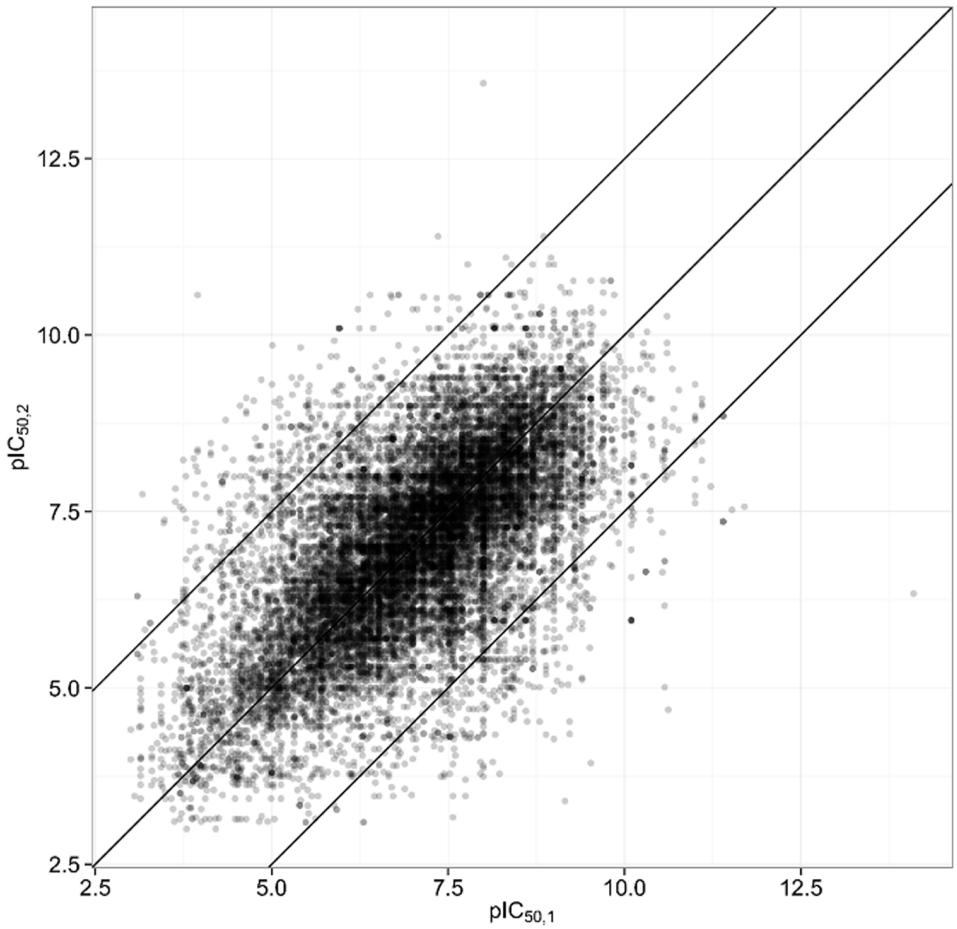

Figure 4. All Pairs of pIC_{50} values extracted from ChEMBL. The two outer diagonal lines indicate the 2.5 log unit threshold, outside which the probability for finding faulty pairs of measurements is very high. The extreme disagreements are all due to clear errors.
doi:10.1371/journal.pone.0061007.g004

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.