



Molecular Predictions Should Remember Temperature, pH, Assay Format, and Biological Context Rather Than Treating Bioactivity as a Context-Free Label

Sara Ben Youssef^{1*}, Amal Trabelsi¹, Karim Boudiaf², Nabil Jebali³

¹Department of Context-Aware Bioactivity Prediction, Faculty of Pharmacy, University of Tunis, Tunis, Tunisia.

²Department of Assay Format and Biological Context Modeling, Faculty of Pharmacy, University of Sousse, Sousse, Tunisia.

³Department of Temperature and pH Effects on Bioactivity, Faculty of Pharmacy, University of Sfax, Sfax, Tunisia.

ABSTRACT

Molecular machine-learning models commonly treat bioactivity as a property attached to a molecular structure, even though experimentally reported activity is produced under particular physicochemical, analytical, and biological conditions. Temperature can alter molecular conformational states and target behavior; pH can redistribute protonation states and modify transport or binding; assay technologies measure different pharmacological levels; and cellular systems can change or confound the response being modeled. This methodological framework argues that these conditions should remain computationally associated with bioactivity rather than disappearing during database harmonization and label construction. The analysis separates molecular identity from experimental context, distinguishes semantic standardization from experimental equivalence, and treats missing contextual metadata as potentially informative rather than automatically neutral. A proposed context-retentive representation is developed in which activity records preserve the measured endpoint together with physicochemical conditions, assay and detection characteristics, biological system, and explicit missingness status. The framework further distinguishes molecule-dependent effects from molecular-by-context interactions and separates generalization to new molecules from prediction under new assays, biological systems, or condition combinations. Its purpose is not to assert that every activity measurement is strongly context sensitive, nor that recording additional metadata guarantees improved prediction. Instead, it provides a structure for determining when contextual information should be represented, tested, withheld, or reported as uncertain. The approach requires prospective evaluation under deliberately shifted conditions and does not establish universal correction rules, mechanistic causality, clinical validity, or deployment readiness.

Key Words: Bioactivity prediction, Assay context, Molecular machine learning, Experimental metadata, Ionization, Domain shift

eIJPPR 2026; 16(1):107-116

HOW TO CITE THIS ARTICLE: Ben Youssef S, Trabelsi A, Boudiaf K, Jebali N. Molecular Predictions Should Remember Temperature, pH, Assay Format, and Biological Context Rather Than Treating Bioactivity as a Context-Free Label. Int J Pharm Phytopharmacol Res. 2026;16(1):107-116. <https://doi.org/10.51847/3jbkdUA9W4>

INTRODUCTION

A molecular structure does not enter a bioassay alone. It is exposed at a particular concentration and temperature, in a specified buffer and detection system, against a particular target preparation, cellular state, or biological model.

Nevertheless, many computational pipelines eventually reduce this experimental event to a molecule-label pair. That reduction is useful for learning, but it can encourage an interpretation in which the measured value appears to belong intrinsically to the molecule. Evidence from public potency data argues against such interchangeability:

Corresponding author: Sara Ben Youssef

Address: Department of Context-Aware Bioactivity Prediction, Faculty of Pharmacy, University of Tunis, Tunis, Tunisia.

E-mail: ✉ sara.benyoussef@inat.ucar.tn

Received: 12 October 2025; **Revised:** 12 January 2026; **Accepted:** 14 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.



combining nominally comparable IC50 or Ki measurements obtained from different sources can introduce substantial additional variability [1]. The methodological problem is therefore not that bioactivity values are unusable, but that their evidential meaning depends partly on how they were generated.

Recent assay-aware modeling provides a complementary computational argument. When biological assay information is represented alongside molecular structure, that information can contribute predictive signal rather than behaving merely as administrative annotation [2]. Dose-response methodology reinforces the same distinction from another direction: quantities such as IC50 and EC50 are experimental summaries whose interpretation depends on how response curves are generated and analyzed [3]. These observations support treating a reported activity as a conditional measurement, not as proof of an immutable molecular property.

The framework proposed here formalizes that distinction without assuming that every contextual variable must influence every prediction. Let molecular identity describe the compound being tested, while experimental context describes conditions under which the activity label becomes observable. The relevant context can include physicochemical state, assay technology, reagents and detection logic, biological system, target state, dose, time, and the quality or absence of metadata. Context may alter the underlying molecular process, alter which process is measured, alter the scale of the reported endpoint, or merely correlate with a source-specific artifact. These possibilities must remain distinguishable.

Accordingly, context retention is not equivalent to adding more covariates indiscriminately. The methodological objective is to preserve enough information to ask whether apparently comparable labels are experimentally compatible, whether context modifies their interpretation, and whether a model is extrapolating beyond conditions represented during training. A context-aware model may ultimately decide that some recorded variables add little information. That conclusion is scientifically stronger when the variables were retained and tested than when they disappeared before modeling began.

How context disappears when experimental records become machine-learning labels

Context loss begins before model fitting. Bioactivity curation routinely requires decisions about inclusion criteria, unit harmonization, duplicate handling, endpoint filtering, qualifiers, and replicate consolidation; recent comparisons show that these decisions change the balance

between data quantity and quality available to prediction models [4]. Large resources such as ChEMBL deliberately integrate heterogeneous activity types, targets, assays, and historical records to make pharmacological evidence computationally reusable [5]. Such integration is indispensable, but harmonized syntax does not itself establish that two measurements represent the same experimental quantity.

A second loss occurs when information present in the experimental record is weakly structured or absent from the machine-readable label. Recent ChEMBL annotation work shows that relevant experimental-method information can be incompletely represented in assay descriptions and may require dedicated annotation to recover what is actually stated [6]. Even after data are standardized, cross-dataset consistency remains a separate problem: integrated molecular-property records can retain outliers, batch effects, and distributional discrepancies that affect downstream prediction [7]. The important distinction is therefore between making records computationally compatible and establishing their experimental comparability.

Reporting frameworks already show that considerably richer experimental descriptions are possible. MICHA organizes chemosensitivity experiments around compounds, biological samples, reagents, procedures, and analysis information rather than preserving only the final response value [8]. Conversely, large chemogenomic resources such as ExCAPE-DB demonstrate why aggregation is attractive: integrated compound–target activity collections create the scale needed for modern computational analysis [9]. The tension is not between curation and context, but between scalable integration and preservation of the conditions required to interpret the integrated label.

The final reduction often occurs at benchmark construction. MoleculeNet helped standardize molecular machine-learning tasks by organizing diverse datasets into tractable prediction problems with consistent representations and evaluation procedures [10]. That abstraction remains useful, but a task label can obscure whether variation within the target variable reflects chemistry, assay configuration, biological state, or some combination. The proposed methodological response is therefore not to abandon standardized benchmarks. It is to retain a recoverable experimental-context layer wherever the underlying evidence permits one.

Table 1 converts these sources of information loss into proposed methodological components for constructing labels without assuming experimental equivalence.

Table 1. Proposed methodological components for retaining experimental meaning during bioactivity-label construction

Method component	What should remain distinct	Purpose in model construction	Failure avoided
Endpoint preservation	IC50, EC50, Ki, Kd, functional and phenotypic outputs	Prevent nominal “activity” from collapsing different measured quantities	Endpoint interchangeability
Assay provenance retention	Source assay, platform, protocol identity, laboratory or dataset origin when available	Permit assessment of source-linked variability	Hidden batch or source effects
Protocol-context retention	Temperature, pH, reagents, timing, concentration regime, detection conditions when reported	Preserve variables capable of changing molecular state or measurement meaning	Context erased before modeling
Biological-context retention	Target preparation, cell type, cellular state, species or other biological system	Separate biological heterogeneity from chemical variation	Context-dependent responses mislabeled as molecule-only effects
Explicit missingness status	Known, unreported, unavailable, inferred, or not applicable contextual fields	Prevent absent information from being treated as biologically neutral	Informative missingness concealed
Compatibility assessment	Semantic normalization separated from evidence of experimental equivalence	Decide whether records should be pooled, stratified, conditioned, or excluded	Standardization mistaken for comparability

Note: The table presents a proposed methodological synthesis rather than a universally validated metadata standard.

Temperature as a determinant of molecular and assay behavior

Temperature is often recorded as a protocol detail, yet it can participate directly in the molecular state being measured. In TRPM4, physiological temperature was shown to influence ligand recognition and channel gating, providing direct evidence that temperature can modify pharmacological behavior rather than merely measurement convenience [11]. The implication is not that temperature exerts a uniform effect across targets. It is that a potency or functional label obtained under one temperature condition cannot automatically be assumed to describe an identical molecular state under another.

Structural experiments illustrate a related issue. Room-temperature crystallography of PTP1B revealed small-molecule fragment-binding behavior that differed from observations made under conventional cryogenic conditions [12]. Experimental temperature can therefore influence which conformational or binding states become visible to an assay or structural method. This matters computationally because a label associated with a molecular graph may implicitly summarize an ensemble generated under a specific thermal regime.

The proposed framework consequently treats temperature in three possible roles. First, it can be state-defining context, where thermal conditions alter conformational or functional populations. Second, it can be measurement context, where the experimental platform observes different ensembles under different conditions. Third, it can be low-information metadata when variation within the relevant range has negligible consequence for the modeled endpoint. These roles should be determined empirically rather than assigned in advance.

Temperature should therefore not be converted into a universal correction factor. Evidence from one ion channel

or one crystallographic target does not establish the magnitude or direction of temperature dependence elsewhere. A condition-aware record instead preserves the temperature when known, permits interaction with molecular and assay representations, and allows evaluation to determine whether that information materially changes prediction or uncertainty.

pH, ionization, and apparent activity

pH is especially important because it links an experimental condition to molecular charge state, transport, binding, and detection. Computational biophysical work shows that protonation kinetics can influence membrane permeation of ionizable molecules [13]. Earlier dynamic-protonation simulations likewise demonstrated that allowing protonation states to change during membrane crossing can produce substantially different permeability behavior from fixed-protonation assumptions [14]. A molecular representation that encodes only a neutral or single canonical structure may therefore omit a condition-dependent component of the experimentally accessible species distribution.

Binding can also be pH sensitive independently of bulk permeability. In a defined enzyme–ligand system, pH was shown to alter intermediate populations along the binding process [15]. This establishes an important methodological boundary: observed activity changes associated with pH need not arise from a single mechanism. They may reflect shifts in ligand protonation, target protonation, membrane transport, binding pathways, solubility, assay chemistry, or combinations of these processes.

The framework proposed here therefore separates recorded pH from predicted ionization descriptors. A calculated pKa or microspecies distribution can help characterize molecular state, but it is not a substitute for the pH under

which an activity was experimentally measured. Conversely, recording assay pH does not establish which protonation event caused a difference in activity.

For modeling, pH should be represented as a contextual variable capable of interacting with molecular features rather than as a universal additive offset. Models should be allowed to learn that a particular chemical series or endpoint is highly pH sensitive while another is comparatively stable. Such interactions remain predictive associations unless independent mechanistic evidence establishes why the activity changed.

Assay format changes what the endpoint means

An activity value cannot be interpreted independently of what the assay actually observes. Target-engagement technologies differ in physical principle, experimental configuration, and proximity to the molecular event of interest [16]. A binding or engagement readout, a biochemical functional assay, a reporter response, and a phenotypic endpoint can therefore describe different levels of pharmacological organization even when they involve the same nominal compound and target.

High-throughput screening further illustrates why the measurement process matters. Machine-learning-assisted hit prioritization can extract useful information from screening data, but the signals available to such models arise from the specific screen, its readouts, and its selection logic [17]. An apparent “active” label is consequently inseparable from detection technology, response thresholding, reagent environment, and the biological or biochemical configuration that generated it. None of these factors implies that the signal is artifactual; they define what was measured.

Public hERG data provide a concrete example of the difficulty created when assay-derived values are aggregated. Heterogeneous publicly accessible hERG measurements complicate predictive modeling because apparently related blockade labels originate from differing experimental sources and conditions [18]. This problem is especially consequential when a model is trained as though those records were exchangeable examples of a single latent property.

The proposed framework therefore separates measurement modality from biological system. Assay technology specifies how a response becomes observable; biological context specifies the system in which that response occurs. Both can modify the meaning of a label, but they should not be conflated. A model may discover that two formats are sufficiently compatible for a particular prediction task, yet compatibility should be demonstrated through analysis rather than assumed during preprocessing. Target engagement must also remain distinct from downstream

functional activity, and neither should be interpreted automatically as therapeutic effect.

Biological system and cellular context

Biological context introduces variation that cannot always be reduced to assay instrumentation. A particularly strong warning comes from recent benchmark analysis showing that cell count alone can predict many small-molecule bioactivity benchmarks surprisingly well [19]. Such performance indicates that models can exploit broad cellular-health signals rather than the intended mechanistic or phenotypic construct. Context-aware modeling must therefore distinguish biologically relevant state from unintended shortcut information.

Differences among biological systems also create domain shift. CODE-AE was developed to improve transfer of drug-response information from cell-line screening to patient-derived settings by separating shared and context-specific representations [20]. The work supports the general methodological proposition that biological domains should not automatically be treated as samples from a single invariant response distribution. It does not establish that computational transfer from cell lines to patients is universally reliable or that predicted response demonstrates clinical benefit.

Cell type itself can be a structured source of predictive variation. PrePR-CT uses cell-type-specific inductive information to model perturbation responses and was evaluated under conditions that include previously unseen cellular contexts [21]. Together with the cell-count result, this creates an important distinction: biological context can contain both scientifically meaningful modifiers and confounding shortcuts. Simply adding a cell identifier is therefore insufficient.

The proposed framework treats cellular and biological metadata as candidate descriptors of the state in which molecular activity is expressed. Depending on the task, relevant variables may include cell type, baseline molecular state, target abundance, genotype, differentiation state, tissue origin, or species, but only information actually available and appropriately defined should enter the record. Model evaluation must then determine whether these variables support transferable interactions or merely memorize domain identity. Biological context is thus not an automatic explanation for activity differences; it is a structured source of heterogeneity that must be represented, tested, and bounded.

Context missingness is not random

Experimental context can be absent for systematic reasons. Assay descriptions may omit methodological details that were never transferred into structured database fields [6],

while minimum-information frameworks exist precisely because compounds, samples, reagents, procedures, and analysis settings are not consistently preserved in reusable screening records [8]. Missing temperature, pH, timing, or biological-state information should therefore not automatically be encoded as though the experiment occurred under an average or biologically neutral condition.

Missingness can also correlate with source, assay generation, publication era, or endpoint type. The heterogeneous provenance documented for public hERG measurements illustrates how missing context may travel

with other source-specific differences [18]. The proposed framework consequently distinguishes *unknown*, *not reported*, *not applicable*, and *explicitly measured* context where the source data permit those distinctions. A missingness indicator is not assumed to explain activity; it allows the model and evaluator to test whether absent context behaves as informative provenance, harmless incompleteness, or a source of unreliable extrapolation.

Table 2 summarizes the proposed workflow for retaining uncertainty about experimental context rather than silently converting absence into a value.

Table 2. Proposed workflow for preserving and interpreting missing experimental context

Workflow step	Record action	Analytical purpose	Decision boundary
Detect	Identify whether a context field is reported	Separate absence from measured values	Never infer an unreported condition as fact
Classify	Mark measured, not reported, unavailable, inferred, or not applicable	Preserve missingness semantics	“Unknown” is not a neutral experimental state
Preserve provenance	Retain assay/source linkage	Test whether missingness tracks source effects	Association with provenance is not mechanism
Model	Supply context and missingness status separately	Permit conditional prediction without forced imputation	Imputed values remain estimates
Stress-test	Evaluate records with withheld or shifted context	Measure sensitivity to information loss	Stable IID accuracy is insufficient
Report	Disclose known and missing context at inference	Bound interpretation of predictions	Unseen context requires explicit qualification

Note: The workflow is proposed here as a modeling discipline rather than an externally validated universal standard.

The operational logic is synthesized in **Figure 1**: recorded experimental conditions may be retained, lost, or missing before model construction, and the proposed architecture

keeps those states distinct before attempting prediction under new conditions.

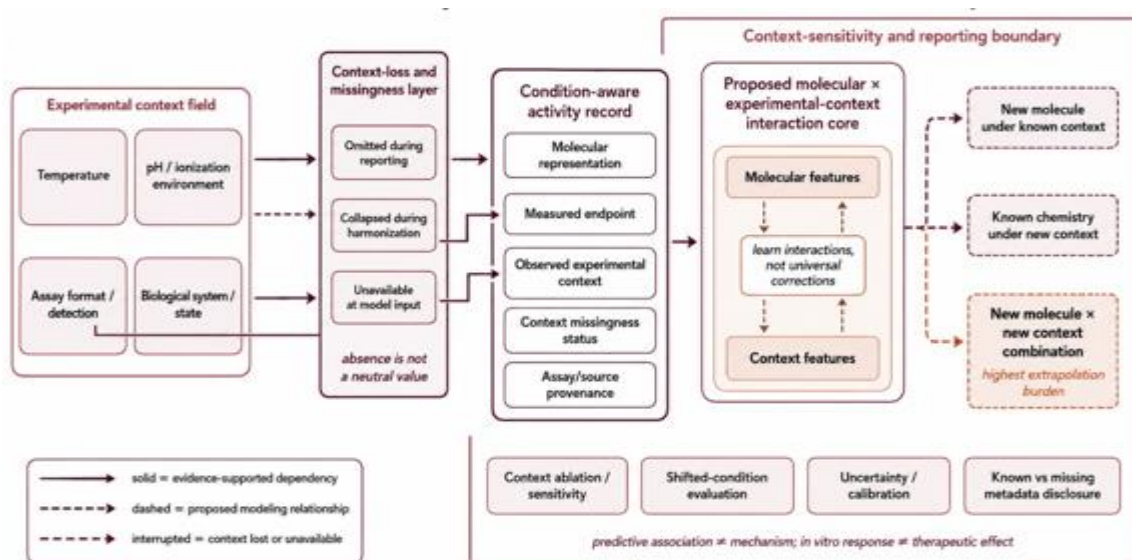


Figure 1. From conditioned bioactivity measurement to bounded new-condition prediction

Building condition-aware bioactivity records

A condition-aware record should preserve molecular structure without forcing experimental context into the

molecule itself. Joint use of chemical structures and assay descriptions provides direct precedent for representing assay information alongside molecular information [22].

The proposed record extends this principle by keeping the measured endpoint, observed conditions, biological system, source provenance, and missingness status as separable fields linked to the same experimental observation.

Context representations need not all have the same form. Graph neural processes demonstrate that molecular prediction can be conditioned on information from related tasks while retaining probabilistic uncertainty [23], whereas heterogeneous drug-discovery measurements can be learned jointly without first collapsing every assay into a single endpoint [24]. The framework therefore permits structured numeric fields, categorical assay descriptors, biological-state embeddings, and text-derived protocol representations, but it does not assume that one representation is optimal. Crucially, inferred context must remain distinguishable from measured context.

The practical record is therefore relational rather than merely concatenative. A canonical molecular graph can remain stable while separate context fields describe the instance under which an endpoint was observed. This arrangement permits repeated measurements of the same molecule to remain distinct when their assay or biological conditions differ. It also allows uncertainty about context to be represented without corrupting the molecular identity itself. Such a schema does not establish which metadata are mechanistically important; it preserves the information required to test that question.

A molecular-experimental context prediction architecture

The proposed architecture contains separate encoders for molecular identity and experimental context, followed by an interaction stage that produces an activity prediction conditional on both. Integrated compound-protein modeling already shows that distinct activity types such as K_i , K_d , EC_{50} , and IC_{50} can be represented within a common learning system without asserting that they are pharmacologically identical [25]. The framework adopts that separation as a design principle rather than pooling all endpoints into one undifferentiated target.

Context conditioning should also accommodate dynamic experimental variables. XPert provides a relevant precedent by modeling cellular perturbation responses across dose, time, and cell state rather than reducing those factors to one static label [26]. Here, the analogous proposal is a modular context encoder whose outputs interact with molecular features before prediction. The architecture is intentionally not tied to transformers, graph networks, or any single algorithm; its defining property is retention of context and explicit testing of molecule-by-context dependence.

The interaction layer may be implemented through cross-attention, gating, conditional normalization, multiplicative

feature interactions, task-conditioned prediction heads, or another architecture capable of representing nonadditivity. However, architectural expressiveness is not evidence that such interactions are scientifically real. The model should be compared against simpler alternatives, including molecule-only models, context-only controls, and additive covariate baselines. A complex context-conditioned model is justified only when deliberately shifted evaluation demonstrates that its added structure contributes transferable information rather than source memorization.

Learning context interactions rather than simple covariates

Treating context as an additive correction assumes that the same temperature, pH, assay, or cell-state effect applies similarly across molecules. The evidence does not justify that simplification. Assay-aware prediction shows that assay information can interact meaningfully with molecular inputs [2], cell-type-specific perturbation models demonstrate that biological state changes predictive relationships [21], and dose-time modeling shows that experimental variables can interact dynamically [26].

The proposed framework therefore favors interaction-capable representations while requiring regularization and ablation tests to prevent context fields from becoming shortcut identifiers. A model should be able to learn both context sensitivity and context invariance. If removing a contextual field leaves predictions unchanged across deliberately shifted evaluations, retaining that field may offer little value for that task.

Conversely, an apparent gain from context does not prove that the modeled variable is mechanistically responsible for the activity difference. A laboratory identifier, assay name, or cell type may correlate with chemistry, endpoint choice, plate design, or dataset-generation history. Context interactions should therefore be interpreted first as conditional predictive structure. Mechanistic interpretation requires orthogonal experimental evidence capable of separating the contextual variable from correlated sources of variation.

Predicting activity under new experimental conditions

Generalization should be described along the axis that is actually new. Few-shot molecular task adaptation demonstrates that models can use limited information from related tasks, but such success depends on the relationship between training and test tasks [23]. Biological-domain transfer similarly shows that prediction across cell-line and patient-derived settings is a distinct problem from prediction of new molecules within one domain [20].

The proposed evaluation taxonomy therefore separates three cases: new molecule under familiar context, familiar

chemistry under new context, and simultaneous molecule-by-context extrapolation. Dose–time and cold-cell evaluations provide precedents for deliberately withholding portions of biological or experimental state space [26]. Performance in the first case must not be presented as evidence for the third.

The distinction is especially important because context space may be sparse and combinatorial. A model can encounter a temperature it has seen previously but not for

the relevant assay; a cell type may be familiar but paired with a new perturbation modality; or the molecule and context may each be individually represented while their combination is absent. These settings create different extrapolation burdens and should be reported separately.

Table 3 contrasts these prediction settings with common alternatives and makes the intended extrapolation claim explicit.

Table 3. Comparison of context-retentive prediction with common molecular-modeling abstractions

Modeling strategy	Information retained	Typical generalization claim	Main unresolved risk
Molecule-only model	Structure plus endpoint label	New molecules within task	Hidden assay or biological heterogeneity
Task/assay identifier model	Structure plus coarse task identity	New molecules within known tasks	Identifier may encode source shortcuts
Structure + assay-description model	Molecular and textual assay information	Context-informed prediction	Text may omit critical experimental conditions
Heterogeneous multitask model	Multiple related assay endpoints	Transfer across observed tasks	Missing values and endpoint semantics remain complex
Proposed context-retentive model	Structure, endpoint, explicit conditions, biological state, provenance, missingness	New molecules, new contexts, or their combination stated separately	Requires prospective shifted-context validation

Note: The final row is the methodological proposal advanced here; the table does not imply demonstrated superiority.

Evaluating context sensitivity

Context-retentive models require evaluation beyond aggregate predictive accuracy. Recent bioactivity work shows that uncertainty can be modeled explicitly through combined Bayesian and evidential approaches [27], while conformal prediction provides another route to prediction-level reliability under its calibration assumptions [28]. Neither approach guarantees valid confidence under arbitrary experimental shift.

Evaluation should therefore pair uncertainty analysis with deliberate context perturbation. UMAP-based clustered splitting illustrates how more demanding distributional separation can expose generalization weaknesses hidden by easier partitions [29]. For context-aware bioactivity modeling, analogous tests should hold out assay formats, biological systems, condition combinations, or metadata strata. Context ablation, counterfactual condition swaps where scientifically defensible, and calibration under shifted conditions can then determine whether a model learned transferable dependencies or merely memorized provenance.

Sensitivity should also be evaluated at the prediction level rather than only through global averages. A model may be context stable for one chemical region yet highly context dependent for another. Reporting the change in prediction, uncertainty, or ranking as context varies can reveal whether the model behaves coherently across experimental states. Such analysis should not be interpreted as a causal effect

of changing context unless the corresponding experimental contrast has been observed.

Minimum metadata for context-aware molecular modeling

The scientific claim attached to a molecular model must remain bounded by the data used to establish it; dataset definition therefore constrains the legitimate QSAR claim [30]. For context-aware modeling, the minimum report should identify molecular representation, endpoint definition, assay or measurement type, biological system, available physicochemical conditions, provenance, and which contextual fields were missing or inferred.

Existing resources show why this requirement must remain pragmatic. ChEMBL provides structured assay and bioactivity information at scale [5], but recent annotation work demonstrates that experimental details can remain incomplete [6]. MICHA shows that richer protocol-level representation is possible for drug-sensitivity experiments [8]. The proposal here is therefore not a universal mandatory field list. It is a reporting principle: models should disclose which experimental dimensions were available during training and inference, which were absent, and whether evaluation crossed any of those dimensions. For deployment, this distinction limits interpretation. A prediction generated for an unrepresented pH, assay format, cell state, or combination of contexts should be identified as such rather than presented as ordinary interpolation. Metadata completeness does not guarantee validity, but without contextual provenance the user may

be unable to determine what kind of extrapolation occurred.

A practical report should also distinguish context that was measured directly from context inferred algorithmically or recovered from textual descriptions. Derived pKa values, text-mined assay labels, or inferred biological states can be useful model inputs, but their uncertainty differs from that of explicitly recorded experimental conditions. Context-retentive modeling therefore requires not only additional fields but provenance for those fields.

Limitations

The framework has several limits. First, the evidence supporting context sensitivity is heterogeneous and frequently system specific. Temperature effects demonstrated for one target, pH-sensitive transport in one molecular class, or biological-state effects in one perturbation dataset cannot establish universal effect sizes or directions. Second, additional context creates sparsity. Many molecule-by-condition combinations will remain unobserved, making a more expressive model easier to overfit.

Third, context itself can become a shortcut. Assay labels, source identifiers, cell types, and missingness patterns may correlate with chemical series or dataset construction rather than transferable pharmacology. The framework therefore cannot treat improved predictive performance as evidence of mechanism. Fourth, historical data may lack the metadata required to reconstruct the proposed record accurately, and retrospective inference cannot restore experimental facts that were never reported.

Finally, no architecture specified here has been prospectively validated across controlled variations of temperature, pH, assay format, and biological state simultaneously. The framework should therefore be evaluated as a methodological hypothesis: retaining context may improve interpretability and generalization when the context contains relevant information, but prospective experiments are required to determine when that proposition holds.

CONCLUSION

Bioactivity is experimentally observed under conditions, even when machine-learning pipelines eventually encode it as a single target value. The framework developed here proposes retaining those conditions long enough to determine whether they matter. Temperature and pH can alter molecular states; assay technologies determine what is observed; biological systems can modify or confound responses; and missing metadata can carry provenance rather than neutrality.

A context-retentive model therefore separates molecular identity, endpoint semantics, experimental conditions, biological state, provenance, and missingness before learning their interactions. It also distinguishes prediction of a new molecule from prediction under a new experimental context and requires that uncertainty and metadata boundaries accompany such claims.

This framework is a methodological proposal, not a validated correction system. Some endpoints may prove robust to contextual variation, some metadata may be redundant, and recorded context can itself become a shortcut. Prospective experiments that deliberately vary or withhold conditions are required to determine when context retention materially improves scientific validity. The central principle is narrower but consequential: a molecular prediction should not forget how its training label came to exist.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] 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
- [2] 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
- [3] Srinivasan B. Dose-response curves and the determination of IC₅₀ and EC₅₀ values. *J Med Chem.* 2024;67(20):17931-4. doi:10.1021/acs.jmedchem.4c02052
- [4] Schiebroek CCG, Landrum GA, Riniker S. Balancing data quantity and quality: Evaluating curation strategies for bioactivity prediction in lead optimization. *J Chem Inf Model.* 2026;66(13):7446-52. doi:10.1021/acs.jcim.6c01018
- [5] Zdrazil B, Felix 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):D1180-92. doi:10.1093/nar/gkad1004
- [6] Smit I, Adasme MF, Manners EJ, Corbett S, Bosc N, Do HMA, et al. Integrating artificial intelligence and

- manual curation to enhance bioassay annotations in ChEMBL. *J Cheminform.* 2026;18(1):24. doi:10.1186/s13321-026-01165-x
- [7] Parrondo-Pizarro R, Menestrina L, Garcia-Serna R, Fernández-Torras A, Mestres J. Enhancing molecular property prediction through data integration and consistency assessment. *J Cheminform.* 2025;17(1):163. doi:10.1186/s13321-025-01103-3
- [8] Tanoli Z, Aldahdooh J, Alam F, Wang Y, Seemab U, Fratelli M, et al. Minimal information for chemosensitivity assays (MICHA): A next-generation pipeline to enable the FAIRification of drug screening experiments. *Brief Bioinform.* 2022;23(1):bbab350. doi:10.1093/bib/bbab350
- [9] Sun J, Jeliaskova N, Chupakhin V, Golib-Dzib JF, Engkvist O, Carlsson L, et al. ExCAPE-DB: An integrated large scale dataset facilitating big data analysis in chemogenomics. *J Cheminform.* 2017;9(1):17. doi:10.1186/s13321-017-0203-5
- [10] 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
- [11] Hu J, Park SJ, Walter T, Orozco IJ, O'Dea G, Ye X, et al. Physiological temperature drives TRPM4 ligand recognition and gating. *Nature.* 2024;630(8016):509-15. doi:10.1038/s41586-024-07436-7
- [12] Skaist Mehlman T, Biel JT, Azeem SM, Nelson ER, Hossain S, Dunnett L, et al. Room-temperature crystallography reveals altered binding of small-molecule fragments to PTP1B. *eLife.* 2023;12:e84632. doi:10.7554/eLife.84632
- [13] Harris J, Chipot C, Roux B. How is membrane permeation of small ionizable molecules affected by protonation kinetics? *J Phys Chem B.* 2024;128(3):795-811. doi:10.1021/acs.jpcc.3c06765
- [14] Yue Z, Li C, Voth GA, Swanson MJ. Dynamic protonation dramatically affects the membrane permeability of drug-like molecules. *J Am Chem Soc.* 2019;141(34):13421-33. doi:10.1021/jacs.9b04387
- [15] Singh K, Muttathukattil AN, Singh PC, Reddy G. pH regulates ligand binding to an enzyme active site by modulating intermediate populations. *J Phys Chem B.* 2022;126(47):9759-70. doi:10.1021/acs.jpcc.2c05117
- [16] 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
- [17] Boldini D, Friedrich L, Kuhn D, Sieber SA. Machine learning assisted hit prioritization for high throughput screening in drug discovery. *ACS Cent Sci.* 2024;10(4):823-32. doi:10.1021/acscentsci.3c01517
- [18] Siramshetty VB, Chen Q, Devarakonda P, Preissner R. The Catch-22 of predicting hERG blockade using publicly accessible bioactivity data. *J Chem Inf Model.* 2018;58(6):1224-33. doi:10.1021/acs.jcim.8b00150
- [19] Seal S, Dee W, Shah A, Cerisier N, Zhang A, Miglietta E, et al. Counting cells can accurately predict small-molecule bioactivity benchmarks. *Nat Commun.* 2026;17(1):2436. doi:10.1038/s41467-026-68725-5
- [20] 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
- [21] Alsulami R, Lehmann R, Khan SA, Lagani V, Maillo A, Gomez-Cabrero D, et al. Predicting and interpreting cell-type-specific drug responses in the small-data regime using inductive priors. *Nat Mach Intell.* 2026;8(3):461-73. doi:10.1038/s42256-026-01202-2
- [22] 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
- [23] García-Ortegón M, Seal S, Rasmussen C, Bender A, Bacallado S. Graph neural processes for molecules: An evaluation on docking scores and strategies to improve generalization. *J Cheminform.* 2024;16(1):115. doi:10.1186/s13321-024-00904-2
- [24] Irwin BWJ, Levell JR, Whitehead TM, Segall MD, Conduit GJ. Practical applications of deep learning to impute heterogeneous drug discovery data. *J Chem Inf Model.* 2020;60(6):2848-57. doi:10.1021/acs.jcim.0c00443
- [25] Gu Y, Xia S, Ouyang Q, Zhang Y. Bioactivity deep learning for complex structure-free compound-protein interaction prediction. *J Chem Inf Model.* 2025;65(19):9910-26. doi:10.1021/acs.jcim.5c00741
- [26] Guo Y, Zhang H, Hu H, Wu J, Cao J, Hsieh CY, et al. Modelling drug-induced cellular perturbation responses with a biologically informed dual-branch transformer. *Nat Mach Intell.* 2026;8:96-112. doi:10.1038/s42256-025-01165-w
- [27] Khalil B, Schweighofer K, Dyubankova N, van Westen GJP, van Vlijmen H. Combining Bayesian and evidential uncertainty quantification for improved bioactivity modeling. *J Chem Inf Model.* 2025;65(24):13057-69. doi:10.1021/acs.jcim.5c01597
- [28] Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications

- in drug discovery. J Cheminform. 2019;11(1):4. doi:10.1186/s13321-018-0325-4
- [29] Guo Q, Hernandez-Hernandez S, Ballester PJ. UMAP-based clustering split for rigorous evaluation of AI models for virtual screening on cancer cell lines. J Cheminform. 2025;17(1):94. doi:10.1186/s13321-025-01039-8
- [30] Nael MA, Alakonda LM, Elokely KM. Defining the data set defines the QSAR claim. J Chem Inf Model. 2026;66(6):2951-4. doi:10.1021/acs.jcim.6c00514