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

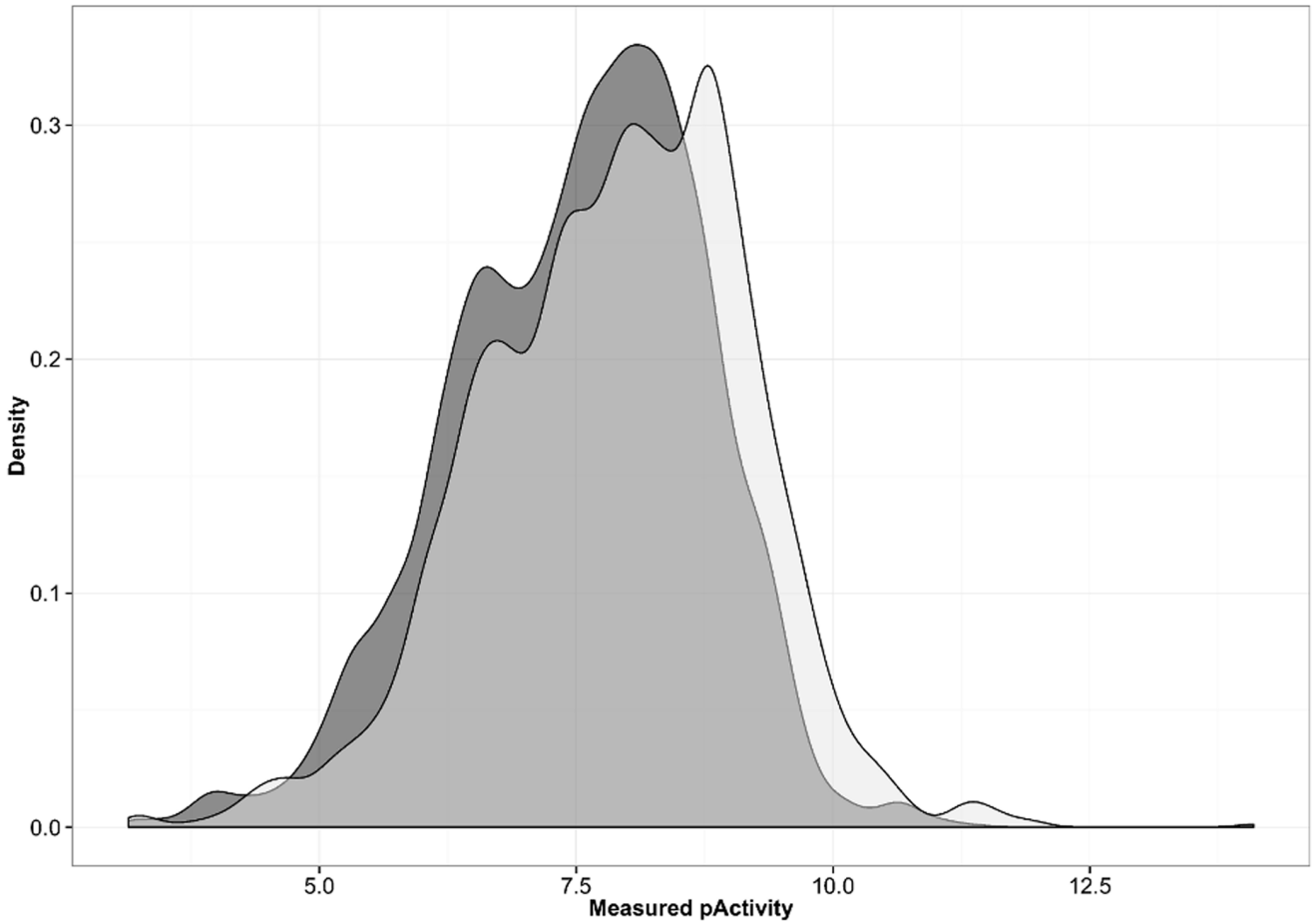

Figure 7. Distribution of published pIC_{50} (dark grey) and pK_i (light grey) values for protein-ligand systems with multiple independent measurements.
doi:10.1371/journal.pone.0061007.g007