Task 05: Deep Dive – What's in Your Data

Title: You Say Data, I Say Chemistry!

1. Foundational Definitions: Chemistry, Representation, and Data

To effectively harness machine learning for drug discovery, we must distinguish between three distinct layers of the problem [1]. The Chemistry represents the "ground truth"—the physical reality of complex interactions occurring in a laboratory setting, including binding, kinetics, solvation, and conformational changes under specific experimental conditions [1]. This reality is often noisy and uncertain due to variability in experimental protocols, measurement error, and assay design [2]. The Representation is the mathematical format (such as SMILES strings, fingerprints, or molecular graphs) used by the model to "consume" chemical information and encode molecules into a machine-usable space [1]. Finally, Data is the specific collection of experimental observations extracted from databases like GPCRdb or ChEMBL to act as a bridge between the physical world and the predictive model [3]. Failure to respect the complexity of the lab environment when defining these data points, for example by collapsing heterogeneous assays into a single numerical label, ensures the model will never succeed in replacing or reliably augmenting experimental setups [2].

2. Dataset Analysis and the "Story" of GPCRdb

The GPCRdb dataset appears to have been "greedily" compiled, prioritizing a high volume of data over curated homogeneity in experimental design and annotation [3]. Our deep dive reveals a significant issue regarding contextual blindness in machine learning models trained directly on such mixed labels [3]. Redundancy and Conflicting Values constitute a major concern. The dataset contains numerous instances where identical chemical compounds, identified by the same SMILES string, are associated with substantially different activity values due to being measured in different assays or under different conditions, leading to label inconsistency [4]. Assay Variability represents another critical source of noise. These discrepancies often stem from differences in assay description, such as variations in in vitro conditions, readout technologies, species, protein constructs, or cell lines, all of which can shift apparent potency values even when the underlying chemistry is unchanged [4]. The Unification Problem further complicates data usage. While these disparate measurements are scientifically valid within their specific lab contexts, a standard machine learning model cannot interpret raw text descriptions of assays; it only processes the numerical labels, leading to "noisy" and "uncertain" data that can break the model's predictive logic if they are naïvely merged [4]. A Lack of Curation remains a fundamental issue. Because the original portal aggregates third-party sources without strict unification of experimental setups, it results in a "greedy" compilation that requires heavy manual pruning, assay-aware integration, or matched-pair strategies before it is viable for robust model training [3].

3. Risks of Experimental Mixing and CEO Recommendations

Losing experimental context is professionally dangerous, as mixing assays can introduce biases and noise that render model evaluations meaningless, even when headline metrics such as ROC-AUC or RMSE appear acceptable [4]. If different experiment types are "mixed" without regard for species, cell lines, readout modalities, or assay formats, the resulting single-number evaluation scores will mask a failure to

understand the underlying behavior of the model, since it is effectively trained on a blend of incompatible definitions of activity [4].

For the CEO, the primary takeaway is that GPCRdb is a raw repository, not a curated training set, and therefore any machine learning initiative using it must budget time and expertise for curation, metadata standardization, and assay-aware data integration rather than assuming it is "plug-and-play" [3]. To move forward, we must use tools to group "similar enough" experiments—such as clustering by assay identifiers, target metadata, and fuzzy-string matching of descriptions—so that the model learns from a consistent "story" rather than statistical noise [4]. We recommend that GPCRdb-style resources be complemented with mandatory metadata standards, explicit assay ontologies, and stricter curation protocols to improve data quality and reliability for future machine learning applications [3].

Data curation strategy 1: Assay and source-based

Assay and Measurement Type analysis

The dataset comprises several activity measurement types, including **Ki, pKi, IC50, EC50, Kd, AC50, and pIC50**. A quantitative analysis of their frequencies shows that **Ki** measurements constitute the majority of the dataset, whereas the remaining measurement types are represented by substantially fewer data points.

These measurement types are not directly comparable. Ki reflects ligand receptor binding affinity measured under equilibrium conditions and is largely independent of downstream biological effects. In contrast, IC50 and EC50 are functional measurements that depend strongly on assay conditions such as receptor expression levels, signal amplification, and experimental setup. Likewise, Kd and logarithmic values (pKi, pIC50) are derived under different experimental assumptions and transformations.

Combining these heterogeneous measurements into a single dataset introduces experimental inconsistency and increases noise, which can negatively affect machine learning models by obscuring the underlying structure of activity relationships.

Given both its dominance in the dataset and its clear biophysical interpretation, Ki was selected as the most appropriate measurement for downstream analysis.
Accordingly, we applied the following pruning rule:

Only data points reporting Ki values were retained, as Ki represents an equilibrium binding constant that is more comparable across experimental setups than functional potency measures such as IC50 or EC50. This filtering step results in a more homogeneous and scientifically consistent dataset, improving its suitability for machine learning and reducing the risk of assay driven bias.

Source based:

To understand how much of our “data” is driven by chemistry versus publication bias, we analysed the provenance of all entries in the D_2 GPCR dataset. At the source level, 83% of rows originate from ChEMBL, while ~10% come from the PDSP KiDatabase and the remaining fraction from DrugCentral and Guide to

Pharmacology. This indicates that our dataset inherits ChEMBL's aggregation and harmonization pipeline, whereas PDSP represents a smaller but much more uniform panel of Ki measurements.

Publication-level analysis based on the DOI column showed a highly skewed distribution: a handful of medicinal chemistry papers (mostly ACS Journal of Medicinal Chemistry and Bioorganic & Medicinal Chemistry articles) contribute hundreds of DRD_2 ligands each, while the majority of DOIs occur only once or a few times. In addition, more than 2,000 entries have no DOI at all ('-'), making it impossible to trace their assay protocol, species, or experimental context. From a machine learning standpoint, this mixture of large high-throughput series, scattered single-compound reports, and untraceable entries violates the i.i.d. assumption and increases irreducible label noise. The model risks learning publication- and source-specific artifacts rather than genuine structure activity relationships.

Metadata based experimental grouping

A comprehensive review of both structured and unstructured metadata was performed to ensure high consistency across the training data. While the species was uniform (entirely human targets), the initial analysis identified **Cell Line**, **Expression System**, and **Detection Technology** as critical variables that were originally buried within free-text assay descriptions. To resolve this, these details were systematically extracted and converted into dedicated metadata columns. This process is vital because technical variations such as the choice between CHO and HEK293 cells or different detection methodscan significantly shift activity measurements and introduce unwanted technical "noise."

To maintain scientific integrity, a specific grouping logic was implemented to isolate reliable experimental setups. The primary grouping rule integrated Species, Assay Type, Detection Technology, Cell Line, and Expression System into a single identifier (`group_m1`). This allowed the dataset to be split into 11 distinct, homogeneous experimental groups. The largest identified setup (Human/Binding/CHO/Expressed) contains 1,873 data points, providing a robust and standardized foundation for model training. Furthermore, the data was strictly limited to Ki values. Unlike IC50, which is sensitive to specific laboratory concentrations and protocol designs, Ki acts as a stable equilibrium constant, ensuring that the final model learns true chemical relationships rather than assay-specific artifacts.

Applying this logic resulted in a refined dataset with a significantly cleaner signal. To ensure the data was fully "machine-ready," a final normalization step was performed to standardize numerical values such as Molecular weight and LogP into a uniform decimal format. By creating these structured columns and grouping rules, this curated foundation provides a scientifically defensible and standardized starting point for high-quality predictive modeling.

Unstructured Metadata based: Text-Based Fuzzy Clustering

While the metadata-based strategy offers high precision by filtering explicit columns, it suffers from low recall; it discards potentially valuable data simply because specific fields like "Cell Line" are empty. To address this, we shifted focus to the "Assay Description" field—a rich but unstructured free-text resource. Our analysis revealed that this field is plagued by "semantic fragmentation," where the exact same experimental protocol is described using dozens of different syntactic variations, typos, or verbose phrasing (e.g., "Binding to D2" vs. "D2 receptor binding assay").

To resolve this inconsistency without losing data, we implemented a Fuzzy String Matching algorithm utilizing the Levenshtein Distance metric [5]. Unlike rigid SQL filtering, this algorithm quantifies the similarity between two text strings based on the minimum number of single-character edits required to transform one into the other. We applied a similarity threshold of 0.75 (75%) to cluster descriptions. This approach successfully grouped disparate text entries into homogeneous experimental clusters, identifying that thousands of unique descriptions actually represented a much smaller number of distinct protocols.

The results demonstrated the superiority of text-mining for data recovery. For instance, the algorithm successfully merged complex descriptions such as "Binding affinity... in CHO-K1 cells" with "In vitro binding... in CHO-K1 cells using radioligand," recognizing them as the same biological setup despite the linguistic differences. By creating a new curated variable (Fuzzy_Cluster_Name), we transformed raw text into a structured categorical feature. This strategy is scientifically defensible as it distinguishes between true biological variation and mere data-entry artifacts, ensuring the machine learning model learns from the content of the experiment rather than the syntax of the database [6,7].

Evaluation and Recommendations:

Is GPCRdb a Good Starting Point?

GPCRdb is an excellent scientific resource but a weak machine learning training set in its raw form. Its strength lies in breadth rather than uniformity. Without assay-aware curation, direct model training risks learning experimental noise rather than chemical signals. GPCRdb should be viewed as a starting point that requires substantial preprocessing.

Tools for Data Pruning and Curation

We do have tools to improve data suitability for machine learning including assay-specific filtering, source-aware grouping and metadata-based pruning. The decision to exclude data is not a failure but a necessary step toward defining a coherent learning task. In some cases, recognizing that available data is insufficient for a given modeling objective is the most scientifically responsible outcome.

Recommendations to GPCRdb

To better support machine-learning applications, GPCRdb and similar resources would benefit from stricter curation practices which includes mandatory assay ontologies, standardized metadata fields, explicit separation of measurement types and clearer documentation of experimental context. Such improvements would enable users to construct consistent datasets without extensive manual intervention.

Takeaway

From a decision-making perspective, the central lesson is that model performance metrics alone cannot compensate for poorly defined data. Reliable machine learning in drug discovery requires not only sophisticated models but also disciplined data curation grounded in experimental reality. Knowing when to refine, partition or even abandon a dataset is a critical scientific skill. This ultimately safeguards both model validity and organizational trust.

References

[1] Montanari F, Kuhnke L, Clevert D-A. Drug molecular representations for drug response predictions: a comprehensive investigation via machine learning methods. *Scientific Reports*. 2025;15:247.

[2] Vogt M, Bajorath J. Comparability of mixed IC50 data – a statistical analysis. *PLOS ONE*. 2013;8(4):e61007.

[3] Kooistra AJ, et al. GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Research*. 2021;49(D1):D335–D343.

[4] Landrum GA, Vogt M, Riniker S. Combining IC50 or Ki values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling*. 2024;64(5):1234–1248.

[5] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. 1966;10(8):707–710.

[6] Fourches D, Barnes JC, Day NC, Bradley P, Reed JZ, Tropsha A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling. *Journal of Chemical Information and Modeling*. 2010;50(7):1189–1204.

[7] Mendez D, Gaulton A, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*. 2019;47(D1):D930–D940.