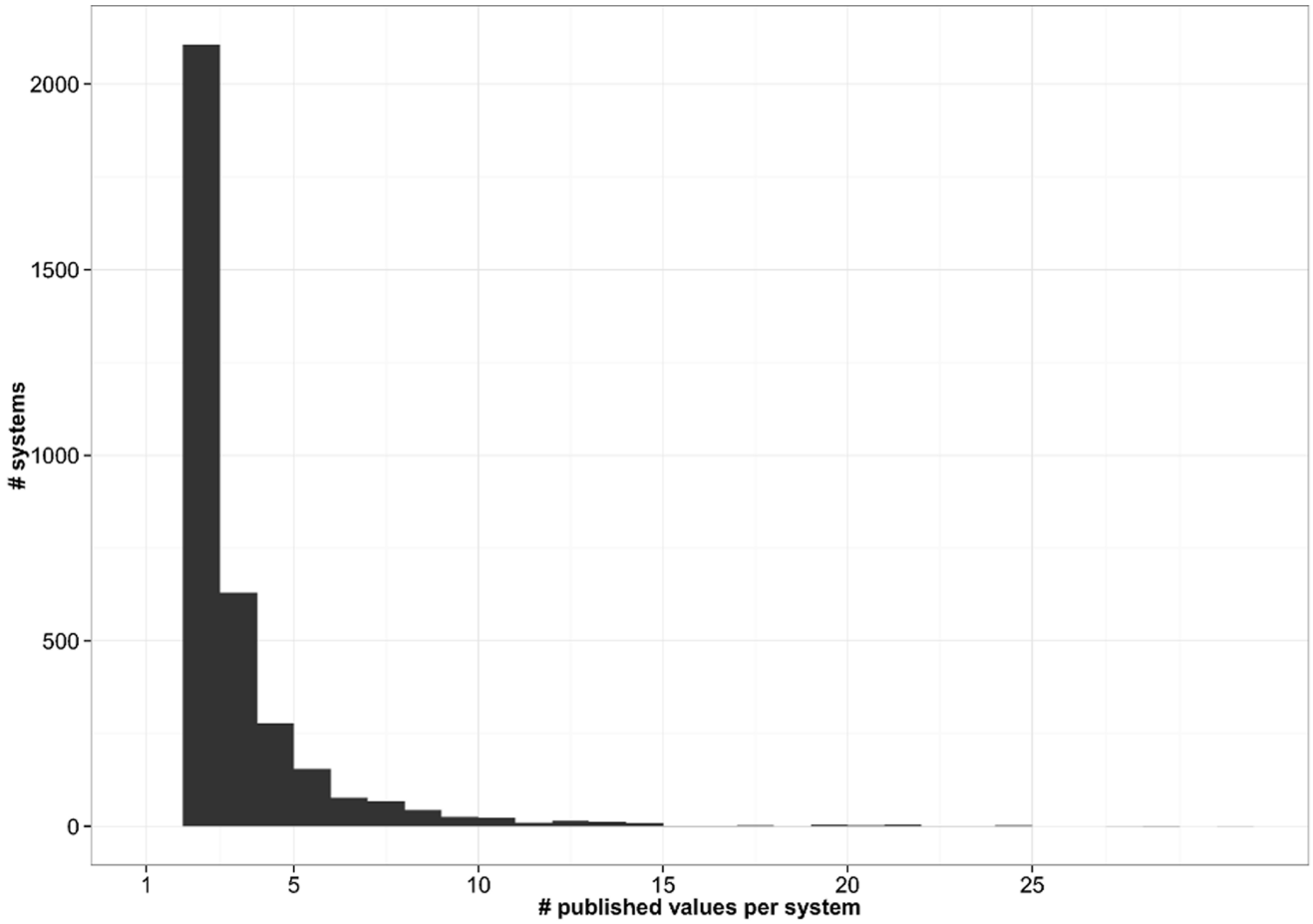

Figure 3. Number of published independent values per protein-ligand system.
doi:10.1371/journal.pone.0061007.g003

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

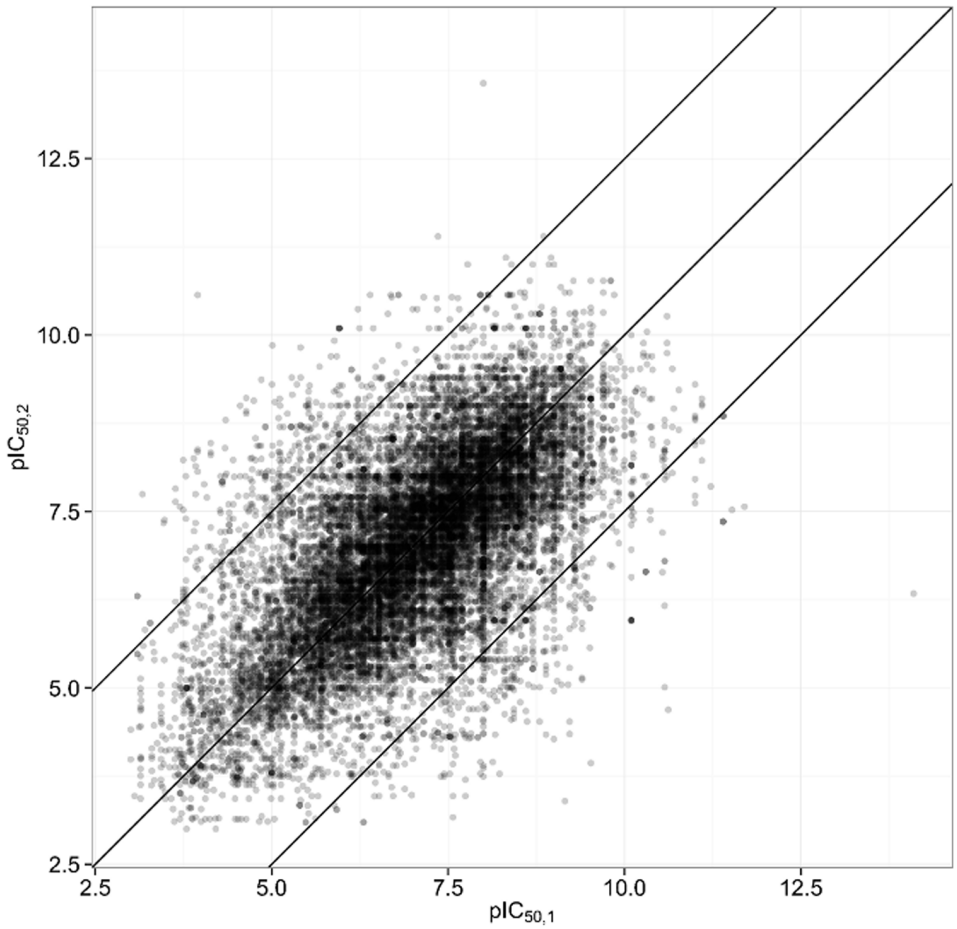

Figure 4. All Pairs of pIC_{50} values extracted from ChEMBL. The two outer diagonal lines indicate the 2.5 log unit threshold, outside which the probability for finding faulty pairs of measurements is very high. The extreme disagreements are all due to clear errors.
doi:10.1371/journal.pone.0061007.g004

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.

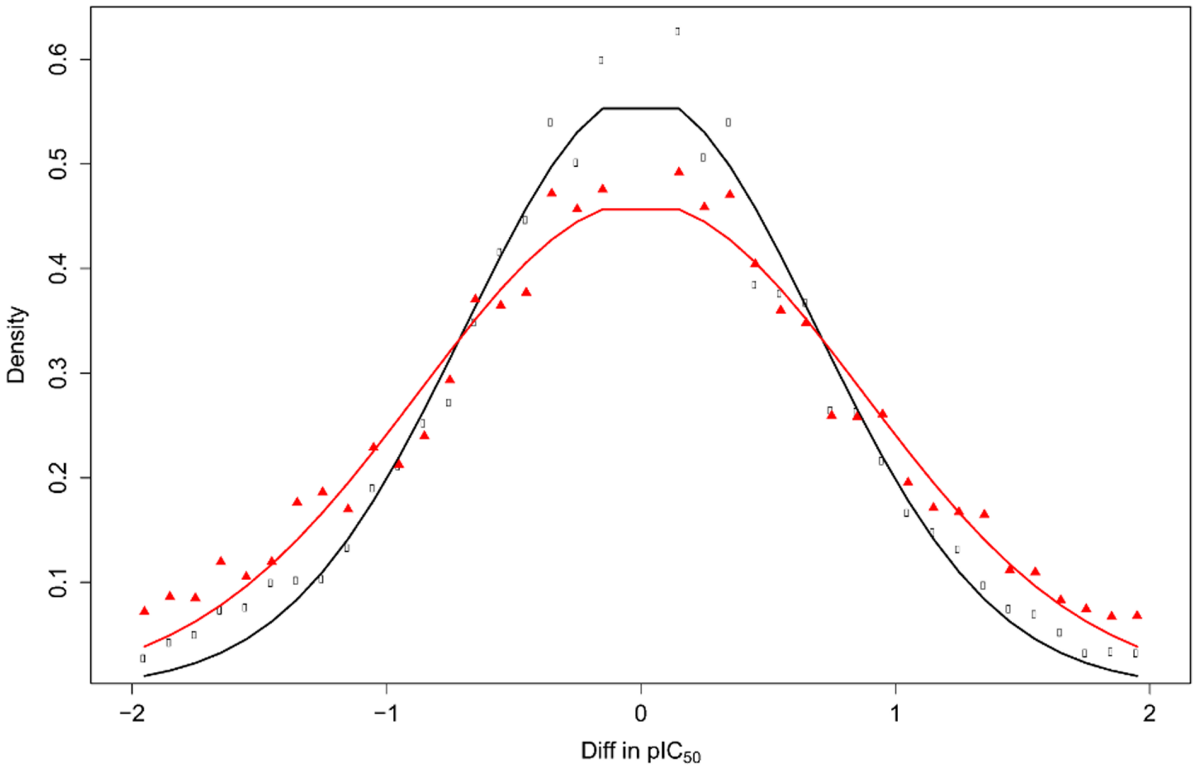

Figure 5. Fitted Gaussian distribution of ΔpIC_{50} (red) and ΔpK_i (black). The Gaussian distributions shown were fitted to all ΔpActivity values with an upper threshold ΔpActivity = 2.0. Standard deviations for the fitted Gaussian distributions are $\sigma_{pIC50} = 0.87$ and $\sigma_{pKi} = 0.69$. 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

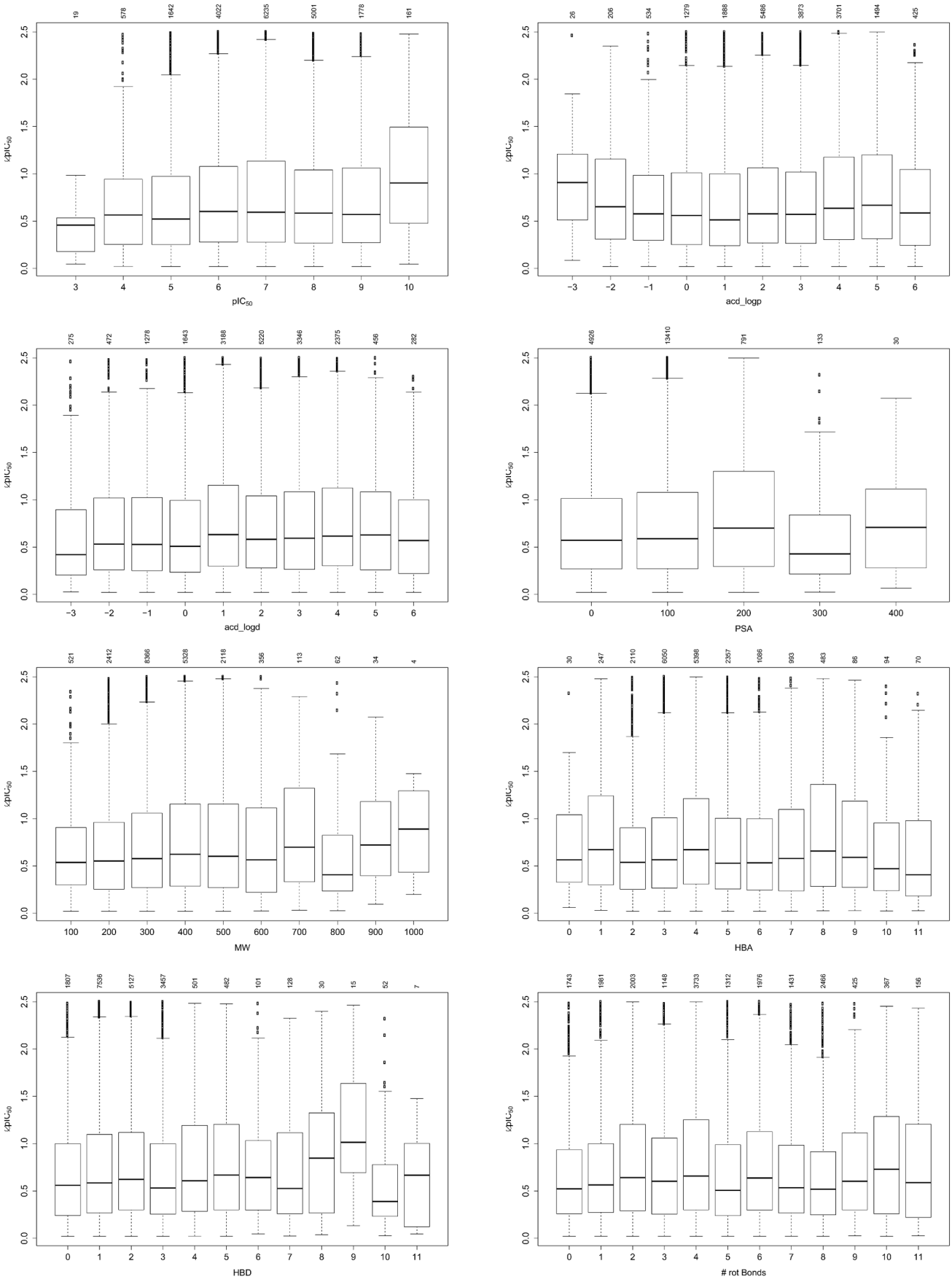
19 578 1642 4022 6235 5001 1778 161
|pIC50|
pIC50
26 206 534 1279 1888 5486 3873 3701 1494 425
|pIC50|
acd_logp
275 472 1278 1643 3188 5220 3346 2375 456 282
|pIC50|
acd_logd
4926 13410 791 133 30
|pIC50|
PSA
521 2412 8366 5328 2118 366 113 62 34 4
|pIC50|
MW
30 247 2110 6050 5398 2357 1086 993 483 86 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

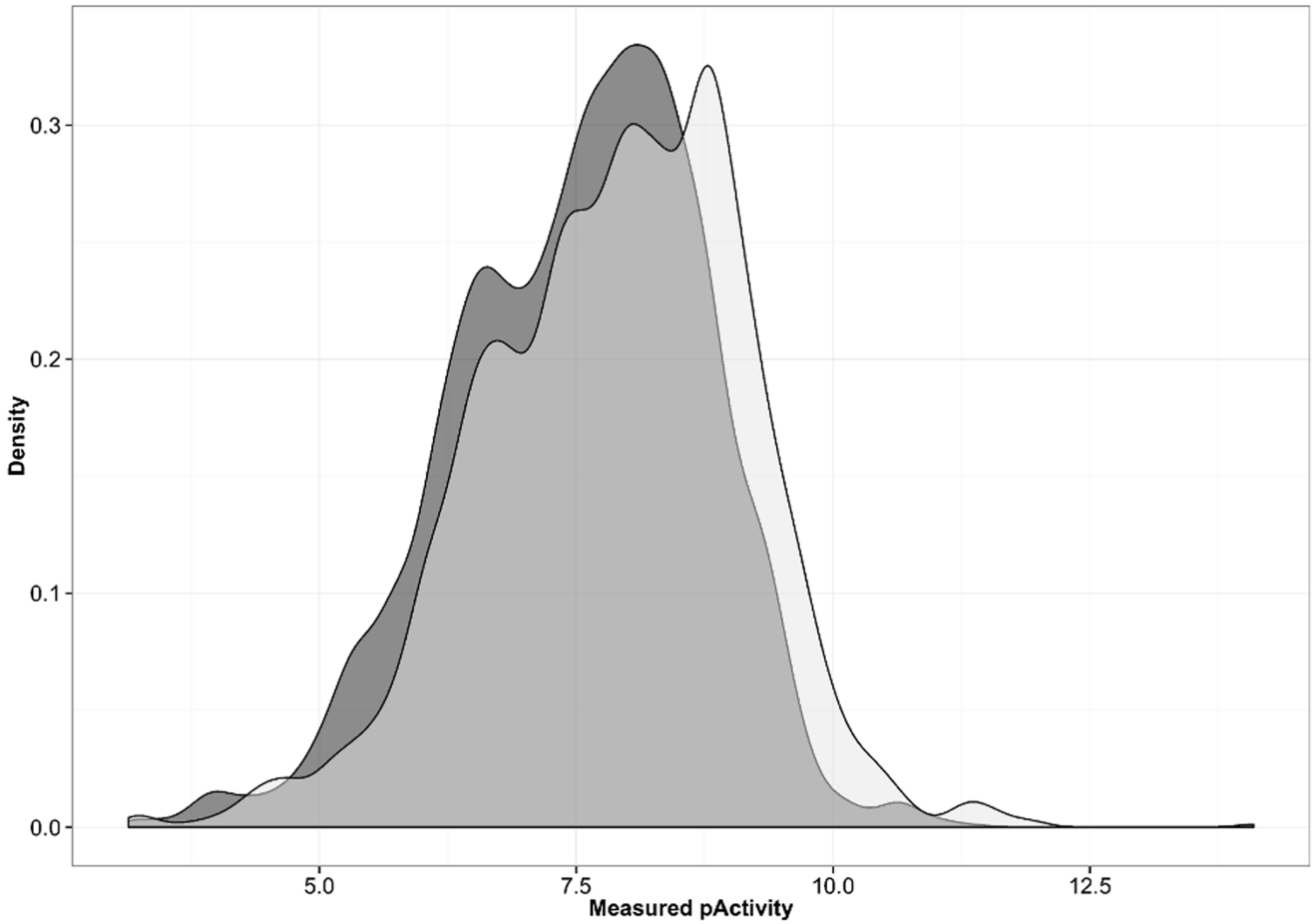

Figure 7. Distribution of published pIC_{50} (dark grey) and pK_i (light grey) values for protein-ligand systems with multiple independent measurements.
doi:10.1371/journal.pone.0061007.g007

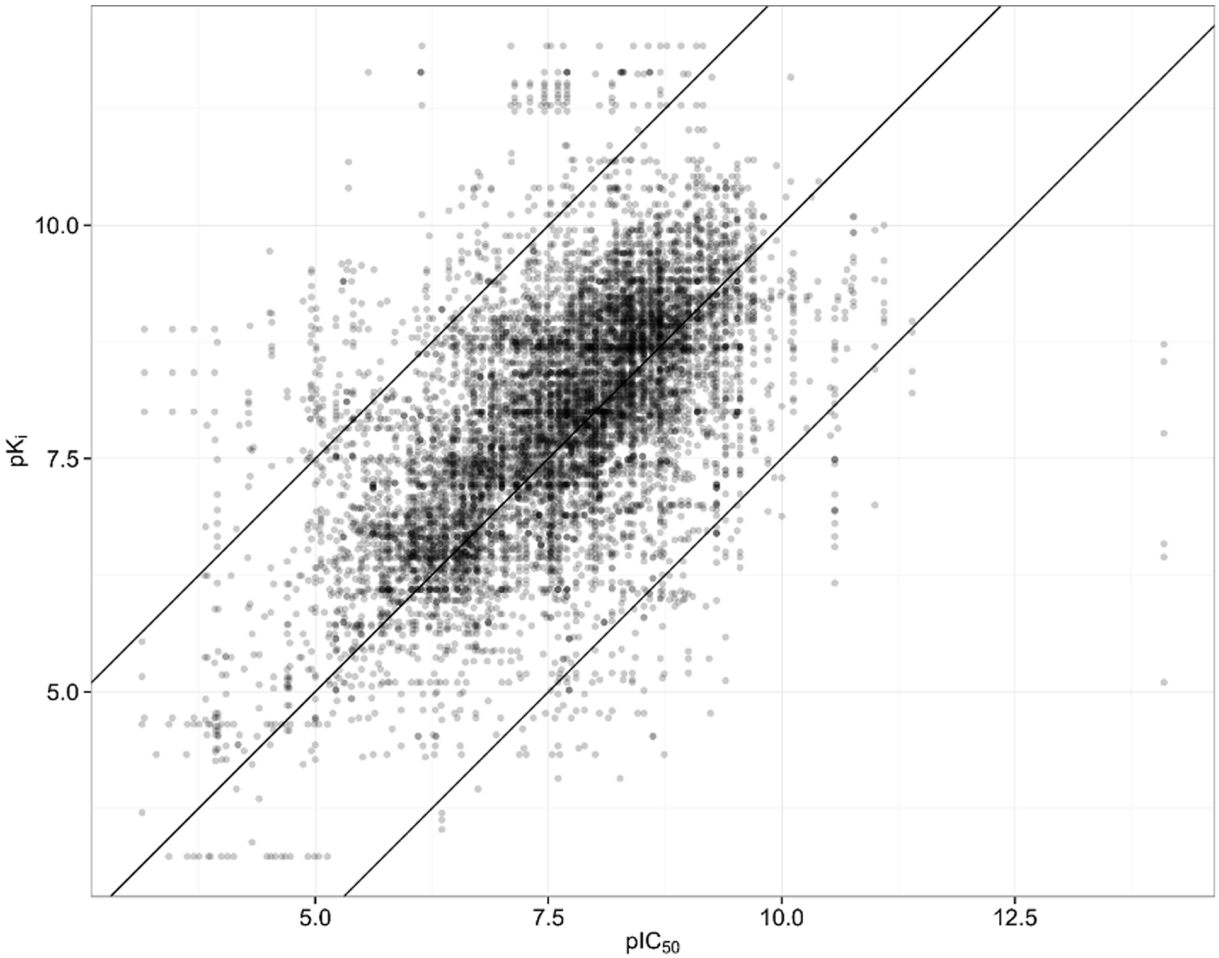

Figure 8. Measured pK_i versus measured pIC_{50} for identical protein-ligand systems.
doi:10.1371/journal.pone.0061007.g008

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

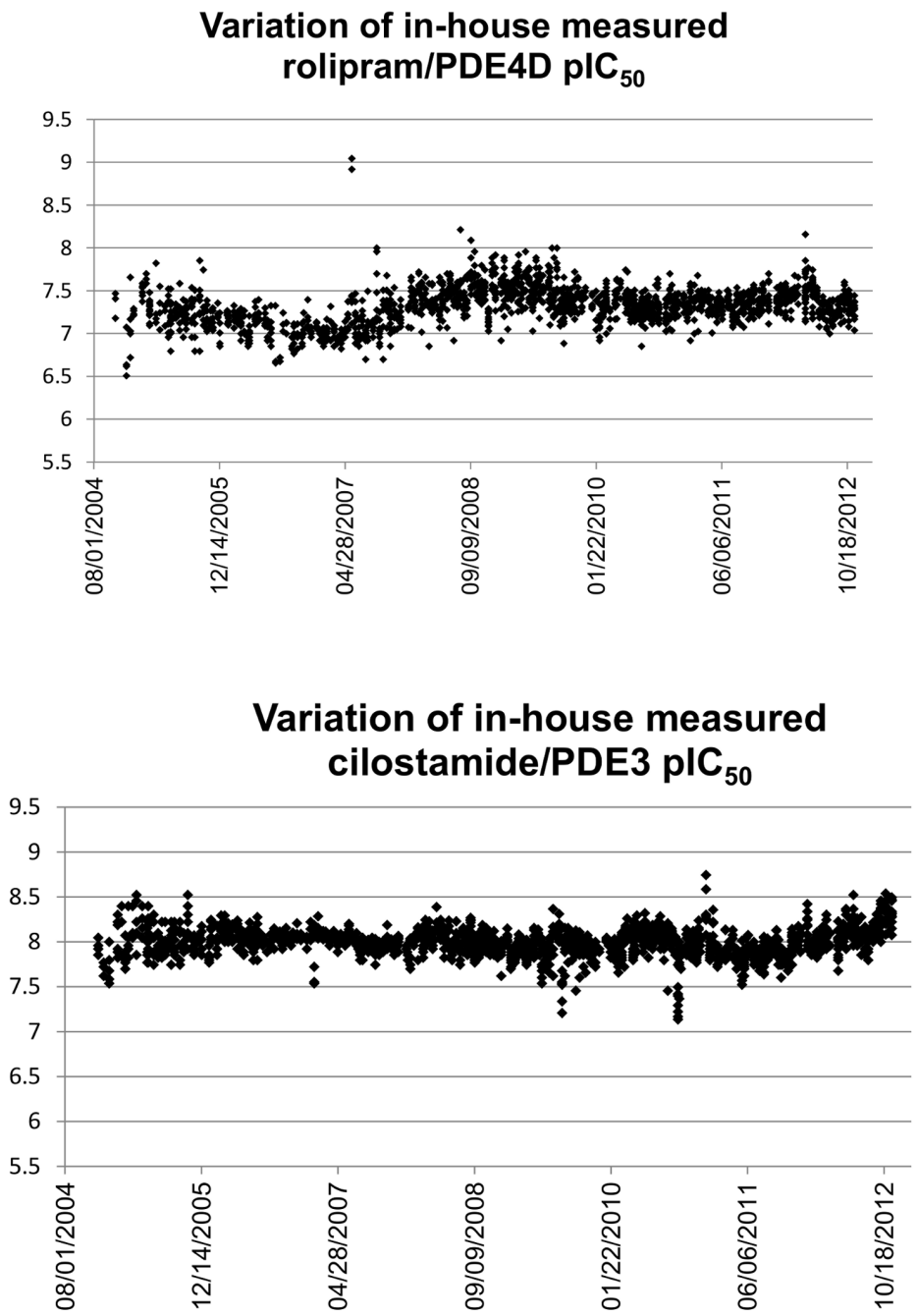

Figure 9. Variation of measured pIC_{50} values over time for rolipram/PDE4D and cilostamide/PDE3.
doi:10.1371/journal.pone.0061007.g009

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]triazino[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.