PLOS ONE

Comparability of Mixed IC$_{50}$ Data – A Statistical Analysis

Tuomo Kalliokoski*, Christian Kramer*, Anna Vulpetti, Peter Gedeck

Global Discovery Chemistry, Novartis Institutes for Biomedical Research, Basel, Switzerland

Abstract

The biochemical half maximal inhibitory concentration (IC$_{50}$) is the most commonly used metric for on-target activity in lead optimization. It is used to guide lead optimization, build large-scale chemogenomics analysis, off-target activity and toxicity models based on public data. However, the use of public biochemical IC$_{50}$ data is problematic, because they are assay specific and comparable only under certain conditions. For large scale analysis it is not feasible to check each data entry manually and it is very tempting to mix all available IC$_{50}$ values from public database even if assay information is not reported. As previously reported for K$_i$ database analysis, we first analyzed the types of errors, the redundancy and the variability that can be found in ChEMBL IC$_{50}$ database. For assessing the variability of IC$_{50}$ data independently measured in two different labs at least ten IC$_{50}$ data for identical protein-ligand systems against the same target were searched in ChEMBL. As a not sufficient number of cases of this type are available, the variability of IC$_{50}$ data was assessed by comparing all pairs of independent IC$_{50}$ measurements on identical protein-ligand systems. The standard deviation of IC$_{50}$ data is only 25% larger than the standard deviation of K$_i$ data, suggesting that mixing IC$_{50}$ data from different assays, even not knowing assay conditions details, only adds a moderate amount of noise to the overall data. The standard deviation of public ChEMBL IC$_{50}$ data, as expected, resulted greater than the standard deviation of in-house intra-laboratory/inter-day IC$_{50}$ data. Augmenting mixed public IC$_{50}$ data by public K$_i$ data does not deteriorate the quality of the mixed IC$_{50}$ data, if the K$_i$ is corrected by an offset. For a broad dataset such as ChEMBL database a K$_i$- IC$_{50}$ conversion factor of 2 was found to be the most reasonable.

Introduction

Public collections of IC$_{50}$ data (the half maximal inhibitory concentrations of ligands on their protein targets) represent a wealth of knowledge on bioactivity with growing importance. One of the major databases of public bioactivities for small molecules is ChEMBL, [1] which currently contains roughly three times more IC$_{50}$ values than K$_i$ values. It has been shown that the gap between the number of IC$_{50}$ and K$_i$ values is still increasing. [2] Proper usage of IC$_{50}$ data facilitates the development of useful methods for drug discovery. Examples of such applications are the global mapping of pharmacological space by Paolini and co-workers, [3] the Similarity Ensemble Approach (SEA), [4] the Bayesian models for adverse drug reactions by Bender and coworkers, [5] the models used for polypharmacological optimization by Hopkins et al., [6] and the kinome-wide activity modeling studies by Schuerer and Muskal. [7] These methods can be used to predict off-target effects based on heterogeneous public activity data and chemical similarity analysis. Usually, public off-target toxicity models like human Ether-à-go-go-Related Gene (hERG) [8] and cytochrome P450 (CYP) models [9,10] are based and validated on mixed public IC$_{50}$ data, since there is not enough public data available that originates from one single assay.

In contrast to K$_i$ values, IC$_{50}$ data is assay specific. For the simplest typical case of competitive monosubstrate enzyme inhibition, K$_i$ can be calculated from the IC$_{50}$ according to the Cheng-Prusoff equation:

$$K_i = \frac{IC_{50}}{1 + \frac{|S|}{K_m}}$$

where |S| is the substrate concentration and K$_m$ is the Michaelis-Menten constant of the substrate. [11] Under the same assay conditions the measured IC$_{50}$ of same inhibitor or two different inhibitors (1 and 2 below) with the same mechanism of action can be compared as

$$\frac{K_{i,1}}{K_{i,2}} = \frac{IC_{50,1}}{IC_{50,2}}$$

The problem is that assay details are not reported in public bioactivity databases. Recently, Zdrazil et al. analyzed human P-glycoprotein bioassay data from the ChEMBL and TP-search databases. [12] They explore the ability of these data, determined in different assays, to be combined with each other. Their study indicates that for inhibitors of human P-glycoprotein this is possible under certain conditions: i.e., data coming from the same type of assay, same cell lines, and also same fluorescent or radiolabeled substrates with overlapping binding sites. However they point out that it is currently not possible to extract such data in automated fashion from the current public databases. Effort in

annotating assay details would increase the capabilities of safe data integration thus increasing the usefulness of those huge data repositories freely available.

In this manuscript we report an estimate of the error introduced by mixing public IC$_{50}$ data from different labs and how this can affect the capability of drawing scientifically sound conclusions from such data. By using the same statistical technique that we have previously introduced to determine the experimental uncertainty of heterogeneous public K$_i$ data [13] we analyze the variability of all pairs of biochemical IC$_{50}$ measurements on the same protein-ligand system independently of assay details.

In the following, we first describe our attempts in extracting a set of at least ten IC$_{50}$ values from ChEMBL that have independently been measured in two comparable assays. Since all sets of identified measurements turn out to be not independent or otherwise faulty, we analyze the standard deviation of all truly independent pairs of IC$_{50}$ values available from ChEMBL. Dubious entries and filters used to spot and remove faulty entries are described in detail. For the remaining pairs of measurements, the original publications of protein-ligand systems showing various ranges of IC$_{50}$ differences were inspected in order to gain an impression of which activity differences are due to database errors and which activity differences are due to the variations in assay conditions. We then fitted a Gaussian distribution to the distribution of IC$_{50}$ differences to estimate the standard deviation of valid pairs of independent IC$_{50}$ measurements. By comparing the IC$_{50}$ standard deviation to the equivalent K$_i$ standard deviation, we can estimate the variability of heterogeneous IC$_{50}$ data. The average difference between K$_i$ and IC$_{50}$ values and their correlation are assessed. Moreover the effect of mixing K$_i$ and IC$_{50}$ values in order to enlarge the data size was evaluated. Lastly, we analyze whether the variability of IC$_{50}$ values depends on simple ligand properties such as molecular weight (MW) and the calculated octanol –water partition coefficient (logP).

Materials and Methods

Dataset Preparation

All measurements were extracted for the ChEMBL database version 14. It is the currently largest public database with bioactivities extracted from the literature. BindingDB [14] is similar in size, but has a significant overlap with ChEMBL with most of the values being copied from ChEMBL.

The raw data was filtered in order to remove erroneous entries as described earlier. [13] Generally, all analyses presented here are based on multiple affinity measurements of the same protein-ligand system. The filtering steps were the following:

1. Remove all data from reviews, since this is not original data.
2. Remove all unclear measurements (i.e. Unit not M, mM, µM, nM, pM, fM; qualified values ("<" or ">"); extremely high (pActivity >15) or extremely low (pActivity <2) values).
3. Remove younger entry for exactly the same value reported twice (younger paper cites older paper).
4. Remove younger entry for very close values reported twice (difference in pActivity <0.02: younger paper cites older paper and rounds).
5. Remove both entries if their difference is exactly 3, 6, or 9. These are citations with unit-conversion errors.
6. Remove entries for which the authors could not be extracted from PubMed.

7. Only keep pairs where the name overlap of the authors is zero to make sure that measurements are from different laboratories.

After each step, protein-ligand systems that had only one measurement entry left (singletons) were removed. All affinity were converted to their negative logarithm pActivity (e.g. pIC$_{50}$ or pK$_i$) with M^{-1} as base unit (e.g. 1 µM is converted to 6 [log Activity units]).

In ChEMBL a confidence score is available for each bioactivity entry. According to the ChEMBL homepage, a confidence score of nine is the highest, a confidence score of four or more indicates a biochemical measurement and a confidence score below four indicates a cellular measurement. For the IC$_{50}$ analysis, two sets of data were generated: Set1 contains all data with a confidence score of four and more, Set2 contains data with the highest confidence score nine only. Since it turned out that there is no difference in variability between Set1 and Set2, here we only report results for Set1.

From the initially available 616.555 IC$_{50}$ values with confidence score greater or equal to four 10.895 IC$_{50}$ values for 3.480 Protein/Ligand systems remained, yielding 20.356 pairs of independent measurements. Overall, the number of both protein/ligand systems and individual IC$_{50}$ data points available for comparisons has been reduced by 94% and 93%. The filtering statistics is shown in Table 1.

Metrics for Evaluating the Distribution of Errors

We analyze the distribution of the differences between two affinity measurements on the same protein-ligand system using the Standard Deviation (σ), the Mean Unsigned (Absolute) Error (MUE), the Median Unsigned Error (M$_{ed}$UE) the squared Pearson's correlation coefficient (R$^2_{pearson}$ = R^2). They are defined as

$$MUE = \frac{1}{n\sqrt{2}} \sum_{i=1}^{n} |y_{pub,i,1} - y_{pub,i,2}|$$

$$M_{ed}UE = \frac{1}{\sqrt{2}} median\{|y_{pub,i,1} - y_{pub,i,2}| \, for \, i \, in \, 1...n\}$$

$$\sigma = \sqrt{\frac{1}{2(n-1)} \sum_{i=1}^{n} \left(y_{pub,i,1} - y_{pub,i,2}\right)^2}$$

$$R^2_{Pearson} = \frac{\sum_{i=1}^{n} \left(y_{pub,i,1} - \bar{y}_{pub,1}\right)\left(y_{pub,i,2} - \bar{y}_{pub,2}\right)}{\sqrt{\sum_{i=1}^{n} \left(y_{pub,i,1} - \bar{y}_{pub,1}\right)^2} \sqrt{\sum_{i=1}^{n} \left(y_{pub,i,2} - \bar{y}_{pub,2}\right)^2}}$$

$$\bar{y}_{pub,1} = \frac{1}{n} \sum_{i=1}^{n} y_{pub,i,1}; \quad \bar{y}_{pub,2} = \frac{1}{n} \sum_{i=1}^{n} y_{pub,i,2}$$

with n being the number of pairs of measurements considered, y$_{pub,i,1}$ and y$_{pub,i,2}$ being the two published values of pair i and $_{pub}$ is the average of all measured values. If more than two measurements are available for a given protein-ligand system, all possible pairs are generated. The order of y$_{pub,i,1}$ and y$_{pub,i,2}$ has to
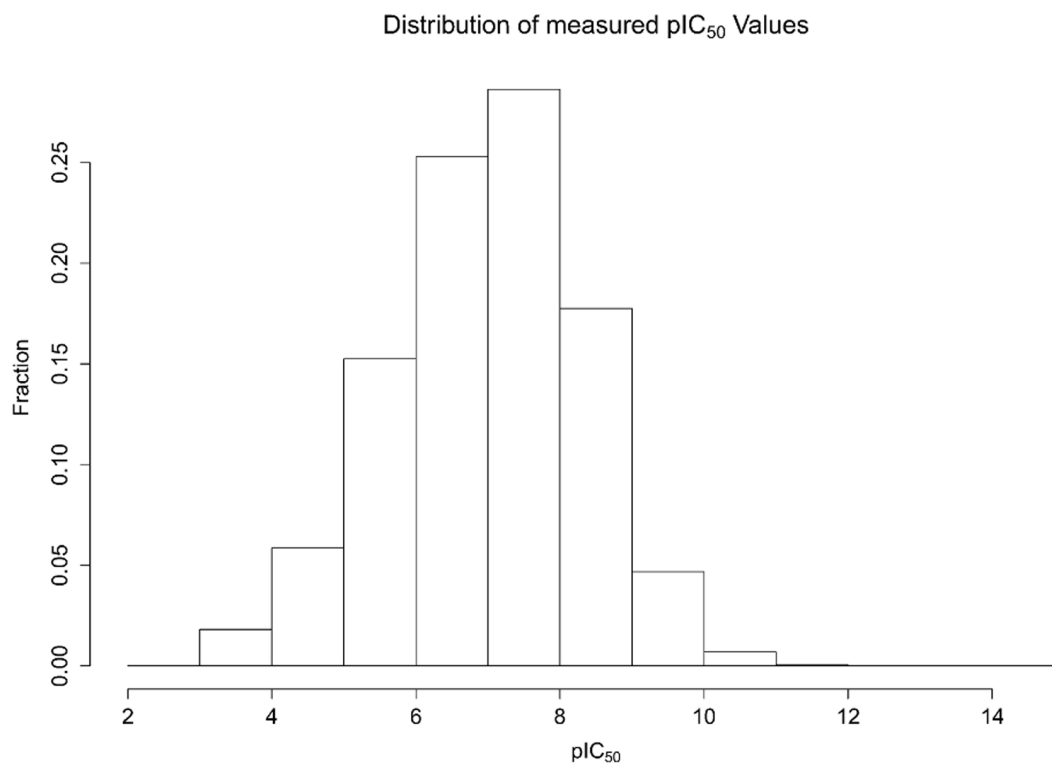
Distribution of measured pIC$_{50}$ Values

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
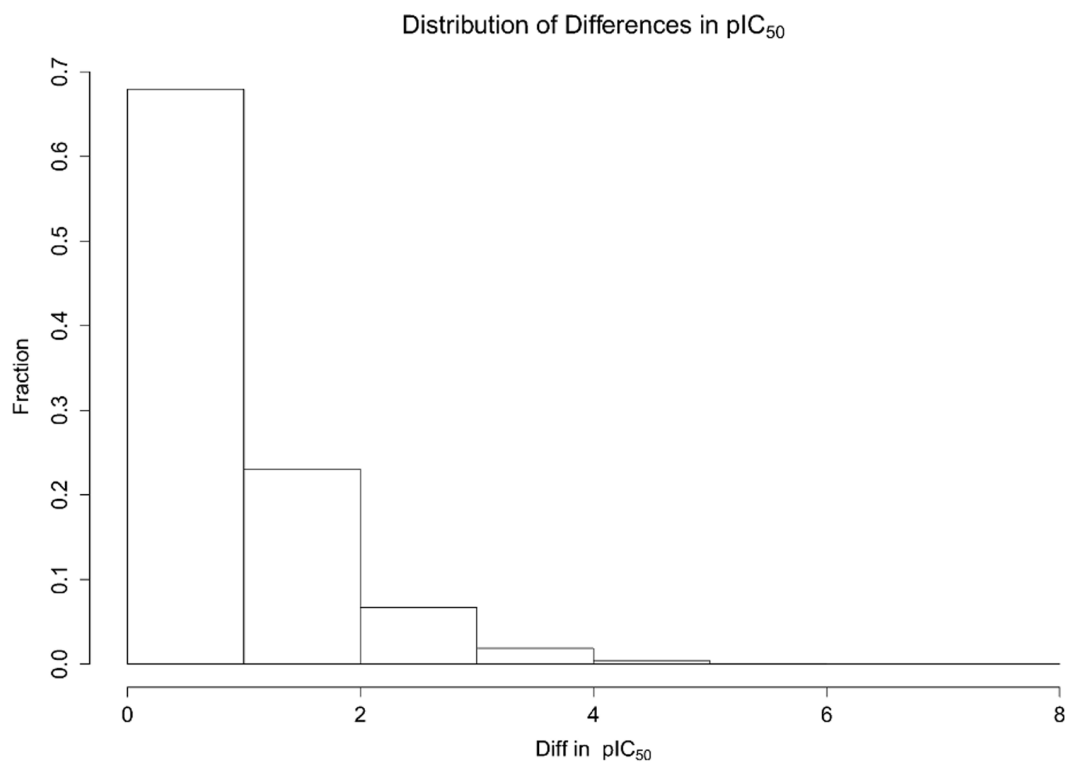
Distribution of Differences in pIC$_{50}$

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.
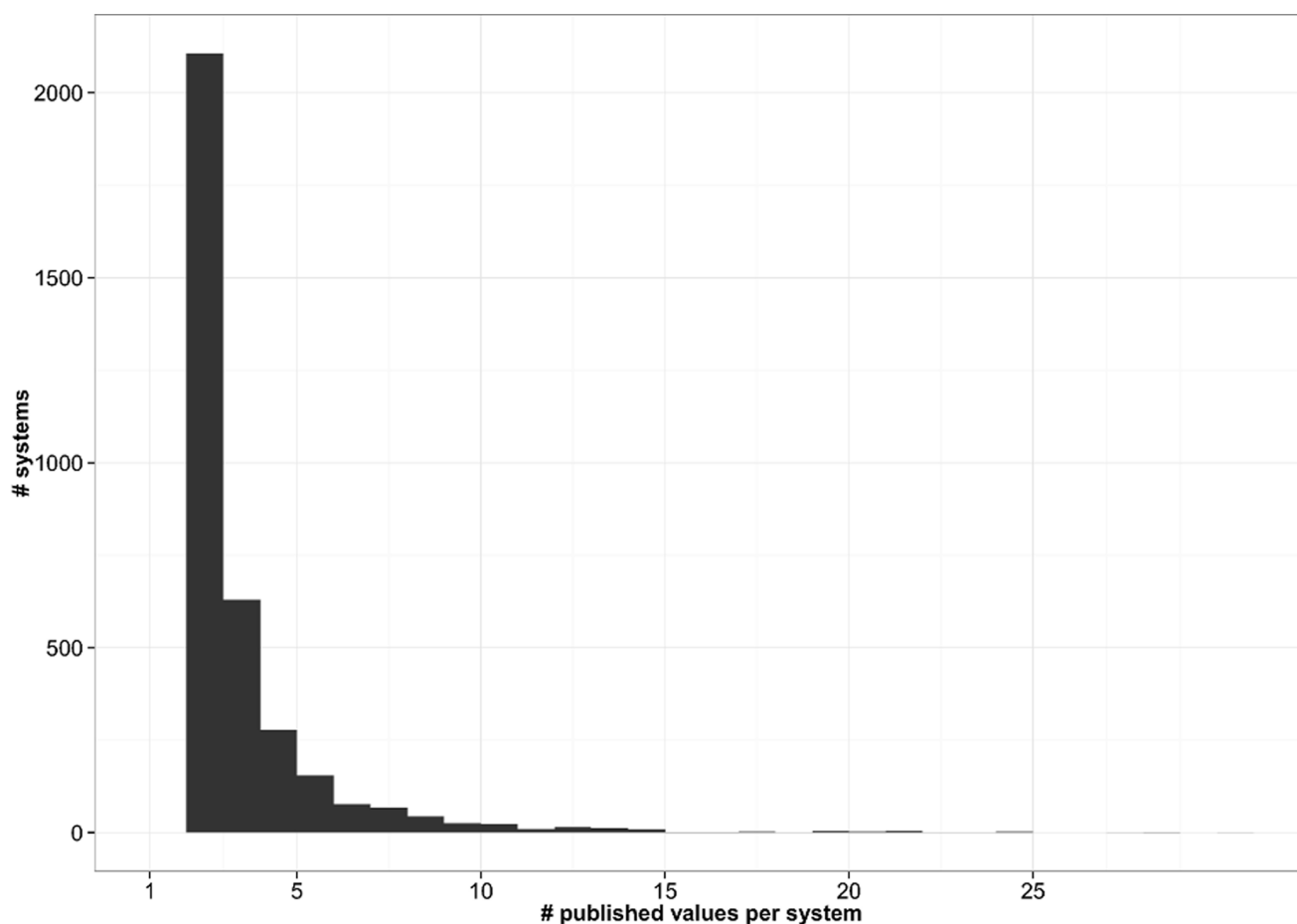doi:10.1371/journal.pone.0061007.g002

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 σ$_{pIC50}$ = 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 ΔpIC$_{50}$ = 2.5. Here, σ$_{pIC50}$ = 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

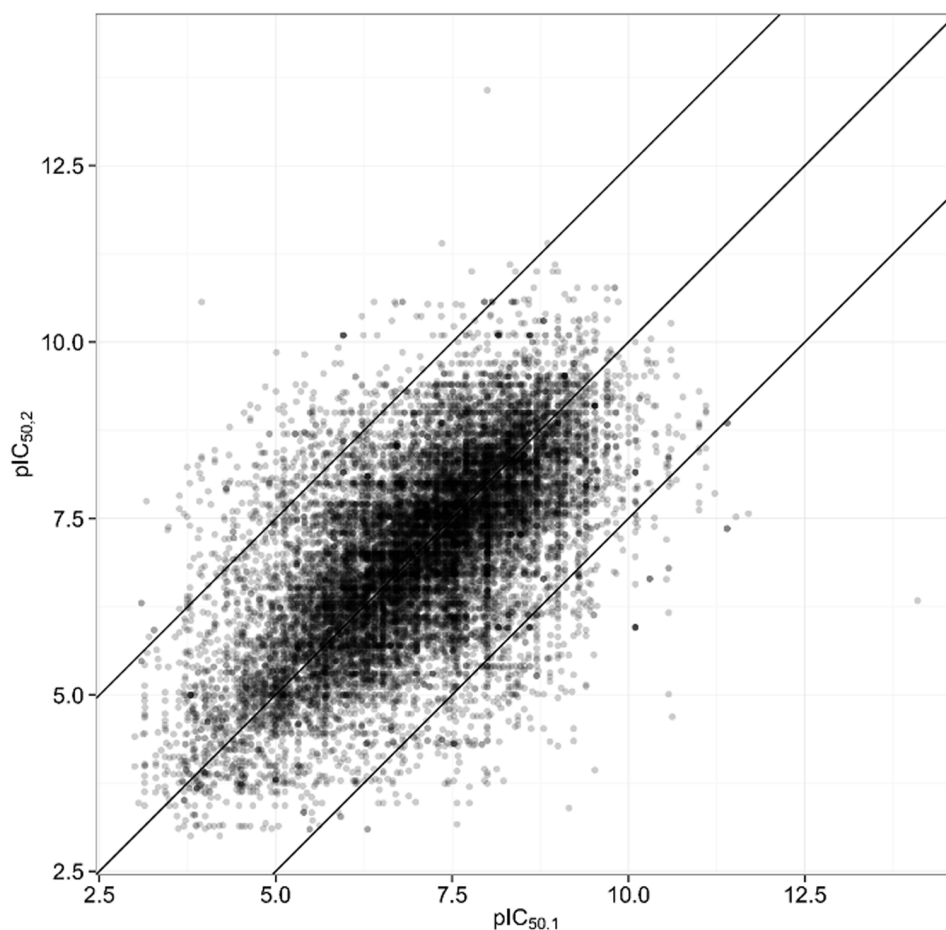doi:10.1371/journal.pone.0061007.t002

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).

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.

Table 3. Standard deviation of a Gaussian distribution fitted to the inner part of the distribution of ΔpIC$_{50}$ and ΔpK$_i$.

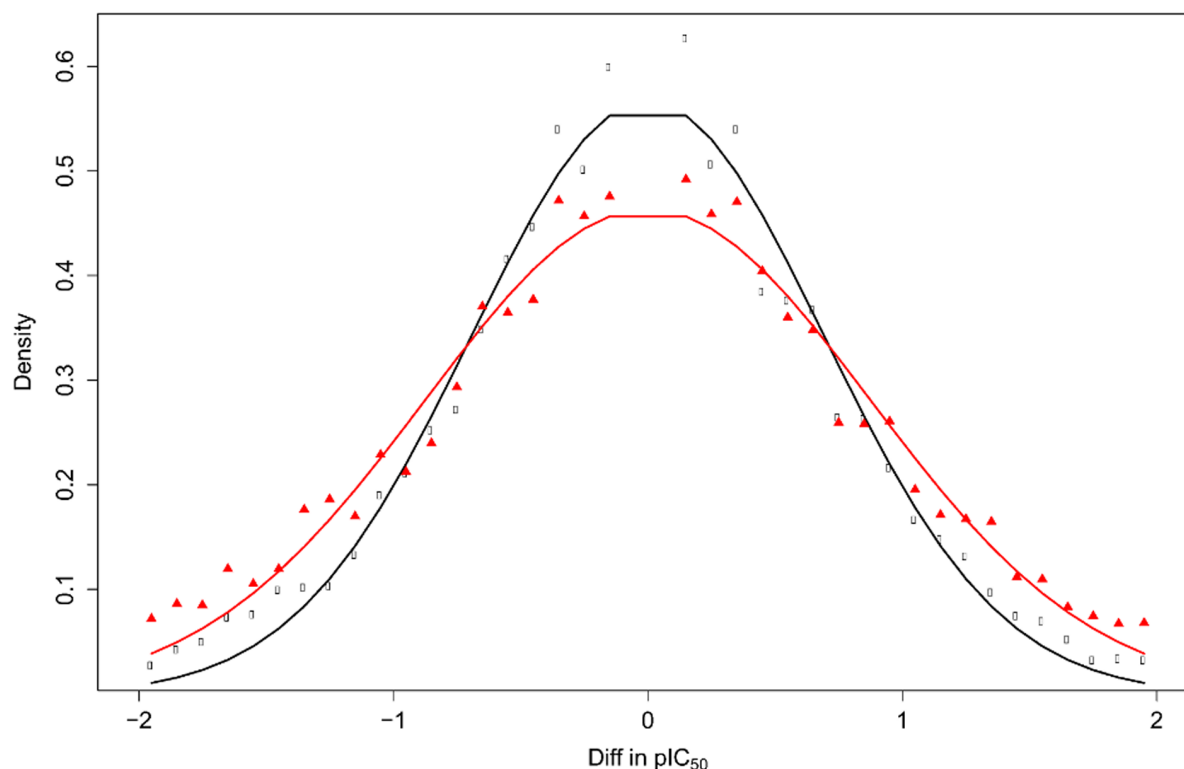| Upper threshold	1.5	2.0	2.5
ΔpIC$_{50}$	$\sigma = 0.80$	$\sigma = 0.84$	$\sigma = 0.86$
ΔpK$_i$	$\sigma = 0.66$	$\sigma = 0.68$	$\sigma = 0.68$

doi:10.1371/journal.pone.0061007.t003

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 σ$_{pIC50}$ = 0.87 and σ$_{pKi}$ = 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 √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 √2, pK$_i$ and pIC$_{50}$ values agree with an R^2 = 0.46, σ = 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 σ$_{pIC50}$ = 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 σ = 0.22 for rolipram/PDE4D and σ = 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
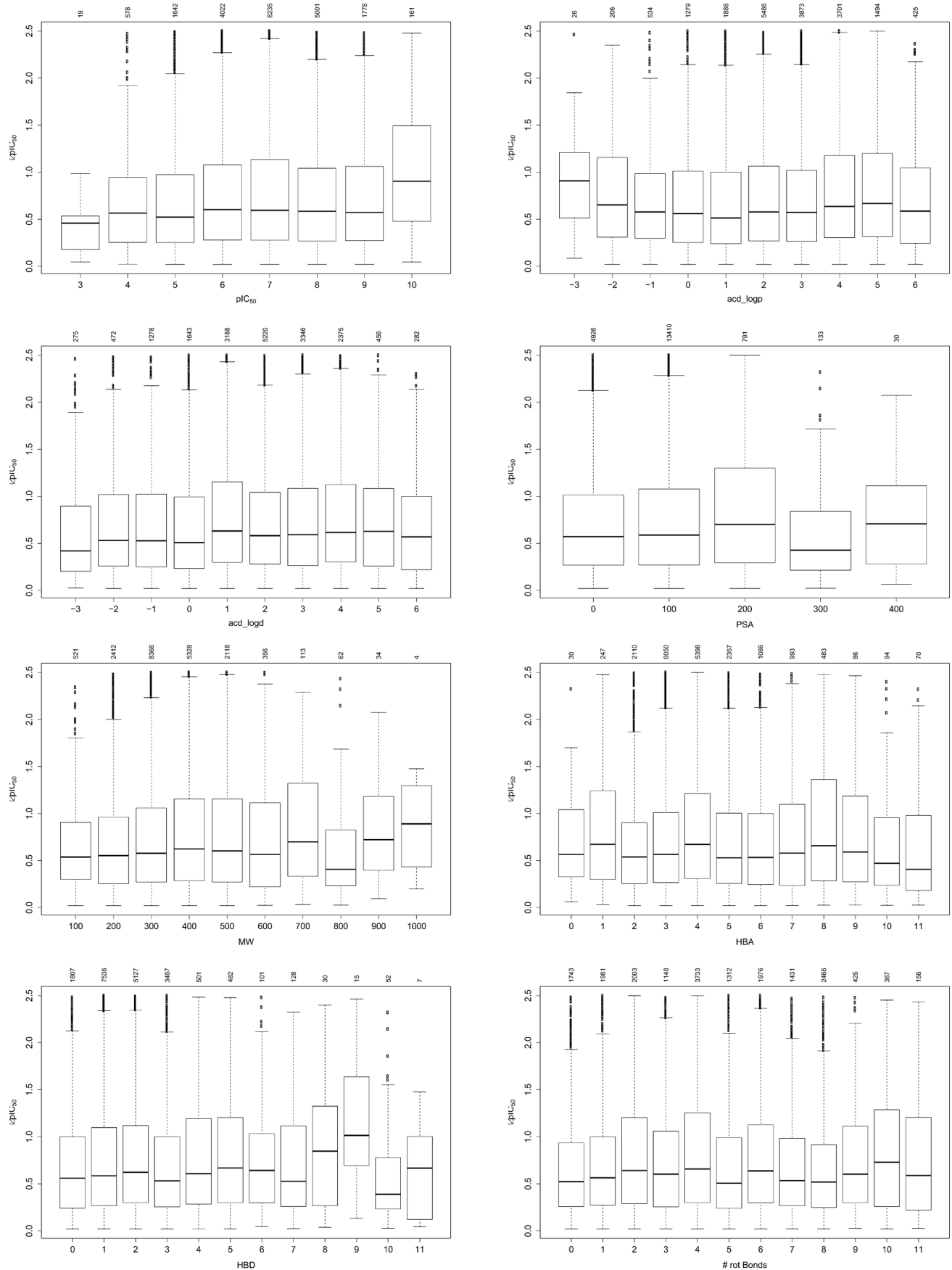
Figure 6. ΔpIC$_{50}$ versus average pIC$_{50}$ measured, logP, logD, polar surface area, molecular weight, number of hydrogen bond acceptors, number of hydrogen bond donors and number of rotatable bonds. The numbers above the boxplot indicate the number of ΔpIC$_{50}$ values falling into the specific bin. Some boxplots are truncated at the very low and high ends because the low number of samples/bin makes the boxplot insignificant.

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
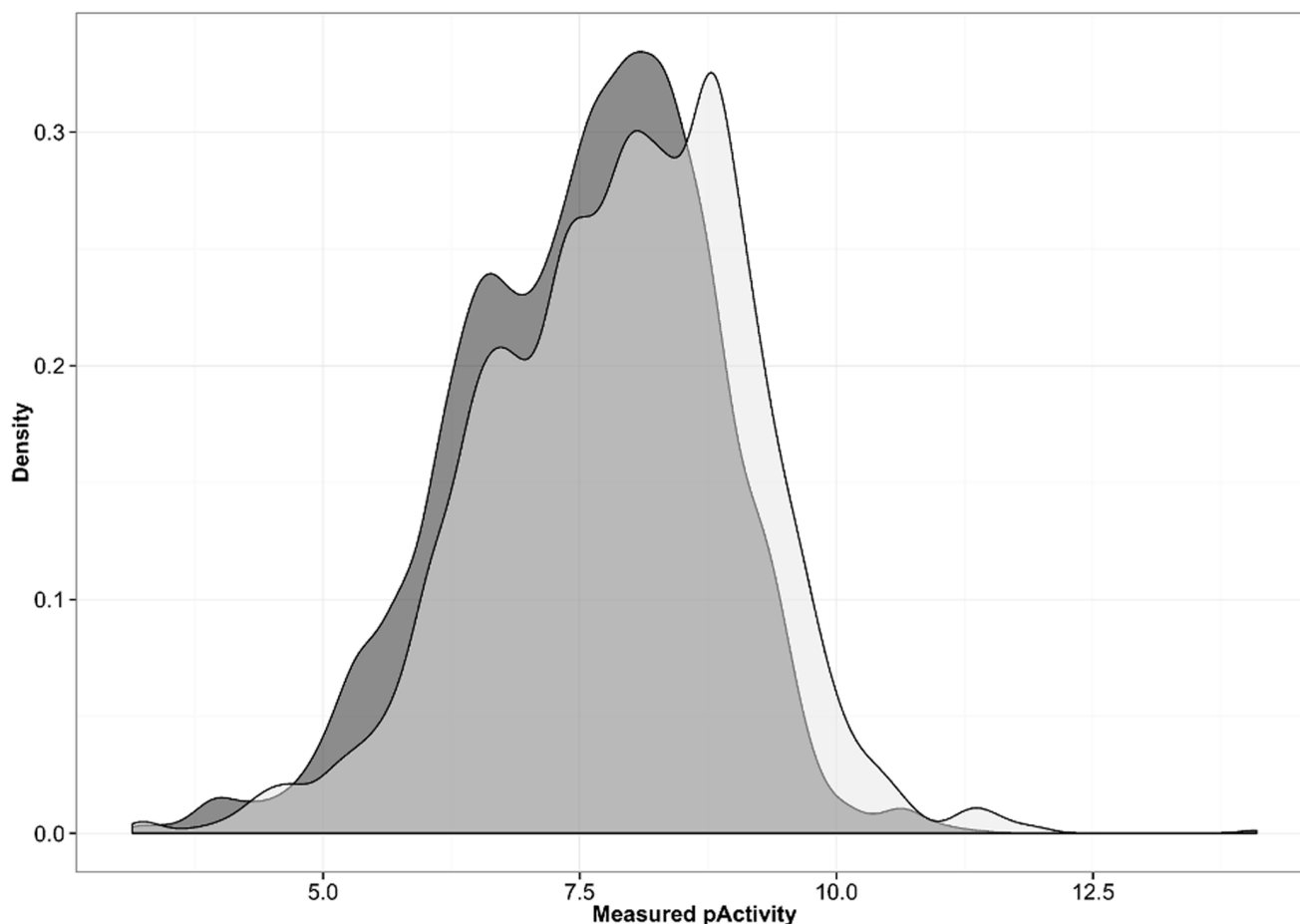

Figure 7. Distribution of published pIC$_{50}$ (dark grey) and pK$_i$ (light grey) values for protein-ligand systems with multiple independent measurements.
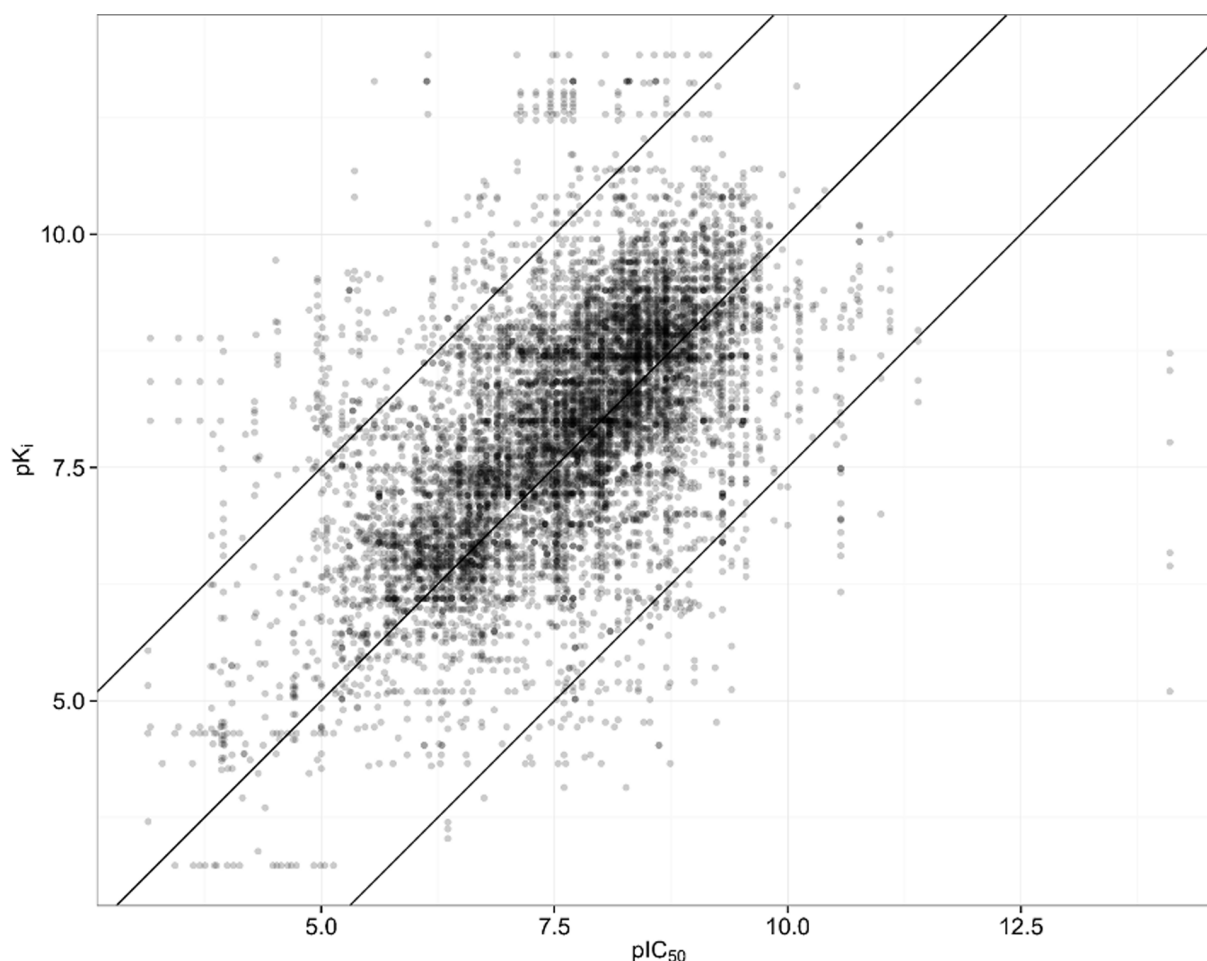
Figure 8. Measured pK$_i$ versus measured pIC$_{50}$ for identical protein-ligand systems.
doi:10.1371/journal.pone.0061007.g008

availability of assay description within ChEMBL would have allowed the analysis of whether specific assay types are statistically better comparable than other assay types or if the variability of pIC$_{50}$ is lower in comparable assays. However, such information is not easily added to the database because this would require detailed assay ontologies and in the original literature assay details are often missing as well.

One might assume that higher IC$_{50}$ values show a larger variability than for example single digit µM IC$_{50}$ values because of solubility limits. However, our analysis shows that on the average this is clearly not the case. Moreover, the variability does not depend on any specific ligand properties such as logP, MW, PSA etc.

While the quality of pure K$_i$ datasets would be reduced by adding IC$_{50}$ data, we have shown that augmenting IC$_{50}$ datasets by K$_i$ data does not deteriorate the quality, if the K$_i$ data is corrected by an offset. We found that pK$_i$ values reported in ChEMBL are on average 0.35 log units higher than pIC$_{50}$ values, which corresponds to a factor of 2.3. The IC$_{50}$ to K$_i$ conversion factor is exactly 2.0 in competitive monosubstrate IC$_{50}$ inhibition assays, if the substrate concentration is set equal to its K$_m$ value. This factor is close to the average difference between pKi and pIC$_{50}$ values in ChEMBL and therefore in absence of any further specific assay knowledge available, a factor of 2.0 is the most probable conversion factor to convert K$_i$ values to IC$_{50}$ values.

Summary and Conclusions

In this contribution, we present an analysis of the comparability of public heterogeneous IC$_{50}$ data. We find that the agreement of independently measured biochemical IC$_{50}$ values is only 23–30% worse than the agreement of pK$_i$ data, irrespective to the used condition and type of assay. For heterogeneous biochemical pIC$_{50}$ data, we find a variability with $\sigma_{pIC50} = 0.68$, $MUE_{pIC50} = 0.55$ and $M_{ed}UE_{pIC50} = 0.43$. Although theoretically IC$_{50}$ values with different assay conditions should not be comparable, this is common practice in analyzing large-scale off-target and toxicity datasets. Our analysis quantitatively assesses the consequence in doing so. We believe that this knowledge should be important for everybody who decides to work with IC$_{50}$ data from various heterogeneous sources. We also show that K$_i$ data can be used to augment IC$_{50}$ datasets without any loss of quality if corrected by a factor of 2, which is the conversion factor most frequently found by comparing the IC$_{50}$/K$_i$ values in ChEMBL for the same protein-ligand systems.

Nevertheless, public IC$_{50}$ data extracted from ChEMBL14 is quite error prone. The most common errors we found are unit conversion errors, receptor subtype errors and errors in mixing up biochemical and cellular assay. The data quality is good enough to build large-scale fishing tools where errors partially cancel each other out, but for detailed SAR analysis and methods based on individual or very few data points like activity cliff or matched pair

Variation of in-house measured rolipram/PDE4D pIC$_{50}$
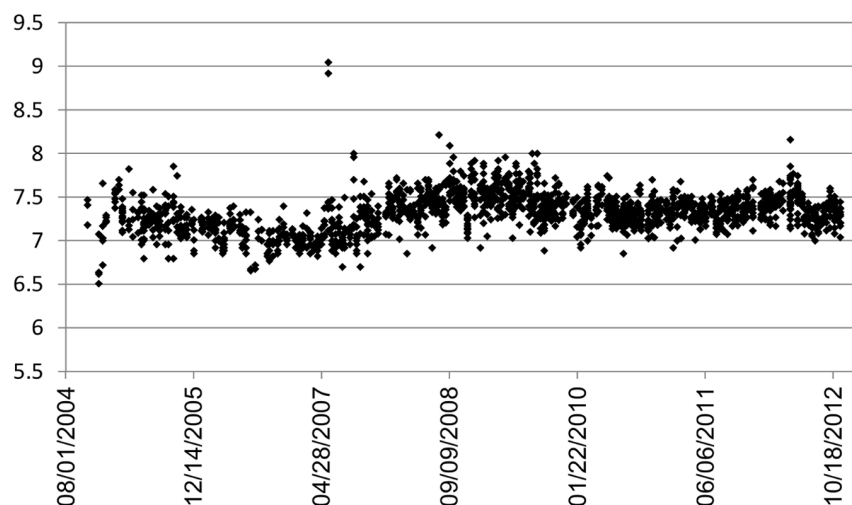
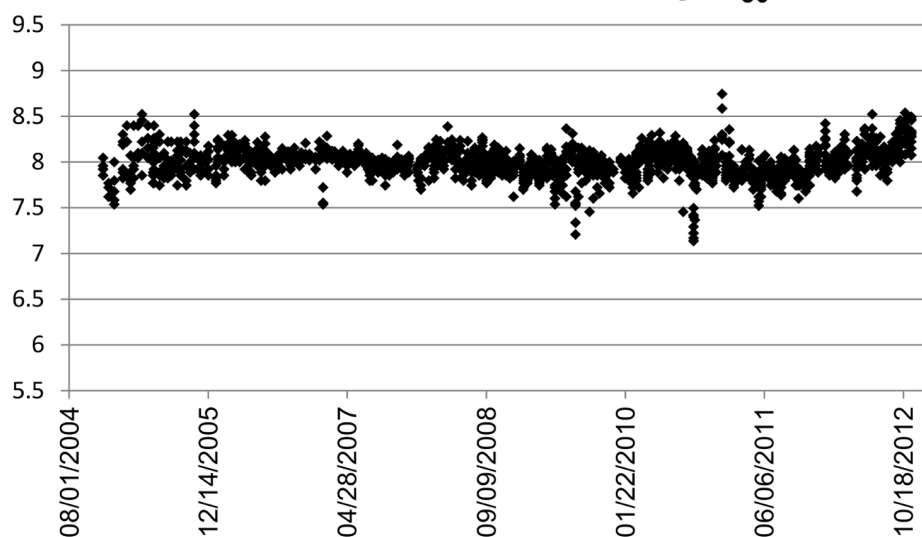## Variation of in-house measured cilostamide/PDE3 pIC$_{50}$

Figure 9. Variation of measured pIC$_{50}$ values over time for rolipram/PDE4D and cilostamide/PDE3.
doi:10.1371/journal.pone.0061007.g009

analysis it is mandatory to take recourse to the original literature and ensure that the values are correctly annotated and comparable.

This work augments our previous work where we focused on the experimental uncertainty of heterogeneous public K$_i$ data. As we have previously stated, it is likely the data quality will rise over time by continuous iterative improvement of the large databases such as ChEMBL and BindingDB. In a different branch of affinity databases, smaller high-quality affinity databases, potentially combined with other physicochemical data or structural knowledge are being built up (see for example the CSARdock challenge

[18,19]). It will also be interesting to see what the reproducibility of such high-quality data is going to be.

It is surprising that we did not find in ChEMBL a single set of at least ten inhibitors for which IC$_{50}$ values on the same target has been independently measured by different laboratories or a scientific contribution in literature addressing the comparison of heterogeneous IC$_{50}$ values. Due to the scarcity of details about the experimental assay setup in both original publications and current large activity databases it is not possible to systematically analyze the comparability of the reproducibility of IC$_{50}$ data for the same assay or various assay types under the same conditions. Using in-house data we were able to estimate the interlab

reproducibility of IC$_{50}$ for the same assay under the same conditions.

We hope that with this article we increase the awareness of noise added during mixing blindly public IC$_{50}$ values during the data selection process for SAR analysis and QSAR models and its impact in limiting the maximal achievable performance of these techniques.

Supporting Information

Figure S1   Agreement of IC$_{50}$ values for two dopamine transporter assays, measured in the same laboratory. Here the pairs of measurements agree quite well with an R^2 of 0.70 and a mean error of 0.29. According to the assay description of the primary literature, the assay conditions have been the same. The same is true for the norepinephrine transporter assay (R^2 = 0.73, MUE = 0.29).
(DOCX)

Figure S2   Agreement of IC$_{50}$ values for two rattus norvegicus dihydrofolate reductase assays, measured in the same laboratory. Although the assays have been run in

the same lab on DHFR from the same species, the IC$_{50}$ values of rattus norvegicus DHFR agree with R^2 = 0.25 and MUE = 0.61.
(DOCX)

Figure S3   Median ΔpIC$_{50}$, binned according to average activity and logP. The numbers indicate the number of entries per bin. We do not see a clear trend in this plot.
(DOCX)

Table S1   All series where more than ten compounds have been measured in two parallel assays.
(DOCX)

Text S1   Closer inspection of Table S1.
(DOCX)

Archive S1   Python- and R-scripts to repeat the analysis.
(GZ)

Author Contributions

Conceived and designed the experiments: TK CK AV PG. Performed the experiments: TK CK. Analyzed the data: TK CK AV PG. Contributed reagents/materials/analysis tools: TK CK. Wrote the paper: TK CK AV PG.

References

1. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40: D1100–D1107.
2. Hu Y, Bajorath J (2012) Growth of Ligand–Target Interaction Data in ChEMBL Is Associated with Increasing and Activity Measurement-Dependent Compound Promiscuity. J Chem Inf Model 52: 2550–2558.
3. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24: 805–815.
4. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, et al. (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25: 197–206.
5. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, et al. (2007) Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. ChemMedChem 2: 861–873.
6. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, et al. (2012) Automated design of ligands to polypharmacological profiles. Nature 492: 215–220.
7. Schürer SC, Muskal SM (2013) Kinome-wide Activity Modeling from Diverse Public High-Quality Data Sets. J Chem Inf Model. 53: 27–38.
8. Kramer C, Beck B, Kriegl JM, Clark T (2008) A Composite Model for hERG Blockade. ChemMedChem 3: 254–265.
9. Kirchmair J, Williamson MJ, Tyzack JD, Tan L, Bond PJ, et al. (2012) Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. J Chem Inf Model 52: 617–648.
10. McCarren P, Bebernitz GR, Gedeck P, Glowienke S, Grondine MS, et al. (2011) Avoidance of the Ames test liability for aryl-amines via computation. Bioorg Med Chem 19: 3173–3182.
11. Cheng Y, Prusoff WH (1973) Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochem Pharmacol 22: 3099–3108.
12. Zdrazil B, Pinto M, Vasanthanathan P, Williams AJ, Balderud LZ, et al. (2012) Annotating Human P-Glycoprotein Bioassay Data. Mol Inform 31: 599–609.
13. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A (2012) The Experimental Uncertainty of Heterogeneous Public Ki Data. J Med Chem 55: 5165–5173.
14. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35: D198–D201.
15. Team RC (2012) R: A Language and Environment for Statistical Computing. Vienna, Austria. Available: http://www.R-project.org.
16. Sahoo PK, Behera P (2010) Synthesis and biological evaluation of [1,2,4]tria-zino[4,3-a] benzimidazole acetic acid derivatives as selective aldose reductase inhibitors. Eur J Med Chem 45: 909–914.
17. Kramer C, Lewis R (2012) QSARs, data and error in the modern age of drug discovery. Curr Top Med Chem 12: 1896–1902.
18. Dunbar JB, Smith RD, Yang C-Y, Ung PM-U, Lexa KW, et al. (2011) CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. J Chem Inf Model 51: 2036–2046.
19. Smith RD, Dunbar JB, Ung PM-U, Esposito EX, Yang C-Y, et al. (2011) CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. J Chem Inf Model 51: 2115–2131.