Combining IC$_{50}$ or $K_i$ Values From Different Sources is a Source of Significant Noise

Gregory A. Landrum, a* Sereina Riniker a*

a Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland. *Email: glandrum@ethz.ch, sriniker@ethz.ch

Abstract

As part of the ongoing quest to find or construct large data sets for use in validating new machine learning (ML) approaches for bioactivity prediction, it has become distressingly common for researchers to combine literature IC$_{50}$ data generated using different assays into a single data set. It is well known that there are many situations where this is scientifically risky thing to do even when the assays are against exactly the same target, but the risks of assays being incompatible is even higher when pulling data from large collections of literature data like ChEMBL. Here, we estimate the amount of noise present in combined data sets by comparing results where results for the same compound are reported in multiple assays against the same target. This approach shows that IC$_{50}$ assays selected using minimal curation settings have poor agreement with each other: almost 65% of the points differ by more than 0.3 log units, 27% differ by more than one log unit, and the correlation between the assays, as measured by Kendall’s $\tau$ is only 0.51. Requiring that most of the assay metadata in ChEMBL matches (“maximal curation”) in order to combine two assays, improves the situation (48% of the points differ by more than 0.3 log units, 13% by more than one log unit, and Kendall’s $\tau$ is 0.71) at the expense of having smaller data sets. Surprisingly, our analysis shows similar amounts of noise when combining data from different literature $K_i$ assays. We suggest that good scientific practice requires careful curation when combining data sets from different assays and hope that our maximal curation strategy will help to improve the quality of the data that is being used to build and validate ML models for bioactivity prediction. To help achieve this, the code and ChEMBL queries that we used for the maximal curation approach are available as open-source software in our GitHub repository, https://github.com/rinikerlab/overlapping_assays.

1 Introduction

Most artificial intelligence/machine learning (AI/ML) methods are very data hungry: They require a large amount of training data in order to build useful predictive models. Additionally, noise in the training data for the models sets an upper limit on the accuracy which can be expected. At the same time, there are not very many large, open data sets available that are applicable to computational drug discovery. Large, consistently measured data sets are typically only available inside of companies and, due primarily to IP concerns, are difficult/impossible to publish in the open scientific literature. There are notable exceptions to this [1, 2], but they are definitely rare. This has consequences for researchers who only have access to public data sources. For example, when extracting data from ChEMBL [3, 4] the only way to be mostly certain that a data set was consistently measured is to only take data from a single assay. Unfortunately, more than 60’000 of the >85’000 IC$_{50}$ assays in ChEMBL32 have data for less than ten distinct compounds, only 650 assays have data for more than 100 distinct compounds, and there are only 54 assays with data for more than 500 distinct compounds (Figure 1). This dearth of large, consistent data sets has led to the common practice of combining results from different assays (measured against the same target) to create data sets for AI/ML applications.
Figure 1: Histograms of the number of compounds per assay in ChEMBL32: IC$_{50}$ assays (left) and K$_i$ assays (right). Only measurements with a non-null pchembl value were included. Assays with 100 or less points are not included in these histograms.

1.1 Compatibility Issues

One source of noise when comparing experimental results from different labs is the inevitable inter-lab variability (note that also the same-lab variability is usually not zero due to inherent experimental errors). This is higher with some assay types than others, for example Caco-2 permeability assays are well-known to have problems with this type of variability due to differences in the cells used in the assay as well as the impossibility of exactly reproducing experimental conditions when working with living systems [5, 6].

Looking beyond lab-to-lab variability of assays that are nominally the same, there are numerous reasons why literature results for different assays measured against the same “target” may not be comparable. Some of these include:

1. Different assay conditions: These can include different buffers, experimental pH, temperature, duration, etc.

2. Substrate identity and concentration: These are particularly relevant for IC$_{50}$ values from competition assays where the identity and concentration of the substrate being competed with play an important role in determining the results. K$_i$ measures the binding affinity of a ligand to an enzyme and so its values are, in principle, not sensitive to the identity or concentration of the substrate.

3. Mode of action for receptors: EC$_{50}$ values can correspond to agonism, antagonism, inverse agonism, etc.

The situation is further complicated when working with databases like ChEMBL, which curate literature data sets:

1. Different targets: Different variants of the same parent protein are assigned the same target ID in ChEMBL

2. Different assay organism or cell types: The target protein may be recombinantly expressed in different cell types (the target ID in ChEMBL is assigned based on the original source of the target) or the assays may be run using different cell types.

3. Any data source can contain human errors like transcription errors or reporting incorrect units. These may be present in the original publication – when the authors report the wrong units or include results from other publications with the wrong units – or introduced during the data extraction process.
In this work, we focus primarily on two of the largest classes of publicly available dose-response bioactivity data: IC$_{50}$ and $K_i$. The conventional wisdom is that it is generally not scientifically valid to combine values from different IC$_{50}$ assays without knowledge of the assay conditions, but that $K_i$ values are more comparable across assays. Reference [7] has a very readable explanation of the relationship between IC$_{50}$ and $K_i$.

### 1.2 Assessing Assay Compatibility

The best way to determine whether or not the results from two different IC$_{50}$ or $K_i$ assays measured on the “same” target are compatible with each other is to read the original publications and directly assess whether all important parameters are the same. However, given the number of available IC$_{50}$ assays for many targets – in ChEMBL32, human CDK2 has 343 assays, human BRD4 has 454 assays, and a common target like hERG has 2020 assays – this is not feasible at any sort of scale, so we need other compatibility metrics. One approach which lends itself to both automation and large-scale analysis is to identify pairs of assays in which the same compound (or multiple compounds) has been tested. Comparing the measured IC$_{50}$ or $K_i$ values for the compound(s) shared between the assays gives a good sense as to whether or not the rest of the results can be compared. Results differing by less than an expected window for experimental error – for example $\Delta$pIC$_{50} < 0.3$ [8–10], approximately a factor of two – support the hypothesis that the assays are compatible.

In this work, we start by estimating the compatibility of IC$_{50}$ and $K_i$ assays for the same target drawn from ChEMBL32. We then develop a curation methodology, which takes advantage of the assay metadata available in ChEMBL to avoid combining results from assays that are clearly incompatible. The impact of this “max curation” scheme on data-set quality and size is estimated and discussed.

### 2 Methods

#### 2.1 Extracting Data From ChEMBL32

Data was extracted from a local copy of ChEMBL32 [4] running in a PostgreSQL database [11] using standard SQL queries within the Jupyter computational notebook environment. The database was constructed directly, without modification, from the PostgreSQL dump provided by the ChEMBL team [12]. All queries used can be found in the Jupyter notebooks in the project GitHub repository: https://github.com/rinikerlab/overlapping_assays.

#### 2.2 Quantifying Assay Compatibility

The compatibility between two assays was measured by comparing pchembl values of overlapping compounds. In addition to plotting the values, a number of metrics were used to quantify the degree of compatibility between assay pairs:

- $R^2$: Coefficient of determination provides a direct measure of how well the “duplicate” values in the two assays agree with each other. Values range from $-1.0$ to $1.0$ with larger values corresponding to higher compatibility.

- Kendall $\tau$: Non-parametric measure of how equivalent the rankings of the measurements in the two assays are. Values range from $-1.0$ to $1.0$ with larger values correspond to higher compatibility.

- $f > 0.3$: Fraction of the pairs where the difference is above the estimated experimental error. Smaller values correspond to higher compatibility.

- $f > 1.0$: Fraction of the pairs where the difference is more than one log unit. This is an arbitrary limit for a truly meaningful activity difference. Smaller values correspond to higher compatibility.

- $\kappa_{bin}$: Cohen's $\kappa$ calculated between the assays after binning their results into active and inactive using bin as the activity threshold. Values range from $-1.0$ to $1.0$ with larger values corresponding to higher compatibility.
• MCC\textsubscript{bin}: Matthews correlation coefficient calculated between the assays after binning their results into active and inactive using \textit{bin} as the activity threshold. Values range from $-1.0$ to $1.0$ with larger values corresponding to higher compatibility.

All metrics were calculated using either scikit-learn [13] version 1.2.2 or scipy [14] version 1.10.1.

2.3 Curation Approaches

Given the obvious scientific problems and amount of noise introduced by combining all IC\textsubscript{50} data (see Results and Discussion below), we explored a number of different strategies for more carefully curating combined IC\textsubscript{50} data sets based purely on the information available in the ChEMBL database.

The curation operations we applied were:

• **Activity curation**: Remove pairs of measurements where the \textit{pchembl} values in the two assays were either exactly the same or differed by 3.0. Given the very low probability of two separate experiments producing exactly the same results, the exact matches are most likely cases where values from a previous paper are copied into a new one, this was discussed in the earlier work by Kramer \textit{et al.} [9] and spot-checked with a number of assay pairs here. The pairs differing by exactly three log units correspond to the same copy action with the twist that a units error was made in either one of the publications or the ingestion into ChEMBL.

• **Duplicate papers**: Remove pairs of measurements where both assays were published in the same document. Having two (or more) IC\textsubscript{50} assays against the same target in the same paper usually only occurs when there is a difference between the two assays: Either they have been run under different conditions, are using different variants of the same protein (ChEMBL’s curation does not always distinguish between variants), etc.

• **Remove mutants**: Because the ChEMBL target metadata does not provide information about variant proteins (still often called “mutants”), different variants of a target protein will share the same target ID as the wild type. However, the assay description field in ChEMBL will often contain some information about which variant was used. Before the release of ChEMBL22, this information was not captured systemically or using a controlled vocabulary. More recent versions of ChEMBL include the \textit{variant\_id} field in the assay metadata, so it is theoretically possible to detect similar variants for more recent assays. We adopt a conservative approach in this curation step and remove any assay which has the text “mutant”, “mutation”, or “variant” in its description or which has a variant ID specified.

• **Assay type**: One of the more important pieces of metadata that ChEMBL provides about assays is the assay type. This can take on values like “Binding”, “Functional”, “Physicochemical”, etc. This curation step removes pairs of assays with different assay types.

• **Assay metadata**: This curation step removes pairs of assays where any of the following assay metadata fields do not match: \textit{assay\_type}, \textit{assay\_organism}, \textit{assay\_category}, \textit{assay\_tax\_id}, \textit{assay\_strain}, \textit{assay\_tissue}, \textit{assay\_cell\_type}, \textit{assay\_subcellular\_fraction}, and \textit{bao\_format}. This list covers almost all of the assay metadata fields available in ChEMBL32 and not already mentioned above.

• **Sources other than documents**: This curation step removes any assay which is from a source that does not have an associated document date. The goal here is to only include data sets from the medicinal chemistry literature and patents, excluding screening data sets or other contributed data sets.

• **Assay size**: By default, any assays which include >100 compounds are removed. The goal of this step is to try and focus attention on the primary literature and ignore sources like review articles. Because the upper limit is a very heuristic threshold, we have also explored (and included the data from) an upper limit of 1000 compounds.
• **Curation confidence:** When this curation step is enabled, any assay which does not have a confidence score value of 9 (indicating that the assay is assigned to a direct single target) is removed.

The impacts of each of these steps individually on the number of IC\textsubscript{50} assay- and compound-pairs from ChEMBL32 is shown in Table S1 in the Supporting Information.

### 2.4 Applying Maximal Curation to Extract Data Sets

The main goal of this work is to identify curation settings for extracting reliable (i.e., less noisy) data from ChEMBL. Once we have identified the appropriate settings, the data sets themselves need to be extracted. This task is easy when doing minimum curation: We simply retrieve all of the IC\textsubscript{50} (or K\textsubscript{i}) data sets for a given target and combine them into a single data set labeled with the target ID. When doing maximal curation, we are more restrictive about which assays are considered:

- Only assays associated with documents are considered.
- Only assays with a curation confidence score of 9 are considered.
- Assays with the text "mutant", "mutation", or ' variant" in their descriptions are removed unless they have a non-null variant\_id.
- If a document contains multiple assays against the same target, only the one with results for the largest number of compounds is retained.

Once we have identified the assays to be considered for a target, we create a "conditions hash" for each one. This is md5 hash of the available assay metadata: assay\_type, assay\_organism, assay\_category, assay\_tax_id, assay\_strain, assay\_tissue, assay\_cell_type, assay\_subcellular\_fraction, bao\_format, and variant\_id. The combination of target ID and conditions hash defines a set of assays which are equivalent as far as we can tell from the information available in ChEMBL32. The final step is to combine these assays together and label them with the target ID and conditions hash.

For both curation settings, only unqualified activity values with nM standard values, non-null pchembl values, and no data\_validity\_comment are used.

### 3 Results and Discussion

#### 3.1 Noise Introduced by Combining Assays

We first looked at the variation in the data sets when IC\textsubscript{50} assays are combined using "only activity" curation (top panels in Figure 2). The noise level in this case is very high: 64% of the \Delta pchembl values are greater than 0.3 and 27% are greater than 1.0. The similar plot for the K\textsubscript{i} data sets is shown in Figure S1 in the Supporting Information. The noise level for K\textsubscript{i} is comparable: 67% of the \Delta pchembl values are greater than 0.3 and 30% are greater than 1.0.
Figure 2: Agreement between duplicate measurements in IC\textsubscript{50} assays on the same target with “only activity” curation (top) and maximal curation (bottom). (Left): Correlation plot between pchembl values from the two assays. The solid black line corresponds to $x = y$, the dot-dashed lines mark a difference of 0.3, and the dashed line marks a difference of 1.0. (Right): Histogram of $\Delta$pchembl, the differences in pchembl values.

The situation for IC\textsubscript{50} improves markedly when using the maximal curation scheme, at the expense of discarding almost 99% of the data (bottom panels in Figure 2). $\tau$ increases from 0.51 to 0.71 and the MAE decreases from 0.50 to 0.27. Note that even with the maximal curation settings 48% of the $\Delta$pchembl values differ by more than 0.3 log units and 13% differ by more than 1.0.

The top panels of Figure 3 show a similar plot for the K\textsubscript{i} data sets with the maximal curation scheme. Here, we have only lost 70% of the data and haven’t improved the quality of the results over activity-only curation: 69% of the $\Delta$pchembl values are greater than 0.3 and 32% are greater than 1.0. Surprisingly, when it comes to the regression parameters presented in Table 1, the maximal curation results are actually worse than the those from activity-only curation. What is happening here?
Figure 3: Agreement between duplicate measurements in $K_i$ assays on the same target with maximal curation (top) and with 239 problematic assays (see text) removed (bottom). (Left): Correlation plot between $pchembl$ values from the two assays. The solid black line corresponds to $x = y$, the dot-dashed lines mark a difference of 0.3, and the dashed line marks a difference of 1.0. The regions outlined with red boxes are discussed in the text. (Right): Histogram of $\Delta pchembl$, the differences in $pchembl$ values.

The top left panel of Figure 3 has two dense clusters of points which are highlighted in red boxes. These points arise from a set of 32 assays reporting $K_i$ values for human carbonic anhydrase I (ChEMBL target ID CHEMBL261). These assays share a corresponding author and include a significant number of overlapping compounds with results that are sometimes inconsistent. The original papers do not provide sufficient information about the sources of the data to understand the source of this variability [15, 16]. Because this is almost certainly artificial variability and not just experimental noise, we removed all data from assays which have more than ten compounds in common with one of these assays (CHEMBL3782909 [17]) from consideration and repeated the statistical analysis. The complete list of 239 assays removed from consideration is reported in the Supporting Information. The bottom panels of Figure 3 show the comparison with these assays removed. Most of the outliers are no longer present and the agreement is significantly better (Table 1). Note that we were only able to be certain that there was a problem with these data by going back to the original publications. Resolving situations like this is a non-trivial curation exercise, which is difficult to automate. We mention it here as an illustration of the kinds of things that can go wrong even after doing maximal curation for “best case” experimental readouts like $K_i$ data. Although we could reasonably expect $K_i$ values to be at least somewhat comparable across laboratories, we were limited in this case by the quality of the data in the primary scientific literature.
3.2 Regression Versus Classification

The previous results demonstrated the amount of noise that activity-only curation introduces to the IC\textsubscript{50} values, which would be used to build a regression model. When building classification models, we work with binned activity data, which raises the question what the impact of the noise is in this case? Table 2 shows κ and MCC values for three activity binning levels commonly used in the literature: \textit{pchembl} = 5 (10 µM), \textit{pchembl} = 6 (1 µM), and \textit{pchembl} = 7 (100 nM). With the activity-only curation setting, the MCC values for all three thresholds are < 0.6. Maximal curation improves the situation somewhat with MCC values ranging from 0.83 to 0.91. Similar improvements are observed for the \textit{K}_i data when using maximal curation combined with the pruning of the suspect assays.

The MCC and Cohen’s κ in Table 2 have very similar values because the confusion matrices are generally quite symmetric [18]. This makes sense given that the ordering of the assays by ChEMBL ID should not introduce any systematic differences in the two \textit{pchembl} values.

### Table 2: Impact of the curation level on classification quality metrics. See the Methods section for a description of the metrics themselves.

<table>
<thead>
<tr>
<th>Readout</th>
<th>Curation level</th>
<th># Assays</th>
<th># Cmpds</th>
<th>\textit{R}^2</th>
<th>τ</th>
<th>MAE</th>
<th>f &gt; 0.3</th>
<th>f &gt; 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC\textsubscript{50}</td>
<td>only activity</td>
<td>1358</td>
<td>38022</td>
<td>0.31</td>
<td>0.51</td>
<td>0.50</td>
<td>0.64</td>
<td>0.27</td>
</tr>
<tr>
<td>IC\textsubscript{50}</td>
<td>maximal</td>
<td>26</td>
<td>340</td>
<td>0.63</td>
<td>0.71</td>
<td>0.27</td>
<td>0.48</td>
<td>0.13</td>
</tr>
<tr>
<td>IC\textsubscript{50} large</td>
<td>only activity</td>
<td>1599</td>
<td>50385</td>
<td>0.32</td>
<td>0.51</td>
<td>0.51</td>
<td>0.65</td>
<td>0.28</td>
</tr>
<tr>
<td>IC\textsubscript{50} large</td>
<td>maximal</td>
<td>44</td>
<td>742</td>
<td>0.60</td>
<td>0.61</td>
<td>0.30</td>
<td>0.51</td>
<td>0.15</td>
</tr>
<tr>
<td>\textit{K}_i</td>
<td>only activity</td>
<td>587</td>
<td>7734</td>
<td>0.13</td>
<td>0.43</td>
<td>0.52</td>
<td>0.67</td>
<td>0.30</td>
</tr>
<tr>
<td>\textit{K}_i</td>
<td>maximal</td>
<td>282</td>
<td>2434</td>
<td>-0.33</td>
<td>0.27</td>
<td>0.47</td>
<td>0.69</td>
<td>0.32</td>
</tr>
<tr>
<td>\textit{K}_i</td>
<td>maximal + pruning</td>
<td>9</td>
<td>115</td>
<td>0.65</td>
<td>0.67</td>
<td>0.45</td>
<td>0.58</td>
<td>0.25</td>
</tr>
<tr>
<td>\textit{K}_i large</td>
<td>only activity</td>
<td>750</td>
<td>9650</td>
<td>0.21</td>
<td>0.46</td>
<td>0.51</td>
<td>0.65</td>
<td>0.28</td>
</tr>
<tr>
<td>\textit{K}_i large</td>
<td>maximal</td>
<td>290</td>
<td>2574</td>
<td>-0.10</td>
<td>0.32</td>
<td>0.47</td>
<td>0.67</td>
<td>0.32</td>
</tr>
<tr>
<td>\textit{K}_i large</td>
<td>maximal + pruning</td>
<td>17</td>
<td>255</td>
<td>0.68</td>
<td>0.71</td>
<td>0.12</td>
<td>0.38</td>
<td>0.21</td>
</tr>
</tbody>
</table>

3.3 Impact of Curation on Data Set Size

As the maximal curation scheme seems to improve data quality, we next investigated its impact on the size and composition of combined data sets from ChEMBL32. We started by using the activity-only curation settings to construct combined data sets for all targets, which contained at least 20 assays and activity values for at least 1000 compounds. This yields 80 targets for IC\textsubscript{50} and 38 targets for \textit{K}_i. The top panels of Figure 4 shows the numbers of compounds in combined data sets using the activity-only and maximal curation settings, whereas the bottom panels shows the number of assays combined into each data set.

---

https://doi.org/10.26434/chemrxiv-2024-2smhk ORCID: https://orcid.org/0000-0003-1893-4031 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0
Figure 4: Number of compounds per combined data set (top) and number of assays per combined data set (bottom) for activity-only curation (left) and maximal curation (right). In each plot, the data sets are sorted by decreasing size. The right panel is truncated at 100 data sets to aid visibility.

Although the maximal curation strategy does reduce the number of larger data sets available to work with, there are still 34 IC\textsubscript{50} data sets and 26 K\textsubscript{i} data sets containing at least 500 compounds. These are composed of data from at least 14 (IC\textsubscript{50}) or 16 (K\textsubscript{i}) assays. These data sets are considerably less likely to contain wildly inconsistent results than those produced by more minimal curation schemes and are better suited to serve as a basis for further analysis or building/validating ML approaches.

4 Conclusion

We have shown that combining literature data from different assays which measure IC\textsubscript{50} values against what is nominally the same target can result in very large amounts of noise. More careful automated curation of the data sets using metadata available in ChEMBL (maximal curation scheme) can substantially reduce the overall noise level in combined data sets with either IC\textsubscript{50} or K\textsubscript{i} as the readout – at the expense of substantially less data points. It is worth pointing out that even with the maximal curation settings, a significant amount of noise remains in the combined data sets.

While doing this work we were surprised by the lack of consistency in the K\textsubscript{i} data sets. We came to the project with the expectation to observe more inter-assay variability in the IC\textsubscript{50} data than in the K\textsubscript{i} data. However, the results did not meet this expectation (particularly before we manually pruned
a large set of the data due to issues with the primary data source). It seems that although there are scientific reasons (like different substrate concentrations) which render the combination of IC50 assays problematic, these are perhaps overwhelmed by practical problems when working with large collections of data drawn from patents and publications.

Good scientific practice requires some level of curation when combining data from different assays into a single data set for analysis (or training of ML models). We have demonstrated here that simplistic exports of data from resources like ChEMBL can result in data sets which combine assays measured against different variants of the same protein or under different conditions. Without the necessary curation, we are left analyzing or building ML models on data sets which, in the best case, contain overwhelming amounts of noise. In the worst case, they do not make scientific sense. Although some level of irreducible noise remains given experimental variability, the inevitable variation between laboratories, errors in the scientific literature, and the limits of what is possible when data sets are manually curated from the literature, we consider the maximal curation settings an important step forward towards high-quality public bioactivity data sets for training or validating ML models.

**Data and Software Availability**

The Jupyter notebooks used for this analysis, as well as the IC50 and Ki data sets discussed in the "Impact of Curation on Data Set Size" section are available under an open-source license in our public GitHub repository: [https://github.com/rinikerlab/overlapping_assays](https://github.com/rinikerlab/overlapping_assays).

**Acknowledgments**

The authors thank Patrick Walters for feedback on a draft of this manuscript and Christian Feldmann for feedback on the RDKit blog post which inspired this paper. The authors also thank Paul Katzberger for pre-submission review of the code and data associated with the manuscript.

**References**


