



Learning Across Bioassays Without Pretending Same Biology via Context-Aware and Assay-Specific Prediction

Omar Zayed^{1*}, Leila Ibrahim¹, Ahmed Mansour²

¹Department of Context-Aware Bioassay Transfer Learning, Faculty of Pharmacy, Cairo University, Cairo, Egypt.

²Department of Assay-Specific Molecular Prediction, Faculty of Pharmacy, Alexandria University, Alexandria, Egypt.

ABSTRACT

Bioactivity modeling often gains statistical power by pooling assays, sharing representations, or transferring parameters across related prediction tasks. These strategies are useful, but they can quietly replace a biological question with a convenience assumption that measurements obtained under different assay conditions represent the same response variable up to noise. This methodological article argues that assay context should instead be treated as part of the prediction target whenever experimental conditions, measurement type, biological system, or provenance can alter the mapping from molecular structure to observed activity. The analysis distinguishes molecular representation from assay representation, related-task transfer from label pooling, and assay-specific prediction from universal endpoint modeling. It proposes a context-aware transfer framework in which molecular features are combined with explicit or learned assay-context information, cross-assay sharing is conditioned on estimated task relevance, and assay-specific outputs are preserved where measurements are not biologically interchangeable. The framework also treats conflicting labels, negative transfer, dataset shift, and context novelty as first-class evaluation problems rather than residual error. Its intended contribution is methodological: to organize existing evidence into a disciplined strategy for learning across bioassays without erasing meaningful heterogeneity. The proposal does not establish a universally optimal architecture, mechanistic interpretation, or prospective performance advantage. Its value therefore depends on future tests that separate chemical novelty from context novelty, compare selective against indiscriminate transfer, and assess calibration under assay families not represented during training.

Key Words: Bioactivity prediction, Assay context, Transfer learning, Multitask learning, Molecular representation, Dataset shift

eIJPPR 2025; 15(3):108-116

HOW TO CITE THIS ARTICLE: Zayed O, Ibrahim L, Mansour A. Learning Across Bioassays Without Pretending Same Biology via Context-Aware and Assay-Specific Prediction. Int J Pharm Phytopharmacol Res. 2025;15(3):108-16. <https://doi.org/10.51847/kUwS08zaQX>

INTRODUCTION

Bioactivity prediction is commonly formulated as a mapping from molecular structure to an activity label, with assay conditions treated as metadata or ignored after curation. That abstraction is often convenient, but it is not always biologically neutral. Assay-aware modeling has emphasized that experimental setting can change what an observed bioactivity value means and therefore may need to enter the modeling problem explicitly [1]. The methodological consequence is not that every protocol variable must become a feature, but that the prediction

target should be defined with enough context to distinguish measurements that are not safely interchangeable.

Recent computational work provides direct evidence that contextual information can contribute predictive signal beyond molecular structure alone. Models that combine chemical structure with bioassay descriptions have shown that assay text can improve molecular property prediction in some settings [2]. Such improvement does not demonstrate that assay descriptions recover causal biology; they may also encode task identity, laboratory conventions, or correlated technical information. Nevertheless, the result weakens the assumption that the molecule alone

Corresponding author: Omar Zayed

Address: Department of Context-Aware Bioassay Transfer Learning, Faculty of Pharmacy, Cairo University, Cairo, Egypt.

E-mail: ✉ omar.zayed@cu.edu.eg

Received: 16 February 2025; **Revised:** 10 June 2025; **Accepted:** 16 June 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.



contains all information needed to interpret an assay-derived label.

Data integration introduces a related problem. Even after chemical standardization and endpoint harmonization, merged datasets can retain distributional and annotation inconsistencies that affect model behavior [3]. These inconsistencies can arise from differences in experimental conditions, source-specific sampling, assay definitions, or curation decisions, and the available evidence does not justify assigning them all to one cause. A context-aware formulation should therefore treat integration as a modeling decision about conditional comparability, not merely as a preprocessing step that automatically creates a larger coherent training set.

Large bioactivity resources make this issue unavoidable because they span multiple activity types, assay formats, biological systems, and time periods [4]. Their value for machine learning is substantial, but scale does not itself establish biological equivalence among records. This article therefore treats an observed label as a relation among a molecule, an assay context, and a measurement process rather than as an intrinsic molecular property. The central proposal is to share information across assays where evidence supports relatedness while preserving assay-specific prediction when context changes the interpretation of the response.

Why cross-assay transfer fails under hidden heterogeneity

The first failure mode is measurement heterogeneity that remains hidden after labels have been placed on a common numerical scale. Direct comparison of IC₅₀ and K_i values aggregated across different sources shows that combining nominally comparable measurements can introduce substantial noise [5]. Such disagreement may reflect protocol differences, source variation, or experimental uncertainty rather than a change in the underlying molecular mechanism. A transfer model that treats these values as identical supervision can therefore learn a blurred target even when its optimization objective is numerically well behaved.

A second failure mode comes from chemistry rather than assay context. Activity-cliff analyses show that structurally similar compounds can display sharply different activities, exposing limits of models that rely heavily on smooth similarity assumptions [6]. This matters because cross-assay transfer can fail for two separable reasons: the assays may not be equivalent, or the local structure–activity relationship may itself be discontinuous. A useful framework must keep those explanations distinct instead of attributing every prediction conflict to contextual biology. A third failure mode is inappropriate source-task selection. Bioactivity tasks differ in hardness and in how useful they are as transfer sources, so transferable knowledge cannot

be assumed to increase monotonically with the amount of auxiliary data [7]. Apparent relatedness may arise from shared chemical series, common targets, overlapping label distributions, or genuine biological similarity. The proposed approach therefore treats transfer eligibility as an empirical and scientific question, not as a default consequence of having two assay columns in the same dataset. In methodological terms, hidden heterogeneity is unmodeled variation that changes the conditional relation between structure and observation; it can be biological, technical, statistical, or curatorial, and its origin should not be assigned before evidence distinguishes among these possibilities.

Representing assay conditions and biological context

Context-aware prediction requires a molecular representation and an assay representation to play different methodological roles. Benchmarking of learned molecular representations has shown that encoder choice can materially affect molecular property prediction [8]. That evidence supports careful molecular featurization, but it does not imply that assay heterogeneity should be absorbed into the same latent vector. The proposal here is to preserve a distinct route by which experimental context can condition the interpretation of molecular features.

The relevant context may be structured, relational, or textual. Methodological analyses of compound–target interaction prediction show that dataset construction, negative definitions, target representation, and validation strategy alter what the prediction task actually means [9]. By extension, assay context should be encoded only through variables with a defensible role in defining the observation, such as target identity, measurement type, biological system, protocol descriptors, or assay-specific examples. Merely adding a dataset identifier can improve fit without establishing that a biologically meaningful context has been represented.

Heterogeneous affinity measurements illustrate the distinction. Multitask bioassay pretraining has explicitly treated IC₅₀, K_i, and K_d measurements as related but distinct tasks rather than assuming full interchangeability [10]. That strategy does not establish that every measurement type needs a separate model, nor does it collapse binding affinity into functional activity. It instead supports a broader methodological principle: context variables can enter as task descriptors, conditioning signals, or output-specific identifiers while molecular information remains shared where such sharing is empirically useful. A practical representation should also distinguish recorded context from latent context: missing protocol details cannot be reconstructed simply because an embedding is expressive. Context encoders should therefore expose missingness, preserve provenance, and

avoid implying mechanistic knowledge absent from the assay record.

Task definitions for context-aware molecular prediction

The task definition should state both what is being predicted and under which context the prediction is intended to hold. Molecular-learning benchmarks have long shown that endpoint choice, split strategy, and evaluation metric are integral to the validity of a prediction task [11]. For context-aware bioactivity modeling, this means distinguishing at least among prediction within a known assay, prediction for a related assay with limited labels, and prediction for an unseen assay or assay family. The article proposes treating these as different generalization problems rather than reporting one pooled metric across all assay records. Formally, the prediction target is better conceptualized as a response conditional on both molecule and assay context, with the admissible context ranging from a stable identifier to a structured descriptor or a small support set. This formulation leaves

open which representation is best while making the source of generalization explicit.

Few-shot assay prediction provides a direct example of the last case. FS-CAP formulates prediction for a previously unseen assay using a small set of assay-specific context compounds and their observed activities [12]. This demonstrates that assay identity need not be represented only by a fixed categorical label; it can also be inferred from local examples that define the task at prediction time. The approach remains dependent on the representativeness of those examples, however, and does not establish that a few measurements recover all relevant biological or technical context.

Taken together, these distinctions define the boundary conditions summarized in **Table 1**. The table is not a taxonomy claimed to be established by prior literature; it is a proposed methodological matrix derived from the preceding argument and constrained by the evidence reviewed above

Table 1. Proposed boundary-condition matrix for defining context-aware bioactivity prediction tasks

Prediction setting	What is held constant	What may change	Main modeling implication	Principal failure if ignored
Known-assay prediction	Assay identity and measurement definition	Molecules	Molecular generalization dominates; assay conditioning may remain fixed	Chemical novelty mistaken for assay shift
Related-assay transfer	Part of biological or measurement context	Molecules and selected assay attributes	Share only information supported as transferable	Negative transfer from superficially related tasks
Novel-assay few-shot prediction	Molecular encoder and prior task experience	Assay identity plus limited contextual observations	Infer task context from support examples and preserve assay-specific outputs	Context set may be unrepresentative
Cross-family or cross-platform prediction	Little beyond broad endpoint intent	Chemistry, assay technology, biology, provenance	Treat as out-of-context generalization requiring explicit validation	Pooled accuracy conceals context failure
Conflicting-measurement setting	Molecule may be identical	Source, measurement type, protocol, or biological state	Preserve disagreement until its origin is modeled or adjudicated	Averaging converts informative heterogeneity into label noise

Note: All rows are proposed methodological synthesis. The matrix does not imply that the listed context dimensions are exhaustive or that any transfer regime is prospectively validated.

Transfer learning across related but nonidentical assays

Transfer learning becomes useful when assays are sufficiently related to share information but sufficiently different that their observations should not simply be pooled. Direct work on transitions between biological assays has shown that information from a source assay can improve prediction for a target assay under appropriate transfer strategies [13]. The result supports reuse of cross-assay information, but it also implies that the transition itself is part of the modeling problem. Similar target names or overlapping compounds do not establish that two assays encode the same response function.

Transfer learning in drug discovery encompasses parameter reuse, representation reuse, multitask learning, and other mechanisms for moving information from source

to target tasks [14]. For bioassays, the proposed decision should precede the transfer operation: identify what is plausibly shared, what remains assay-specific, and what evidence indicates relatedness. Useful signals may include overlap in molecular space, target or pathway relationships, measurement definitions, platform characteristics, or demonstrated predictive benefit on target-context validation. None should be interpreted automatically as biological equivalence.

The methodological distinction is between transfer and pooling. Pooling asks one model to explain observations as instances of a common label; transfer allows one task to inform another while preserving separate conditional targets. The latter is preferable when the scientific question requires assay identity to remain recoverable. This article

therefore proposes selective transfer as the default hypothesis: shared information is permitted where it improves context-appropriate validation, but assay-specific pathways remain available when sharing degrades performance or obscures conflicting observations. This proposal requires empirical testing rather than assuming that greater parameter sharing is intrinsically better. Relatedness should also be assessed against the target context, because a source task can appear useful in aggregate while contributing little to a particular assay, endpoint definition, or chemical region. Transfer decisions should therefore be revisable as target-context evidence accumulates.

Assay-specific heads, conditioning, and hierarchical models

A context-aware architecture needs a place where information can be shared and a place where assay-specific distinctions can survive. Multitask QSAR studies demonstrate that shared neural representations can exploit information across related endpoints, while gains depend on the structure and data available across tasks [15]. A shared molecular trunk is therefore defensible as one architectural option, but it should not be interpreted as evidence that the tasks possess one identical biological label. Sharing is a statistical mechanism whose scientific meaning depends on the tasks being connected.

Large-scale profile-QSAR provides a complementary precedent. All-Assay-Max2 pQSAR used information across a large assay portfolio while retaining assay-specific activity predictions [16]. The important principle for the present argument is not its reported industrial performance

but its separation of shared cross-assay information from endpoint-specific outputs. An assay-specific head can preserve distinct calibration, label distributions, and response mappings even when upstream molecular information is shared.

Sparse kinase modeling further shows why architecture and validation have to be considered together. Multitask models can outperform single-task alternatives in related sparse kinase settings, yet performance is weaker when testing is organized around chemically dissimilar clusters rather than random partitions [17]. Shared learning can therefore be useful without resolving extrapolation. Hierarchical models may additionally share information at several levels—for example among assays within a target family before sharing across broader contexts—but such hierarchies are proposed design choices and require evidence that the chosen grouping reflects transferable structure rather than convenient metadata.

The resulting architecture is best understood as modular: molecular encoding supplies reusable chemical information; context conditioning modifies how that information is interpreted; assay-specific heads retain endpoint identity; and a transfer gate or hierarchy regulates cross-task sharing. **Table 2** converts these components into validation requirements rather than treating architectural complexity as evidence of readiness. Conditioning also creates a falsifiable architectural test: if removing or perturbing assay context leaves predictions unchanged, the model may be ignoring context, or the supplied context may add no information beyond molecular and task identifiers.

Table 2. Proposed validation-readiness matrix for modular assay-aware transfer architectures

Component	Intended role	Evidence anchor	Readiness test	Failure interpretation
Shared molecular trunk	Reuse chemical features	Multitask QSAR	Compare with assay-isolated encoders	Gain may reflect shared chemistry, not shared biology
Context conditioner	Modify molecular representation by assay information	Distinct-task conditioning	Ablate or corrupt context fields	No change suggests context is unused or redundant
Assay-specific head	Preserve endpoint-specific mapping	Profile-QSAR	Compare with one pooled output	Pooled gain may conceal endpoint miscalibration
Hierarchical sharing	Borrow within related task groups	Sparse kinase multitask modeling	Hold out groups and dissimilar chemistry	Within-family success does not establish cross-family transfer
Transfer gate	Restrict harmful sharing	Proposed synthesis	Compare selective with indiscriminate sharing	Benefit must survive target-context validation

Note: Architecture rows are methodological synthesis derived from the completed argument. Evidence anchors support the stated modeling precedents; they do not validate this combined architecture.

Negative transfer, dataset shift, and conflicting labels

Cross-assay learning becomes scientifically unsafe when transfer remains beneficial on pooled validation data but harms the target context of interest. Task-oriented transfer

methods support the premise that source information should be selected for relevance rather than transferred indiscriminately [18]. In the proposed framework, negative transfer is therefore not limited to a lower aggregate score.

It also includes degradation confined to a particular assay, biological context, chemical region, or label type that would be hidden by averaging across tasks.

Dataset shift creates a related but distinct problem. Chemical-toxicity modeling has shown that data drift can materially alter model performance over time or across data distributions [19]. Although the magnitude of that result should not be transplanted quantitatively to bioassays, it establishes the methodological need to distinguish historical fit from performance after the data-generating process changes. For bioactivity prediction, shift may arise from new chemistry, changed assay technology, altered target constructs, different laboratories, or changed sampling practices.

Conflicting labels should consequently remain diagnostically visible. A disagreement may represent

cross-source measurement noise, a real difference between measurement types, an activity cliff, contextual biology, or erroneous annotation. Automatically averaging conflicts assumes the explanation before testing it. The proposed model should instead retain provenance and allow disagreement to influence transfer decisions, calibration, or abstention. This separation is operationally important because the appropriate response differs by cause: measurement noise may call for replication, contextual disagreement may justify separate targets, and chemical discontinuity may require local modeling rather than assay redefinition. **Figure 1** summarizes this manuscript-derived logic as a distributed set of heterogeneity signals surrounding a compact transfer decision core.

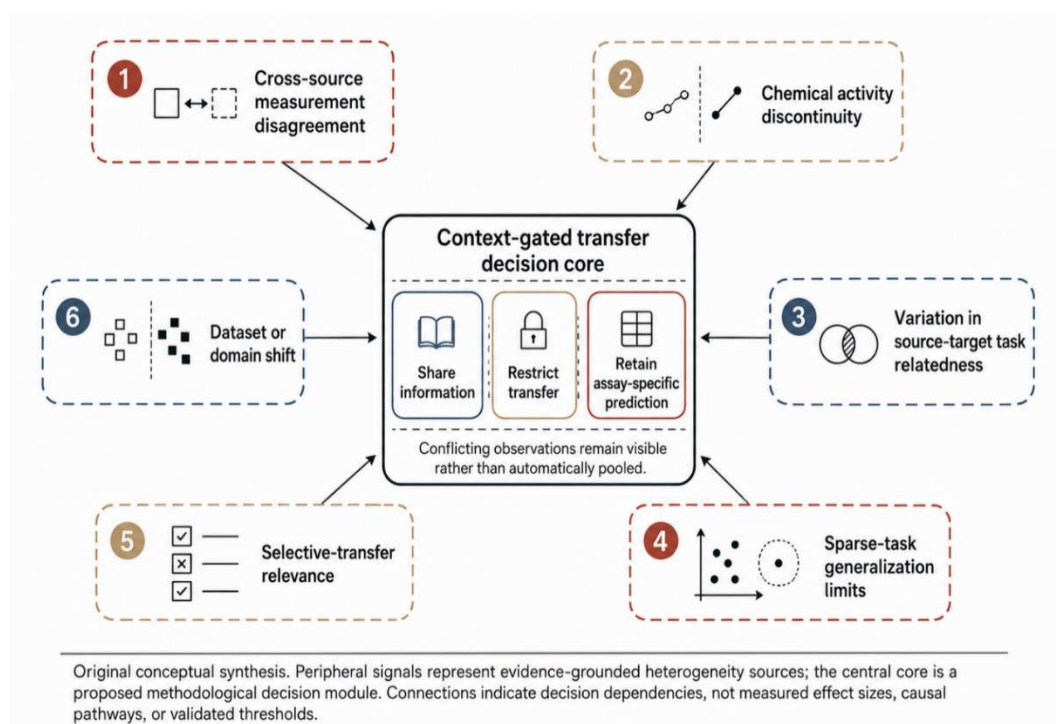


Figure 1. Context-gated transfer under assay heterogeneity

Evaluation across assay families and novel contexts

Evaluation should ask where generalization occurs. Few-shot drug-discovery benchmarking shows that low-data tasks require evaluation procedures designed for learning from limited task-specific information [20]. In assay-aware prediction, a model evaluated only by randomly holding out molecules within already represented assays answers a narrower question than a model challenged with a new assay. Good performance in the first regime should therefore not be reported as evidence of context transfer.

Large-scale ChEMBL benchmarking likewise demonstrates that validation design can materially influence estimated drug-target prediction performance

[21]. Chemical-series separation is especially important because a model can appear to transfer across assays while succeeding primarily through close analogs already represented during training. Context-aware evaluation should consequently cross two dimensions: chemical novelty and assay novelty. A useful design can test known-assay/novel-molecule, novel-assay/familiar-chemistry, and joint novel-assay/novel-chemistry conditions separately.

Assay-family evaluation adds a further level. Holding out one assay from a well-represented family tests interpolation across related contexts; holding out an entire family, platform, or measurement regime asks whether the

model learned a more transportable relationship. Neither setting by itself proves prospective generalization, because laboratory and temporal shifts may remain unrepresented. The goal is diagnostic decomposition: determine whether an apparent gain originates from molecular learning, context adaptation, or leakage between them. Reporting should therefore provide assay-resolved results alongside

aggregate metrics and should identify which contexts were represented during training. A global score can otherwise reward frequent or easier assays while masking failure on precisely the novel contexts that motivated transfer. **Table 3** identifies data-quality conditions that can otherwise make those sources of performance difficult to distinguish.

Table 3. Data-quality and evaluation-consequence matrix for cross-assay generalization claims

Data condition	Diagnostic check	Consequence if ignored	Permissible interpretation
Chemical-series overlap	Cluster- or scaffold-aware holdout	Memorized analog relationships inflate transfer estimates	Performance applies to represented chemistry
Sparse novel assay	Few-shot context evaluation	Pooled testing hides adaptation difficulty	Report support-set-dependent generalization
Temporal or domain drift	Shift-stratified validation	Historical accuracy overstates stability	Restrict claims to tested distributions
Conflicting source measurements	Preserve provenance rather than average	Disagreement becomes unexplained noise	Treat comparability as unresolved
Activity-cliff regions	Analyze local discontinuities	Smooth similarity assumptions obscure hard cases	Separate chemical from contextual failure
Sparse multitask matrix	Compare random and dissimilarity-aware splits	Shared-task gains may mask extrapolation limits	Attribute gains to tested split regime only

Note: The matrix synthesizes evidence-supported failure modes into proposed reporting boundaries; it does not prescribe universal split thresholds.

A context-aware transfer framework

The proposed context-aware transfer framework joins the preceding distinctions into a sequence of modeling decisions rather than a single required network. First, the molecular encoder represents chemical structure independently of assay identity. Second, an assay-context representation captures recorded information relevant to interpreting the observation, while explicitly marking missing or uncertain context. Third, a relatedness mechanism estimates whether knowledge from candidate source tasks should be made available to the target. Fourth, an assay-specific prediction pathway preserves the possibility that the same molecule maps differently under different experimental contexts.

Evidence from few-shot nuclear-receptor modeling shows that related biological tasks can provide useful information in low-data settings [22]. The framework uses this as support for structured borrowing, not as evidence that target families define universally valid transfer groups. Biological family membership may be informative, but similarity in target annotation can coexist with differences in constructs, readouts, cellular systems, or chemistry.

In-context bioactivity prediction offers another route: task information can be supplied through examples rather than only through fixed identifiers or textual metadata [23]. The framework therefore allows assay context to be represented by structured descriptors, learned descriptions, support compounds, or combinations of these sources. Context representations should be compared by ablation because an apparently informative embedding can encode

dataset identity without capturing the variables responsible for assay behavior.

Selective auxiliary-task transfer further supports making relevance explicit. Transfer from activity-cliff prediction to drug-target interaction prediction illustrates that a deliberately chosen auxiliary task can provide useful information for another prediction objective [24]. The proposed framework generalizes this principle cautiously: source tasks should compete for inclusion through scientifically defensible relatedness and target-context validation. A source that improves pooled performance but degrades the target assay should be down-weighted or excluded. A model should also be permitted to abstain from transfer when context is too novel or poorly documented. These rules define a testable methodological architecture; they do not establish that the resulting system will outperform well-tuned single-task models in every assay regime. Importantly, the framework permits the simplest model to win: where context contributes no reproducible information, additional conditioning or hierarchy should be rejected rather than retained for conceptual completeness.

Implications for bioactivity data integration

Context-aware transfer changes the objective of bioactivity integration. The goal is no longer to maximize the number of measurements assigned to a common endpoint, but to create a representation in which records can be shared without deleting distinctions required for later modeling. Papyrus demonstrates the value of extensive curation,

normalization, and standardized identifiers for large-scale bioactivity prediction resources [25]. Such standardization is necessary for computational interoperability, yet it cannot recover protocol variables that were never recorded or make biologically different assays equivalent.

PubChem likewise integrates chemical and bioassay information at repository scale across highly diverse submitted experiments [26]. For context-aware learning, that heterogeneity should be represented through provenance and assay-level identifiers rather than treated purely as an obstacle to cleaning. A useful integrated record should preserve, when available, measurement type, target or biological system, experimental modality, source, units and transformations, activity qualification, and links to the originating assay. Missingness in these fields is itself relevant because a model cannot reliably condition on context that is absent.

Label construction deserves equal attention. Experiments with stochastic negative examples show that the definition of negative data can alter molecular bioactivity prediction [27]. A presumed negative, an explicitly measured inactive compound, and an untested compound therefore should not be silently collapsed into the same semantic state. The same logic applies to censored potency values, qualifiers, and contradictory measurements. Context-aware integration favors typed observations over premature consensus labels.

This perspective also changes how duplicates are treated. Multiple records for the same molecule–target pair are not automatically redundant when assay conditions or measurement processes differ. Some can reveal reproducibility limits, source effects, or genuine context dependence. The proposed integration workflow would standardize chemistry and units while retaining an observation layer that preserves provenance and context before any model-specific aggregation. Aggregation may still be appropriate when equivalence is justified, but it becomes an explicit, auditable decision. In this way, integrated databases can support both cross-assay learning and assay-specific prediction without pretending that data harmonization has solved the biological comparability problem.

Limitations

The proposed framework is methodological and has not been prospectively validated as a unified architecture. Its usefulness will depend on the availability and quality of assay metadata, the chemical coverage of source and target tasks, and whether recorded context variables correspond to differences that actually influence the observation. Some relevant conditions may remain undocumented, while others may act as proxies for laboratories or datasets rather than biology.

Reliability estimation also remains conditional. Conformal prediction provides a principled route for attaching calibrated predictive regions or confidence information under its assumptions [28], but calibration achieved within one context need not survive a substantial assay or chemical shift. Context-aware models therefore require calibration studies aligned with the intended deployment regime rather than a generic uncertainty layer.

Uncertainty methods themselves should not be treated as interchangeable assurances of correctness. Comparative benchmarking frameworks show that uncertainty-quantification approaches can behave differently across molecular prediction settings [29]. Future validation should therefore test point accuracy, calibration, transfer benefit, and failure detection separately. Prospective studies should withhold entire assay families or platforms, include deliberately unrelated source tasks as negative-transfer controls, ablate context representations, and evaluate joint chemical and contextual novelty. Such experiments could support, refine, or reject the proposed framework. External replication across organizations would also be needed to separate architecture-specific gains from local assay conventions, curation practices, and portfolio composition.

CONCLUSION

Learning across bioassays is valuable precisely because assays contain partially reusable information, but reuse should not require pretending that their measurements are biologically identical. This article proposes a context-aware methodological framework that separates molecular representation from assay representation, distinguishes transfer from label pooling, preserves assay-specific outputs, and makes negative transfer and context novelty explicit validation targets.

The framework also reframes bioactivity integration as preservation of conditional comparability rather than maximal endpoint consolidation. Conflicting measurements, missing context, sparse tasks, and distribution shift are treated as information about the limits of transfer rather than inconveniences to be erased automatically. Its central claims remain methodological: no architecture, context representation, relatedness criterion, uncertainty method, or transfer gate is asserted to be universally superior.

The decisive next step is empirical falsification. Models should be challenged under separate and joint chemical and assay novelty, with selective transfer compared against pooled, single-task, and conventional multitask alternatives. Until such evidence exists, context-aware transfer is best regarded as a disciplined way to formulate the prediction problem and expose where cross-assay

assumptions require validation, not as a guarantee of better bioactivity prediction.

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] 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
- [3] 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
- [4] Zdrazil B, Felix E, Hunter F, Mannes 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
- [5] Landrum GA, Riniker S. Combining IC50 or Ki 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
- [6] 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
- [7] 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
- [8] 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
- [9] Atas Guvenilir H, Doğan T. How to approach machine learning-based prediction of drug/compound-target interactions. *J Cheminform*. 2023;15(1):16. doi:10.1186/s13321-023-00689-w
- [10] Yan J, Ye Z, Yang Z, Lu C, Zhang S, Liu Q, et al. Multi-task bioassay pre-training for protein-ligand binding affinity prediction. *Brief Bioinform*. 2024;25(1):bbad451. doi:10.1093/bib/bbad451
- [11] 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
- [12] Eckmann P, Anderson J, Yu R, Gilson MK. Ligand-based compound activity prediction via few-shot learning. *J Chem Inf Model*. 2024;64(14):5492-9. doi:10.1021/acs.jcim.4c00485
- [13] Arvidsson McShane S, Ahlberg E, Noeske T, Spjuth O. Machine learning strategies when transitioning between biological assays. *J Chem Inf Model*. 2021;61(7):3722-33. doi:10.1021/acs.jcim.1c00293
- [14] Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, et al. Transfer learning for drug discovery. *J Med Chem*. 2020;63(16):8683-94. doi:10.1021/acs.jmedchem.9b02147
- [15] Xu Y, Ma J, Liaw A, Sheridan RP, Svetnik V. Demystifying multitask deep neural networks for quantitative structure-activity relationships. *J Chem Inf Model*. 2017;57(10):2490-504. doi:10.1021/acs.jcim.7b00087
- [16] Martin EJ, Polyakov VR, Zhu XW, Tian L, Mukherjee P, Liu X. All-Assay-Max2 pQSAR: Activity predictions as accurate as four-concentration IC50s for 8558 Novartis assays. *J Chem Inf Model*. 2019;59(10):4450-9. doi:10.1021/acs.jcim.9b00375
- [17] Luukkonen S, Meijer E, Tricarico GA, Hofmans J, Stouten PFW, van Westen GJP, et al. Large-scale modeling of sparse protein kinase activity data. *J Chem Inf Model*. 2023;63(12):3688-96. doi:10.1021/acs.jcim.3c00132
- [18] Duan Y, Yang X, Zeng X, Wang W, Deng Y, Cao D. Enhancing molecular property prediction through task-oriented transfer learning: Integrating universal structural insights and domain-specific knowledge. *J Med Chem*. 2024;67(11):9575-86. doi:10.1021/acs.jmedchem.4c00692
- [19] Morger A, Garcia de Lomana M, Norinder U, Svensson F, Kirchmair J, Mathea M, et al. Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data. *Sci Rep*. 2022;12(1):7244. doi:10.1038/s41598-022-09309-3
- [20] Vella D, Ebejer JP. Few-shot learning for low-data drug discovery. *J Chem Inf Model*. 2023;63(1):27-42. doi:10.1021/acs.jcim.2c00779
- [21] 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
- [22] Torres LHM, Arrais JP, Ribeiro B. Combining graph neural networks and transformers for few-shot nuclear receptor binding activity prediction. J Cheminform. 2024;16(1):109. doi:10.1186/s13321-024-00902-4
- [23] Schimunek J, Luukkonen S, Klambauer G. MHNfs: Prompting in-context bioactivity predictions for low-data drug discovery. J Chem Inf Model. 2025;65(9):4243-50. doi:10.1021/acs.jcim.4c02373
- [24] Ibragimova R, Iliadis D, Waegeman W. Enhancing drug-target interaction prediction through transfer learning from activity cliff prediction tasks. J Chem Inf Model. 2025;65(13):6558-67. doi:10.1021/acs.jcim.5c00484
- [25] Béquignon OJM, Bongers BJ, Jespers W, IJzerman AP, van der Water B, van Westen GJP, et al. Papyrus: A large-scale curated dataset aimed at bioactivity predictions. J Cheminform. 2023;15(1):3. doi:10.1186/s13321-022-00672-x
- [26] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373-80. doi:10.1093/nar/gkac956
- [27] 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
- [28] Arvidsson McShane S, Norinder U, Alvarsson J, Ahlberg E, Carlsson L, Spjuth O. CPSign: Conformal prediction for cheminformatics modeling. J Cheminform. 2024;16(1):75. doi:10.1186/s13321-024-00870-9
- [29] Lanini J, Huynh MTD, Scebbba G, Schneider N, Rodríguez-Pérez R. UNIQUE: A framework for uncertainty quantification benchmarking. J Chem Inf Model. 2024;64(22):8379-86. doi:10.1021/acs.jcim.4c01578