Journal of Cheminformatics

BRIEF REPORT

Open Access

Matched pairs demonstrate robustness against inter-assay variability

Jochem Nelen[1,2], Horacio Pérez-Sánchez[1], Hans De Winter[3*] and Dries Van Rompaey[4]

Abstract

Machine learning models for chemistry require large datasets, often compiled by combining data from multiple assays. However, combining data without careful curation can introduce significant noise. While absolute values from different assays are rarely comparable, trends or differences between compounds are often assumed to be consistent. This study evaluates that assumption by analyzing potency differences between matched compound pairs across assays and assessing the impact of assay metadata curation on error reduction. We find that potency differences between matched pairs exhibit less variability than individual compound measurements, suggesting systematic assay differences may partially cancel out in paired data. Metadata curation further improves inter-assay agreement, albeit at the cost of dataset size. For minimally curated compound pairs, agreement within 0.3 pChEMBL units was found to be 44–46% for K_i and IC_{50} values respectively, which improved to 66–79% after curation. Similarly, the percentage of pairs with differences exceeding 1 pChEMBL unit dropped from 12 to 15% to 6–8% with extensive curation. These results establish a benchmark for expected noise in matched molecular pair data from the ChEMBL database, offering practical metrics for data quality assessment.

Keywords Matched structural pairs, Assay noise, Data curation, ChEMBL, Machine learning

Introduction

In recent work, Landrum and Riniker demonstrated that combining IC_{50} or K_i measurements from different experiments can introduce substantial noise to datasets [1]. They demonstrated this by comparing assay measurements from the ChEMBL32 database [2] across different assays for the same target. Similarly, previous studies have highlighted the challenges posed by assay variability and prediction errors in biological datasets, emphasizing their impact on model reliability and the interpretation of results [3, 4]. Landrum and Riniker did report that data curation can partially help to mitigate these effects, but that the overall amount of noise remains high. In the current era of machine learning (ML) and artificial intelligence (AI), where the quality of training data is crucial for building accurate and robust models, understanding and quantifying the noise introduced by combining data from multiple sources is essential. Noisy data can lead to reduced model performance and unreliable predictions, underscoring the importance of identifying best practices for data collection and integration. A common assumption is that within-assay comparisons between different compounds are more reliable than direct between-assay comparisons, a rationale which is also used for Matched Molecular Pair Analysis (MMPA) [5]. MMPA can be used to propose new compounds or to calculate the potential effect of a given modification to a compound. The MMPA

*Correspondence:
Hans De Winter
hans.dewinter@uantwerpen.be
[1] Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), HiTech Innovation Hub, UCAM Universidad Católica de Murcia, 30107 Murcia, Spain
[2] Health Sciences PhD Program, Universidad Católica de Murcia UCAM, 30107 Murcia, Spain
[3] Department of Pharmaceutical Sciences, Faculty of Pharmaceutical, Biomedical and Veterinary Sciences, University of Antwerp, Universiteitsplein 1A, 2610 Wilrijk, Belgium
[4] Drug Discovery Data Sciences, Janssen Pharmaceutica NV, Turnhoutseweg 30, 2340 Beerse, Belgium

paradigm has also given rise to various machine learning approaches which aim to improve upon the performance of standard MMPA [6–8]. However, the impact of assay variability on these methods remains underexplored. In this study, we expand on Landrum and Riniker's previous research by identifying and leveraging structural analogs within the ChEMBL32 database to further investigate the relative variability in bioactivity measurements across different assays. By focusing on matched molecular pairs, we aim to provide a quantified estimate of the expected noise, which is helpful to contextualize the performance of models trained on such matched pair data.

Method

We analyzed pairwise differences in K_i and IC_{50} values among structural analogs across assays to evaluate the consistency of trends in the ChEMBL database [2]. We built our data curation process upon the established workflow of Landrum and Riniker [1], which investigated assay noise by comparing affinity measurements directly across different assays. For ease of comparison, we also used ChEMBL32 as the primary database, and we performed a similar minimal and maximal curation procedure for all K_i and IC_{50} data in ChEMBL, in order to investigate if more careful curation can improve overall trends and data quality. We note that the extension of our work to new versions of ChEMBL is straightforward.

The minimal curation procedure involved a relatively lenient filtering process, where assays were considered comparable if they were measured for the same protein target and had at least five compounds in common. Additionally, activity curation was applied to filter out compounds with identical activity values. This process also removed compounds with activity values that differed by exactly 3.0 log units to account for a unit error, such as incorrect annotation of units (e.g., μM instead of nM or vice versa). The aim of activity curation was to avoid accidental duplicates in the results, as these could greatly impact the findings. By incorporating these specific criteria and filtering steps, we ensured that the curation process mostly removed true duplicates, as it is improbable for different measurements to coincidentally report the exact same affinity value.

The maximal curation (maxcur) procedure also applied the same basic curation steps as the minimal curation process (mincur), including activity curation and requiring an overlap of five or more compounds between two assays. Furthermore, several additional filtering steps were included before considering two assays to be compatible for the maxcur procedure. In short, only compounds with a high confidence score as assigned by ChEMBL were kept, as well as data that came directly from published papers. Additionally, any assays that were associated with mutant/variant proteins or had multiple assays for the same target were filtered out. Finally, assays

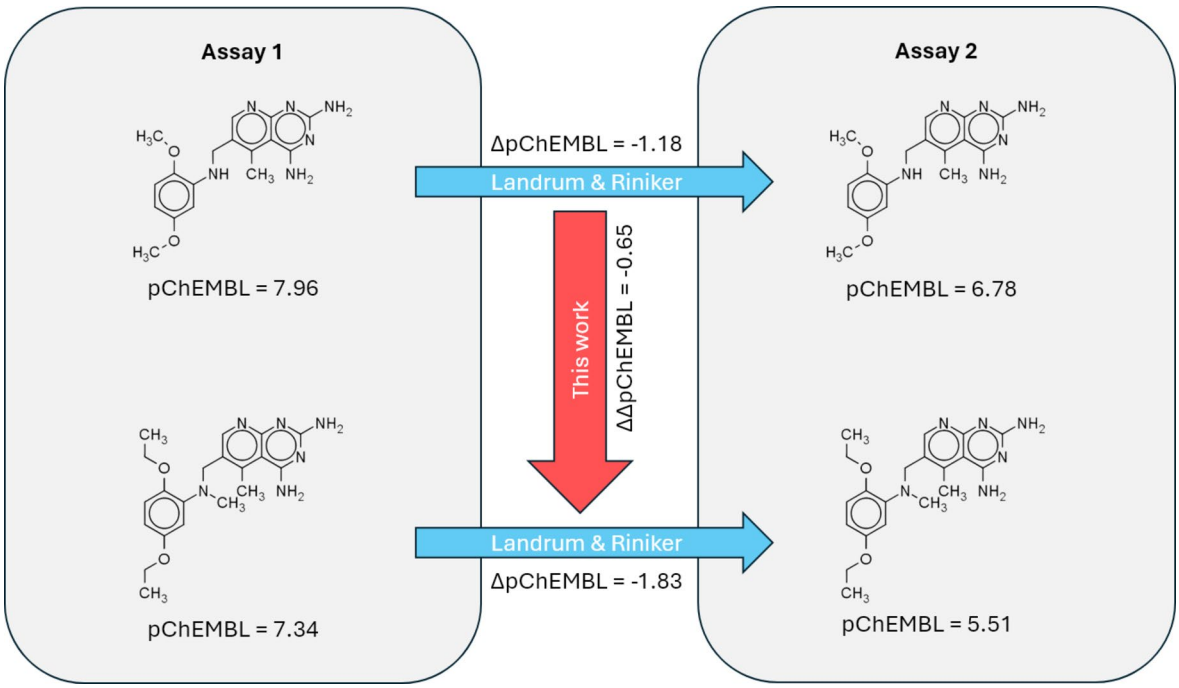

Fig. 1 Schematic Overview of ΔΔpChEMBL Calculation for Matched Molecular Pairs

Table 1 Pairwise metrics and dataset characteristics for the matched pairs of structural analogs

Curation level	MAE	F > 0.3	F > 1.0	Assay pairs	Compound pairs
IC_{50} minimal	0.33	0.54	0.12	1,338	372,683
IC_{50} maximal	0.18	0.34	0.06	24	1,773
K_i minimal unpruned	0.36	0.56	0.18	584	36,189
K_i maximal unpruned	0.40	0.62	0.43	279	2,900
K_i minimal pruned	0.35	0.55	0.15	219	32,363
K_i maximal pruned	0.03	0.21	0.08	8	311

Summary of pairwise metrics (Median Absolute Error, fractions > 0.3 and 1.0) and dataset characteristics (assay pairs and compound pairs) for all the tested datasets and curation procedures. The pruned datasets have all assays associated with Carbonic Anhydrase filtered away

were filtered based on their conditions metadata such as the type of assay (e.g. binding assay, functional assay, ...) and the target organism, and only assays that shared identical assay conditions metadata were considered compatible. For a more in-depth overview and reasoning of all of the individual curation steps, we refer to the original manuscript, where the same procedure was used [1]. We also investigated some potential "medium" curation procedures, aiming to balance the number of retained pairs with a low noise level. Starting from the maximal curation settings, we sequentially disabled individual settings which we deemed to be retain scientific validity (e.g. relaxing strict requirements on matching exact assay conditions or including lower-confidence results). However, none of the tested configurations consistently improved both IC_{50} and K_i datasets. Typically, relaxing the curation settings resulted in either minimal gains in new datapoints or a substantial increase in noise, with none of the options examined herein offering a suitable middle ground. Additional details can be found in the Supplementary Information (SI): Figure S1 presents the IC_{50} data, Figure S2 shows the K_i data, and Table S1 summarizes the performance statistics for both.

Following the data curation process, we identified structural analogs between compatible assays for the minimal and maximal curated IC_{50} and K_i data. This identification process involved two stages. Pairs within the same assay were identified using rdRascalMCES, an RDkit [9] implementation of RASCAL [10], which is an algorithm that uses bond matching to efficiently identify the Maximum Common Edge Subgraphs (MCES) between two molecules. The rdRascalMCES method was selected for its ability to directly identify the largest common scaffold and quickly terminate calculations for pairs with initial low Johnson similarity estimates, making it well-suited for analyzing large datasets. For the specific rdRascalMCES settings, we used 0.6 as the Johnson similarity threshold and allowed partial aromatic rings to match. We refer to Figure S3 in the supplementary information for concrete examples. Subsequently, we compared the identified intra-assay pairs with their compatible assays, as determined by the initial curation step. If an exact match between two pairs across assays was found, it was considered a matched pair of structural analogs. All relevant data such as the associated ligand and assay ChEMBL IDs, reported pChEMBL affinity values and protein target ID were retained for subsequent analysis. For each matched pair, we calculated the difference in pChEMBL values within the same assay, representing the relative potency difference between the two molecules. We then compared the intra-assay differences for the matched pair across the two assays, defining this difference between the intra-assay differences as the ΔΔpChEMBL. Figure 1 provides a schematic overview of the calculation of the ΔΔpChEMBL for a matched pair.

The final part of the workflow involved conducting a statistical analysis based on earlier work by Landrum and Riniker [1]. The Median Absolute Error (MAE), as well as the fractions of values exceeding 0.3 (F > 0.3) and 1.0 (F > 1.0) were computed for each dataset. In contrast to the work of Landrum and Riniker, the R^2 and Kendall Tau metrics were not included, as these correlation metrics would be sensitive to both the arbitrary assignment of the direction of the transformation (e.g. depending on the direction of the transformation, signs may be positive or negative), as well as the range spanned by the assay measurements. Additionally, we generated a histogram of the ΔΔpChEMBL values to provide a complementary overview of the overall distribution of differences. RDKit version 2024.03.1 was used for pair analytics and cheminformatics. Data processing was performed using pandas 2.2.2, and plots were generated with matplotlib 3.8.4.

For the K_i datasets, we also applied data pruning inspired by the original manuscript's findings [1], which identified that several Carbonic Anhydrase (CA) assays were responsible for a substantial part of the noise observed. However, instead of only pruning away the assays indicated in their work, we decided to remove all data associated with CA assays. This decision was made after manually inspecting the pruned data, which revealed that a cluster of large outliers associated with CA assays persisted even after excluding the assays indicated in the original paper (Figure S4). The Supplementary Information includes comparative figures (Figure S5-6) and details regarding the pruned assays for each procedure. Although both methods produce similar trends and figures, the more extensive pruning yields

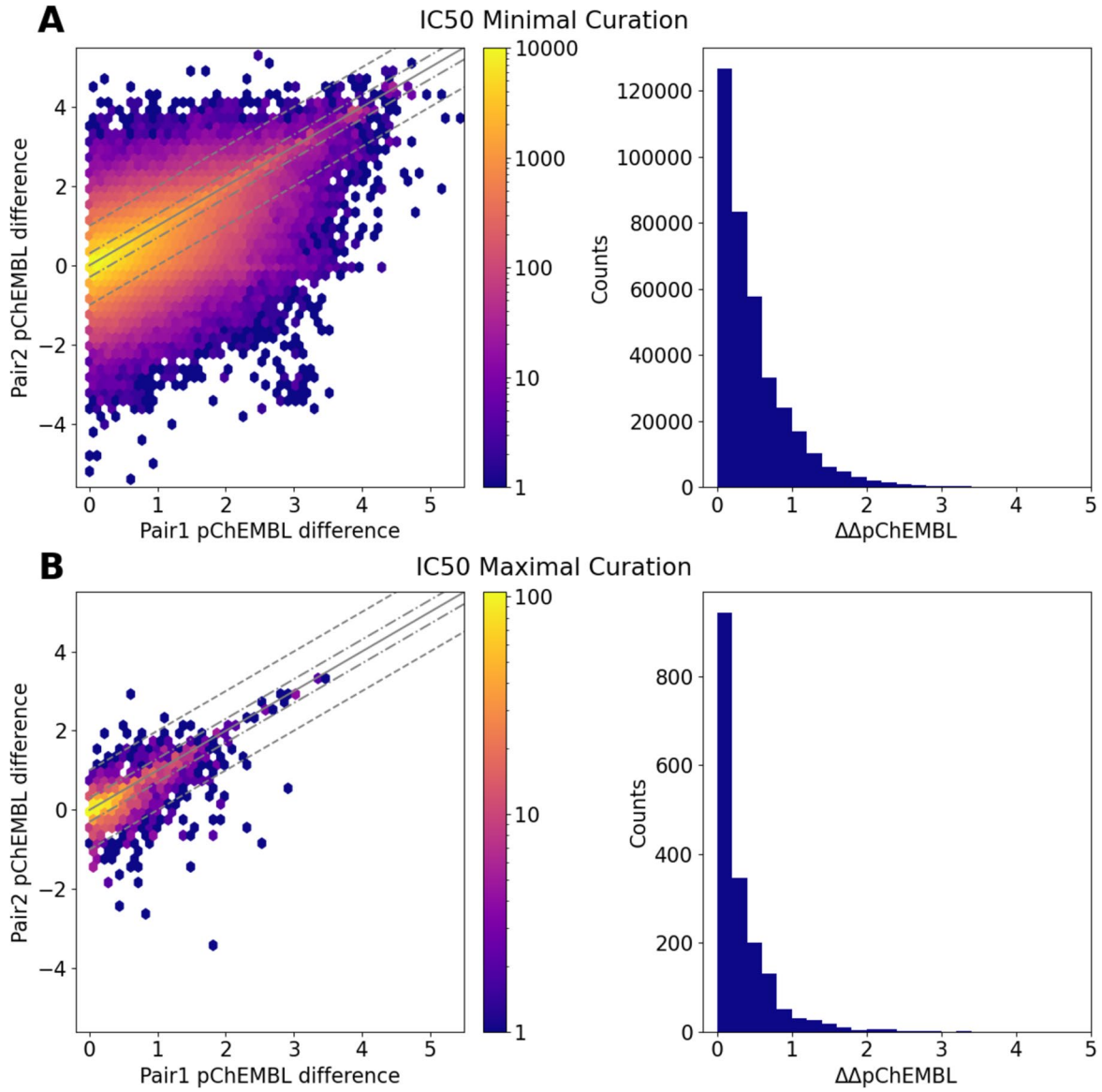

Fig. 2 Hexbin plots (left) and histograms (right) for the ΔΔpChEMBL IC_{50} data. **A** Minimal curation procedure applied to the IC_{50} data resulting in 372,683 data points. **B** Maximal curation procedure performed for the IC_{50} data resulting in 1,773 data points. The colors indicate the number of datapoints in that area. To highlight the 0.3 and 1.0 log difference, dotted lines are drawn as guidelines

better results, with improved pairwise metrics and fewer outliers.

Results and discussion

We first sought to estimate the inter-assay differences in potency between the differences of matched compound pairs present in combined data sets. Table 1 summarizes the pairwise metrics for the mincur and maxcur procedures applied to both the IC_{50} and K_i data. Specifically, we evaluated the noise of each dataset using three metrics: the Median Absolute Error (MAE), and the fraction of values exceeding 0.3 and 1.0 log units. These metrics were chosen based on the original manuscript's thresholds, which consider differences smaller than 0.3 log units to be within acceptable experimental noise limits, and differences greater than 1 log unit to be indicative of significant discrepancies. Additionally, the number of assay pairs, and total compound pairs (data points) are provided to give context regarding the dataset sizes.

The IC_{50} data, also visualized in Fig. 2, demonstrates the impact of curation on data quality. The minimal curation

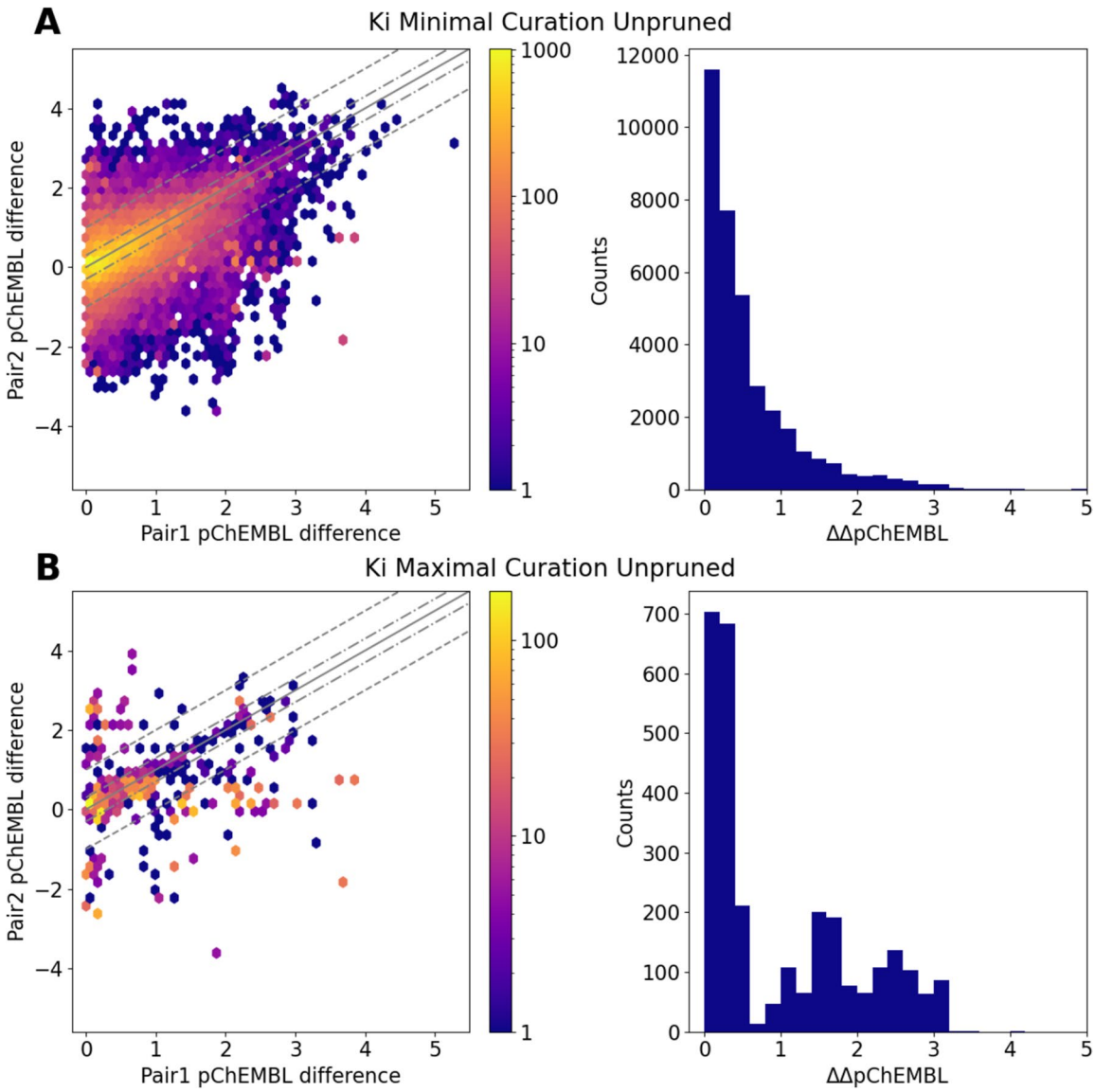

Fig. 3 Hexbin plots (left) and histograms (right) for the unpruned ΔΔpChEMBL K_i data. **A** Minimal curation procedure applied to the K_i data resulting in 36,189 data points. **B** Maximal curation procedure performed for the K_i data resulting in 2,900 data points. The colors indicate the number of datapoints in that area. To highlight the 0.3 and 1.0 log difference, dotted lines are drawn as guidelines.

dataset (Fig. 2A) contains just under 372,700 data points, which is reduced to around 1,770 with maxcur (Fig. 2B). As shown in Table 1, this reduction is accompanied by substantial improvements in data quality, including a 45% reduction in MAE, a 37% decrease in F > 0.3, and a 50% decrease in F > 1.0.

Interestingly, the unpruned K_i data (Fig. 3) show a different trend. The metrics worsen when going from the minimal to the maximal curation procedure (Table 1). This is especially apparent for F > 1.0, which saw an increase from 18 to 43%. Both datasets are visualized in Figs. 3A and B respectively. The histogram in Fig. 3B also displays an irregular pattern, indicating the presence of a substantial amount of noise in the data. This observation prompted us to perform data pruning as described in the methods section.

After pruning away all data associated with Carbonic Anhydrase assays, substantial improvements are observed (Fig. 4). The pruned minimal curation data (Fig. 4A) show minimal changes in pairwise metrics

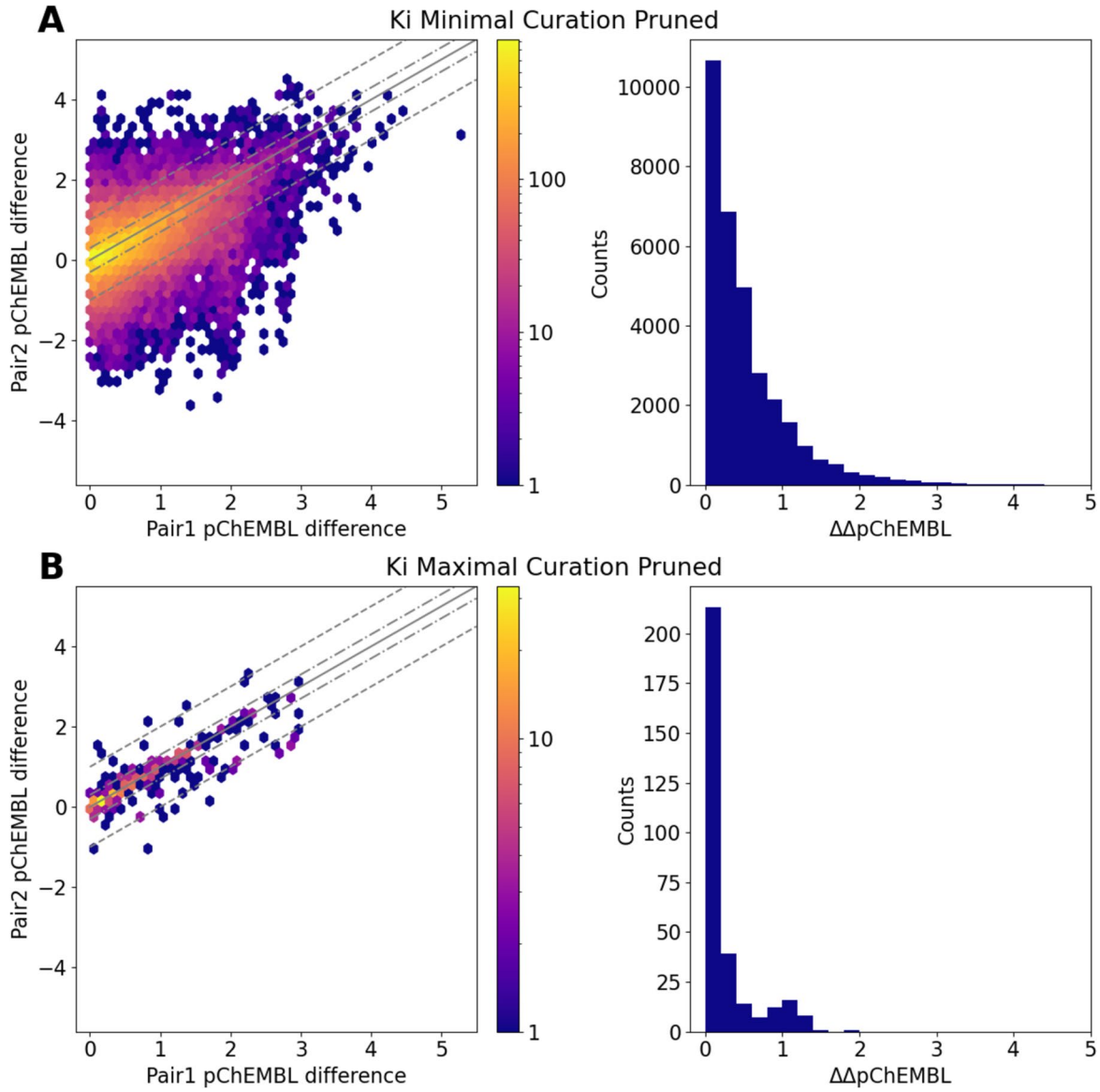

Fig. 4 Hexbin plots (left) and histograms (right) for the unpruned ΔΔpChEMBL K_i data. **A** Minimal curation procedure performed for the K_i data resulting in 32,363 data points. **B**: Maximal curation procedure applied to the K_i data resulting in 311 data points. The colors indicate the number of datapoints in that area, while lines indicating the 0.3 and 1.0 log difference are drawn as dotted lines

compared to the unpruned minimal curation data. In contrast to the unpruned datasets, the curation of the pruned K_i data (Fig. 4B) does lead to substantial improvements. As quantified in Table 1, the MAE sees more than a 90% reduction, the $F > 0.3$ is reduced by over 60%, and $F > 1.0$ is halved. However, the pruned K_i maxcur data are reduced from 32,363 to only 311 data points.

Table 2 presents a summary and comparison of our matched pair data (ΔΔpChEMBL) with the results reported in the original paper [1], which directly compared assay measurements (ΔpChEMBL). The errors observed by Landrum and Riniker are a combination of annotation errors, intra-assay variability (i.e. normal experimental variability) and inter-assay differences (e.g. different types of readout technology, different concentrations of medium or substrate, etc.).

Our data shows improved pairwise metrics in all but one comparison—in which noise from a single target had a drastic impact—indicating that comparing pairs of structural analogs reduces noise. Inter-assay differences, such as different substrate concentrations or lipophilicity-driven interactions with assay medium or

Table 2 Comparison of matched pair data with the original manuscript results [1]

Curation level	Type	MAE	F > 0.3	F > 1.0	Assay pairs	Compound pairs
IC_{50} minimal	Ref	0.50	0.64	0.27	1,358	38,022
	Pair	0.33	0.54	0.12	1,338	372,683
IC_{50} maximal	Ref	0.27	0.48	0.13	26	340
	Pair	0.18	0.34	0.06	24	1,773
K_i minimal	Ref	0.52	0.67	0.30	587	7,734
	Pair	0.36	0.56	0.18	584	36,189
K_i maximal	Ref	0.47	0.69	0.32	282	2,434
	Pair	0.40	0.62	0.43	279	2,900
K_i maximal pruned	Ref	0.45	0.58	0.25	9	115
	Pair	0.03	0.21	0.08	8	311

Note that the compound pairs for *Pair-type* datasets refers to cross-assay matched structural pairs, rather than pairs of individual compounds as in the *Ref-type* datasets. Additionally, reference pruning was less strict, leaving some Carbonic Anhydrase assays. These were removed in our workflow after identifying a cluster of outliers (Figure S4)

assay materials [11] will often have similar effects on the observed potencies of structurally similar compounds, resulting in cancellation of these sources of error when looking at the difference in potency. Consequently, assessing pairwise differences appears more robust than comparing direct assay measurements, though some variability still persists as shown in our analysis.

The single metric on which the matched pair data performed worse was the F > 1.0 for the unpruned K_i maxcur data. This anomaly is due to the large fraction of noisy samples present in that dataset, which is amplified further due to the formation of pairs between them. It is also worth noting that while the pruning methods employed differ slightly between the two studies, the trends remain consistent, and we provide the data using the same pruning procedure in the Supplementary Information (Table S2) for comparison.

Conclusions

In conclusion, our study finds that analyzing matched pairs is more robust to inter-assay variability than directly comparing assay measurements across assays, likely driven by the cancellation of systematic differences between assays. For minimally curated compound pairs, 44–46% exhibited a difference within 0.3 pChEMBL units, compared to only 33–36% for direct assay measurements. Similarly, the proportion of compound pairs with differences exceeding 1 pChEMBL unit was substantially lower for matched pairs (12–15%) compared to direct assay comparisons (27–30%). Additionally, our results show that careful data curation can help to mitigate noise introduced by combining datasets. Following curation, 66–79% of the pairs agreed within 0.3 pChEMBL units, up from 42–52% for direct assay comparisons. Furthermore, the proportion of compound pairs with differences exceeding 1 pChEMBL unit (F > 1.0) was reduced to 6–8% after curation, compared to 13–25% for direct comparisons.

Datasets with less noise enable the development of more performant machine learning models. However, stricter data curation also results in discarding a substantial amount of data. Practitioners should carefully consider this tradeoff when assembling datasets for MMPA or deep learning MMP methods. To facilitate further exploration, we provide all datasets used in this study, along with a Jupyter notebook that enables researchers to experiment with different curation settings and assess the impact of their choices on noise levels and robustness. Finally, our analysis reaffirms that careful examination of datasets for potential artifacts can improve data quality, as evidenced by the differences in pairwise metrics between the unpruned and pruned datasets.

Abbreviations

AI	Artificial intelligence
CA	Carbonic anhydrase
MAE	Median absolute error
Maxcur	Maximal curation
MCES	Maximum common edge subgraphs
Mincur	Minimal curation
ML	Machine learning
MMPA	Matched molecular pair analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00956-y.

Supplementary material 1.

Acknowledgements

We would like to thank Dr. Greg Landrum and Dr. Jeremy R. Ash for their thoughtful comments and insightful suggestions which helped to improve our methodology, results, and conclusions.

Author contributions
JN conducted the formal analysis, created the figures, and drafted the original manuscript. HDW and DVR conceptualised and supervised the study, and revised the manuscript to enhance its quality. HPS contributed to the refinement and finalization of the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding
Jochem Nelen was funded by Cátedra Villapharma-UCAM.

Availability of data and materials
The code and datasets used in this analysis are publicly available on GitHub at https://github.com/Jnelen/ChEMBL_MatchedPairsAnalysis.

Declarations

Competing interests
The authors declare no competing interests.

Received: 7 August 2024 Accepted: 11 January 2025
Published online: 20 January 2025

References

1. Landrum GA, Riniker S (2024) Combining IC50 or Ki values from different sources is a source of significant noise. J Chem Inf Model 64:1560–1567. https://doi.org/10.1021/acs.jcim.4c00049
2. Mendez D, Gaulton A, Bento AP et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930–D940. https://doi.org/10.1093/nar/gky1075
3. Kramer C, Dahl G, Tyrchan C, Ulander J (2016) A comprehensive company database analysis of biological assay variability. Drug Discov Today 21:1213–1221. https://doi.org/10.1016/j.drudis.2016.03.015
4. Brown SP, Muchmore SW, Hajduk PJ (2009) Healthy skepticism: assessing realistic model performance. Drug Discov Today 14:420–427. https://doi.org/10.1016/j.drudis.2009.01.012
5. Kramer C, Fuchs JE, Whitebread S et al (2014) Matched Molecular Pair Analysis: Significance and the Impact of Experimental Uncertainty. J Med Chem 57:3786–3802. https://doi.org/10.1021/jm500317a
6. Jin W, Yang K, Barzilay R, Jaakkola T (2019) Learning Multimodal Graph-to-Graph Translation for Molecular Optimization
7. Fralish Z, Chen A, Skaluba P, Reker D (2023) DeepDelta: predicting ADMET improvements of molecular derivatives with deep learning. J Chemin 15:101. https://doi.org/10.1186/s13321-023-00769-x
8. Fralish Z, Skaluba P, Reker D (2024) Leveraging bounded datapoints to classify molecular potency improvements. RSC Med Chem. https://doi.org/10.1039/D4MD00325J
9. RDKit. In: RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/. Accessed 13 Jun 2024
10. Raymond JW, Gardiner EJ, Willett P (2002) RASCAL: calculation of graph similarity using maximum common edge subgraphs. Comput J 45:631–644. https://doi.org/10.1093/comjnl/45.6.631
11. Palmgrén JJ, Mönkkönen J, Korjamo T et al (2006) Drug adsorption to plastic containers and retention of drugs in cultured cells under in vitro conditions. Eur J Pharm Biopharm 64:369–378. https://doi.org/10.1016/j.ejpb.2006.06.005

Publisher's Note