# IC50, EC50, Ki, and Kd Units for Machine Learning Datasets

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · ic50 · ec50 · ki · kd · cheng-prusoff · pic50 · chembl · bioactivity data · qsar · drug discovery machine learning


## Executive Summary

Machine learning datasets for drug discovery, built from public repositories such as **ChEMBL**, **BindingDB**, **PubChem BioAssay**, **PDBbind**, **DAVIS**, and **KIBA**, routinely pool four related but non-interchangeable pharmacology measurements: **IC50** (half maximal inhibitory concentration), **EC50** (half maximal effective concentration), **Ki** (inhibition constant), and **Kd** (dissociation constant). IC50 and EC50 are functional, assay-condition-dependent concentration-response endpoints, while Ki and Kd describe underlying equilibrium binding constants (www.ncbi.nlm.nih.gov). The standard bridge between them, the Cheng-Prusoff equation, converts a measured IC50 into an estimated Ki, but only holds for reversible, competitive, single-site binding at true equilibrium with a known substrate concentration relative to the Michaelis constant (Km); the choice of substrate concentration itself measurably changes the observed IC50 (cdnsciencepub.com), a relationship first formalized by Cheng and Prusoff in a 1973 paper on **"the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction"** (www.semanticscholar.org).

Because concentration-based endpoints span many orders of magnitude, the field commonly uses a negative-log molar scale—pIC50, pKi, pEC50, and pKd. ChEMBL's **pChEMBL** field places specified response, potency, and affinity measures on that scale; a 1 nM IC50 corresponds to a pChEMBL value of 9 (chembl.gitbook.io). As of its ChEMBL_37 release (29 May 2026), ChEMBL contains 24,527,044 bioactivity records across 2,921,148 compounds (ftp.ebi.ac.uk), while BindingDB separately holds roughly 3.24 million binding measurements for 11,509 proteins (www.bindingdb.org), and PubChem BioAssay preserves submitter-reported active/inactive calls as a primary archive rather than a normalized value (pubchem.ncbi.nlm.nih.gov). Kinase benchmarks diverge sharply in strategy: DAVIS reports a single measurement type across 72 inhibitors and 442 kinases (www.nature.com), while KIBA deliberately fuses IC50, Ki, and Kd into one composite score across 246,088 values (www.ebi.ac.uk).

The comparability problem is not theoretical. Bioactivity labels can arrive right-censored (reported only as "greater than" a tested concentration); a 2025 methods paper reports that approximately one-third or more of experimental labels are censored in real pharmaceutical settings (www.sciencedirect.com). Cell-based versus biochemical assay formats can also diverge because of membrane permeability, efflux, and off-target cellular effects that a purified system cannot exhibit (www.aatbio.com). ML-specific benchmarks such as **MoleculeACE** address the resulting risk of spurious or masked activity cliffs, structurally similar compounds with disproportionate potency differences (pure.tue.nl), while the **Polaris Hub** consortium and the MoleculeNet benchmark's recommendation to use scaffold rather than random data splits (arxiv.org) both push toward standardized, documented curation rather than ad hoc pooling (polarishub.io).

This report concludes that a defensible bioactivity data pipeline for machine learning must normalize units explicitly, treat censored values as inequality constraints rather than discarding or mis-averaging them, match assay metadata before merging duplicate measurements, distinguish measured Ki from Cheng-Prusoff-derived Ki, and evaluate models against the field's own documented measurement noise rather than an assumption of noise-free labels. Life-sciences data and AI consultancies such as **IntuitionLabs**, an adjacent advisor rather than a bioactivity-database vendor, frame this class of heterogeneous-data engineering, preserving provenance, governance, and quality through pipeline transformations, as a core capability for pharmaceutical organizations building AI and analytics infrastructure (intuitionlabs.ai).

## Introduction and Background

Machine learning models for drug discovery are trained on bioactivity numbers pulled from public repositories such as **ChEMBL**, **BindingDB**, **PubChem BioAssay**, **PDBbind**, and kinase-specific benchmarks like **DAVIS** and **KIBA**. Those numbers are reported under four different labels, half maximal inhibitory concentration (**IC50**), half maximal effective concentration (**EC50**), inhibition constant (**Ki**), and dissociation constant (**Kd**), and it is common practice to pool them into a single training column. That pooling is a source of quiet, hard-to-diagnose error: these four quantities measure different biological events (inhibition versus activation versus direct binding), rest on different mathematical assumptions, and depend on experimental conditions, most importantly substrate or ligand concentration, that a downstream modeler rarely sees.

This report is a technical reference for practitioners who curate bioactivity datasets for **quantitative structure-activity relationship (QSAR)** modeling, molecular deep learning, or drug-target-interaction (DTI) prediction. It defines each quantity precisely, using the U.S. National Institutes of Health (NIH) Assay Guidance Manual and peer-reviewed pharmacology literature as of September 2026 (www.ncbi.nlm.nih.gov); walks through the unit conventions used to make them comparable, molar concentration and its negative-log transform (**pIC50**, **pKi**,

**pEC50**, **pKd**); explains the **Cheng-Prusoff equation**, the standard method for converting a measured IC50 into an estimated Ki, with the assumptions under which that conversion holds; and surveys how the major public databases and machine learning benchmarks actually handle the mixing problem. It closes with a set of concrete, sourced data-cleaning rules: unit normalization, handling of censored ("greater than"/"less than") values, duplicate-measurement resolution, activity-cliff sensitivity, and data-leakage risk across train/test splits.

Life-sciences and AI consultancies such as **IntuitionLabs**, an adjacent advisor to pharmaceutical and biotechnology organizations rather than a bioactivity-database vendor, describe building the data-engineering layer, pipelines, quality checks, and integration work, that sits underneath exactly this kind of heterogeneous scientific data before it reaches an analytics or machine learning team (intuitionlabs.ai). The remainder of this report focuses on the scientific and statistical substance of that curation problem: what each measurement is, what a database actually stores versus what an assay actually measured, and where the defensible limits of "comparability" lie.

# Defining IC50, EC50, Ki, and Kd: A Taxonomy of Bioactivity Measurements

## IC50: Half Maximal Inhibitory Concentration

**IC50** is a functional, empirical potency measure: the concentration of a compound that reduces a measured signal, enzyme activity, receptor binding, or cell viability, to half of its uninhibited value. The NIH Assay Guidance Manual defines the "absolute" IC50 used in radioligand-binding assays as **"the molar concentration of a substance that reduces the specific binding of a radioligand to 50% of the maximum specific binding"** (www.ncbi.nlm.nih.gov). Critically, IC50 is not a fixed physical constant of the compound-target pair: a 2022 peer-reviewed enzymology paper defines it explicitly as **"the concentration of inhibitor that causes 50% inhibition under certain experimental conditions"** (pmc.ncbi.nlm.nih.gov), meaning the number is tied to the specific substrate, its concentration, and the assay format used to generate it.

## EC50: Half Maximal Effective Concentration

**EC50** is an assay-specific concentration-response endpoint: the concentration producing a response halfway between the assay baseline and maximum for the specified endpoint. It is not inherently an activation measure. In whole-cell assays, the Assay Guidance Manual says the result-type label should reflect the perceived pharmacology regardless of whether the raw signal increases or decreases with test-substance concentration (www.ncbi.nlm.nih.gov). IC50 and EC50 therefore should not be treated as interchangeable labels without the endpoint and assay context.

## Ki: The Inhibition Constant

**Ki** is a binding-affinity constant, in principle a property of the compound-target interaction rather than of the specific assay's substrate concentration. Ki may be experimentally determined or, in an appropriate competitive assay with the needed inputs, estimated from IC50 using the Cheng-Prusoff relationship (FDA guidance). The Assay Guidance Manual notes that **"Ki carries the same prefix as the IC50 from which it is derived"** (www.ncbi.nlm.nih.gov); preserve measurement provenance where available.

### Kd: The Dissociation Constant

**Kd** is the equilibrium dissociation constant of a direct binding interaction, most often measured for the radiolabeled tracer ligand itself in a binding assay, or for a test compound in a direct biophysical binding experiment (surface plasmon resonance, isothermal titration calorimetry, or a direct radioligand-saturation assay). A peer-reviewed *British Journal of Pharmacology* review defines it as **"Kd, which is a measure of the tendency of the receptor-ligand complex to dissociate"** (pmc.ncbi.nlm.nih.gov). Kd is generally the least assay-condition-dependent of the four quantities because it is measured at equilibrium under conditions designed specifically to isolate the binding event, but it still depends on temperature, buffer, and whether the measurement is made on a purified protein or a cell surface receptor.

*Table 1* below summarizes the four quantities side by side, including the biological event each one measures and its typical assay dependency.

| Quantity | What it measures | Typical assay | Condition-dependence | Typical reporting units |
|---|---|---|---|---|
| **IC50** | Concentration producing 50% inhibition of a functional or binding signal (defined above) | Enzyme inhibition, competition binding, cell viability | High: scales with substrate/ligand concentration relative to Km via Cheng-Prusoff (cdnsciencepub.com) | M, then ~mM to nM |
| **EC50** | Concentration producing a half-maximal response for the specified endpoint | Functional concentration-response assays | High: depends on endpoint, cell system, and readout (www.ncbi.nlm.nih.gov) | M, then ~mM to nM |
| **Ki** | Equilibrium inhibition constant, in principle substrate-independent for competitive inhibitors (defined above) | Competition binding (often converted from IC50) | Lower in principle, but inherits IC50 error when derived | M, then ~mM to nM |
| **Kd** | Direct equilibrium dissociation constant of a binding complex (defined above) | Direct binding (SPR, ITC, radioligand saturation) | Lowest: designed to isolate the binding event at equilibrium | M, then ~mM to nM |

*Table 1* shows why these four labels cannot be treated as one column of "potency" without qualification: two of them (IC50, EC50) are read directly off a dose-response curve and are inherently tied to the exact assay that produced them, while the other two (Ki, Kd) describe an underlying equilibrium constant that is only assay-independent under specific mechanistic assumptions, which the next two sections examine in detail.

## Units, Logarithms, and the pX Convention

These concentration-valued measurements are commonly expressed in molar units (M), with practical values often spanning millimolar to picomolar. A negative base-10 logarithm places concentration-based endpoints on a common numerical scale: **pIC50 = -log10(IC50 in M)**, and analogously **pKi**, **pEC50**, and **pKd**. The transform does not make IC50 or EC50 equilibrium binding constants; affinity/free-energy interpretation belongs to Ki and Kd when their equilibrium assumptions are satisfied.

**ChEMBL**, the most widely used public bioactivity database for machine learning, formalizes this as a dedicated field. Its documentation states: **"pChEMBL is defined as: -Log(molar IC50, XC50, EC50, AC50, Ki, Kd or Potency)"** (chembl.gitbook.io), giving the worked example that **"an IC50 measurement of 1nM would have a pChEMBL value of 9"** (chembl.gitbook.io). The single pChEMBL/pX value is the closest thing the field has to a unified target label across measurement types, and it is why most ML papers regress on pIC50 or pKi rather than raw IC50 or Ki.

There are two independent reasons the field prefers the pX transform over raw concentration for machine learning:

- **Statistical**: raw IC50/Ki/Kd values are extremely right-skewed (a handful of very weak binders can dominate a linear scale), while pX values are closer to normally distributed, which is friendlier to standard regression loss functions and error metrics.
- **Physical**: binding affinity is linearly related to the standard free energy of binding, not to concentration itself. A 2026 methods paper deriving the classical drug-receptor occupancy relationship shows that **"each additional 1.4 kcal mol -1 of binding free energy therefore corresponds to a roughly ten-fold gain in affinity"** at physiological temperature (www.ncbi.nlm.nih.gov), meaning a one-unit change in pKd corresponds to an equal, fixed increment of binding energy everywhere on the scale, while the same one-unit change in raw Kd represents wildly different energy increments depending on where on the potency range it occurs.

Unit normalization is also a practical curation problem, not just a statistical nicety. ChEMBL's curation team reports having **"enabled the conversion of activity values recorded in the literature with 133 different concentration units to consistent nM values"** (www.ncbi.nlm.nih.gov), a figure that illustrates how much of the published literature reports IC50/EC50 in inconsistent or non-standard units (ug/mL, percent, or ambiguous abbreviations) before any curation happens. The same curators flag likely unit-entry errors using a simple heuristic: records **"whose activity values differ by exactly 3 or 6 orders of magnitude, thus indicating a likely error in the units"** (e.g., uM mistakenly entered as nM, exactly a 1,000-fold, or three-order-of-magnitude, gap) (www.ncbi.nlm.nih.gov). Any practitioner assembling a training set from multiple literature sources should reproduce this exact check before trusting a merged value.

## The Cheng-Prusoff Equation: Converting IC50 to Ki, and When That Conversion Is Defensible

The **Cheng-Prusoff equation**, published by Yung-Chi Cheng and William Prusoff in *Biochemical Pharmacology* in 1973 in a paper on **"the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction"** (www.semanticscholar.org), is the standard bridge between an experimentally convenient IC50 and the more mechanistically meaningful Ki. In its most common competition-binding form, the *British Journal of Pharmacology* states that **"the equilibrium inhibitor constant Ki is calculated from such experiments using the Cheng-Prusoff transformation"**, expressed as Ki = IC50 / (1 + [L]/KdL), where [L] is the concentration of labeled radioligand used in the assay and KdL is that radioligand's own dissociation constant (pmc.ncbi.nlm.nih.gov).

For uncompetitive or complex (ill-defined) binding mechanisms, report an IC50 rather than representing Ki as the raw IC50; the Cheng-Prusoff equation assumes a competitive, bimolecular interaction (NIH Assay Guidance Manual). A 2022 peer-reviewed derivation makes this dependence explicit for enzyme kinetics: the apparent

inhibition constant only equals IC50 directly in the non-competitive special case; for a competitive inhibitor, IC50 instead scales upward with the ratio of substrate concentration to the Michaelis constant (Km). A 2025 methods paper states this plainly: **"The choice of substrate concentration will affect the observed IC50 values as governed by the Cheng-Prusoff equations"** (cdnsciencepub.com). This is the single most important practical fact for anyone merging IC50 values across studies: two laboratories can report genuinely different IC50s for the identical compound against the identical target, with no error by either party, simply because they ran the assay at different substrate concentrations.

The Cheng-Prusoff correction rests on several assumptions that do not always hold, and its failure modes are well documented in the pharmacology literature:

- **Reversible, competitive, single-site binding.** The standard correction assumes the inhibitor competes directly and reversibly with a single substrate or radioligand binding site. It does not apply cleanly to irreversible (covalent) inhibitors, to inhibitors with multiple binding sites, or to allosteric modulators.
- **True equilibrium.** The *British Journal of Pharmacology* review warns that **"inappropriate experimental design may result in ligand depletion and non-attainment of equilibrium, distorting the calculation of Kd and KdA"** (pmc.ncbi.nlm.nih.gov), meaning a Ki calculated from a non-equilibrium IC50 curve inherits that distortion.
- **No allosteric mechanism.** The classic Cheng-Prusoff correction assumes the unlabeled competitor drives specific binding of the radioligand toward zero at high concentration. The same review notes this assumption breaks for allosteric interactions, where displacement instead **"reaches a non-zero value if the interaction is allosteric"** (pmc.ncbi.nlm.nih.gov), so applying the standard formula to an allosteric modulator's IC50 produces a Ki with no clear physical meaning.
- **Known, correctly reported substrate concentration and Km.** The correction requires an accurate [S] and Km for the specific assay; when these are missing, mis-transcribed, or measured under different conditions than the inhibition assay itself, the resulting Ki is only as reliable as those inputs.

The practical implication is that a Ki value in a public database that was back-calculated from an IC50 is not an independent, assay-agnostic measurement; it is a derived value that carries forward the assumptions, and any errors, in the substrate concentration, Km, and inhibition mechanism used in the original conversion. Reviewing whether the underlying inhibition mechanism was actually competitive before trusting a converted Ki is a defensible, reproducible check, not an optional refinement.

## Why Assay Conditions Break Cross-Study Comparability

Even setting aside the Cheng-Prusoff correction, IC50 and EC50 values for the same compound-target pair are measurably noisy across independent measurements. A 2013 statistical analysis of pooled ChEMBL and literature IC50 data found that **"68.2% of all IC50 measurements agree within a factor of 4.8"** between independent sources for the same compound-target pair (pmc.ncbi.nlm.nih.gov), meaning roughly one-third of matched pairs disagree by more than a factor of ~5 even before any assay-format difference is considered. The same paper ties part of that scatter directly back to Cheng-Prusoff, noting that a Ki-to-IC50 conversion factor of 2 corresponds to the specific, and not universally used, condition where **"substrate concentration is equal to the Km value"** (pmc.ncbi.nlm.nih.gov).

An inter-laboratory study spanning pharmaceutical and contract-research labs measuring P-glycoprotein inhibitor IC50s for identical compounds found spreads between the lowest and highest reported value that **"ranged from a minimum of 20- and 24-fold"** for the least variable compounds **"to a maximum of 407- and 796-fold"** for the most variable ones (pmc.ncbi.nlm.nih.gov), and the same study attributed most of that scatter to genuine interlaboratory variability rather than a systematic difference between test systems. A separate line of evidence comes from comparing assay formats directly: a peer-reviewed comparison of cell-based and cell-free (biochemical) binding measurements reports that **"in-cell Kd values can differ by up to 20-fold"** from cell-free values for the same interaction, even after known confounders are controlled for (pmc.ncbi.nlm.nih.gov), because intracellular crowding, pH, ionic strength, and viscosity differ from the simplified buffers used in plate-based biochemical assays. A widely referenced pharmacology methods FAQ attributes cell-based versus biochemical IC50/EC50 divergence specifically to a test compound being **"unable to penetrate the cell membrane or the cell may be pumping the"** compound back out, alongside off-target cellular effects that a purified biochemical system cannot exhibit (www.aatbio.com).

Taken together, this evidence supports a specific, falsifiable claim: **IC50 comparability requires matched assay conditions**, same assay format (cell-based versus biochemical), same substrate or ligand concentration relative to Km, and ideally the same laboratory or protocol, not merely the same compound and target. A 2024 curation methods paper quantifies how much this matters for machine learning specifically: with minimal curation, **"almost 65% of the points differ by more than 0.3 log units"** when IC50/Ki measurements are pooled across ChEMBL assay records for the same compound-target pair, a fraction that falls only when curation matches assay type, organism, category, and confidence score (pmc.ncbi.nlm.nih.gov). What a database stores is a number and a set of metadata fields; what an assay measured is a specific, conditional event. Collapsing the two is the single most common source of unexplained label noise in bioactivity ML datasets.

## How Major Bioactivity Databases and ML Benchmarks Handle Mixed Units

Public bioactivity resources differ sharply in how much normalization they perform before a modeler ever sees the data, and it matters for a practitioner which resource, and which release, they are using.

**ChEMBL**, maintained by the European Bioinformatics Institute (EMBL-EBI), is the most heavily curated general-purpose bioactivity database used in ML pipelines. As of release ChEMBL_37 (29 May 2026), the database records 24,527,044 bioactivity records ("activities") covering 2,921,148 distinct compounds and 18,552 targets (ftp.ebi.ac.uk), figures independently confirmed by ChEMBL's live API status endpoint (www.ebi.ac.uk). ChEMBL's schema standardizes reported activity labels into a fixed vocabulary, described as a **"standardised version of the published_activity_type"** so that "Ic-50", "IC50", and "ic-50" are all mapped to one canonical type (ftp.ebi.ac.uk), and it records censored values through a dedicated relation field described as the **"symbol constraining the activity value (e.g. >, <, =)"** (www.ebi.ac.uk). Rather than silently discarding questionable measurements, ChEMBL flags them with a controlled data-quality vocabulary that includes categories such as **"potential author error"** (ftp.ebi.ac.uk).

**BindingDB** takes a complementary approach, focused specifically on protein-ligand binding data. As of its current status page, BindingDB **"currently contains about 3,241,782 binding data for 11,509 proteins and over 1,440,011 drug-like molecules"** (www.bindingdb.org), and it explicitly imports **"selected PubChem confirmatory**

**BioAssays, and ChEMBL entries for which a well defined protein target"** is identifiable (www.bindingdb.org), alongside its own literature and patent curation, with recent growth attributed chiefly to curating US patent literature.

**PubChem BioAssay**, by contrast, is deliberately a primary archive rather than a normalized database: its own documentation states **"as a primary archive, the provenance of all records belongs to the submitter"** (pubchem.ncbi.nlm.nih.gov), and its core curated field is a binary expert call rather than a normalized potency value, since **"each test result must provide an expert opinion on the overall activity of the substance"**, typically active or inactive (pubchem.ncbi.nlm.nih.gov). This is an important distinction for ML practitioners: PubChem preserves what a specific screen reported, while ChEMBL and BindingDB perform additional standardization on top of the same underlying literature.

**PDBbind** takes yet another approach, linking experimentally measured binding affinities (Ki, Kd, or IC50, deliberately mixed) directly to solved three-dimensional protein-ligand structures from the Protein Data Bank. The current PDBbind version is 2025 and reports 35,924 binding data entries, including 29,001 protein-ligand complexes (PDBbind+). Because PDBbind mixes measurement types by design (its structure-affinity pairs come from whatever the original crystallography paper reported), it is commonly used for structure-based scoring-function benchmarks with the explicit caveat that its affinity labels are not measured on a single common assay.

Kinase-focused ML benchmarks handle the mixing problem two different ways. **DAVIS**, from Davis et al. (2011, *Nature Biotechnology*), reports a single measurement type throughout: the paper describes testing **"the interaction of 72 kinase inhibitors with 442 kinases covering >80% of the human catalytic protein kinome"** (www.nature.com), using dissociation-constant-style Kd values throughout, avoiding the mixed-type problem by construction. **KIBA**, from Tang et al. (2014, *Journal of Chemical Information and Modeling*), takes the opposite strategy and deliberately fuses heterogeneous measurement types: the authors describe **"a model-based integration approach, termed KIBA"** (www.ebi.ac.uk) that combines Ki, Kd, and IC50 measurements into a single unified score, producing **"52,498 chemical compounds and 467 kinase targets, including a total of 246,088 KIBA scores"** (www.ebi.ac.uk). Any model trained on the KIBA score is therefore learning a statistically fused proxy, not any single physical quantity, a distinction that is frequently lost when KIBA is described simply as a "bioactivity" benchmark.

Two more recent, ML-specific efforts address the mixing problem at the dataset-construction stage rather than leaving it to individual modelers. **Papyrus** (Béquignon et al.) is a large-scale aggregation explicitly built for ML use, drawing primarily on ChEMBL and the ExCAPE-DB dataset: it **"is comprised of around 60 million"** data points (pmc.ncbi.nlm.nih.gov) aggregated down to **"1,270,570 unique two-dimensional compound structures and 6926 proteins"** after processing. Papyrus deliberately restricts its normalized potency column to Ki, Kd, IC50, and EC50 values, whose **"logarithm transforms were considered if expressed in molar concentrations"** (pmc.ncbi.nlm.nih.gov), explicitly excluding less comparable endpoints rather than silently pooling everything. **Polaris Hub**, built through a cross-industry consortium, positions itself as **"the benchmarking platform for drug discovery"** (polarishub.io) and publishes **"guidelines for dataset curation and method evaluation & comparison"** (polarishub.io) intended to standardize exactly the kind of assay-metadata matching this report describes, rather than leaving each research group to invent its own rules.

*Table 2* below summarizes how these seven resources differ in scope, normalization philosophy, and documented scale.

| Resource | Primary purpose | Documented scale (as of source date) | Handling of mixed measurement types |
|---|---|---|---|
| **ChEMBL** | General bioactivity database | 24,527,044 activities, 2,921,148 compounds (ChEMBL_37, 29 May 2026) (ftp.ebi.ac.uk) | Standardizes activity type, relation, and units; computes unified pChEMBL only for defined types |
| **BindingDB** | Protein-ligand binding data | ~3.24 million binding data, 11,509 proteins (www.bindingdb.org) | Imports ChEMBL/PubChem data plus own literature and patent curation; retains original measurement type per record |
| **PubChem BioAssay** | Primary screening archive | Millions of deposited assay records (submitter-defined) | Preserves submitter's own active/inactive call; minimal cross-record normalization (cited above) |
| **PDBbind** | Structure-affinity pairs | 35,924 binding data entries, including 29,001 protein-ligand complexes (v2025) | Deliberately mixes Ki/Kd/IC50 tied to solved 3D structures |
| **DAVIS** | Kinase selectivity benchmark | 72 inhibitors x 442 kinases (~80% of catalytic kinome) (www.nature.com) | Single measurement type (Kd-style) throughout; avoids mixing by design |
| **KIBA** | Kinase DTI benchmark | 246,088 KIBA scores, 52,498 compounds, 467 targets | Statistically fuses Ki, Kd, and IC50 into one composite score |
| **Papyrus** | ML-ready bioactivity aggregation | ~60 million points aggregated to 1,270,570 compounds, 6,926 proteins | Restricts normalized column to Ki/Kd/IC50/EC50 on a common molar scale |

The pattern across *Table 2* is that databases built primarily as archives (PubChem) preserve heterogeneity and leave normalization to the user, databases built as curated hubs (ChEMBL, BindingDB) perform partial standardization but still require assay-metadata matching for reliable comparison, and resources purpose-built for ML (Papyrus, KIBA, Polaris) make an explicit, documented choice about how mixing is handled, either by restricting scope or by openly fusing measurement types into a composite score.

## Practical Data-Cleaning Rules for Machine Learning Practitioners

Based on the documented curation practices above, a defensible bioactivity data-cleaning pipeline for QSAR or DTI modeling should apply the following steps, each traceable to a specific, sourced practice:

- **Normalize units to a single molar-derived scale before merging any two sources.** Convert every value to nM (or directly to pX) rather than trusting the units column as reported, and re-derive pX values rather than trusting a pre-computed one if the source units are ambiguous (www.ncbi.nlm.nih.gov).
- **Screen for exact-order-of-magnitude outliers.** Flag any duplicate compound-target measurement pair whose values differ by exactly 1,000-fold (three orders of magnitude) or 1,000,000-fold (six orders of magnitude), the signature of a uM/nM or mM/uM unit transcription error, rather than a genuine biological discrepancy (www.ncbi.nlm.nih.gov).
- **Treat censored ("greater than"/"less than") values as inequality constraints, not point estimates.** A 2025 methods paper reports that **"approximately one-third or more of experimental labels are censored"** in real pharmaceutical settings (www.sciencedirect.com). Standard curation practice in many

published discovery pipelines is simply to exclude such qualified records from training sets entirely; more sophisticated pipelines instead model them explicitly as one-sided constraints so the information they carry (a compound is inactive at least up to some tested concentration) is not simply thrown away.

- **Match assay metadata before averaging duplicate measurements.** A 2024 curation methods paper shows that requiring matched assay_type, assay_organism, assay_category, and a **ChEMBL confidence score of 9** (direct single-protein-target assignment) meaningfully reduces cross-source noise, since **"any assay that does not have a confidence score value of 9... is removed"** in its recommended pipeline (pmc.ncbi.nlm.nih.gov), and that this stringent matching substantially reduces the roughly 65% cross-source disagreement rate noted above.
- **Keep, but flag, converted Ki values separately from directly measured ones.** Since a Ki calculated via Cheng-Prusoff inherits the substrate concentration, Km, and mechanism assumptions of the original IC50 assay, a defensible pipeline distinguishes "measured Ki" from "converted Ki" rather than treating the pooled column as homogeneous.
- **Check for activity-cliff sensitivity before splitting data.** The **MoleculeACE** benchmark defines **"activity cliffs, pairs of molecules that are highly similar in their structure but exhibit large differences in potency"** (pure.tue.nl), operationalized as a structural **"similarity larger than 90%"** paired with at least a 10-fold potency difference (pure.tue.nl), and its curation pipeline removes a molecule outright when replicate measurements disagree by more than one log unit, since **"the corresponding molecule was removed"** in that case (pure.tue.nl); the benchmark's authors caution that public potency data used this way is itself **"affected by undetectable experimental noise"** (pure.tue.nl), so mixing measurement types without normalization can manufacture false activity cliffs (a Ki-vs-IC50 artifact masquerading as a genuine structure-activity discontinuity) or, just as damaging, mask a genuine one. A related peer-reviewed definition describes activity cliffs as **"the pinnacle of structure-activity relationship (SAR) discontinuity"** (pmc.ncbi.nlm.nih.gov), detrimental to quantitative SAR predictions and exactly the pattern a poorly curated, mixed-unit dataset can spuriously create or destroy.
- **Use scaffold or time-based splits, not random splits, for evaluation.** MoleculeNet states plainly that **"random splitting of molecular data isn't always best for"** evaluating ML methods (arxiv.org), recommending scaffold splitting, which **"attempts to separate structurally different molecules into different subsets"** (arxiv.org) so near-duplicate structures do not leak potency information across the split and inflate apparent performance.

## Data Analysis and Evidence

The quantitative record on cross-assay and cross-laboratory variability is unusually well documented for a data-quality issue, because pharmacology and cheminformatics researchers have specifically studied it as a modeling confound. *Table 3* collects the key figures cited throughout this report in one place.

| Comparison | Documented figure | Source |
|---|---|---|
| Agreement of independently reported IC50s for the same compound-target pair | 68.2% agree within a factor of 4.8 | Kalliokoski et al., PLOS ONE 2013 (cited above) |
| Inter-laboratory spread for P-glycoprotein inhibitor IC50s (least to most variable compounds) | 20 to 24-fold at minimum, up to 407 to 796-fold at maximum | AAPS/DMD inter-laboratory study 2013 (cited above) |

| Comparison | Documented figure | Source |
|---|---|---|
| In-cell versus cell-free binding affinity divergence | Up to 20-fold | Peer-reviewed cellular-binding comparison (cited above) |
| Matched-pair IC50/Ki measurements differing by more than 0.3 log units under minimal ChEMBL curation | Almost 65% | 2024 J. Chem. Inf. Model. curation study (cited above) |
| Share of experimental bioactivity labels reported as censored (">"/"<") rather than exact | Approximately one-third or more in real pharmaceutical settings | 2025 machine learning methods paper (cited above) |
| ChEMBL_37 scale (as of 29 May 2026) | 24,527,044 activities, 2,921,148 compounds, 18,552 targets | ChEMBL release notes (cited above) |
| BindingDB scale (current status page) | ~3.24 million binding data, 11,509 proteins, 1.44 million compounds | BindingDB info page (cited above) |
| Papyrus aggregated scale | ~60 million raw points, 1,270,570 compounds, 6,926 proteins | Béquignon et al. 2023 (cited above) |

*Table 3* makes a specific point that is easy to lose in qualitative discussion: pooling IC50 values across sources without matching assay conditions can introduce substantial measurement variation, from roughly a factor of 4 to 5 at the median to several hundred-fold in documented worst cases for certain transporter targets. A regression model reporting a mean absolute error of 0.3 to 0.5 log units on pooled, poorly curated bioactivity data may simply be reproducing measurement noise rather than learning structure-activity relationships, since the underlying labels themselves disagree by a similar or larger margin. This is also why activity-cliff benchmarks such as MoleculeACE are structured the way they are: a model that appears to fail on a "true" activity cliff may instead be correctly reproducing the fact that the two potency values being compared were never measured under comparable conditions in the first place.

## Implications and Future Directions

The practical direction of travel in the field is toward explicit, machine-readable assay metadata rather than a single pooled "activity" column. ChEMBL's confidence-score and data-validity-comment fields, BindingDB's provenance tagging, and Papyrus's restriction to a defined set of comparable endpoint types are all instances of the same underlying response: making the conditions under which a number was produced a first-class part of the dataset, not an afterthought buried in a linked publication. Benchmark efforts such as Polaris Hub extend this further by publishing shared curation guidelines intended to apply consistently across research groups rather than leaving each modeling team to rediscover the same pitfalls independently.

For organizations building bioactivity data infrastructure on top of these public resources, or integrating them with proprietary assay data from a lab notebook or LIMS, the discipline required is closer to enterprise data integration than to model architecture research: matching schemas, propagating censoring and confidence flags rather than discarding them, and preserving assay-condition metadata through every transformation. Life-sciences data and integration specialists, including consultancies such as IntuitionLabs, describe this class of problem, connecting heterogeneous scientific data while preserving integrity and governance, as a core service line for pharmaceutical organizations building AI infrastructure (intuitionlabs.ai): the hardest part of bioactivity machine learning is often not the model, but knowing what each training-set number actually represents.

Looking forward, the datasets and benchmarks that explicitly document their handling of mixed measurement types, rather than silently pooling them, are likely to remain the more defensible foundation for new model development, and the growing use of censored-data-aware statistical methods (treating ">"/"<" values as inequality constraints rather than discarding them) suggests the field is slowly moving away from simply excluding censored labels.

## Frequently Asked Questions (FAQs)

**What is the difference between IC50 and EC50?** IC50 is a concentration associated with a defined inhibitory response, whereas EC50 is the concentration associated with a half-maximal response for the specified assay endpoint. Either label must be interpreted with its assay definition, direction of response, and conditions; numerical similarity alone does not establish comparability (www.ncbi.nlm.nih.gov).

**How is IC50 different from Ki and Kd?** IC50 and EC50 are read directly off a dose-response curve and are tied to the specific assay conditions that produced them. Ki and Kd describe underlying equilibrium constants: Kd is typically measured directly, while Ki may be experimentally determined or, in an appropriate competitive assay with the needed inputs, estimated from IC50 using Cheng-Prusoff (FDA guidance). Preserve whether a Ki was measured or estimated.

**How do you convert IC50 to Ki?** Using the Cheng-Prusoff equation, Ki = IC50 / (1 + [L]/KdL) for competition binding. The conversion is only valid for reversible, competitive, single-site binding at true equilibrium; it does not apply cleanly to allosteric or irreversible inhibitors, per the assumptions and failure modes described above.

**Why is pIC50 used instead of raw IC50 in machine learning?** pIC50 is a negative-log representation of a concentration-response endpoint, which places values on a logarithmic numerical scale. ChEMBL uses pChEMBL to compare roughly comparable half-maximal response, potency, and affinity measures on that scale; it does not make every pX value a binding-free-energy measure (chembl.gitbook.io).

**Why are IC50 values not comparable across different assays or laboratories?** Because IC50 depends on assay-specific conditions, principally substrate or ligand concentration relative to Km, plus assay format (cell-based versus biochemical). Documented studies show spreads as large as several-hundred-fold for the same compound-target pair in the most variable cases, and even typically, roughly a third of independently reported IC50 pairs disagree by more than a factor of ~5, per the data analysis section above.

**How should nanomolar and micromolar values be handled when merging datasets?** Convert every value to a single consistent unit (nM is the field's de facto convention) before any comparison or merge, and specifically check for values that differ by exactly 1,000-fold or 1,000,000-fold between duplicate records, the signature of a uM/nM unit transcription error rather than a genuine biological difference, per the unit-normalization practice described above.

## Conclusion

IC50, EC50, Ki, and Kd are four related but distinct quantities: IC50 and EC50 are functional, assay-condition-dependent concentration-response endpoints whose direction and meaning depend on the assay definition, while Ki and Kd describe equilibrium binding constants under their relevant mechanistic assumptions. The Cheng-Prusoff equation bridges IC50 and Ki, but only within those assumptions, and fails for allosteric, irreversible, or poorly

characterized mechanisms. Documented studies show that pooling these values without matching assay conditions introduces noise on the order of a factor of five at the median and up to several hundredfold in worst-case examples.

Major public resources, ChEMBL, BindingDB, PubChem BioAssay, PDBbind, DAVIS, KIBA, Papyrus, and Polaris Hub, have each made an explicit, documented choice about how to handle this heterogeneity, ranging from preserving raw submitter data unmodified to statistically fusing multiple measurement types into one composite score. None of these choices eliminates the underlying problem; they simply make it visible and, in the better-documented cases, traceable. For a machine learning practitioner, the defensible response is not to treat "bioactivity value" as a single homogeneous column, but to normalize units explicitly, preserve and reason about censoring and confidence metadata, match assay conditions before merging duplicate measurements, and evaluate models using splits and error thresholds calibrated to the field's own documented measurement noise, rather than to a noise-free label that does not, in practice, exist.

---