Table 1. Filtering statistics for extracting independent pairs of IC_{50} measurements on identical systems.

Filter	# protein/ligand systems remaining	# IC_{50} data points remaining
Systems with multiple measurements only	54.505	137.043
Remove multiple values from identical publications	18.804	85.705
Remove exact duplicate values	8.387	33.187
Remove pairs with unit errors	8.141	22.770
Remove duplicates with rounding errors	7.263	19.487
Remove unrealistic values	7.228	19.383
Remove pairs with overlapping authors	3.480	10.895

doi:10.1371/journal.pone.0061007.t001

be scrambled in order to not bias the calculation of $R^2_{Pearson}$ and σ. As we have shown earlier, [13] MUE, $M_{ed}UE$ and σ calculated from pairs of measurements are overestimated by a factor of $\sqrt{2}$. Therefore MUE, $M_{ed}UE$ and σ calculated from pairs of measurements were divided by $\sqrt{2}$.

Raw data was extracted from ChEMBL14 using MySQL statements. Filtering and pairing of measurements were done using Python 2.7. The statistical analysis was carried out using R version 2.15.1. [15] All R-, Python- and MySQL-scripts used including detailed instructions on how to repeat the work can be found in the Archive S1.

Results

In order to assess the comparability of IC_{50} values, we first extracted all series of compounds that have been measured against the same protein target in two independent assays from whole ChEMBL. There were twelve series of ten or more compounds whose activity on the same target has been measured in different assays. An overview of the different series is given in Supporting Information (Table S1, Text S1–S2 and Figures S1–S2). However, eleven out of twelve series had overlapping authors and the single independently measured series was incorrectly annotated into the database.

Since it is not possible to find independently measured sets of at least ten IC_{50} values for the same target, the IC_{50} variability was determined differently. In the following, we analyze the IC_{50} data using an approach that we have previously introduced for analyzing the reproducibility of heterogeneous K_i data. All pairs of identical protein-ligand systems with independently measured IC_{50} values were extracted from ChEMBL and the variability of the differences between the pairs of measurements was calculated.

The distribution of pIC_{50} values is shown in Figure 1. The distribution of measured values is slightly skewed to the left with a maximum of roughly 30% of all pIC_{50} values reported between 7.0 and 8.0.

The distribution of ΔpIC_{50} values and the distribution of the number of independent measurements per protein-ligand system are shown in Figures 2 and 3. Roughly 70% of all ΔpIC_{50}'s are smaller than one log unit.

Most systems with multiple independent measurements have two or three independent measurements. The most frequently measured system is celecoxib on cyclooxygenase-2 with 30 independently measured IC_{50} values.

Sets of ten pairs of measurements for seven ranges of ΔpIC_{50} were closely inspected. The selected ranges of ΔpIC_{50} for the inspected ten cases span the whole range of ΔpIC_{50} (see Figure 2). The values of 3.2 and 1.1 were selected to avoid pairs which could contain combinations of citation of previous values and unit transcription errors. The findings are summarized in Table 2.

We found that very high differences in pIC_{50} ($\Delta pIC_{50}>2.5$) were in most cases due to annotation errors. Some measurements had wrong units assigned (unit error). The receptor subtype was sometimes incorrectly assigned or not assigned at all (receptor subtype error). Other errors come from wrong stereoisomers of ligands (stereochemistry error), cellular assays assigned as biochemical assays (cellular assay error), incorrect target annotations (target error) and erroneous values extracted from original publications (value error).

Unit errors are the most common error. Receptor subtype errors occur most often for older publications (e.g., papers from the 1980's with published IC_{50} values for dopamine receptors, opioid receptors, and mono-amino oxidases in general, i.e. without distinguishing the subtypes). This data is mixed with the subtype specific data in ChEMBL. Stereochemistry errors occur when the stereochemistry is wrongly extracted from the original literature. Cellular assay errors occur when the reported IC_{50} values have been measured in a cellular assay, despite being associated with a confident score greater than four (see Dataset preparation section).

Pairs with small ΔpIC_{50}'s can also be composed of erroneously reported IC_{50} data. For example, the group of pairs with $\Delta pIC_{50} = 0.05$ contains one case where the IC_{50} extracted from the literature is incorrect as in the original manuscript there is an activity range given, whereas in the ChEMBL database only one threshold of the range is reported with an equal sign. Another smaller set of problems come from retracted original publications (for example, the original publication [16], publishing an IC_{50} value for the compound with ChEMBL ID CHEMBL266497 on aldose reductase (CHEMBL2622), was retracted). Considering the number of invalid pairs out of the ten inspected for the seven ΔpIC_{50} ranges there is a high probability that pairs with $\Delta pIC_{50} \geq 2.5$ contains errors in the database or in the original publication.

A plot of all pairs of pIC_{50} values is shown in Figure 4. The correlation coefficient for the raw extracted data is $R^2 = 0.40$. Excluding a major part of the invalid pairs by removing all pairs with $\Delta pIC_{50} \geq 2.5$, the correlation coefficient becomes $R^2 = 0.53$.

We also calculated the standard deviation σ of all ΔpIC_{50} and ΔpK_i values between 0.05 (lower threshold) and a variable upper threshold (1.5, 2.0 and 2.5) by fitting the data to a Gaussian distribution. The lower threshold of 0.05 was selected to remove pairs which were just rounded duplicates. The standard deviations obtained for the ΔpIC_{50} and ΔpK_i distributions are shown in Table 3. The fitted Gaussian and the raw distributions for

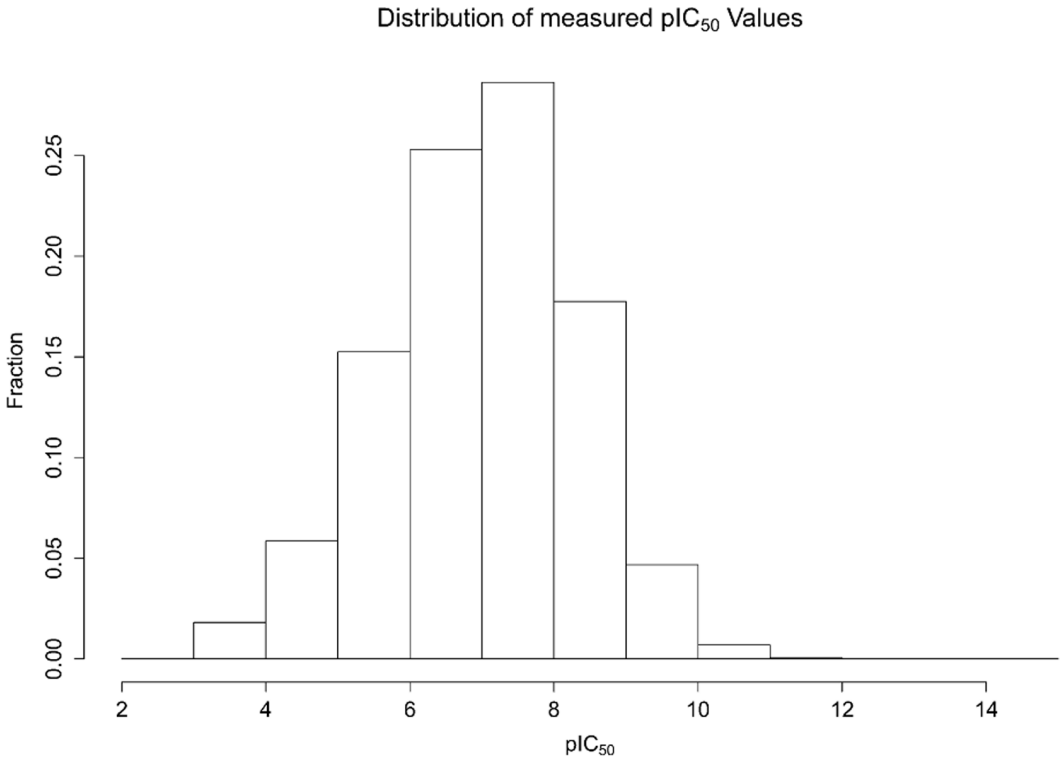

Figure 1. Distribution of the 9.465 pIC_{50} values for protein-ligand systems with independent multiple measurements.
doi:10.1371/journal.pone.0061007.g001

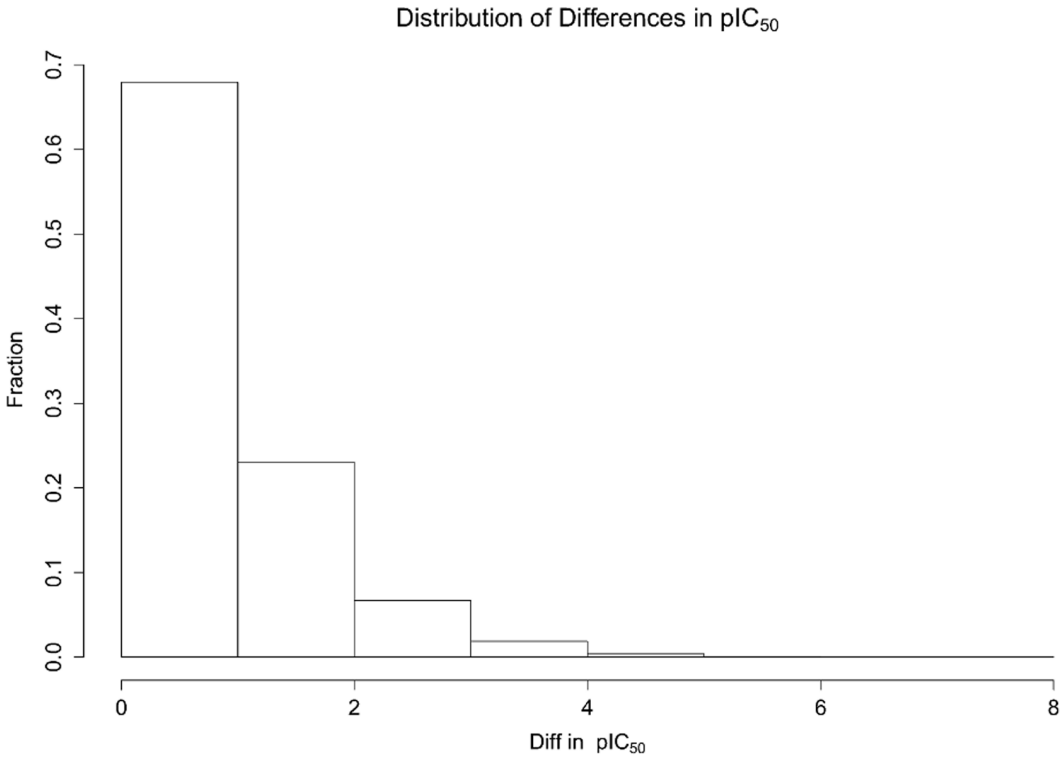

Figure 2. Distribution of the 16.844 pairs of ΔpIC_{50} values for protein-ligand systems with independent multiple measurements.
The largest ΔpIC_{50} is 7.7 log units.
doi:10.1371/journal.pone.0061007.g002

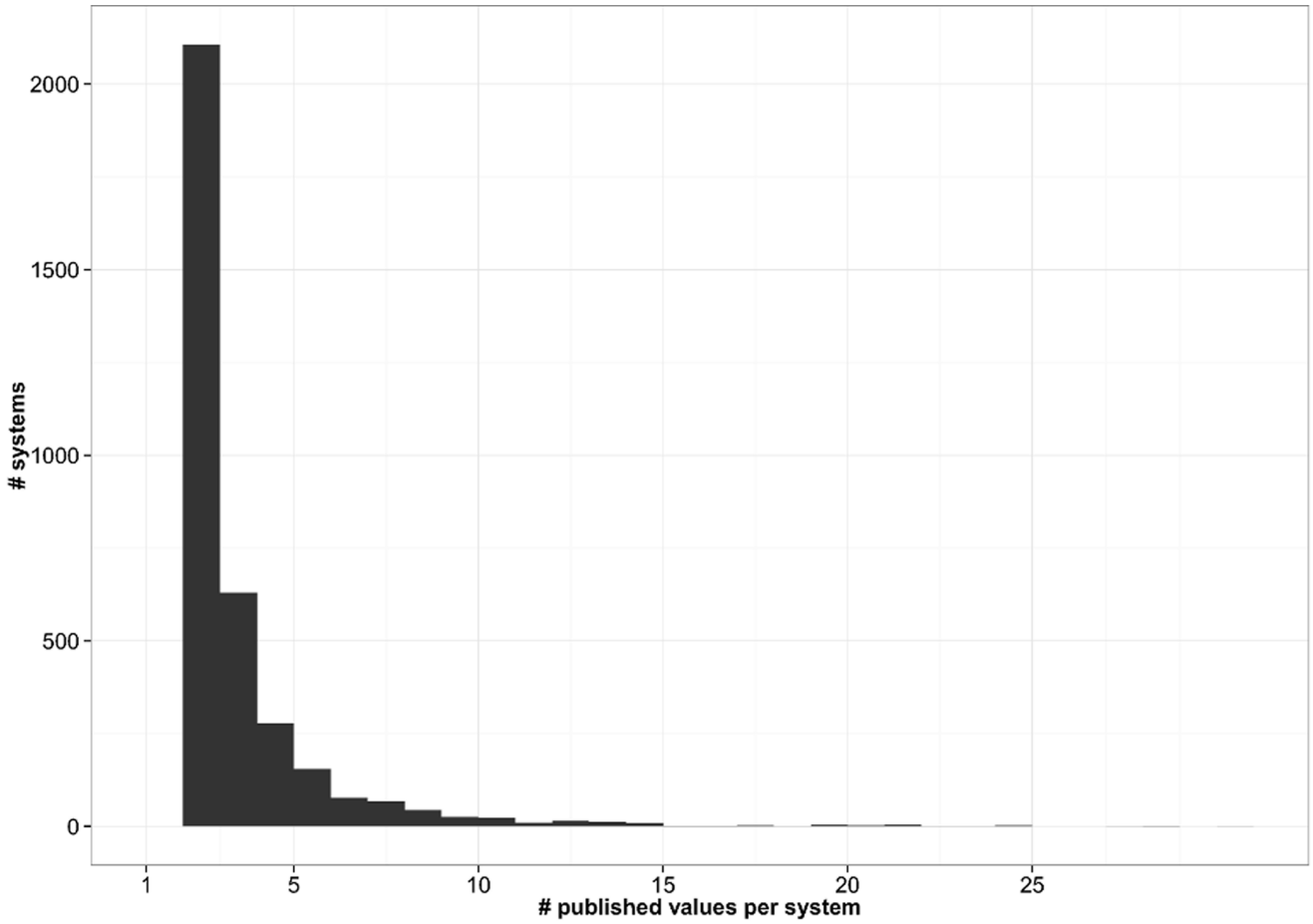

Figure 3. Number of published independent values per protein-ligand system.
doi:10.1371/journal.pone.0061007.g003

ΔpIC_{50}'s and ΔpK_i's with an upper threshold of 2.0 are shown in Figure 5.

The standard deviations of the ΔpIC_{50} data is constantly 21–26% larger than the standard deviation of the ΔpKi data. After dividing by $\sqrt{2}$, the σ for the Gaussian distribution fitted to all ΔpK_i values <2.5 then becomes 0.47 (a bit lower than the σ value of 0.54 previously calculated for heterogeneous pK_i data from ChEMBL version 12 data without upper threshold for ΔpKi data. [13] Since σ, MUE, and $M_{ed}UE$ are proportional to each other in Gaussian distributions, we can estimate σ, MUE and MedUE for the IC_{50} data to be 21–26% larger than the same metrics for pK_i data, yielding $\sigma_{pIC50} = 0.68$, $MUE_{pIC50} = 0.55$ and $M_{ed}UE_{pIC50} = 0.43$ (when using a factor of +25% for converting pK_i data to pIC_{50} data).

In order to test the alternative approach of directly obtaining quality metrics from the data, we calculated the quality metrics from the ΔpIC_{50} data with an upper threshold of $\Delta pIC_{50} = 2.5$. Here, $\sigma_{pIC50} = 0.68$, $MUE_{pIC50} = 0.54$ and $M_{ed}UE_{pIC50} = 0.43$ are obtained. These values are very similar to the values obtained from comparing fitted Gaussian distributions and indicate that the

Table 2. Errors found for samples of pairs of measurements with specific differences in measured pIC_{50}.

ΔpIC_{50}	# invalid pairs out of 10	Error types found
From 4.7 to 7.8	9	unit error, receptor subtype error, stereochemistry error, cellular assay error
3.2	10	unit error, cellular assay error, target error, value error
2.5	8	unit error, receptor subtype error, value error
1.5	6 (+2 dubious)	unit error, cellular assay error, receptor subtype error, value error
1.1	1 (+2 dubious)	cellular assay error, receptor subtype error
0.05	1 (+1 dubious)	value error, different assay conditions
0.02	0 (+4 dubious)	original paper retracted, data cited from third source which is not available any more, receptor subtype error

doi:10.1371/journal.pone.0061007.t002

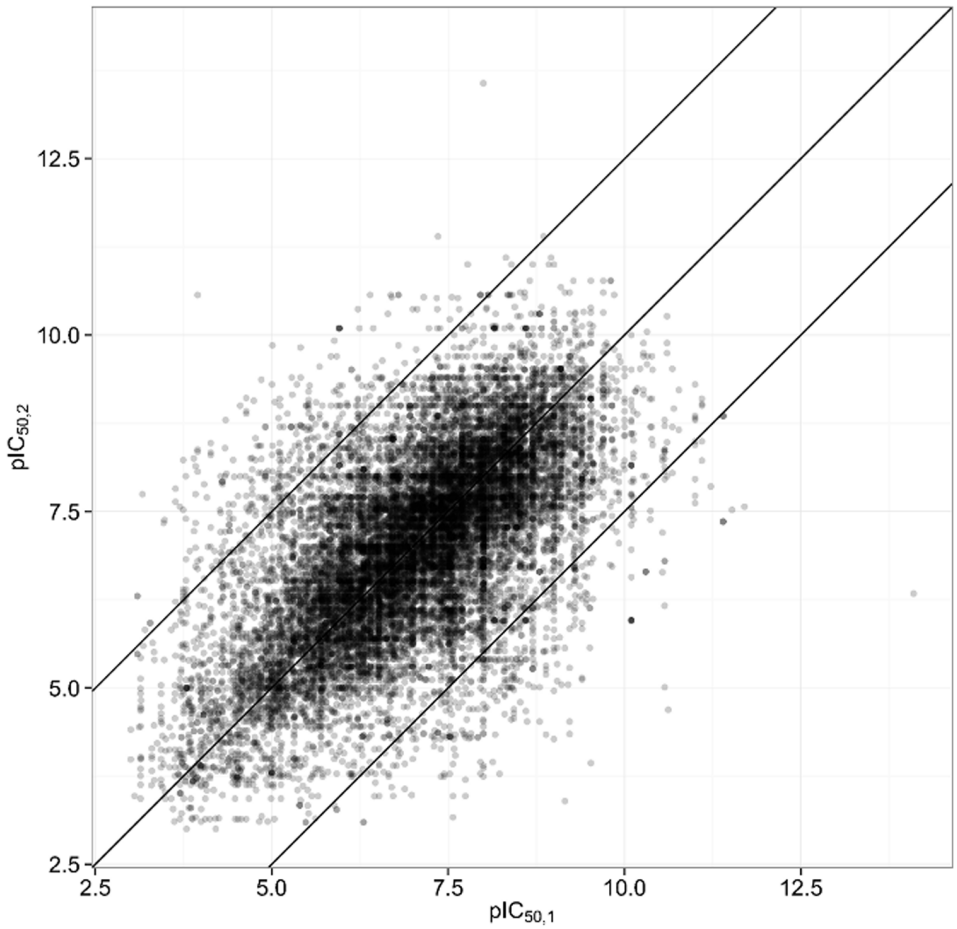

Figure 4. All Pairs of pIC_{50} values extracted from ChEMBL. The two outer diagonal lines indicate the 2.5 log unit threshold, outside which the probability for finding faulty pairs of measurements is very high. The extreme disagreements are all due to clear errors.
doi:10.1371/journal.pone.0061007.g004

erroneous pairs of measurements do not have a large effect on the overall result.

Similar performance was obtained considering only IC_{50} data with ChEMBL confidence score of nine (data not shown). As ChEMBL contains data from both human input and automatic extraction processes, we also looked if there was a difference between the two. Equally to the confidence score filtering, the results were similar with both data types.

We checked whether the ΔpIC_{50} depends on the overall activity measured or on physicochemical ligand properties like logP, logD, molecular weight (MW), polar surface area (PSA), the number hydrogen bond acceptors (HBA), the number hydrogen bond donors (HBD) or the number of rotatable bonds. Boxplots of all those properties versus the ΔpIC_{50} are shown in Figure 6. The ΔpIC_{50}'s depend neither on the average measured pIC_{50} nor on any of the ligand properties examined.

We also examined whether the ΔpIC_{50} depends on the combination of average activity and logP, since one might expect large deviations in measured pIC_{50}'s for compounds with low activity and high logP due to solubility issues. Here we also did not find a clear trend (Figure S3).

Table 3. Standard deviation of a Gaussian distribution fitted to the inner part of the distribution of ΔpIC_{50} and ΔpK_i.

Upper threshold	1.5	2.0	2.5
ΔpIC_{50}	$\sigma = 0.80$	$\sigma = 0.84$	$\sigma = 0.86$
ΔpK_i	$\sigma = 0.66$	$\sigma = 0.68$	$\sigma = 0.68$

doi:10.1371/journal.pone.0061007.t003

Can ChEMBL K_i and IC_{50} Data be Mixed?

Empirical statistical models and SAR interpretations improve with the amount of data. Above, we have shown that the variability of heterogeneous IC_{50} data is roughly 25% worse than that of K_i data. Therefore it is not recommendable to add IC_{50} data to K_i data as this would lower the quality of the data. However, since there is much more IC_{50} data than K_i data available, it is interesting to see what happens by augmenting the IC_{50} dataset with additional K_i data. Figure 7 shows the distribution of pK_i and pIC_{50} data extracted from ChEMBL with the filters mentioned in Table 1. Overall, pIC_{50} and pK_i data show a similar distribution with the pK_i data slightly shifted towards higher values.

For identical protein-ligand systems, we extracted all pairs of pK_i and pIC_{50} data that have passed the filters individually. This yields 11.556 pairs of measurements on 670 protein-ligand systems. A plot of measured pIC_{50} versus pK_i is shown in Figure 8.

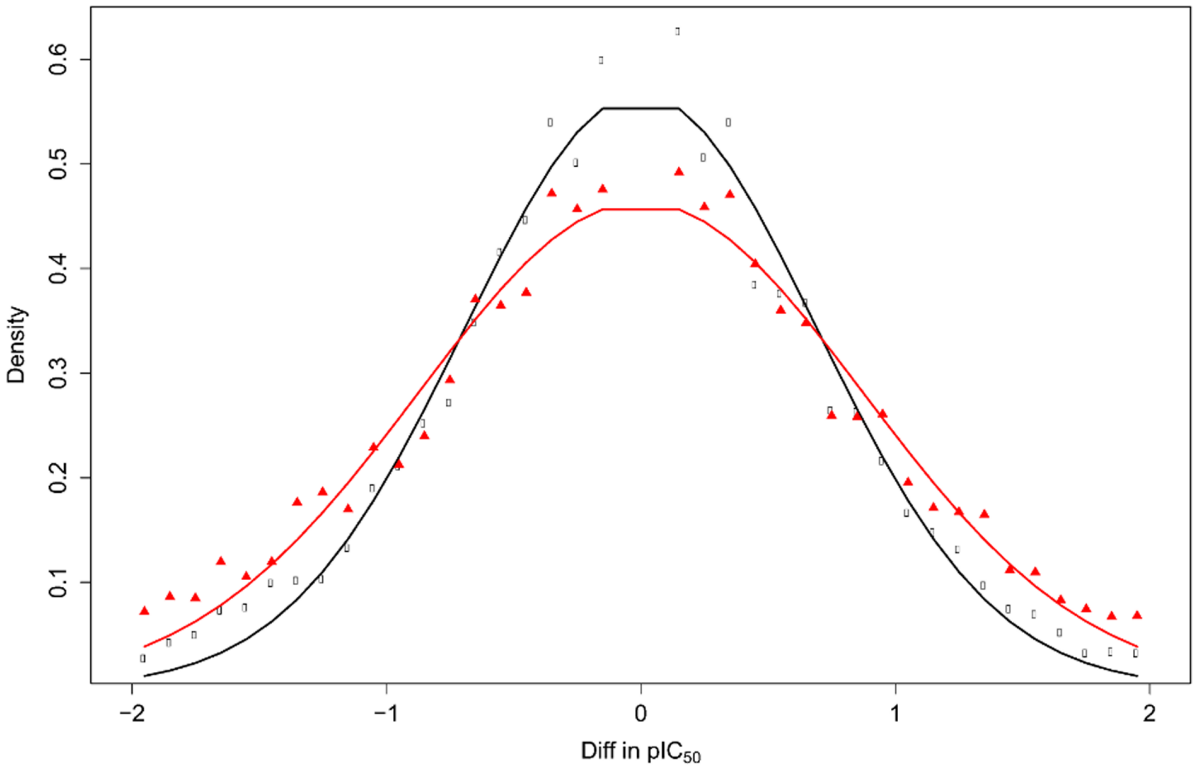

Figure 5. Fitted Gaussian distribution of ΔpIC_{50} (red) and ΔpK_i (black). The Gaussian distributions shown were fitted to all ΔpActivity values with an upper threshold ΔpActivity = 2.0. Standard deviations for the fitted Gaussian distributions are $\sigma_{pIC50} = 0.87$ and $\sigma_{pKi} = 0.69$. Note that since the σ here is calculated from pairs of measurements each containing experimental uncertainty and other sources of variability, it has to be divided by $\sqrt{2}$ in order to obtain the true σ of the individual measurements [13].
doi:10.1371/journal.pone.0061007.g005

Based on the Cheng-Prusoff equation and under the assumption of a competitive mechanism of action, pK_i values are larger or equal to pIC_{50} values. However due to unknown mechanism, experimental uncertainty and some database annotation errors in the data, there are a significant number of pairs where the pIC_{50} is larger than the pK_i. On average, the measured pK_i values are 0.355 log units larger than the measured pIC_{50} values, corresponding to a factor of 2.3. A factor of 2 is in agreement with a balanced assay condition in which the substrate concentration is equal to the K_m value. This is often used in order to allow the detection of inhibitors with different mechanism of action.

After subtracting 0.35 log units from the pK_i values and correcting by $\sqrt{2}$, pK_i and pIC_{50} values agree with an $R^2 = 0.46$, $\sigma = 0.68$, $MUE = 0.54$ and $M_{ed}UE = 0.43$. The standard deviations of Gaussian distributions fitted to the inner part with an upper threshold of 1.5, 2.0 and 2.5 ΔpActivity units are 0.79, 0.83, and 0.85.

Overall, this is close to or even slightly better than the agreement obtained for pIC_{50} values with themselves. Therefore we can conclude that pK_i values can be used to augment pIC_{50} values without any loss of quality, if they are corrected by an offset. In the absence of assay information, the best guess for the conversion factor between K_i into IC_{50} is extrapolated from the average offset calculated from the heterogeneous ChEMBL data, i.e. a factor of 2.3, corresponding to 0.35 pActivity units.

Discussion

In this contribution we show how the comparability of IC_{50} data can be analyzed using the public ChEMBL database. We find that when comparing all independently measured pIC_{50} data, the variability found for pIC_{50} data is approximately 25% larger than the variability found for pK_i data, with $\sigma_{pIC50} = 0.68$, $MUE_{pIC50} = 0.55$ and $M_{ed}UE_{pIC50} = 0.43$. These values correspond to the most probable variability of pIC_{50} data mixing from different (unknown) assays.

We want to stress that pIC_{50} data from different assays can only be compared under certain conditions. However, as discussed in the introduction, this is often done in large-scale data analysis. A standard deviation of 0.68 corresponds to a factor of 4.8, meaning that 68.2% of all IC_{50} measurements agree within a factor of 4.8, even when measured in different laboratories under potentially different assay conditions. One reason why the variability of IC_{50} data is found only moderately higher than the variability of K_i data might be that practically most of the IC_{50} assays may have been run using very similar assay protocols. Unfortunately, the assay descriptions available within ChEMBL are too terse to permit analyzing this any further.

IC_{50} values measured in the same laboratory usually show a better reproducibility. From our in-house database, we extracted series of reference pIC_{50} values measured for assay standards. The plots in Figure 9 show the pIC_{50} values measured for rolipram on PDE4D and cilostamide on PDE3. The standard deviation of the pIC_{50} values are $\sigma = 0.22$ for rolipram/PDE4D and $\sigma = 0.17$ for cilostamide/PDE3.

There is some variation over time which could indicate changes in the assay conditions and solution handling. We also tried to find public series of at least ten compounds that have been measured in independent parallel assays. However, such series did not exist within ChEMBL as all the series we found were either measured in

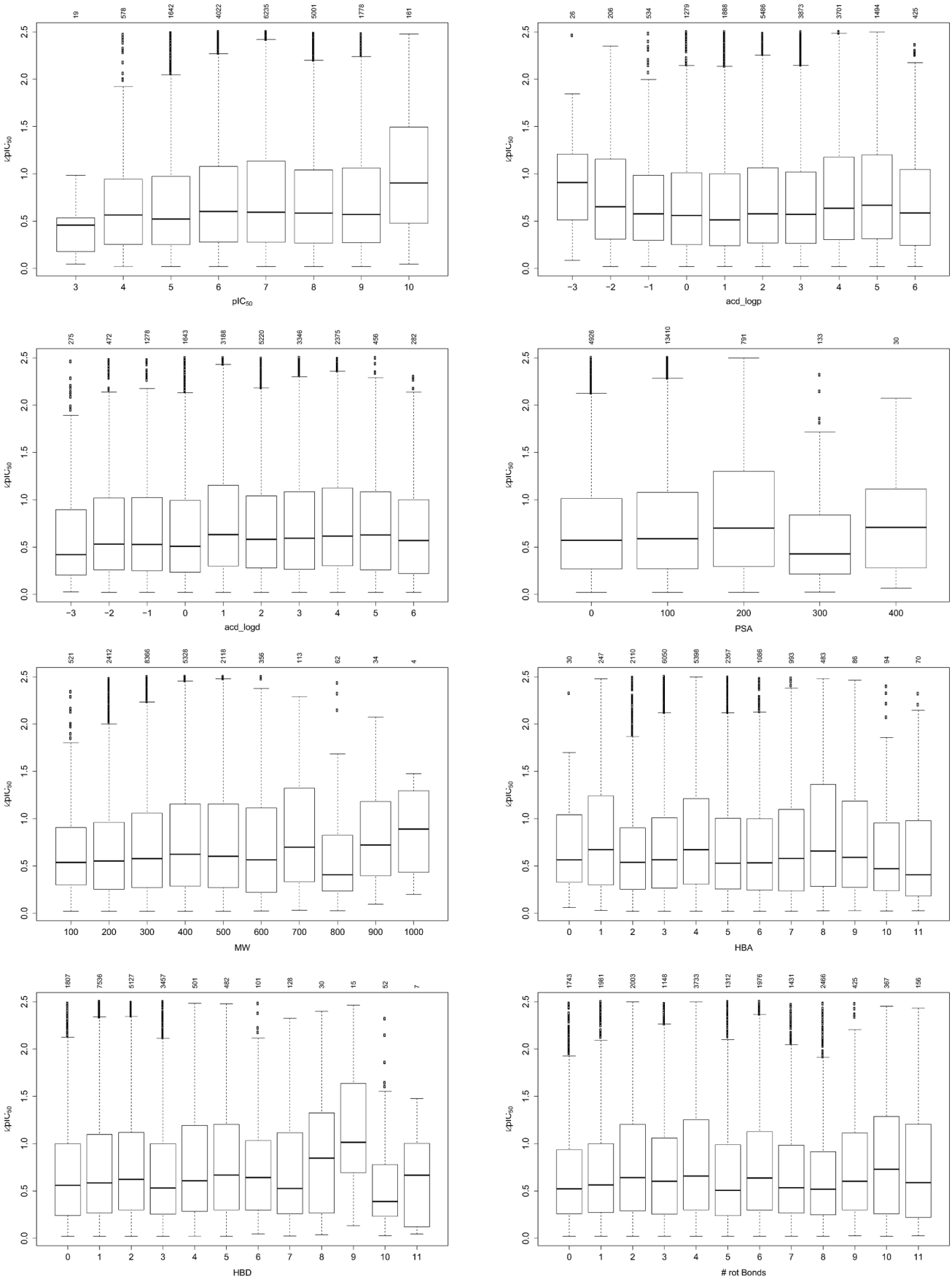
19 578 1642 4022 6235 5001 1778 161
|pIC50|
pIC50
26 206 534 1279 1888 5486 3873 3701 1494 425
|pIC50|
acd_logp
275 472 1278 1643 3188 5220 3346 2375 456 282
|pIC50|
acd_logd
4926 13410 791 133 30
|pIC50|
PSA
521 2412 8366 5328 2118 366 113 62 34 4
|pIC50|
MW
30 247 2110 6050 5398 2357 1086 993 483 86 94 70
|pIC50|
HBA
1807 7536 5127 3457 501 482 101 128 30 15 52 7
|pIC50|
HBD
1743 1981 2003 1148 3733 1312 1976 1431 2466 425 367 156
|pIC50|
rot Bonds

Figure 6. ΔpIC_{50} versus average pIC_{50} measured, logP, logD, polar surface area, molecular weight, number of hydrogen bond acceptors, number of hydrogen bond donors and number of rotatable bonds. The numbers above the boxplot indicate the number of ΔpIC_{50} values falling into the specific bin. Some boxplots are truncated at the very low and high ends because the low number of samples/bin makes the boxplot insignificant.
doi:10.1371/journal.pone.0061007.g006

the same laboratory or the target protein was mistakenly annotated.

For extracting the pairs of IC_{50} data, which are indeed independently measured on the same protein-ligand system, we applied a set of filters that we have previously applied to filter and analyze K_i data. Here, the filters removed more than 90% of the IC_{50} data erroneously assumed to be independent measurements on the same protein-ligand system. When inspecting the remaining 20.356 pairs of measurements from 3.480 protein-ligand systems, we found that there are still a number invalid pairs, especially but not limited to the pairs with larger ΔpIC_{50}. The main errors we found were unit transcription errors, wrong annotation of the receptor subtype, and annotation of cellular assays as biochemical assays. More rarely occurring errors were wrongly assigned stereochemistry, values and protein targets. These errors cannot be automatically detected and have to be manually curated out of the database over time [17].

In contrast to our previous study of K_i values, we observed a larger number of invalid pairs even for smaller ΔpIC_{50} approximately 2.5. To reduce the impact of these hard to find cases, we applied a different strategy to find the variability of the true pairs. By fitting a Gaussian distribution to the central part of the distribution we were able to compare the variability of the pIC_{50} data to the variability of the pK_i data. We found that the ratio between pK_i and pIC_{50} variability is relatively stable between 21 and 26% when varying the upper threshold for fitting the Gaussian distribution between 1.5 and 2.5 ΔpActivity units. Using this approach, we were able to estimate the variability of the IC_{50} data from the variability of the K_i data.

ChEMBL has a confidence score assigned for each activity value. The confidence score indicates how much the ChEMBL authors trust the value reported. Confidence scores below four indicate that the assay was a cellular assay, whereas confidence scores between four and nine indicate biochemical assays. In this study, we used all values that had a confidence score of at least four. The most confident data with a confidence score of nine was also exclusively used, but the results did not change. We also examined, whether there is a difference in data annotated as "autocurated" and data annotated as "expert" data. In this experiment, we also did not find any significant difference. The

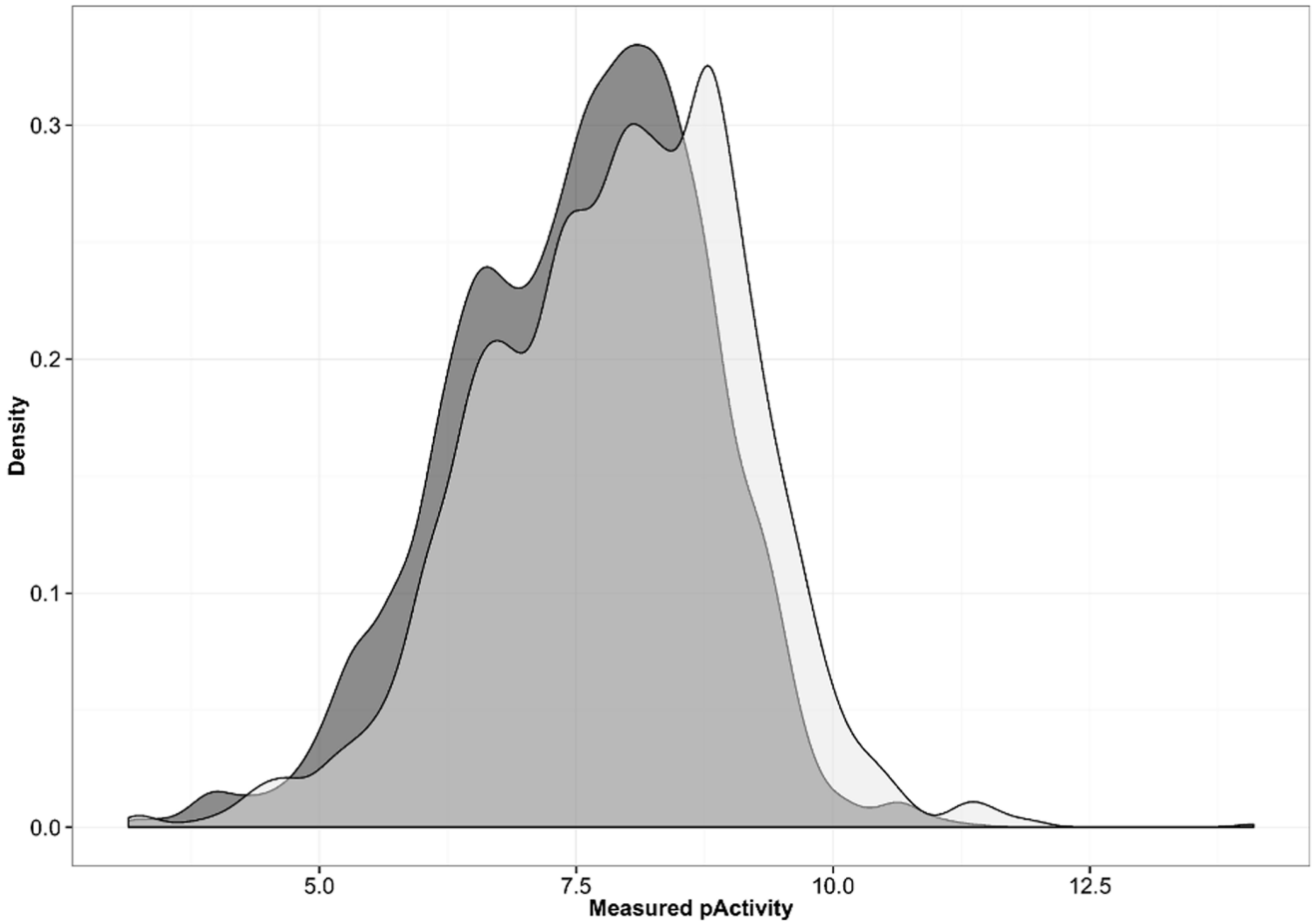

Figure 7. Distribution of published pIC_{50} (dark grey) and pK_i (light grey) values for protein-ligand systems with multiple independent measurements.
doi:10.1371/journal.pone.0061007.g007