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