufficiently reliable frozen predictor and declining to make an unsupported prediction. The object of adaptation is not novelty itself but a detectable mismatch for which unlabeled test-time information contains useful structure and for which updating can be constrained against degradation.

This distinction matters chemically. A compound can be structurally unfamiliar yet lie on a locally predictable property surface; conversely, a close analogue can encounter an activity cliff, changed assay context, or target-domain discontinuity. Model confidence, chemical distance, and distributional novelty therefore cannot be treated as interchangeable signals. The proposed framework integrates these signals into a conditional adaptation architecture while preserving a strong frozen baseline, explicit uncertainty, and the option not to adapt. Its purpose is not to claim prospective superiority, but to specify what a scientifically defensible molecular test-time adaptation system would have to detect, use, constrain, evaluate, and refuse.

Chemical distribution shift in prospective drug discovery

Chemical distribution shift is broader than movement in fingerprint space. Medicinal-chemistry datasets evolve

through compound selection, synthesis priorities, assay campaigns, target programs, endpoint definitions, and data-acquisition practices. Increasing the scale of training data can strengthen molecular models, but scaling does not remove the generalization problem when new chemistry or prediction contexts differ systematically from the data that determined the fitted parameters [4]. Likewise, molecular benchmarking has long shown that conclusions depend on task definition, splitting strategy, endpoint type, imbalance, and evaluation metric rather than on model architecture alone [5]. A realistic definition of prospective shift must therefore describe what changed, relative to which training support, and at what analytical level.

This article treats distribution shift as a change in the relationship among molecular inputs, prediction targets, and the process generating observations. Structural displacement concerns movement into sparsely represented chemotypes or substructures. Property-support displacement occurs when test compounds occupy regions of the response landscape poorly sampled during training, even when structural novelty is moderate. Target- or assay-domain displacement arises when biological or measurement context changes sufficiently that a previously learned structure–label relationship may no longer transport cleanly. Data-process displacement covers temporal, curation, portfolio, or laboratory changes that alter the observed distribution without necessarily introducing a new scaffold.

These shifts can coexist, and none by itself proves that a prediction is wrong. Their value is diagnostic: they identify reasons why an IID performance estimate may no longer describe the current inference problem. **Table 1** separates the principal shift constructs by what changes, what becomes uncertain, and what adaptation must not assume.

Table 1. Distribution-shift constructs relevant to prospective molecular prediction

Construct	What changes at inference	Primary computational concern	What cannot be assumed
Structural shift	Local chemotypes, scaffolds, or substructures	Representation support and extrapolation	Structural novelty implies prediction failure
Property-support shift	Region of the response landscape occupied by test molecules	Extrapolation beyond observed target support	Nearby structures have smooth property values
Target/assay-domain shift	Biological target, endpoint, or measurement context	Transportability of learned structure–label relations	The same molecular pattern retains the same functional meaning
Data-process shift	Portfolio, time, curation, or acquisition process	Dataset comparability and calibration	Distributional change is purely chemical
Mixed shift	Multiple structural, property, biological, or process changes	Identifying which signal should govern adaptation	A single similarity or uncertainty score explains the mismatch

The broader argument is summarized in **Figure 1**: chemical shift initiates a diagnostic problem, but adaptation becomes eligible only after the available

inference-time information, update risks, uncertainty, and validation boundary are considered together.

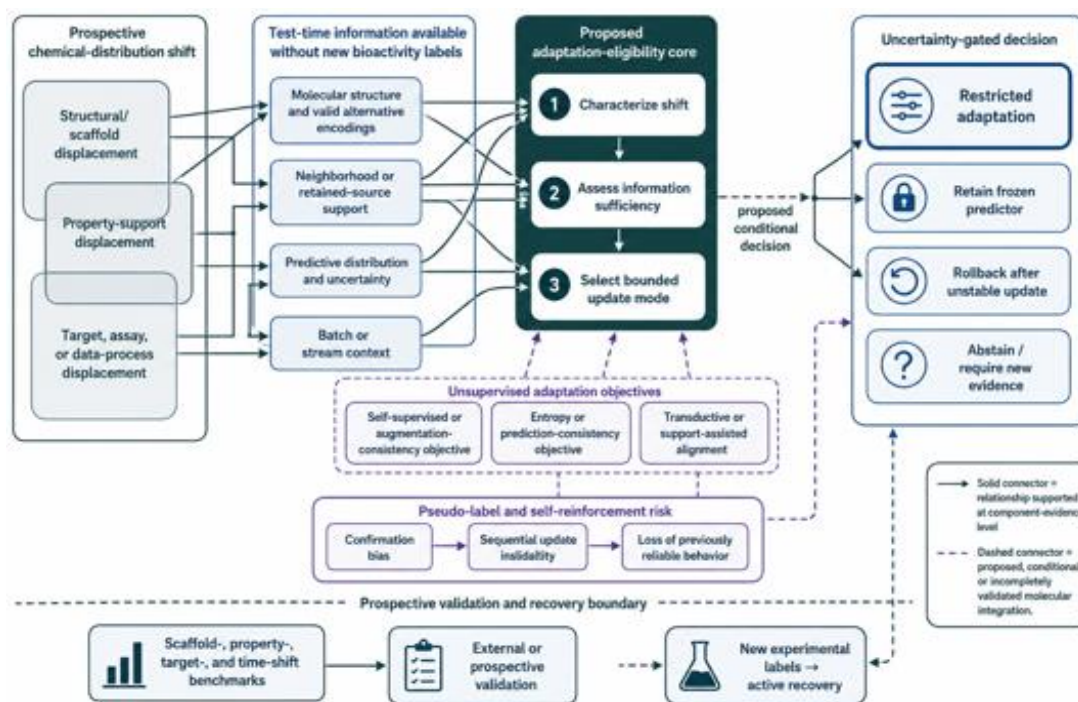


Figure 1. From prospective chemical shift to bounded molecular test-time adaptation

Scaffold shift, property shift, and target-domain shift

Scaffold shift is the most intuitive molecular form of distribution change, but it is not sufficient as a complete deployment definition. Learned molecular representations have shown task- and dataset-dependent generalization across both public and proprietary prediction problems, indicating that representation quality cannot be separated from the regime in which it is evaluated [6]. Structural departure therefore matters because the learned representation may no longer organize test compounds in a way that preserves the relationships useful for the target task. Yet a novel scaffold can remain predictable if the relevant property relation is smooth, while a familiar scaffold can be difficult if small modifications produce steep response changes.

Target-domain shift introduces a different problem. Virtual-screening research has shown that benchmark construction can encode strong biases in active/inactive discrimination, making apparent generalization partly a property of how compounds and targets were sampled [7]. A change in target or assay domain can therefore modify what the label means, how negatives arise, or which structural cues remain informative. The resulting failure should not automatically be attributed to “chemical novelty,” because altered biological or measurement context may be the more consequential source of mismatch.

Property shift further complicates any similarity-only criterion. Activity-cliff analyses demonstrate that closely related molecules can exhibit large changes in measured activity and can remain difficult for molecular machine-

learning models [8]. Chemical distance and property distance are therefore related imperfectly, particularly in rugged local SAR regions. The three constructs are best viewed as overlapping diagnostics rather than mutually exclusive classes. A prospective compound may combine a new scaffold, an extreme property region, and a changed assay context, leaving the adaptation problem to determine which evidence can legitimately modify the model rather than merely identify unfamiliarity.

Why static molecular models cannot correct themselves

A static molecular predictor may be sophisticated without being adaptive. Training-time strategies can improve invariance, representation quality, and OOD robustness before deployment; recent molecular OOD work, for example, shows that more consistent semantic representations can improve performance under predefined distribution shifts [9]. Such methods are valuable because many prospective failures should be prevented during training rather than repaired later. They nevertheless leave a basic computational fact unchanged: once the fitted model is frozen, the arrival of a new unlabeled chemical distribution does not alter its parameters, internal statistics, or task representation.

That limitation is most consequential when the new distribution exposes regions where the learned structure–property relation is poorly supported. Molecular property landscapes can differ substantially in roughness, and greater landscape roughness has been associated with poorer out-of-sample modellability [10]. A frozen model can report a prediction and perhaps an uncertainty estimate

in such a region, but without an update mechanism it cannot use repeated unlabeled evidence from the emerging test distribution to reorganize its representation or decision surface. Whether it should do so is a separate question.

The proposed framework therefore does not treat adaptivity as an intrinsic advantage. Updating introduces new failure modes that a frozen model does not possess, including confirmation bias, sequential dependence, and forgetting of reliable behavior. A defensible molecular TTA method must consequently beat a strong frozen comparator under the same shift while preserving calibration and behavior on chemistry for which the original model remains trustworthy. “Cannot correct itself” should thus be interpreted mechanically, not normatively: a static predictor lacks inference-time plasticity, but plasticity is justified only when the available evidence makes controlled updating preferable to preserving the original model.

What information is actually available at test time?

The most important constraint on molecular test-time adaptation is that the true property or bioactivity of a newly encountered compound is ordinarily unavailable at the moment adaptation is proposed. What remains available can nevertheless be richer than a single query structure. A model may observe the molecular graph or string, valid stochastic or conformational encodings, distances or neighborhoods in representation space, predictions from an ensemble, repeated views of the same molecule, and the composition of an arriving batch or sequential stream. If deployment permits retention of source statistics, prototypes, or model-state summaries, those can supply an additional reference against which the new distribution is interpreted.

Uncertainty belongs in this information layer, but it is not a ground-truth error signal. Comparative molecular studies show that neural uncertainty methods differ in how effectively their scores correspond to predictive error, spatially resolved approaches demonstrate that uncertainty can sometimes be localized rather than represented only as one global value, and conformalized molecular prediction provides a distinct coverage-oriented reliability construction [11–13]. Together these approaches support using uncertainty as diagnostic evidence, not as an unquestioned command to adapt.

The proposed test-time information hierarchy therefore distinguishes query-intrinsic evidence, population evidence, and retained-source evidence. Query-intrinsic evidence comes from the molecule and the model’s response to it; population evidence arises when several unlabeled test compounds reveal a coherent emerging distribution; retained-source evidence permits explicit comparison with training-time support. Adaptation should become more permissive only when the information

required by the chosen objective is genuinely available. When it is not, the scientifically appropriate action may remain frozen prediction, widened uncertainty, or abstention.

Unsupervised adaptation without new bioactivity labels

Test-time adaptation is meaningful only if the update can be defined without pretending that experimental labels are already known. General TTA methodology provides a family of candidate objectives in which unlabeled target examples influence model state during inference, including entropy-related objectives, consistency under valid transformations, self-supervised prediction, normalization or feature-statistic updates, and transductive exploitation of the test distribution [14]. These methods establish computational possibilities, not direct evidence that any particular objective improves prospective molecular prediction.

The information retained from model development materially changes the problem. Source-free domain adaptation explicitly addresses cases in which the original training data are unavailable during adaptation, whereas support-assisted settings may preserve source examples, prototypes, sufficient statistics, or other anchors against which an arriving test distribution can be compared [15]. For molecular deployment this distinction is consequential: proprietary, computational, governance, or privacy constraints may prevent access to original medicinal-chemistry data even when the trained predictor remains deployable. The proposed framework therefore treats reference-free and support-assisted adaptation as different operating modes rather than variants that can be evaluated interchangeably.

Unlabeled updating also creates a circularity that molecular models cannot ignore. If predictions are converted into pseudo-labels and then used to train the same model, initial errors may be reinforced instead of corrected. Sequential test-time training studies in other machine-learning domains demonstrate that naive self-training can suffer confirmation bias as adaptation proceeds through a shifted stream [16]. This evidence does not establish the frequency or magnitude of the problem in medicinal chemistry, but it identifies a credible methodological hazard. Molecular TTA should therefore prefer objectives anchored in chemically valid self-supervision, transformation consistency, test-distribution structure, or carefully filtered pseudo-labels, and expose pseudo-label-dependent updates to independent uncertainty and stability checks.

A molecular test-time adaptation framework

The proposed molecular test-time adaptation framework begins only after shift is detected or credibly suspected. Its first operation is diagnosis, not updating: determine

whether arriving chemistry differs structurally, occupies poorly supported property regions, changes target or assay context, or reflects a data-process shift. Unlabeled test distributions can contain information useful for molecular OOD prediction [17], while molecular transferability varies with source–target relatedness rather than data availability alone [18]. Chemical language models further show that unlabeled molecular representations encode reusable structural regularities [19]. These findings support ingredients for adaptation, not prospective validation of the integrated framework.

The framework separates three decisions. An eligibility decision asks whether the shift is coherent and whether available information can support an unlabeled objective.

A mode decision selects reference-free adaptation when only the deployed model and target molecules are available, or support-assisted adaptation when legitimate source summaries, prototypes, or examples remain accessible. A safety decision limits what may change and tests whether post-update behavior remains acceptable. All three decision rules are proposed here.

Candidate updates may use self-supervised consistency, target-distribution structure, or filtered pseudo-label information, but they must compete against retaining the frozen predictor. Post-update checks determine whether to retain a restricted adapted state, revert, or abstain. **Table 2** organizes the resulting operating distinctions without treating adaptation as the default outcome.

Table 2. Proposed operating distinctions for molecular test-time adaptation

Decision dimension	Reference-free mode	Support-assisted mode	Proposed safety implication
Available evidence	Model plus unlabeled target molecules	Target molecules plus retained source support	Never assume unavailable training information
Shift comparison	Internal uncertainty, neighborhood structure, stream coherence	Target-to-source comparison supplements internal signals	More reference evidence may justify narrower adaptation
Candidate objective	Self-supervision, valid-view consistency, transductive structure	Same objectives plus source anchoring	Objective must match available information
Update scope	Prefer minimal or restricted changes	Restrict changes against source behavior	Broad updating needs stronger justification
Failure response	Freeze, rollback, or abstain	Re-anchor, restrict, rollback, or abstain	Adaptation is not the default success state
Recovery boundary	Pseudo-labels do not constitute new experimental evidence	Retained source labels do not constitute new target evidence	New target labels begin recovery

Note: The distinctions in this table are proposed framework elements rather than validated universal operating rules.

121

Figure 2 then represents the framework as a conditional progression in which an eligible update must still pass uncertainty and stability checks before being retained.

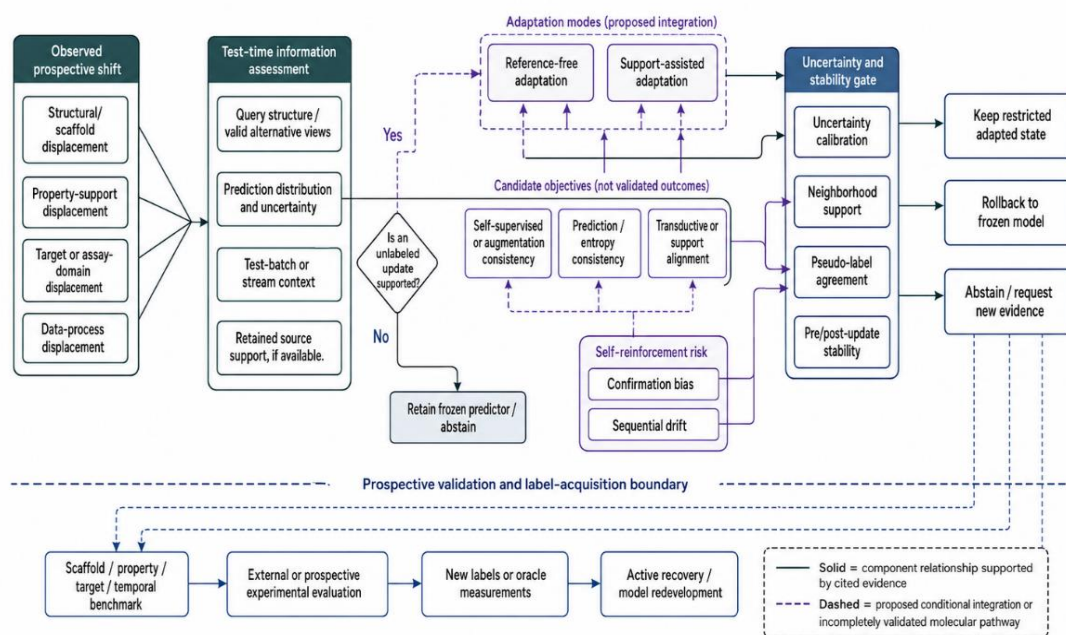


Figure 2. Conditional progression from molecular shift detection to adaptation, rollback, or recovery

Preventing pseudo-label drift and self-reinforcing error

A dangerous adaptation failure occurs when an initially plausible prediction becomes training material for its own reinforcement. Sequential test-time training in other domains shows that naive self-training can amplify confirmation bias [16]. In molecular prediction, that risk is especially concerning around sparse chemotypes, activity cliffs, or property extremes, where an erroneous pseudo-label could reshape the model in the region with weakest support.

Molecular uncertainty has been used to guide property prediction and candidate prioritization [20], but it cannot certify a pseudo-label. The proposed framework therefore requires pseudo-label-dependent updates to combine acceptable uncertainty with local or source support, agreement across valid molecular views or perturbations, and limited parameter movement. Failure of any criterion does not prove intrinsic unpredictability; it weakens the case for updating.

Online TTA benchmarks also show sensitivity to batch size, stream order, optimization, and stability controls [21]. Molecular deployment should therefore preserve a recoverable frozen state, compare pre- and post-update behavior, limit cumulative update magnitude, and rollback when stability cannot be demonstrated. A successful update is thus not defined merely by movement of parameters or increased confidence, but by controlled improvement without evidence of destabilizing previously reliable chemistry.

Uncertainty-gated adaptation

Uncertainty should regulate adaptation eligibility rather than automatically trigger it. Molecular graph models can estimate multiple uncertainty components, including epistemic and aleatoric contributions [22]. Such distinctions are useful because lack of support for a chemotype differs from noisy measurement, although uncertainty decomposition does not reveal the true causal source of error.

Recent molecular analyses also show that common failure modes can weaken the correspondence between uncertainty scores and actual errors [23]. The framework therefore proposes a multi-signal gate in which uncertainty is considered jointly with structural or representation-space support, prediction consistency, test-stream coherence, and retained-source comparison when available. No universal scalar threshold is assumed.

Low uncertainty should not authorize unrestricted adaptation under evident shift, and high uncertainty should not force updating when the model lacks information from which to learn. The gate may instead narrow the update, favor the frozen model, increase monitoring, or recommend abstention. Uncertainty is therefore treated as

one component of an evidential decision layer rather than a surrogate for ground truth.

Deciding when not to adapt

A molecular model should not adapt merely because a compound is distant from its training examples. Distance-aware prediction shows that location in nonlinear structure–property space can provide applicability information [24], but the meaning of distance depends on representation, endpoint, and local property geometry. Distance is evidence about support, not a universal cutoff. The proposed framework identifies three reasons not to adapt. Information insufficiency occurs when unfamiliarity is detected but the arriving chemistry offers too little coherent unlabeled structure for a defensible objective. Update instability occurs when candidate updates materially change predictions without improving reliability. Scientific discontinuity occurs when assay change, activity-cliff behavior, or extreme property displacement leaves unlabeled evidence unable to resolve what relation should be learned.

In such states, retaining a calibrated frozen prediction may be preferable; when even that prediction lacks defensible support, abstention is preferable to false precision. “Do not adapt” is therefore an explicit framework outcome rather than a failure of the adaptive system. This distinction is important because the objective is reliable inference under shift, not maximal parameter plasticity.

Prospective benchmarks that simulate real chemical shift

Random splitting alone cannot validate molecular TTA. OOD benchmarking shows that robustness conclusions depend strongly on how distributional difference is constructed, prospective industrial ADME studies demonstrate the value of later-arriving compounds, and public-to-internal Caco-2 transfer supplies a concrete external-domain stress test [3, 25, 26]. These regimes answer complementary questions rather than forming one benchmark hierarchy.

The proposed evaluation architecture separates scaffold shift, property-support extrapolation, target or assay transfer, temporal evolution, external organizational shift, and sequential stream effects. Every adaptive model should be compared with the frozen predictor and a strong training-time robustness method. Evaluation should examine task performance together with calibration, uncertainty–error association, abstention behavior, update stability, degradation on previously reliable chemistry, stream-order sensitivity, and update burden.

Table 3 organizes the principal benchmark dimensions without implying that any one split reproduces prospective discovery.

Table 3. Proposed benchmark dimensions for molecular adaptation under realistic shift

Shift dimension	Benchmark construction	Primary question	Required comparator or safeguard
Scaffold	Hold out coherent chemotypes or structural families	Can the method handle structural displacement?	Frozen model plus strong OOD-trained baseline
Property support	Test sparsely represented response regions	Does adaptation help beyond observed target support?	Assess predictive error and calibration together
Target/assay	Separate biological or measurement contexts	Does structure–label information transport?	Preserve assay and target interpretation
Temporal	Train earlier and evaluate later arrivals	Can the method tolerate evolving chemistry?	Preserve chronology
External/internal	Separate organizational data sources	Does adaptation survive operational change?	Document curation and measurement differences
Sequential stream	Present shifted compounds in realistic order	Does updating accumulate benefit or drift?	Test batch size, order sensitivity, and rollback

Note: These dimensions constitute a proposed combined validation architecture; they do not represent validated universal benchmark requirements or performance thresholds.

Model recovery after encountering novel chemical space

Test-time adaptation should end when its information assumptions end. Atomic uncertainty methods show that molecular failures can sometimes be localized to critical structural regions rather than represented only by global confidence [27]. Localization can identify where a deployed model struggles, but it does not determine which update will repair the problem or whether unlabeled information is sufficient.

Once new experimental labels or reliable oracle evaluations are obtained, the methodological problem changes. Closed-loop active learning demonstrates that new molecular evaluations can update predictive models and guide search beyond the initially characterized region [28]. This supports a distinct recovery phase in which data acquisition, model redevelopment, and deeper retraining become legitimate.

The proposed framework therefore separates reversible, label-free adaptation from label-acquiring recovery. Blurring the two would allow pseudo-labels to masquerade as experimental evidence and would make claimed recovery difficult to interpret. Recovery is not a more aggressive form of TTA; it is a transition to a different evidence regime in which newly measured target information can legitimately alter the learned structure–property relationship.

Deployment risks and boundaries

Deployment risk begins with misidentifying the source of shift. Activity cliffs can undermine similarity-based reasoning [8], conformal uncertainty depends on assumptions that may become fragile under distribution change [13], and online adaptation can depend on test-stream composition and ordering [21]. A safe system must therefore distinguish chemical novelty, measurement change, calibration failure, and prediction error rather than collapse them into one OOD score.

Molecular uncertainty itself can be unreliable in difficult regimes [23]. The proposed gate consequently requires

endpoint- and representation-specific stress testing rather than assuming that one calibrated confidence score transports across targets, assays, model architectures, or changing chemical programs. Prospective industrial validation remains important because retrospective results may not capture evolving chemistry [25], while external transfer illustrates how assay and organizational differences can create additional applicability questions [26].

The framework is not a claim of discovery effectiveness. Successful adaptation on a shifted benchmark would not establish target engagement, pharmacological mechanism, in-vivo exposure, therapeutic benefit, or regulatory acceptability. When unlabeled evidence remains inadequate, crossing into label-acquiring recovery is more defensible than continuing uncontrolled adaptation [28]. The deployment boundary is therefore both computational and scientific: a model should know not only when its statistical environment has changed, but when available evidence is insufficient to justify altering its own behavior.

Limitations

This framework has several important limitations. First, its adaptation logic is synthesized from molecular OOD, uncertainty, transferability, prospective-validation, and general test-time adaptation evidence rather than from a prospective medicinal-chemistry trial of the integrated architecture. The proposed gates, update restrictions, rollback rules, and operating modes therefore remain hypotheses to be tested rather than established performance-improving procedures.

Second, “distribution shift” can conceal heterogeneous causes. Structural novelty, assay drift, target change, temporal portfolio evolution, label noise, and property-landscape discontinuity may generate similar diagnostic signals while requiring different responses. The framework can organize these possibilities but cannot infer their causal origin from unlabeled molecules alone. Third, its usefulness depends on what is available during

deployment: single-query inference provides less population evidence than coherent batches or streams, while source-free settings remove potentially valuable anchors.

Finally, uncertainty, distance, and prediction consistency are model-dependent proxies rather than universal measures of epistemic adequacy. A system could therefore adapt confidently and still move in the wrong direction. Prospective testing must treat nonadaptation, rollback, and abstention as legitimate outcomes rather than failures to use the framework.

CONCLUSION

Molecular prediction increasingly operates where inference chemistry does not remain stationary. The framework developed here treats this neither as a mandate for continuous retraining nor as a reason to accept frozen predictions uncritically. It proposes that a model characterize the shift, identify what information is genuinely available, decide whether that information supports a bounded unlabeled update, and preserve the option to remain frozen, rollback, or abstain.

The central distinction is between detecting that the world has changed and possessing evidence sufficient to change the model safely. Uncertainty, chemical support, test-stream structure, retained source information, and update stability contribute different evidence and should not be collapsed into one confidence score. Label-free adaptation must also remain separate from recovery after new experimental measurements become available.

The framework therefore defines a validation agenda rather than an implementation claim. Its value depends on prospective testing across structural, property, target, assay, temporal, and operational shifts with strong frozen comparators and explicit monitoring of calibration, drift, and abstention.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] 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
- [2] Tossou P, Wognum C, Craig M, Mary H, Noutahi E. Real-world molecular out-of-distribution: Specification and investigation. *J Chem Inf Model.* 2024;64(3):697-711. doi:10.1021/acs.jcim.3c01774
- [3] 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
- [4] Chen J, Chung Y, Tynan J, Cheng C, Yang S, Cheng AC. Data scaling and generalization insights for medicinal chemistry deep learning models. *J Chem Inf Model.* 2025;65(12):5887-98. doi:10.1021/acs.jcim.5c00538
- [5] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci.* 2018;9(2):513-30. doi:10.1039/C7SC02664A
- [6] Yang K, Swanson K, Jin W, Coley CW, 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
- [7] Tran-Nguyen VK, Jacquemard C, Rognan D. LIT-PCBA: An unbiased data set for machine learning and virtual screening. *J Chem Inf Model.* 2020;60(9):4263-73. doi:10.1021/acs.jcim.0c00155
- [8] van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. *J Chem Inf Model.* 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
- [9] 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
- [10] Aldeghi M, Graff DE, Frey N, Morrone JA, Pyzer-Knapp EO, Jordan KE, et al. Roughness of molecular property landscapes and its impact on modellability. *J Chem Inf Model.* 2022;62(19):4660-71. doi:10.1021/acs.jcim.2c00903
- [11] 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
- [12] Heid E, Schörghuber J, Wanzenböck R, Madsen GKH. Spatially resolved uncertainties for machine learning potentials. *J Chem Inf Model.* 2024;64(16):6377-87. doi:10.1021/acs.jcim.4c00904
- [13] 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
- [14] Liang J, He R, Tan T. A comprehensive survey on test-time adaptation under distribution shifts. *Int J Comput*

- Vis. 2025;133(1):31-64. doi:10.1007/s11263-024-02181-w
- [15] Fang Y, Yap PT, Lin W, Zhu H, Liu M. Source-free unsupervised domain adaptation: A survey. *Neural Netw.* 2024;174:106230. doi:10.1016/j.neunet.2024.106230
- [16] Su Y, Xu X, Li T, Jia K. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering regularized self-training. *IEEE Trans Pattern Anal Mach Intell.* 2024;46(8):5524-40. doi:10.1109/TPAMI.2024.3370963
- [17] 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
- [18] 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
- [19] Skinnider MA, Stacey RG, Wishart DS, Foster LJ. Chemical language models enable navigation in sparsely populated chemical space. *Nat Mach Intell.* 2021;3(9):759-70. doi:10.1038/s42256-021-00368-1
- [20] 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
- [21] Wang Z, Luo Y, Zheng L, Chen Z, Wang S, Huang Z. In search of lost online test-time adaptation: A survey. *Int J Comput Vis.* 2025;133(3):1106-39. doi:10.1007/s11263-024-02213-5
- [22] Jiang S, Qin S, Van Lehn RC, Balaprakash P, Zavala VM. Uncertainty quantification for molecular property predictions with graph neural architecture search. *Digit Discov.* 2024;3(8):1534-53. doi:10.1039/D4DD00088A
- [23] 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
- [24] Kim JY, Vlachos DG. Distance-aware molecular property prediction in nonlinear structure-property space. *J Chem Inf Model.* 2025;65(13):6744-56. doi:10.1021/acs.jcim.5c01037
- [25] Fang C, Wang Y, Grater R, Kapadnis S, Black C, Trapa P, et al. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *J Chem Inf Model.* 2023;63(11):3263-74. doi:10.1021/acs.jcim.3c00160
- [26] Wang D, Jin J, Shi G, Bao J, Wang Z, Li S, et al. ADMET evaluation in drug discovery: 21. Application and industrial validation of machine learning algorithms for Caco-2 permeability prediction. *J Cheminform.* 2025;17(1):3. doi:10.1186/s13321-025-00947-z
- [27] Qian J, Luo Y, Zhou Q, Guan X. Atomic uncertainty pinpoints critical failure structures for trustworthy molecular property prediction. *J Chem Inf Model.* 2026. doi:10.1021/acs.jcim.6c00973
- [28] Antoniuk ER, Li P, Keilbart N, Weitzner S, Kailkhura B, Hiszpanski AM. Active learning enables generation of molecules that advance the known Pareto front. *NPJ Comput Mater.* 2026;12(1):81. doi:10.1038/s41524-025-01924-8