IntuitionLabs

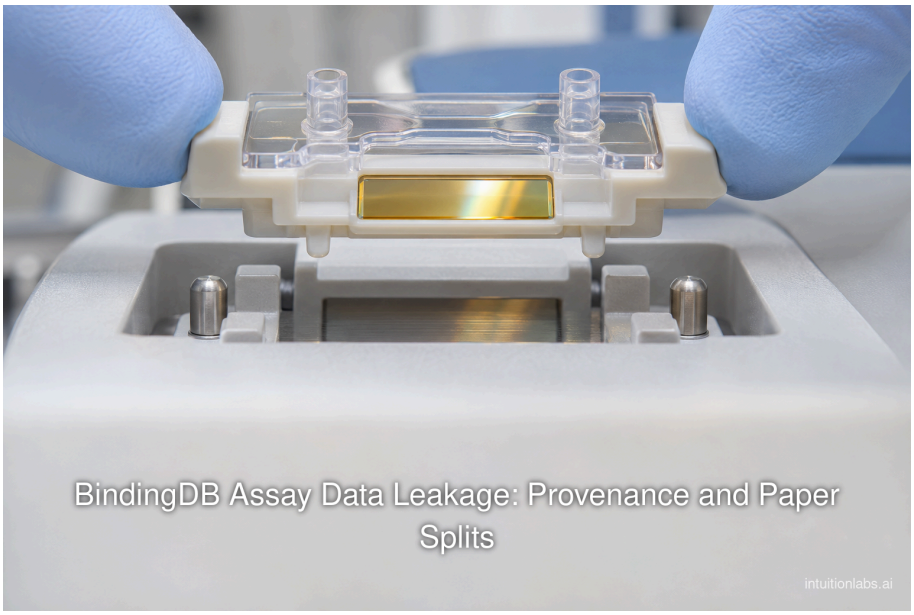
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

BindingDB Assay Data Leakage: Provenance and Paper Splits

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-11-02 · bindingdb assay data leakage · bindingdb data leakage affinity prediction · bindingdb assay provenance · paper-level split · binding affinity dataset · bindingdb download · assay-aware bindingdb · group split by publication · drug target affinity benchmark · bindingdb curation

Original article: https://intuitionlabs.ai/articles/bindingdb-assay-provenance-paper-level-leakage



In this article

1. Executive Summary
2. Introduction and Background
3. What a Provenance-Aware Dataset Must Establish
4. Freezing Exports and Running the Three-File Join
5. Assay Context, Censoring and Conflicting Repeats
6. Paper-, Scaffold- and Time-Level Evaluation
7. The Assay-Aware BindingDB Preprint: Findings and Limits
8. Data Analysis and Evidence
9. A Go/No-Go Evidence Pack for an Internal Pilot
10. Frequently Asked Questions (FAQs)
11. Implications and Future Directions
12. Conclusion

Executive Summary

BindingDB assay data are usable for an internal affinity-model pilot only if every measurement row can be traced to its source paper or patent and to the assay description that produced it. BindingDB's download page, read on November 2, 2026, lists **3,241,782 measurements, 1,440,011 compounds and 11,509 targets** in its 202611 release (www.bindingdb.org). The main TSV (tab-separated values) archive keeps one row per binding measurement, but the plain-text assay descriptions are deliberately stored elsewhere because they are "bulky and repetitive". Recovering them takes a documented two-step join: Reaction Set ID to Entry ID plus Assay ID, then Entry ID plus Assay ID to the assay text.

The reason this matters for machine learning is evaluation leakage. BindingDB states that an entry is typically tied to a single article or patent (academic.oup.com), and that its data generally come from a single institution measured in the same manner. Rows from one paper therefore share assay protocol, target and chemotype series, which lets a model recognise the paper instead of learning structure-activity relationships. A **September 27, 2026** arXiv preprint on assay-aware BindingDB (not peer reviewed) makes this explicit: under a pair-level split, an assay-context embedding "would act as a near-unique identifier" of the source paper (arxiv.org), so the authors evaluate on paper-level held-out splits.

The preprint's headline numbers are bounded. Across **65,169 pairs** with uncensored affinities, adding assay context moved mean squared error (MSE) from **1.27 to 1.19** and Pearson correlation from **0.64 to 0.67** (arxiv.org). These are the authors' own unreplicated results. Isothermal titration calorimetry (ITC) showed no improvement (arxiv.org), censored affinities were excluded, and the model sits on weights already pretrained on BindingDB.

Independent literature supports the broader caution. Grouping by assay is an established design in ChEMBL benchmarks (pmc.ncbi.nlm.nih.gov), and different labs' measurements of the same compound disagree: **roughly 65%** of same-target IC50 pairs from minimally curated ChEMBL assays differ by more than 0.3 log units (pubmed.ncbi.nlm.nih.gov). Strict curation of ChEMBL IC50 data reduces that disagreement but discards **almost 99%** of the data (pmc.ncbi.nlm.nih.gov).

The practical recommendation is a go/no-go evidence pack: frozen, checksummed exports; a join audit with denominators; an assay-context table with human-reviewed samples; and split-overlap counts for random, paper-level, scaffold and time splits. No audit counts were measured for this report; join cardinalities are quoted from documentation or labelled as checks to run. Licensing statements also differ between BindingDB pages, so counsel should confirm terms before any redistribution (academic.oup.com).

Introduction and Background

BindingDB is a public database of measured protein-small molecule binding affinities, curated from articles and patents and supplemented by imports such as ChEMBL (academic.oup.com). Computational chemistry teams use it as a training and benchmark source because it is large, downloadable monthly, and covers Ki, IC50, Kd and EC50 endpoints. The curators note strong interest in the dataset for AI methods (academic.oup.com). The curators report that most stored affinity data are IC50 values, at **1.8 million** (academic.oup.com). About half of the measurements were extracted from US patents and scientific articles, and most of the remainder were imported from ChEMBL (academic.oup.com).

That scale invites a shortcut: download the main TSV, convert units, deduplicate molecules, split rows at random, and train. Unit conversion and molecular identity are necessary hygiene, but they leave provenance untouched. A BindingDB row is a measurement made in a particular experiment, reported in a particular document. If the evaluation split ignores that structure, test rows can sit next to near-identical training rows from the same paper.

This report is an explainer and acceptance reference for teams deciding whether a BindingDB-derived set is fit for an internal model pilot, and whether further assay curation is worth funding. It documents the **three-file join**, specifies the provenance and assay-context fields to retain, maps split designs to the leakage each one tests, summarises a recent preprint on assay-aware BindingDB, and proposes an evidence pack for a go/no-go decision.

Three boundaries apply. Measured counts of unmatched mappings, multiple descriptions, missing context or censored rows are not reported here; they appear only as audit steps to run on a team's own export. The preprint is not peer reviewed and was not replicated here. Nothing here is legal advice or a regulatory opinion. IntuitionLabs is a consultancy, not a data vendor or software provider; the evidence cited here comes from independent public sources.

What a Provenance-Aware Dataset Must Establish

A provenance-aware BindingDB dataset answers five questions for every row that enters a training or evaluation set. Where did the measurement come from? Which assay produced it? Under what conditions? Is the value exact or censored? Which other rows share its source?

Source and measurement unit

The TSV format documentation says each row contains information for one binding measurement (www.bindingdb.org). The first column is a stable internal identifier for the binding reaction (www.bindingdb.org), which the download page calls the Reaction Set ID. A Curation/DataSource field records

provenance and is typically one of BindingDB, ChEMBL, PubChem or PDSP Ki (www.bindingdb.org). That split matters: BindingDB says data it extracts itself typically include more detail on experimental conditions than imported data (www.bindingdb.org), and it imports from ChEMBL only data that have a well-defined protein target (academic.oup.com).

An Institution column records where the measurement was made, usually a university or company (www.bindingdb.org). For data drawn from PubChem there is a separate assay identifier column.

Why the assay is the unit that matters

Affinity values are not property constants. The NCBI Assay Guidance Manual notes that substrate-to-Km ratio and the amount of product formed have the greatest effect on measured IC50 (www.ncbi.nlm.nih.gov), and that concentration-response experiments run at constant enzyme and substrate concentrations (www.ncbi.nlm.nih.gov). An IC50 is therefore a function of the assay design. BindingDB's own curators write that different measurement procedures can introduce different biases (pmc.ncbi.nlm.nih.gov), and the CARA benchmark authors make the same point about assays (pmc.ncbi.nlm.nih.gov).

Public databases make this hard to control. A 2013 statistical analysis of ChEMBL data observed that assay details are not reported in public bioactivity databases (pmc.ncbi.nlm.nih.gov). The ChEMBL curators likewise describe assay annotation as a filter that removes assays that are clearly not comparable (link.springer.com), and still observe significant variability between measurements from different labs (link.springer.com).

Evidence classes in a provenance-aware set

- **Source span.** The exact sentence or table cell in the assay description or paper that supports a field value.
- **Human-reviewed value.** A value an assay scientist checked against the source and signed off.
- **Missing or not applicable.** A value that the source does not state, or a field that does not apply to the technique. The TSV documentation leaves missing information as a blank cell, so a blank cannot distinguish "absent from the source" from "not applicable".
- **Model or rule inference.** A value produced by an extraction model or keyword rule. These must be labelled as inference, never mixed with reviewed values.

Freezing Exports and Running the Three-File Join

Freeze the inputs first

BindingDB files change. The download page says the files are updated when new data are added, usually monthly (www.bindingdb.org), and file names carry the release tag, such as the **202611** main TSV archive, listed at 566.75 MB and updated October 31, 2026. The release tag changes with each update, so a 202610 file and a 202611 file are different snapshots; the page read on November 2 listed 202611. Anyone who pins only "the latest download" cannot reproduce a result a month later.

The release notes quantify the churn. The October 5, 2026 notes report **7,831** new affinity data (www.bindingdb.org), the August 22 notes report 10,190 (www.bindingdb.org), and the February 5 notes report 7,233 (www.bindingdb.org). The release notes list counts rather than schema or mapping changes, so they do

not show whether column definitions or mappings changed between releases.

For documentable AI training, BindingDB recommends a single archived version from its long-term archive, with quarterly deposits at the University of California San Diego library (www.bindingdb.org). A pilot should therefore record, for each input file: the versioned file name, byte size, the published MD5 checksum, the download date, and the archive handle if one is used. The download page links an MD5 file beside each archive.

The documented join

Table 1 below summarises the three files and the keys that connect them. Key names and descriptions come from BindingDB's download page and TSV documentation; the cardinality column states what the documentation implies and what an audit must confirm, because no export was profiled for this report.

File (202611 name)	Role and key	Join key to next file	Cardinality to verify	Schema or change risk
Main affinity TSV (BindingDB_All_202611_tsv.zip)	One row per binding measurement; first column is the Reaction Set ID	Reaction Set ID	One row per measurement; confirm the Reaction Set ID is unique per row	Column count may vary from row to row for multichain targets (verify when parsing); tab-delimited because no field contains a tab
Reaction-set mapping (BindingDB_rsid_eaids_202611_tsv.zip)	Maps Reaction Set ID to Entry ID concatenated with Assay ID (EntryID_AssayID)	EntryID_AssayID	Expected many measurements to one Entry-Assay pair; audit unmatched Reaction Set IDs and any Reaction Set ID with more than one pair	Mapping file is refreshed with each release; compare row counts across releases
Assay descriptions (BindingDB_Assays_202611_tsv.zip)	Maps Entry ID plus Assay ID to the plain-text description of each assay (www.bindingdb.org)	Entry ID plus Assay ID	Expected one description per Entry-Assay pair; audit missing and duplicated descriptions	Free text; wording style varies by curator and by source (article, patent, import)

The first interpretive point is that the assay is attached through an intermediate key, not a column in the main file. The documentation states the order explicitly: use the mapping file to go from the Reaction Set ID to the corresponding Entry ID and Assay ID, then refer to the Assays file for the plain-text description (www.bindingdb.org). The cardinalities in the table are expectations, not measurements. A defensible pipeline reports each as a count with a denominator.

Join audit counts to report

Script the audit, store its output with the dataset, and report each item with its denominator:

- **Input rows.** Rows in the main TSV after any filter, and rows dropped by each filter.
- **Distinct Reaction Set IDs.** Compare with input rows to confirm uniqueness.

- **Unmatched mappings.** Reaction Set IDs absent from the mapping file, as a count and as a percentage of input rows.
- **Multiple pairs or descriptions.** Reaction Set IDs mapped to more than one Entry-Assay pair, and Entry-Assay pairs with more than one description.
- **Missing assay text.** Mapped pairs with an empty or absent description.
- **Source distribution.** Rows by Curation/DataSource value, since imported rows may carry less condition detail.
- **Censored rows.** Rows whose affinity carries an inequality qualifier (see the next section).

The SDfile specification adds a caution for deduplication: it does not carry all the detail about each affinity measurement that BindingDB stores (www.bindingdb.org).

Assay Context, Censoring and Conflicting Repeats

The context fields worth retaining

BindingDB's technique list spans enzyme inhibition and kinetics, isothermal titration calorimetry, nuclear magnetic resonance, and radioligand and competition assays (www.bindingdb.org). A field schema that suits one technique may not apply to another, which is why a not-applicable state must exist alongside a missing state. *Table 2* proposes a retention schema. It is a design proposal, not a BindingDB format.

Field group	Examples	Likely source span	Human review	Missing / not applicable handling	Inference label
Provenance	Entry ID, Assay ID, article or patent identifier, institution, Curation/DataSource	Mapping file, main TSV columns	Spot check identifiers against the source document	Blank cell means absent in export	None; copy as exported
Target construct	Species, isoform, mutation, truncation, tagged or full-length	Assay description; source paper methods	Required for a sample per target family	Mark "not stated" separately from "wild type"	Mark any value inferred from a UniProt mapping
Assay technique	Radioligand binding, fluorescence polarization, surface plasmon resonance, ITC, enzyme inhibition	Assay description keywords	Reviewer confirms technique on a sample	Not applicable fields (for example, immobilization for ITC) set explicitly	Keyword rule output is inference
Experimental conditions	pH, temperature, buffer, substrate or radioligand concentration, incubation time	Assay description or paper methods	Required, since IC50 depends on these	Blank where the source is silent	LLM or regex extraction is inference
Endpoint and relation	Ki, IC50, Kd, EC50; exact versus qualified value	Main TSV value columns	Reviewer checks a sample of qualified values	Do not coerce qualifiers into point values	None

The table's central point is the separation of evidence classes. An extraction model can fill many cells quickly, but a cell filled by a model is a hypothesis. Human-reviewed values should stay distinct from extracted ones.

Censored values and qualifiers

The TSV documentation states that Ki, IC50, Kd and EC50 values are "occasionally reported as" greater-than values, "following the source format" (www.bindingdb.org). In practice, a qualified value is not a point estimate; it is a bound. Dropping such rows is a common simplification, and the assay-aware preprint did so, excluding pairs whose reported affinity is censored (arxiv.org). Dropping is defensible for a regression baseline but it changes the label distribution, so the pilot should report how many rows were removed and in which assays.

There is an alternative. A 2024 arXiv paper on censored regression adapted standard model types to censored labels using the Tobit model from survival analysis (arxiv.org). Whether that helps a given project is an empirical question, and it is future work rather than a finding here. The Papyrus resource illustrates the scale of what is at stake: of its compound-protein pairs, **2,585,248** carry exact bioactivity values and **354,981** carry censored data (jcheminf.biomedcentral.com).

Conflicting repeats and cross-source duplicates

The same compound-target pair can appear several times with different affinities. The Therapeutics Data Commons (TDC) documentation for BindingDB says many drug-target pairs share sequence information but have different binding affinities because of different experimental assays (tdcommons.ai), and offers a helper that retains either the maximum or the mean for duplicated pairs (tdcommons.ai). Either choice discards information about which assay produced which value, and both reduce a distribution to a point.

Two further traps are worth stating. First, BindingDB imports from ChEMBL, and an earlier analysis found that BindingDB overlaps heavily with ChEMBL with most values copied from it (pmc.ncbi.nlm.nih.gov). Training on BindingDB and testing on a ChEMBL-derived set can therefore be leakage across databases. Second, apparently separate assays on one target are often not independent: of twelve ChEMBL series measured in two assays, eleven had overlapping authors (pmc.ncbi.nlm.nih.gov). Document-level grouping is a safer proxy for independence than assay identifiers alone.

Unit errors also persist. BindingDB's curators note that the most common curation error is a factor-of-1000 error from miscuration (academic.oup.com), and ChEMBL flags entries that differ by exactly 3 or 6 orders of magnitude (chembl.gitbook.io). Cross-source filters have likewise removed pairs differing by exactly 3, 6 or 9 log units (pmc.ncbi.nlm.nih.gov). Unit hygiene complements, but does not replace, provenance tracking.

Paper-, Scaffold- and Time-Level Evaluation

Each split design tests a different kind of leakage.

What each split tests

- **Random pair-level split.** Tests interpolation within a known mixture of papers, targets and chemotypes. It is the least informative about new chemistry or new documents.

- **Paper- or document-level group split.** Keeps all rows from a source paper or patent on one side of the split. It removes same-paper analog series, shared assay protocol and any paper fingerprint a model might learn. It does not by itself remove similar compounds or the same target appearing in other papers.
- **Assay-level group split.** Similar to paper-level, but the group is the assay. The CARA benchmark defines an assay as a set of compound activities against a target measured under a specific experimental condition (pmc.ncbi.nlm.nih.gov) and splits training and evaluation at the level of assays (pmc.ncbi.nlm.nih.gov). Because assays in one paper often share authors and conditions, paper-level grouping is stricter.
- **Scaffold split.** Keeps compounds with the same core scaffold together. It tests chemotype extrapolation, but a similarity analysis found that, even under a scaffold split, **56%** of MoleculeNet HIV test molecules still have a training neighbour with Tanimoto similarity above 0.4 (arxiv.org).
- **Time split.** Trains on earlier measurements and tests on later ones, which approximates prospective use. Time-split validation is broadly recognised as the gold standard for validating predictive models in medicinal chemistry (jcheminf.biomedcentral.com), but the LP-PDBBind authors argue it is still not an ideal solution on its own, because similar proteins or ligands can remain across the cut (arxiv.org).
- **Cold-target or cold-compound splits.** These hold out whole targets or whole compounds. A study of four evaluation settings, which included new-drug and new-target cases, found that more realistic problem formulations may lead to drastically decreased prediction accuracies (pmc.ncbi.nlm.nih.gov).

The grouping tool is simple. The scikit-learn documentation advises group-wise cross-validation when the generative process has a group structure (scikit-learn.org), so that validation-fold samples come from groups not represented in the training fold (scikit-learn.org), and its GroupKFold places each group in the test set exactly once (scikit-learn.org). In a BindingDB pipeline, the group key is the source paper or patent identifier recovered from the join.

Leakage also has levels. The simplest form is the same samples appearing in multiple folds, which deduplication removes (www.nature.com); the harder forms need grouping or similarity-aware splitting.

A calculation: source-paper overlap in a pair-level split

For a pair-level random split with training fraction p (here 0.8), a test row from a paper contributing n rows shares its paper with at least one training row with probability 1 minus (1 minus p) to the power of (n minus 1). This is arithmetic, not a measurement of any BindingDB file:

- **n = 2, 3, 5 and 10: 0.80, 0.96, 0.9984 and 0.9999995** respectively. With n = 1 the value is 0.0, since no other row exists to leak from.

BindingDB's database paper reports averages of **160** data per patent and **40** per article (academic.oup.com). Under these averages, almost every test row of an average article or patent has same-paper training neighbours in a pair-level random split. In a paper-level split this overlap fraction is zero by construction. The audit should therefore compute, for each proposed split, the fraction of test rows whose source paper also appears in training, the fraction with a training neighbour above a chosen Tanimoto threshold, and the fraction of test targets present in training.

(Hypothetical Example)

Suppose a pilot extracts 30,000 uncensored Ki rows for 40 kinase targets from 1,200 papers. A random 80/20 split might leave 99% or more of test rows sharing a paper with training rows; a paper-level split leaves 0%. The numbers are illustrative and were not computed from real data.

What the literature says about how much the split matters

- **Benchmark leakage can be large.** In a PDBbind analysis, about 600 train-test similarities involved **49%** of all CASF complexes (pmc.ncbi.nlm.nih.gov). This concerns structure-based data, not BindingDB, but the mechanism is the same.
- **Random splits of benchmark data.** The LP-PDBBind authors state that a random train, validation and test split of PDBbind will inevitably introduce significant data leakage (arxiv.org).
- **Random versus temporal splits differ in magnitude.** Papyrus reports average Matthews correlation coefficients of 0.51, 0.61 and 0.60 under random splits (pmc.ncbi.nlm.nih.gov), but they plummeted to **0.20, 0.29 and 0.30** for QSAR, PCM and DNN models under a temporal split (pmc.ncbi.nlm.nih.gov). Pooled metrics also flatter: in CARA's simulated example, average per-assay correlation was only **0.28**, below the all-assay value (pmc.ncbi.nlm.nih.gov). The Papyrus authors chose temporal splits partly because that approach minimizes congeneric series being split between training and test sets (pmc.ncbi.nlm.nih.gov).

No document-level versus random experiment on BindingDB affinity regression was identified among the sources cited here. The evidence above comes from adjacent benchmarks, so the size of the effect on a team's own targets must be measured, not assumed.

The Assay-Aware BindingDB Preprint: Findings and Limits

What the preprint reports

The preprint "Assay-Aware BindingDB: Curating Experimental Context for Binding Affinity Prediction" was submitted to arXiv on September 27, 2026 (arxiv.org). It is labelled a preprint throughout this report because it has not been peer reviewed, and arXiv lists only version 1 at the time of reading. The authors assign assay type by keyword matching on BindingDB's description field, then extract context with two language-model agents instantiated with Qwen3.5-27B. They retain **74,425** protein-ligand pairs across ITC, surface plasmon resonance (SPR), radioligand binding assays (RBA) and fluorescence polarization assays (FPA), from **5,991** unique papers (arxiv.org).

This article's *Table 3* reproduces the preprint's Table 2 (dataset composition) with attribution. Values are transcribed from the preprint's Table 2 as it appeared on November 2, 2026 (arxiv.org). The text states that the counts are computed over the final retained pairs.

Assay type	Retained pairs	Unique targets	Unique compounds	Literature retrieved (papers)
ITC	1,898	272	1,382	346

Assay type	Retained pairs	Unique targets	Unique compounds	Literature retrieved (papers)
SPR	4,253	424	3,351	590
RBA	49,291	593	30,136	4,343
FPA	18,983	373	11,349	712
Total as printed	74,425	1,662	46,218	5,991

Source: Wu et al. preprint, Table 2 (arXiv 2609.34001v1), reproduced for analysis. Because a target or compound can be measured by more than one technique, the target and compound totals might be expected to be non-additive across slices. The arithmetic shows they are additive: 272, 424, 593 and 373 sum to 1,662 targets, 1,382, 3,351, 30,136 and 11,349 sum to 46,218 compounds, and papers (346, 590, 4,343 and 712) sum to 5,991. Exact additivity implies either that no target, compound or paper appears in more than one slice, or that totals were summed from per-slice counts. The preprint text does not say which, so an independent check against the BindingDB release is needed. These sums are this report's arithmetic on the printed table.

Results and caveats

The model results come from a paired comparison with and without assay context. On the combined regime of 65,169 pairs (arxiv.org), the authors report MSE falling from 1.27 to 1.19 and Pearson correlation rising from 0.64 to 0.67, as mean values over 10 independent paper-level splits. Results by assay type differ: SPR shows the largest single-cohort improvement (arxiv.org), the RBA MSE change is described as not significant (arxiv.org), and ITC is the lone exception (arxiv.org) (per-assay values appear under Data Analysis).

The caveats are the substance of the decision:

- **Unreplicated and author-reported.** These are the authors' own experiments. This report did not reproduce them, and a benchmark reproduction is future work.
- **Censored affinities excluded.** The combined regime removes pairs whose reported affinity is censored, so 74,425 pairs shrink to 65,169.
- **Pretraining exposure.** Both configurations are evaluated on top of Boltz-2 weights, which the authors state are pretrained on BindingDB (arxiv.org).
- **Coverage.** The paper covers four assay types and says other formats are out of scope, and some extraction failures reflect under-specified source descriptions.
- **Extraction accuracy is sampled.** Extraction is evaluated on **30** sampled papers per assay type, **120** in total (arxiv.org). That is a sample, not a full human review.
- **Leakage limits not addressed.** Splits are by paper only. The limitations text does not discuss compound or target overlap across train and test, so targets and compounds can recur. The results should not be read as evidence of generalisation to new chemistry or new targets.
- **Dataset introduced, but no repository or licence located.** The abstract says the authors "introduce Assay-Aware BindingDB", but the arXiv abstract page and the main text read list no paper-specific code repository, release location or separate dataset or code licence, and the HTML version was truncated in Appendix B, so whether the dataset or code is publicly released is unresolved; locate and check any release before relying on it.

What the preprint does and does not imply for a pilot

It supports one narrow idea: assay context may carry signal for some assay types, and testing that requires a paper-level split, since the context embedding can act as a paper identifier otherwise. The authors state that such a model could recognise the paper at test time and predict within its range, inflating metrics without genuine context-aware generalization (arxiv.org). It does not show that the gains transfer to a biotech's own targets, nor that language-model-extracted metadata are ground truth, so assay scientists should review a sample before any reliance.

Data Analysis and Evidence

This section collects the quantitative evidence behind the acceptance criteria.

Reported results by assay type (preprint, author-reported)

The preprint's Table 3 (arxiv.org) reports mean MSE without and with assay context over 10 paper-level splits. Reading the values as transcribed from that table:

- **SPR (3,650 pairs):** MSE 2.47 to 2.11, Pearson 0.50 to 0.59, the largest single-cohort improvement.
- **FPA (15,535 pairs):** MSE 1.06 to 0.94, Pearson 0.69 to 0.73.
- **RBA (44,223 pairs):** MSE 1.17 to 1.15, Pearson 0.59 to 0.61; the MSE change is reported as not significant.
- **ITC (1,761 pairs):** MSE 1.35 to 1.47, Pearson 0.49 to 0.41; context made results worse.
- **All (65,169 pairs):** MSE 1.27 to 1.19, Pearson 0.64 to 0.67.

The cohort pair counts are smaller than the dataset totals in this article's *Table 3* (the preprint's Table 2) because censored and unprocessable pairs were excluded. Standard deviations across splits are reported in the paper but not reproduced here, and the text reports no confidence intervals or random seeds. With only 1,761 ITC pairs, the ITC result may reflect sample size as well as technique, which this report cannot separate.

Noise floors that bound achievable accuracy

- **Independent Ki measurements.** An analysis of public Ki data estimated a median error of **0.34 pKi units** (pubmed.ncbi.nlm.nih.gov) and cautioned that models with errors smaller than the experimental uncertainty are necessarily overtrained (pubmed.ncbi.nlm.nih.gov).
- **Cross-assay IC50 pairs.** In ChEMBL data with minimal curation, **almost 65%** of points differ by more than 0.3 log units (pubmed.ncbi.nlm.nih.gov) and **27%** differ by more than one log unit (pubmed.ncbi.nlm.nih.gov). With maximal curation, **48%** still differ by more than 0.3 log units (pubmed.ncbi.nlm.nih.gov).
- **Cost of curation.** Maximal curation of IC50 data discards almost 99% of the data (pmc.ncbi.nlm.nih.gov). One of the curation steps removes measurement pairs whose assays were published in the same document (pmc.ncbi.nlm.nih.gov).

For scale, ChEMBL 33 bioactivities were measured in more than **1.6 million** assays (pmc.ncbi.nlm.nih.gov), so manual comparability checks do not scale. These figures come from ChEMBL, not BindingDB exports, but BindingDB marks the ChEMBL-curated rows it imports (academic.oup.com), so the comparison is relevant though not identical.

A Go/No-Go Evidence Pack for an Internal Pilot

Decision and audience

The decision is whether a BindingDB-derived affinity set is fit for an internal model pilot, and whether assay curation is worth funding. Regulatory frameworks supply vocabulary, not necessarily requirements. The FDA's January 2025 draft guidance proposes a risk-based credibility framework (www.fda.gov) built around a defined context of use (www.fda.gov). The draft defines reliable data as accurate, complete and traceable (www.fda.gov). It also states that it does not address AI used in drug discovery (www.fda.gov), and it was distributed for comment only (www.fda.gov). The European Medicines Agency's reflection paper, adopted in final form on **September 9, 2024** (www.ema.europa.eu), says AI in drug discovery can be of low regulatory impact (www.ema.europa.eu), while still warning that the risk of direct or indirect data leakage must be considered (www.ema.europa.eu). EMA and FDA jointly published ten principles for good AI practice in medicine development (www.ema.europa.eu), including documented data source provenance (www.fda.gov). The NIST AI Risk Management Framework likewise ties maintained training-data provenance to attribution and transparency (nvlpubs.nist.gov).

An early-stage internal pilot is probably outside those documents' scope, but if results later enter regulatory evidence, the sponsor should be able to show how development data were collected, processed and annotated (www.fda.gov) and keep data sources and processing steps traceable (www.ema.europa.eu). A datasheet practice makes this concrete: Gebru and coauthors propose that every dataset be accompanied by a datasheet documenting its motivation and other characteristics (arxiv.org), and the FAIR principles require that data be associated with detailed provenance (www.nature.com).

Acceptance matrix

Use the following as pass criteria, and record the measured value beside each:

- **Reproducibility.** Versioned file names, MD5 checksums and download date stored; the build reruns to the same counts.
- **Join integrity.** Unmatched Reaction Set IDs, multi-pair mappings and missing descriptions reported as counts and percentages. Threshold for a pass is a team decision; this report offers none.
- **Provenance retained.** Every training row carries Entry ID, Assay ID, source document identifier and Curation/DataSource.
- **Censoring handled.** Qualified rows counted, either excluded with a count or modelled explicitly.
- **Conflicts documented.** Rule for repeated compound-target pairs stated (retain all, mean, maximum, or per-assay), with the number of affected pairs.
- **Context review.** An assay scientist reviews a stratified sample of extracted fields by technique and by source type, and records agreement per field. The sample size is a team decision.

- **Split audit.** Source-paper overlap fraction, target overlap and nearest-neighbour similarity reported for random, paper-level, scaffold and time splits. Primary reported metric is from the strictest split the use case permits.
- **Baselines.** A simple nearest-neighbour or ligand-only baseline run on the same split, so a model is compared against memorisation.
- **Licence review.** Terms confirmed for each source, as set out below.

Licensing is stated differently on different pages

Terms deserve a separate check because the sources disagree. The BindingDB information page states that staff-curated data are under Creative Commons Attribution 3.0 (www.bindingdb.org) and ChEMBL-imported data under the ShareAlike 3.0 license (www.bindingdb.org). The database paper states that BindingDB-curated data are shared under CC BY 4.0 (academic.oup.com), and a 2024 University of California San Diego archive record likewise says staff-curated data are CC BY 4.0 (library.ucsd.edu). ChEMBL itself publishes its data under the ShareAlike 3.0 license (www.ebi.ac.uk). The pages are undated or differ in date, so which is current is unresolved here. The Creative Commons BY-SA 4.0 deed notes that ShareAlike requires that contributions built on the material be distributed under the same license (creativecommons.org), though the same deed says commercial sharing is permitted (creativecommons.org). The ChEMBL license is version 3.0, whose legal text was not read. No source read addresses whether purely internal model training, or distribution of trained weights, counts as such an adaptation, so counsel should decide. The preprint's CC BY 4.0 license covers the paper itself (arxiv.org), not any dataset or code; no dataset or code licence was identified in the abstract page or main text read, and the full paper should be checked.

Funding assay curation: a framing

Curation is worth funding when the pilot's decision depends on assay-conditional values, such as potency ranking within a target across technologies. It is lower priority for early triage on a single in-house assay, where BindingDB serves only as pretraining. A staged approach limits cost: run the audit and split comparison first, then fund human review where split-corrected performance remains useful.

Frequently Asked Questions (FAQs)

What is BindingDB assay data leakage?

It is information from the evaluation set leaking into training through shared structure. For BindingDB, the most direct form is same-paper leakage: rows from one article or patent share conditions and series, and BindingDB states an entry typically maps to one article or patent. A related form is leakage across databases, since many BindingDB values come from ChEMBL.

How can a team get assay descriptions from a BindingDB download?

Map each Reaction Set ID in the main TSV to an Entry ID and Assay ID with the reaction-set mapping file, then look up the description in the Assays file.

Does BindingDB data leakage affect affinity prediction results?

It can. The effect depends on split design and has been demonstrated on adjacent benchmarks: for example, metrics dropped sharply under temporal splits in Papyrus (pmc.ncbi.nlm.nih.gov). No BindingDB-specific document-level versus random study was found, so a team should measure the gap on its own data.

What do BindingDB assay conditions include?

The main TSV documents fields such as institution and data source, and details are recorded in plain-text descriptions. Data extracted by BindingDB staff typically include more experimental-condition detail than imported data (www.bindingdb.org).

What is a paper-level split, and how is one built?

It keeps all rows from the same source paper or publication on one side of the split, which is a group split by publication. Recover a paper or patent identifier through the join, then pass it as the group key to a grouped splitter such as scikit-learn's GroupKFold.

How are Ki, IC50 and Kd handled in BindingDB curation?

They are separate endpoints with different counts, and IC50 is the most common. TDC splits BindingDB into Kd, IC50 and Ki datasets of **52,284, 991,486 and 375,032** pairs (tdcommons.ai), and inequality qualifiers can appear in values.

What preprocessing does a BindingDB machine-learning dataset need?

Pin and checksum the release, join the assay text, parse qualifiers, resolve repeated pairs with a documented rule, check unit outliers, and choose a split matching the use case, with a count and denominator at each step.

Implications and Future Directions

For computational chemistry leads, the main implication is a change in what counts as an acceptable dataset. A BindingDB extract is a collection of documents as much as of molecules, so the acceptance question becomes whether the team can show, row by row, which paper and assay each label came from, and whether reported performance survives a split that separates those papers.

For assay scientists, the implication is a role in the pipeline. Extraction by a language model is cheap, but whether a field such as substrate concentration was captured correctly can only be settled against the source. The preprint's 120-paper sampled evaluation shows how small a review can start, though it cannot establish rates for rare fields.

For scientific-AI platform owners, the architecture implication is versioned lineage: every derived table should record its input export, join logic and split definitions. BindingDB's own recommendation of an archived, citable version for AI training points the same way. The checklist above can be applied by internal staff or an outside adviser.

Open questions remain. A document-level versus random comparison on BindingDB affinity regression, replicated by an independent group, would clarify the size of the effect. Whether assay-context features generalise to new targets is untested in the preprint, since splits are by paper only. Handling of censored data in training has methods but thin BindingDB evidence. And each new BindingDB release changes the inputs, so this reference should be updated on each export release used, with the audit counts rerun.

Conclusion

BindingDB can support an internal affinity-model pilot, provided provenance is treated as part of the data and not as an afterthought. The export design places assay text outside the main file, so a documented three-file join is the first requirement. The second is evaluation design: because an entry typically maps to one article or patent, pair-level random splits place same-paper rows on both sides of the split, and paper-level grouping removes that overlap by construction.

The September 2026 assay-aware preprint is a useful signal that assay context may help for some techniques, with author-reported gains under paper-level splits and no improvement for ITC. It is unreplicated and does not test generalisation to new targets or chemistry. It should shape what a pilot measures, not what a pilot assumes.

The recommended decision path is staged: freeze and checksum inputs, run the join and overlap audit, compare random, paper-level, scaffold and time splits on a simple baseline, review a sample of assay context with an assay scientist, confirm licensing, and only then decide whether broader curation is worth funding. All audit counts for a team's own data remain to be measured.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.bindingdb.org/rwd/bind/chemsearch/marvin/Download.jsp
2. https://academic.oup.com/nar/article/53/D1/D1633/7906836
3. https://arxiv.org/html/2609.34001v1
4. https://arxiv.org/abs/2609.34001
5. https://pmc.ncbi.nlm.nih.gov/articles/PMC11150475
6. https://pubmed.ncbi.nlm.nih.gov/38394344
7. https://pmc.ncbi.nlm.nih.gov/articles/PMC10934815
8. https://www.bindingdb.org/rwd/bind/chemsearch/marvin/BindingDB-TSV-Format.pdf
9. https://www.bindingdb.org/rwd/bind/info.jsp
10. https://www.ncbi.nlm.nih.gov/books/NBK92007
11. https://pmc.ncbi.nlm.nih.gov/articles/PMC11701568
12. https://pmc.ncbi.nlm.nih.gov/articles/PMC3628986
13. https://link.springer.com/article/10.1007/s10822-015-9860-5
14. https://www.bindingdb.org/rwd/bind/downloads/rnotes/BDB_rn2026-10-05.txt

15. https://www.bindingdb.org/rwd/bind/downloads/rnotes/BDB_rn2026-08-22.txt
16. https://www.bindingdb.org/rwd/bind/downloads/rnotes/BDB_rn2026-02-05.txt
17. https://www.bindingdb.org/rwd/bind/chemsearch/marvin/BindingDB-SDfile-Specification.pdf
18. https://arxiv.org/html/2409.04313v1
19. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-022-00672-x
20. https://tdcommons.ai/multi_pred_tasks/dti
21. https://chembl.gitbook.io/chembl-interface-documentation/frequently-asked-questions/chembl-data-questions
22. https://arxiv.org/html/2310.06399
23. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-023-00787-9
24. https://arxiv.org/html/2308.09639
25. https://pmc.ncbi.nlm.nih.gov/articles/PMC4364066
26. https://scikit-learn.org/stable/modules/cross_validation.html
27. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html
28. https://www.nature.com/articles/s41467-025-58606-8
29. https://pmc.ncbi.nlm.nih.gov/articles/PMC12552120
30. https://pmc.ncbi.nlm.nih.gov/articles/PMC9824924
31. https://pubmed.ncbi.nlm.nih.gov/22643060
32. https://pmc.ncbi.nlm.nih.gov/articles/PMC10767899
33. https://www.fda.gov/media/184830/download
34. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf
35. https://www.ema.europa.eu/en/news/ema-fda-set-common-principles-ai-medicine-development-0
36. https://www.fda.gov/media/189581/download
37. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
38. https://arxiv.org/abs/1803.09010
39. https://www.nature.com/articles/sdata201618
40. https://library.ucsd.edu/dc/object/bb1479731z
41. https://www.ebi.ac.uk/chembl/about
42. https://creativecommons.org/licenses/by-sa/4.0

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.