



What Should Molecular Models Learn from Inactive Compounds When Absence of Activity Reflects Biology, Assay Design, Exposure, or Measurement Limits?

Lukas Weber^{1*}, Sophie Janssen², Jonas Richter¹, Elena Fischer³

¹Department of Inactive Compound Learning and Assay Design, Faculty of Pharmacy, University of Stuttgart, Stuttgart, Germany.

²Department of Biological Absence and Exposure Limits, Faculty of Pharmaceutical Sciences, Eindhoven University of Technology, Eindhoven, Netherlands.

³Department of Measurement Limits and Model Training, Faculty of Pharmacy, University of Basel, Basel, Switzerland.

ABSTRACT

Molecular bioactivity models typically treat inactive compounds as a homogeneous negative class, yet inactivity can arise from multiple biologically and experimentally distinct causes. A compound may genuinely fail to engage its target, reach insufficient effective concentration, fall below detection thresholds, encounter unfavorable cellular states, or remain untested rather than being truly negative. This theory article examines the consequences of collapsing these different evidential states into a single computational label and argues that inactivity should be interpreted primarily as an observation generated under specified experimental conditions rather than as an intrinsic molecular property. Drawing on evidence from assay context, target engagement, intracellular exposure, solubility and permeability, experimental uncertainty, phenotypic state, negative sampling, and model evaluation, we develop a latent-cause framework for inactive bioactivity labels. This framework distinguishes relatively certain negative evidence from ambiguous negative evidence and unobserved activity, while preserving the possibility that genuinely inactive compounds provide valuable structural information. The proposed interpretation does not require uncertain negatives to be relabeled as active; instead, it motivates models and benchmarks that explicitly represent the provenance and confidence of negative evidence. Such treatment may improve the scientific meaning of molecular prediction and candidate deprioritization. However, the theory does not establish that any specific uncertainty-aware algorithm will outperform conventional binary learning. Its practical value therefore depends on richer assay metadata, orthogonal measurements, context-aware evaluation, and prospective validation.

Key Words: Bioactivity prediction, Inactive compounds, Label uncertainty, Assay context, Molecular machine learning, Negative evidence

eIJPPR 2025; 15(5):101-110

HOW TO CITE THIS ARTICLE: Weber L, Janssen S, Richter J, Fischer E. What Should Molecular Models Learn from Inactive Compounds When Absence of Activity Reflects Biology, Assay Design, Exposure, or Measurement Limits? Int J Pharm Phytopharmacol Res. 2025;15(5):101-110. <https://doi.org/10.51847/vn3MiOdt9m>

INTRODUCTION

Molecular machine-learning datasets often encode bioactivity as a binary distinction: active versus inactive. Computationally, this representation is convenient because each observation can be converted into a target variable and optimized through conventional classification

objectives. Biologically, however, the label is more conditional. Assay-aware modeling has shown that biological and experimental context can contribute information not contained in molecular structure alone [1]. Integrated experimental–computational workflows similarly emphasize that model interpretation depends on how bioactivity measurements were produced and

Corresponding author: Lukas Weber

Address: Department of Inactive Compound Learning and Assay Design, Faculty of Pharmacy, University of Stuttgart, Stuttgart, Germany.

E-mail: ✉ lukas.weber@uni-stuttgart.de

Received: 01 June 2025; **Revised:** 09 August 2025; **Accepted:** 12 August 2025

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.



contextualized [2]. An inactive label therefore does not simply describe a molecule; it describes the relationship among a molecule, a biological system, an assay configuration, a concentration regime, and a decision rule. The problem becomes sharper when measurements from different sources are combined. Cross-source aggregation of IC₅₀ and K_i values can introduce substantial noise because superficially similar potency measurements are not automatically interchangeable [3]. Prediction behavior can also change when confidence and curation rules are altered, demonstrating that the training signal depends partly on decisions made before modeling begins [4]. These observations do not imply that bioactivity databases are unreliable. They instead indicate that data provenance and confidence are part of the scientific meaning of a label. Large bioactivity resources necessarily integrate heterogeneous assay types, targets, activity measures, and experimental records [5]. Converting these records into binary outcomes can therefore erase information about why a compound crossed—or failed to cross—a chosen activity threshold. This article proposes that molecular models should treat inactivity as evidence with a cause and a degree of certainty, rather than as a context-free molecular attribute. The proposal is deliberately narrower than claiming that inactive labels are generally wrong. Many inactives are likely to be informative negatives. The theoretical problem is that biologically credible inactivity, experimentally unresolved inactivity, and merely unobserved activity can become computationally indistinguishable once compressed into the same zero.

The many biological meanings hidden inside a zero label

A zero label can occupy different positions in the pharmacological sequence between molecular encounter and measured response. Target-engagement assays are

specifically designed to determine whether a compound interacts with its intended target in a biologically relevant environment, distinguishing engagement from downstream functional readouts [6]. Experimental approaches such as CeTEAM further demonstrate that target engagement can be linked to subsequent pharmacological response without assuming that the two measurements are equivalent [7]. Consequently, absence of a downstream effect cannot, by itself, identify whether the compound failed at target recognition, cellular access, engagement, signal propagation, or the measured endpoint. At the same time, inactivity can be scientifically informative. InertDB demonstrates that inactive small molecules can be curated as an explicit resource rather than discarded as failed observations [8]. Structure–inactivity relationship analysis likewise emphasizes that molecular changes associated with loss of activity can constrain medicinal-chemistry reasoning and define boundaries around productive chemical space [9]. A negative observation should therefore not automatically be discounted simply because it lacks a positive phenotype.

A further distinction arises when “negatives” are constructed for computational purposes. Virtual-screening workflows frequently rely on decoys whose role is to contrast known actives with compounds selected to behave as plausible nonbinders within a benchmark [10]. Such examples can be methodologically useful without being equivalent to experimentally demonstrated target inactivity. The important theoretical distinction is thus not active versus inactive alone, but what type of evidence produced the inactive state and what conclusions that evidence can support.

The resulting distinction between established experimental interpretations and the extension proposed here is summarized in **Table 1**.

Table 1. From observed inactivity to evidentially qualified negative labels: established interpretations and the proposed modeling extension

Observed negative state	What existing evidence permits	What the observation does not establish	Proposed modeling extension
Failure to engage a target	Supports negative target-engagement evidence when the assay is appropriate	Universal absence of affinity across conditions	Treat as relatively strong negative evidence, conditional on assay adequacy
No downstream pharmacological response	Supports absence of the measured functional response	Which upstream step produced the nonresponse	Preserve functional nonresponse separately from inferred nonbinding
Curated inactive annotation	Provides reusable negative bioactivity information	Context-independent mechanistic inactivity	Retain provenance and confidence rather than collapse to an unqualified zero
Structure–inactivity relationship	Can identify structural changes associated with activity loss	That every inactive analogue failed for the same mechanism	Use inactivity as informative local chemical evidence where experimental context is compatible
Computational decoy	Provides contrast for virtual-screening evaluation or training	Experimentally demonstrated biological inactivity	Encode as constructed negative evidence rather than measured inactivity

Note: The final column is a proposed theoretical extension developed in this article; it is not presented as an experimentally validated negative-label taxonomy.

True target inactivity

A defensible theory of ambiguous inactivity must preserve the possibility of genuine negative evidence. If a compound is tested under conditions that provide a credible opportunity for target engagement and nevertheless fails to engage, the observation can reasonably carry more negative evidential weight than an untested or artificially generated compound [6]. Likewise, when target engagement and functional pharmacology are measured separately, concordant absence across appropriately designed measurements provides a stronger basis for deprioritization than downstream nonresponse alone [7]. “True target inactivity” should therefore be understood here as an evidential category, not as an absolute metaphysical claim about a molecule.

This distinction matters because inactivity can contain useful chemical information. Curated inactive datasets demonstrate that negative observations can be retained systematically [8], while structure–inactivity relationships can reveal where molecular modifications are associated with loss of measured activity [9]. Such information may help define local boundaries in chemical space, particularly when compared within a coherent assay series. Treating every negative as suspect would therefore discard precisely the counterexamples needed to learn specificity.

The proposed theory instead separates strong negative evidence from weakly interpretable negative evidence. A computational decoy selected to challenge a model is useful for a different reason than an experimentally reproduced nonengager [10]. Neither should be erased, but neither should be assumed to mean the same thing. The relevant question for molecular learning is not whether inactivity is “real” in the abstract; it is how strongly the available experiment constrains the hypothesis that the molecule can produce the target or functional activity of interest under the conditions being modeled.

Insufficient exposure masquerading as lack of activity

A molecule may be intrinsically capable of engaging a target yet fail to generate a cellular response because an adequate concentration never reaches the relevant intracellular compartment. Direct measurements of intracellular exposure have shown that accounting for compound access can help bridge discrepancies between target-based potency and cell-based pharmacology [11]. This creates an important interpretive asymmetry: a negative cellular outcome may provide evidence against effective cellular activity without providing equally strong evidence against intrinsic target recognition.

Permeability is one component of that distinction. Cell-free permeability systems are used precisely because molecular access across membrane-like barriers is a developability variable rather than a guaranteed consequence of

extracellular dosing [12]. More direct analytical approaches can go further: in-cell NMR has been used to quantify both cellular permeability and target engagement for middle-sized molecules, demonstrating experimentally that access and engagement are separable properties that can be interrogated within cells [13]. Accordingly, failure of a cell-based assay should not automatically be translated into a structure-level negative label when target access has not been established.

Solubility introduces another constraint on effective concentration. Analyses of intrinsic-solubility modeling show that estimates are sensitive to the quality and consistency of the underlying experimental data [14]. Nominal concentration therefore should not be treated as equivalent to freely available concentration without qualification. The theory proposed here consequently treats exposure-limited inactivity as a candidate explanation, not a default rescue hypothesis for every negative result. Cellular exposure, permeability, solubility, nonspecific binding, target abundance, and assay duration may interact, and the present evidence does not justify attributing a specific inactive observation to one of these causes without additional measurements. What molecular modeling should preserve is the distinction between “no effect despite credible exposure” and “no effect under unresolved exposure conditions.”

Assay sensitivity and censoring

Binary labels also compress uncertainty created by the assay itself. A measured response close to an activity threshold is not equivalent to an observation far from that threshold, yet a hard cutoff can assign both the same class. Probabilistic modeling of bioactivity measurements has shown that explicitly incorporating experimental uncertainty can improve prediction near classification thresholds [15]. This does not imply that every near-cutoff inactive is secretly active; it means that the label carries less certainty than its binary encoding suggests.

Assay validity creates a separate problem. Observed screening signals can be compromised by assay interference, artefactual hits, or nuisance chemistry and therefore require validity assessment before mechanistic interpretation [16–18]. Much of this literature concerns false-positive signals, so it should not be used to infer that false-negative inactivity is equally prevalent. Its relevance here is epistemic: an assay readout is not automatically identical to the underlying molecular event.

For inactive labels, censoring therefore has at least two meanings. A compound may fall below a predefined detection or activity boundary, and the experiment may leave the magnitude or cause of that nonresponse unresolved. The proposed theory treats such cases as measurement-qualified negatives.

Cell context and pathway state as sources of conditional inactivity

Cellular inactivity is conditional on the state of the biological system in which it is measured. Context-aware deconfounding of cell-line drug response shows that cellular covariates can materially affect predictive relationships [19]. Phenotypic profiles likewise provide information that complements chemical structure, indicating that the biological state produced by a compound need not be recoverable from molecular representation alone [20]. A compound that is inactive in one cellular configuration may therefore supply evidence about that configuration without establishing context-independent pharmacological inactivity.

Morphological profiling makes this distinction operational. Cell Painting-based prediction has been used to improve screening prioritization in tested high-throughput settings [21]. Transcriptomic approaches extend the same principle to molecular state: chemical-induced expression profiles can encode perturbational information relevant to phenotype-based discovery, and generative models can predict context-dependent transcriptional responses to new perturbations. Cellular-profile studies show that compound-induced biological state can provide activity information beyond a structure-only representation [21–23]. These profiles do not establish target mechanism.

A negative functional label can therefore be conditional on cell lineage, pathway configuration, target abundance, compensatory signaling, dose, and time. When such contextual variables are available, they should be treated as part of the observation rather than as irrelevant metadata.

Negative labels created by experimental design

Some negative labels are not experimental nonresponses at all. Drug–target interaction datasets often contain known positive edges and a much larger set of unknown relationships; computational construction of the learning task determines which of those relationships are treated as negative examples [24]. In such settings, “negative” may mean “not known to interact,” which is epistemically different from “tested and found not to interact.”

Positive–unlabeled learning was developed for precisely this asymmetry: positives are observed while the remaining examples are not assumed to be confirmed negatives [25]. Molecular bioactivity studies have also shown that adding stochastic negative examples can change predictive performance [26]. These approaches demonstrate that negative-class composition is a modeling choice as well as a biological fact. They do not establish that unlabeled examples are likely active, only that unknown status should not be mistaken for experimentally demonstrated inactivity.

Not all measured inactives are equally useful either. Data-valuation approaches in high-throughput screening indicate that training examples can differ in informational value [27], while models that combine molecular structures with bioassay descriptions show that assay context can add predictive information beyond structure alone [28].

Table 2 organizes the architecture that follows directly from these distinctions.

Table 2. Proposed construct architecture for negative evidence in molecular bioactivity learning

Evidence construct	Defining observation	Evidential interpretation	Appropriate modeling stance
Measured strong negative	Credible assay opportunity with reproducible nonactivity [6]	Relatively strong evidence against activity under matched conditions	Preserve as informative negative; retain assay scope
Measurement-qualified negative	Nonresponse near a threshold or under unresolved assay sensitivity [15]	Negative evidence with reduced certainty	Represent uncertainty or confidence where available
Context-qualified negative	Nonresponse under a defined cellular, dose, or pathway state [19]	Evidence against activity in that state	Condition the label on biological context
Exposure-qualified negative	Functional nonresponse without established effective target-site exposure [11]	Evidence against observed effect, not necessarily intrinsic target recognition	Preserve exposure metadata; avoid mechanistic overinterpretation
Unobserved relationship	Compound–target pair not experimentally resolved [25]	Missing status rather than demonstrated inactivity	Treat as unlabeled unless justified otherwise
Constructed negative	Decoy or sampled nonpositive generated for learning [26]	Computational contrast, not empirical inactivity	Encode provenance and evaluate sensitivity to sampling

Note: The constructs and recommended modeling stances are proposed in this article; supporting evidence for their component distinctions derives from the preceding experimental and computational literature.

The relationships among observed labels, latent explanations, and defensible model interpretations are shown in **Figure 1**.

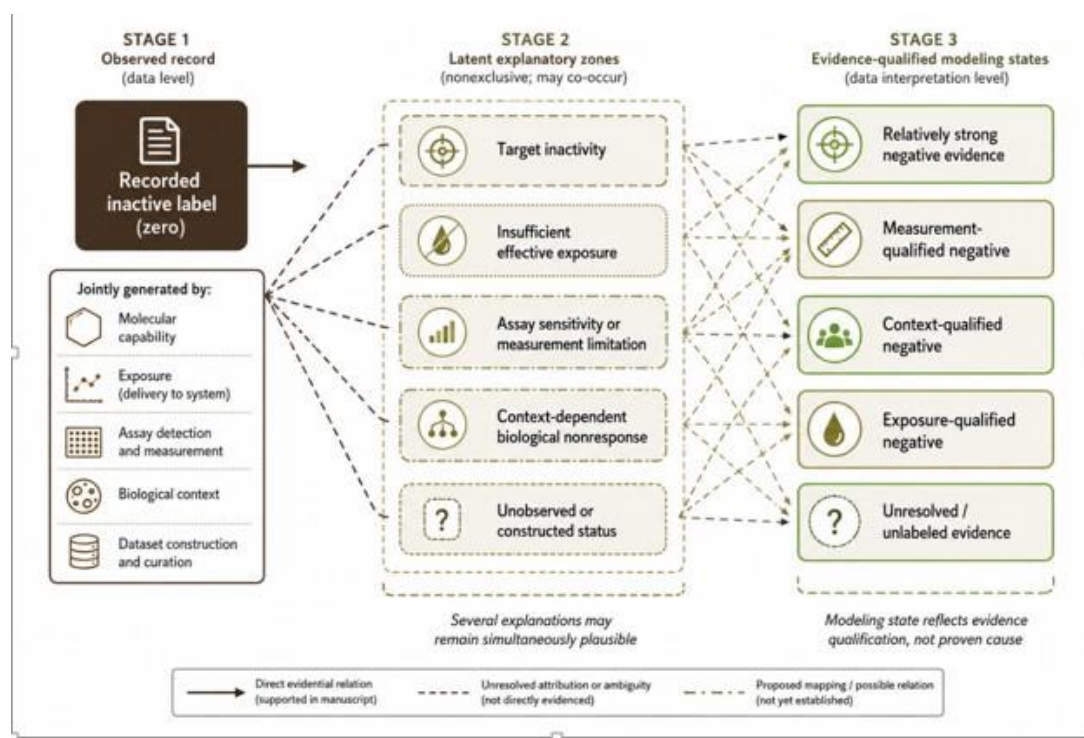


Figure 1. From a recorded zero to an evidentially qualified negative: a proposed latent-cause model of inactive bioactivity labels

Why treating all inactives equally distorts molecular learning

Collapsing heterogeneous negatives into one class changes what a classifier is asked to learn. Dose-dependent transcriptomic modeling shows that biological response can vary with concentration [29], whereas activity-cliff benchmarks demonstrate that small structural changes can correspond to large changes in measured activity and remain difficult for molecular models [30]. A binary inactive class can contain both without indicating which uncertainty dominates.

The problem is not solved simply by choosing a more powerful algorithm. Large-scale comparisons across thousands of bioactivity datasets show substantial task-dependent variation in model performance [31]. Analyses of transfer learning similarly indicate that bioactivity tasks differ in intrinsic hardness and transferability [32]. These findings caution against interpreting one classifier's success as evidence that the negative class has coherent biological meaning.

The theoretical claim is therefore about supervision, not architecture. When certain negatives, uncertain observations, context-specific nonresponses, and constructed negatives are optimized as interchangeable zeros, the loss function treats evidentially different observations as equally authoritative. That can encourage a model to learn assay provenance, chemical sampling rules, or dataset-specific shortcuts rather than a generalizable boundary of molecular inactivity. This

article does not claim that such distortion always decreases conventional accuracy. A heterogeneous label can still be highly predictable. The concern is that predictability and scientific meaning may diverge.

A latent-cause theory of inactive bioactivity labels

The proposed latent-cause theory begins from a distinction between the observed label and the unobserved reason that produced it. Let the recorded outcome be an inactive observation under a particular assay context. Behind that observation may lie several nonexclusive states: genuine lack of target interaction, inadequate exposure, a detection-limited response, context-dependent absence of downstream effect, or absence of an experiment from which activity could have been observed. Existing evidence establishes these component possibilities separately; the theory proposed here integrates them as alternative explanations of the same computational symbol.

Two dimensions should then be separated. The first is latent cause: *why* activity was not observed. The second is evidential confidence: *how strongly* the available experiment constrains the activity hypothesis. A reproduced target-engagement failure may carry high negative confidence even when the molecular reason remains unknown. Conversely, an untested drug–target pair has low negative confidence because there is no direct negative experiment. An exposure-limited cellular result can be experimentally reliable yet mechanistically weak

evidence about intrinsic target binding. Cause and confidence are therefore not the same variable.

This distinction also sets an identifiability boundary. Molecular structure alone cannot generally reveal which latent state produced a zero when distinct experimental processes yield the same label. Assay-aware modeling [1] and the added predictive value of assay descriptions [28] support the importance of contextual information, but they do not establish that metadata can recover every latent cause. Some ambiguity is irreducible without orthogonal measurements.

The theory therefore proposes an evidence-aware target representation rather than mandatory relabeling. A dataset may retain the observed bioactivity class while attaching provenance, confidence, exposure status, assay context, or missingness indicators where available. The aim is not to turn ambiguous negatives into positives. It is to prevent absence of evidence, evidence of absence, and evidence obtained under constrained observation from being treated as logically identical.

Modeling certain negatives, ambiguous negatives, and unobserved activity

Once negative evidence is separated conceptually, several modeling strategies become possible without prescribing a universal algorithm. Large-scale ChEMBL benchmarking has long shown that model performance depends on the task, representation, and algorithmic regime [33]. More recent virtual-screening evaluation work similarly argues that model assessment should reflect the retrieval objective and class distribution rather than rely on a single generic accuracy paradigm [34].

Relatively certain negatives can remain conventional negative examples. Ambiguous negatives may instead receive uncertainty-aware likelihoods, confidence weights, interval or censored treatment when measurements justify it, or separate auxiliary indicators. Unobserved pairs may be handled through positive–unlabeled methods rather than automatically assigned to the negative class [25]. Constructed negatives may remain useful for discrimination, but their provenance should be retained and

sensitivity to sampling assessed [26]. Data-valuation methods provide another possibility when observations contribute unequally to learning [27].

These are candidate implementations, not validated consequences of the latent-cause theory. Their suitability will depend on the information actually available, the prediction task, and the cost of different errors. A dataset with only structures and binary endpoints cannot recover experimental context that was never recorded. In that case, explicit acknowledgement of label ambiguity may be more scientifically defensible than an elaborate model that implies unavailable certainty.

Learning from inactivity without inflating false positives

Recognizing ambiguity must not become a license to discount every negative. Out-of-distribution evaluations show that molecular models can lose robustness when test chemistry departs from the training distribution [35]. If uncertain inactives are simply softened or relabeled without constraint, a model may increase apparent sensitivity by declaring chemically unfamiliar compounds plausible actives, thereby shifting rather than solving the error problem.

A safer interpretation is asymmetric. Strong negative evidence should continue to oppose unsupported positive predictions. Ambiguous negatives should contribute less certainty about inactivity, not positive evidence of activity. Unobserved pairs should remain unresolved unless additional information changes their status. This preserves the distinction between “the experiment did not establish activity” and “the molecule is likely active.”

For screening applications, that distinction matters because false positives consume experimental capacity while false negatives can remove useful chemistry. The appropriate balance depends on the intended decision and class prevalence, which is why virtual-screening evaluation should be tied to retrieval objectives rather than generic classification performance [34]. The proposition matrix in **Table 3** translates the theory into testable, falsifiable modeling expectations rather than implementation claims.

Table 3. Testable propositions for learning from heterogeneous negative evidence

Proposition	Predicted consequence if the theory is useful	Required test	Falsifying or limiting result
P1: Preserve provenance	Models can distinguish measured from constructed or unobserved negatives	Matched benchmark with provenance retained versus collapsed	No reproducible benefit or improved validity
P2: Represent uncertainty	Near-threshold or context-limited negatives need not exert full binary certainty	Compare hard labels with calibrated uncertainty-aware treatment	Worse calibration or unacceptable false-positive inflation
P3: Condition on context	Biological and assay metadata may improve interpretation of conditional nonresponse	Context-held-out and chemistry-held-out evaluation	Gains vanish outside dataset-specific shortcuts
P4: Protect strong negatives	Ambiguity-aware learning should not erase well-supported inactivity	Error analysis stratified by negative-evidence strength	Strong negatives are systematically promoted as positives

P5: Separate missingness	Unobserved pairs should be evaluated separately from tested negatives	Positive–unlabeled or explicit-missingness benchmark	Conventional negative assignment performs equally well under robust external tests
---------------------------------	---	--	--

Note: These propositions are proposed for validation; no threshold, expected effect size, or performance gain is asserted.

Evaluation when the negative class is heterogeneous

A model cannot validate the proposed theory if its benchmark already hides the relevant distinctions. Benchmark redundancy, chemical bias, and evaluation design can inflate or misrepresent apparent molecular-model performance [36–38]. Consequently, random or similarity-leaking splits are insufficient tests of whether evidence-aware negative handling improves generalization.

Evaluation should separate at least three questions. First, does the model discriminate within familiar chemistry? Second, does it remain calibrated when chemical or assay context changes? Third, does its treatment of ambiguous negatives alter the operational error balance relevant to screening? Explicit out-of-distribution testing is particularly important because strong in-distribution performance does not guarantee robustness under distribution shift [35]. Virtual-screening metrics should likewise reflect the practical retrieval objective [34].

Where annotation permits, errors should also be stratified by negative-evidence status rather than pooled into one confusion matrix. Performance on experimentally supported negatives, threshold-adjacent observations, context-qualified nonresponses, constructed negatives, and unlabeled pairs can then be examined separately. Such analysis would test whether a model learns useful distinctions or merely exploits provenance. Calibration, enrichment, ranking, and false-positive burden may all be informative, but no single metric is inherently privileged. The evaluation design must match the scientific claim, consistent with broader guidance for machine learning in chemical sciences [38].

Implications for dataset construction and benchmarking

The theory shifts dataset construction from label cleaning alone toward evidence preservation. At minimum, bioactivity records should retain assay identity, measurement type, activity threshold or quantitative value where available, concentration regime, provenance, and confidence information rather than immediately converting every observation to an anonymous binary endpoint. Assay-aware modeling [1], cross-source potency analysis [3], curation studies [4], database design [5], and assay-description modeling [28] collectively indicate why those contextual fields can matter, although their usefulness will vary by task.

Negative provenance should also remain explicit. Experimentally measured inactives, curated inactive

resources [8], computational decoys [10], unknown drug–target pairs [24], and sampled negatives [26] should not become indistinguishable after preprocessing. This does not require excluding any category. It requires recording enough information to evaluate how each category influences a model. Datasets that cannot reconstruct such provenance should state the limitation rather than infer missing experimental histories.

For benchmarking, the central comparison should be between models trained on the same underlying chemistry but given different representations of negative evidence. Prospective or reconstructed datasets could pair bioactivity outcomes with orthogonal information on target engagement, intracellular exposure, assay reproducibility, concentration, solubility, permeability, and biological context where feasible. The proposed framework would be supported if evidence-aware representations improve calibration or decision-relevant generalization without unacceptable false-positive inflation; it would be weakened if the distinctions add no reproducible value beyond ordinary labels.

Candidate deprioritization is the most consequential application. Structure–inactivity relationships show that genuine negatives can be highly informative [9], and data valuation indicates that examples need not contribute equally to learning [27]. The practical recommendation is therefore not to rescue inactive compounds indiscriminately. It is to align the strength of deprioritization with the strength of the negative evidence. A compound repeatedly inactive despite credible exposure and target opportunity warrants a different interpretation from one that was never tested, poorly exposed, or observed only under a context known to constrain the readout. Whether this evidence-aware strategy improves discovery efficiency remains an empirical question.

CONCLUSION

“Inactive” is often computationally represented as a molecular property even though it is experimentally produced as an observation. The distinction matters because absence of activity can reflect genuine target inactivity, inadequate exposure, assay sensitivity, biological context, or missing experimentation, and these states do not carry equivalent evidential meaning.

This article proposes a latent-cause theory in which the recorded label is separated from both the reason for nonobservation and the confidence of the negative

evidence. The framework preserves genuine inactives as valuable counterexamples while distinguishing them from measurement-qualified, context-qualified, exposure-qualified, constructed, and unobserved negatives. It does not require uncertain negatives to be relabeled as active and does not establish that uncertainty-aware learning will improve every prediction task.

The central methodological implication is that molecular models should learn from the provenance of inactivity whenever the relevant evidence exists and acknowledge nonidentifiability when it does not. Testing that proposition requires benchmarks that preserve assay and biological context, challenge chemical generalization, stratify errors by evidential status, and measure false-positive consequences. Until such tests are performed, the framework should be treated as a theory for designing more interpretable questions about negative data—not as a validated algorithm or universal rule for candidate selection.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] 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
- [2] Charest N, Sinclair G, Eytcheson SA, Chang DT, Martin TM, Lowe CN, et al. Combined in vitro and in silico workflow to deliver robust, transparent, and contextually rigorous models of bioactivity. *J Chem Inf Model*. 2025;65(9):4426-41. doi:10.1021/acs.jcim.5c00713
- [3] Landrum GA, Riniker S. Combining IC₅₀ or K_i values from different sources is a source of significant noise. *J Chem Inf Model*. 2024;64(5):1560-7. doi:10.1021/acs.jcim.4c00049
- [4] Xerxa E, Vogt M, Bajorath J. Influence of data curation and confidence levels on compound predictions using machine learning models. *J Chem Inf Model*. 2024;64(24):9341-9. doi:10.1021/acs.jcim.4c01573
- [5] Zdrazil B, Félix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res*. 2024;52(D1). doi:10.1093/nar/gkad1004
- [6] St John-Campbell S, Bhalay G. Target engagement assays in early drug discovery. *J Med Chem*. 2025;68(12):12331-68. doi:10.1021/acs.jmedchem.4c03115
- [7] Valerie NCK, Sanjiv K, Mortusewicz O, Zhang SM, Alam S, Pires MJ, et al. Coupling cellular drug-target engagement to downstream pharmacology with CeTEAM. *Nat Commun*. 2024;15(1):10347. doi:10.1038/s41467-024-54415-7
- [8] An S, Lee Y, Gong J, Hwang S, Park IG, Cho J, et al. InertDB as a generative AI-expanded resource of biologically inactive small molecules from PubChem. *J Cheminform*. 2025;17(1):49. doi:10.1186/s13321-025-00999-1
- [9] López-López E, Fernández-de Gortari E, Medina-Franco JL. Yes SIR! On the structure-inactivity relationships in drug discovery. *Drug Discov Today*. 2022;27(8):2353-62. doi:10.1016/j.drudis.2022.05.005
- [10] Victoria-Muñoz F, Menke J, Sánchez-Cruz N, Koch O. Efficient decoy selection to improve virtual screening using machine learning models. *J Cheminform*. 2025;17(1):165. doi:10.1186/s13321-025-01107-z
- [11] Mateus A, Gordon LJ, Wayne GJ, Almqvist H, Axelsson H, Seashore-Ludlow B, et al. Prediction of intracellular exposure bridges the gap between target- and cell-based drug discovery. *Proc Natl Acad Sci U S A*. 2017;114(30). doi:10.1073/pnas.1701848114
- [12] Jacobsen AC, Visentin S, Butnaru C, Stein PC, di Cagno MP. Commercially available cell-free permeability tests for industrial drug development: Increased sustainability through reduction of in vivo studies. *Pharmaceutics*. 2023;15(2):592. doi:10.3390/pharmaceutics15020592
- [13] Sukigara Y, Kamoshida H, Tokunaga Y, Fujisaki M, Imai M, Kobayashi Y, et al. Cell permeability and target engagement of middle-sized molecules quantified by in-cell NMR. *Anal Chem*. 2025;97(35):18955-64. doi:10.1021/acs.analchem.5c01494
- [14] Zhao J, Hermans E, Sepassi K, Tistaert C, Bergström CAS, Ahmad M, et al. Effect of data quality and data quantity on the estimation of intrinsic solubility: Analysis based on a single-source data set. *Mol Pharm*. 2024;21(10):5261-71. doi:10.1021/acs.molpharmaceut.4c00685
- [15] Mervin LH, Trapotsi MA, Afzal AM, Barrett IP, Bender A, Engkvist O. Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account

- experimental uncertainty. *J Cheminform.* 2021;13(1):62. doi:10.1186/s13321-021-00539-7
- [16] Aldrich C, Bertozzi C, Georg GI, Kiessling L, Lindsley C, Liotta D, et al. The ecstasy and agony of assay interference compounds. *J Med Chem.* 2017;60(6):2165-8. doi:10.1021/acs.jmedchem.7b00229
- [17] Alves VM, Yasgar A, Wellnitz J, Rai G, Rath M, Braga RC, et al. Lies and liabilities: Computational assessment of high-throughput screening hits to identify artifact compounds. *J Med Chem.* 2023;66(18):12828-39. doi:10.1021/acs.jmedchem.3c00482
- [18] Hu H, Yi X, Xue L, Baell JB. A collection of useful nuisance compounds (CONS) for interrogation of bioassay integrity. *JACS Au.* 2024;4(12):4883-91. doi:10.1021/jacsau.4c00851
- [19] He D, Liu Q, Wu Y, Xie L. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nat Mach Intell.* 2022;4(10):879-92. doi:10.1038/s42256-022-00541-0
- [20] Moshkov N, Becker T, Yang K, Horvath P, Dancik V, Wagner BK, et al. Predicting compound activity from phenotypic profiles and chemical structures. *Nat Commun.* 2023;14(1):1967. doi:10.1038/s41467-023-37570-1
- [21] Fredin Haslum J, Lardeau CH, Karlsson J, Turkki R, Leuchowius KJ, Smith K, et al. Cell Painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. *Nat Commun.* 2024;15(1):3470. doi:10.1038/s41467-024-47171-1
- [22] Tong X, Qu N, Kong X, Ni S, Zhou J, Wang K, et al. Deep representation learning of chemical-induced transcriptional profile for phenotype-based drug discovery. *Nat Commun.* 2024;15(1):5378. doi:10.1038/s41467-024-49620-3
- [23] Qi X, Zhao L, Tian C, Li Y, Chen ZL, Huo P, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nat Commun.* 2024;15(1):9256. doi:10.1038/s41467-024-53457-1
- [24] Veleiro U, de la Fuente J, Serrano G, Pizurica M, Casals M, Pineda-Lucena A, et al. GeNNius: An ultrafast drug-target interaction inference method based on graph neural networks. *Bioinformatics.* 2024;40(1). doi:10.1093/bioinformatics/btad774
- [25] Li F, Dong S, Leier A, Han M, Guo X, Xu J, et al. Positive-unlabeled learning in bioinformatics and computational biology: A brief review. *Brief Bioinform.* 2022;23(1). doi:10.1093/bib/bbab461
- [26] Cáceres EL, Mew NC, Keiser MJ. Adding stochastic negative examples into machine learning improves molecular bioactivity prediction. *J Chem Inf Model.* 2020;60(12):5957-70. doi:10.1021/acs.jcim.0c00565
- [27] Hesse J, Boldini D, Sieber SA. Machine learning-driven data valuation for optimizing high-throughput screening pipelines. *J Chem Inf Model.* 2024;64(21):8142-52. doi:10.1021/acs.jcim.4c01547
- [28] Schuh MG, Boldini D, Sieber SA. Synergizing chemical structures and bioassay descriptions for enhanced molecular property prediction in drug discovery. *J Chem Inf Model.* 2024;64(12):4640-50. doi:10.1021/acs.jcim.4c00765
- [29] Godinez WJ, Trifonov V, Fang B, Kuzu G, Pei L, Guiguemde WA, et al. Compound activity prediction with dose-dependent transcriptomic profiles and deep learning. *J Chem Inf Model.* 2024;64(7):2695-704. doi:10.1021/acs.jcim.3c01855
- [30] 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
- [31] Lane TR, Foil DH, Minerali E, Urbina F, Zorn KM, Ekins S. Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery. *Mol Pharm.* 2021;18(1):403-15. doi:10.1021/acs.molpharmaceut.0c01013
- [32] Fooladi H, Hirte S, Kirchmair J. Quantifying the hardness of bioactivity prediction tasks for transfer learning. *J Chem Inf Model.* 2024;64(10):4031-46. doi:10.1021/acs.jcim.4c00160
- [33] 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
- [34] Wellnitz J, Jain S, Hochuli JE, Maxfield T, Muratov EN, Tropsha A, et al. One size does not fit all: Revising traditional paradigms for assessing accuracy of QSAR models used for virtual screening. *J Cheminform.* 2025;17(1):7. doi:10.1186/s13321-025-00948-y
- [35] 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
- [36] 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
- [37] Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine

learning in structure-based virtual screening. J Chem Inf Model. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712

- [38] 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