Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

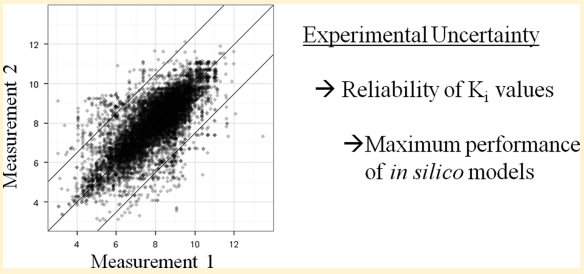
The Experimental Uncertainty of Heterogeneous Public K_i Data

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

Novartis Institutes for BioMedical Research, Novartis Pharma AG, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland

Supporting Information

ABSTRACT: The maximum achievable accuracy of in silico models depends on the quality of the experimental data. Consequently, experimental uncertainty defines a natural upper limit to the predictive performance possible. Models that yield errors smaller than the experimental uncertainty are necessarily overtrained. A reliable estimate of the experimental uncertainty is therefore of high importance to all originators and users of in silico models. The data deposited in ChEMBL was analyzed for reproducibility, i.e., the experimental uncertainty of independent measurements. Careful filtering of the data was required because ChEMBL contains unit-transcription errors, undifferentiated stereoisomers, and repeated citations of single measurements (90% of all pairs). The experimental uncertainty is estimated to yield a mean error of 0.44 pK_i units, a standard deviation of 0.54 pK_i units, and a median error of 0.34 pK_i units. The maximum possible squared Pearson correlation coefficient (R^2) on large data sets is estimated to be 0.81.



INTRODUCTION

Knowing the experimental uncertainty of the biological measurements from which models are derived and validated is necessary in order to judge the quality of in silico models.[1] The experimental uncertainty sets the upper limit of performance of in silico models that can be achieved. Recent review papers on the state of molecular modeling have pointed out that the experimental uncertainty in public data is an important factor that needs to be assessed more thoroughly.[2,3] Although this is a simple fact, a systematic estimate on measured experimental uncertainties in the context of in silico models of biological data has never been published.

Biochemical studies often report a standard error of the measurements. This error commonly expresses the *repeatability*, i.e., the variability of the measurements obtained by one person repeatedly measuring one system using the same experimental setup. For judging models based on heterogeneous data sets, it is important to know the variability of the measurements caused by differences in operator behavior, lab conditions, or experimental setup. This is defined as *reproducibility*. The International Union of Pure and Applied Chemistry's (IUPAC) definition of reproducibility is: "The closeness of agreement between independent results obtained with the same method on identical test material but under different conditions."[4] Although data published by different laboratories has usually not been measured by the same method on identical test material, in silico models and large-scale SAR analyses rarely differentiate between the sources of biological material and the experimental methods used. In this context, "identical material" has to be interpreted as "identical target protein" and "same method" as "method yielding the same physical constant".

From a mathematical point of view, reproducibility is the variability of the average values obtained by several operators while measuring the same item. For example, the median standard deviation reported for the 343 measurements in the CSARdock 2010 benchmark data is 0.05 pK_i units (repeatability error) and it can be as low as 0.001 pK_i units.[5] Such precision is typically not achieved when the assay is run in different laboratories under different assay conditions or even by different methods. In those cases, the standard deviation between measurements is usually much larger. The variability in affinity-based biosensor studies was recently explored in a global benchmark: 150 participants from 20 countries were given the same protein samples and asked to determine kinetic rate constants for the selected protein−protein interaction case study.[6] The study yielded an average rate constant of 0.62 nM with a standard deviation of 0.98 nM obtained by independent investigators using various biosensor technologies. Because of the experimental flexibility that was given to the participants, the reported variability was larger than the variability obtained in previous benchmark studies where participants were asked to run the experiment on their own instruments while using a detailed fixed protocol.[7] In those cases, the variability observed in the association and dissociation rate constants varied from less than 14% to ~20% or ~40% depending on the specific case study.[8,9]

In silico model generation and validation usually faces a trade-off between small, highly comparable data sets and large, heterogeneous data sets. Experiments are regarded as more comparable if they have been carried out in the same laboratory under the same conditions. Models and validations however get better and more reliable with increasing data set size.[10] For most interesting biological targets, there are only small consistently measured data sets available in the literature. Therefore, data used for in silico models is usually heterogeneous (i.e., various laboratories, assay conditions, assay methods) and reproducibility is the relevant measure. Data from different assays can only

Received: January 30, 2012
Published: May 29, 2012

dx.doi.org/10.1021/jm300131x | J. Med. Chem. 2012, 55, 5165−5173

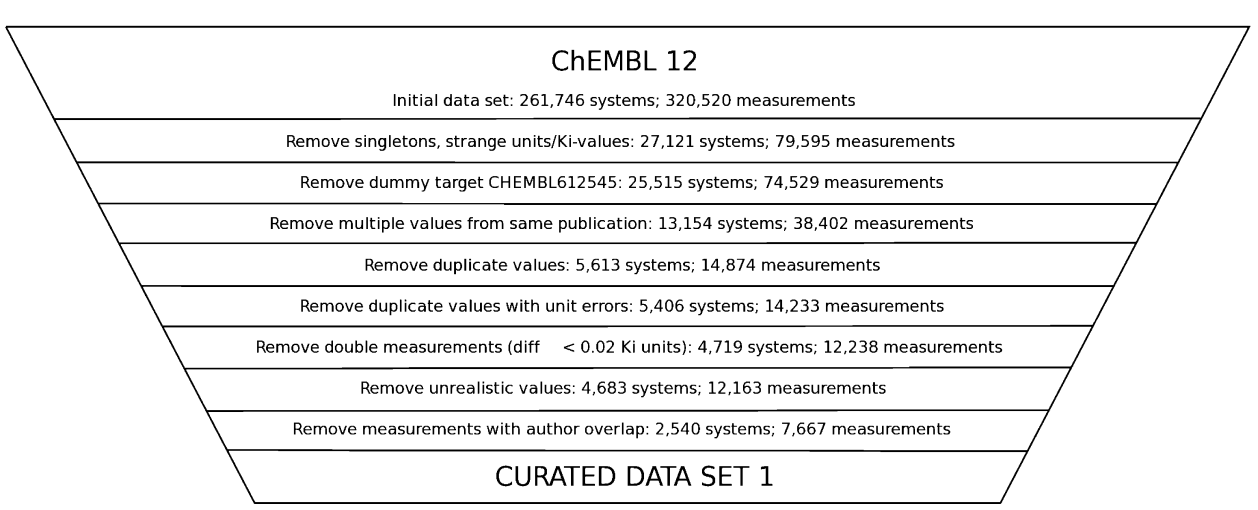


Figure 1. The data curation process. System is a specific protein−ligand complex.

be compared if it is reported as K_i or K_d data. IC_{50}, EC_{50}, percent inhibition, etc. data is assay-specific and can not be compared reliably.[11]

In recent years, a number of public databases have been set up with the goal of collecting and making available all the biological activity data published in the scientific literature and patents. The most prominent among these databases are: PubChem,[12] ChEMBL,[13] Binding DB,[14] PDBbind,[15] and BindingMOAD.[16] These databases have a substantial overlap because usually they are derived from the same journals. For example, ChEMBL forms a large subset of the activity data in BindingDB.[17]

The databases allow large-scale data analysis and build the foundation of chemogenomics,[18] novel scoring functions,[19−21] and large-scale SAR analysis.[22] However, it is also well-known by the community that the data quality in the databases is questionable. For example, some entries contain unclear stereoisomer annotations, wrong structures, unrealistic values, and wrongly assigned activity units that can cause significant issues in modeling.[17,23] In this contribution, the major pitfalls that occur when extracting data from ChEMBL for multiple measurements are highlighted and a number of filters for the identification of a set of truly independent measurements are proposed. An estimate for the worst and the best case of experimental uncertainties of biological K_i data is calculated. By extrapolation, this is suggested to be applicable to all those cases in which multiple independent measurements are not available and can be used as a reference value to compare actual predictions against.

DATA SOURCE

ChEMBL version 12 was downloaded from the European Bioinformatics Institute's FTP-site as a MySQL dump (Accessed Dec 8, 2011) and imported into a local database server running MySQL version 5.0.77. All K_i measurements were extracted from the database into a text file. The lists of authors for the publications were obtained using a Python script that linked the bibliographical information available within ChEMBL to PubMed. A similar script was developed to filter out review papers from the data set by removing measured values from publications that had the "Review" tag in PubMed. This script also checked for retracted publications, which were not detected in this case. See Supporting Information for the SQL query and all scripts used.

DATA CURATION

The raw K_i measurements were filtered using a Python-script (Figure 1). After each removal step, the target/compound systems that only have one single measurement were discarded. The filtering steps were the following:

(1) Measurements for a certain system of target and compound were grouped together by their ChEMBL target/compound identifier (ID) numbers. All singletons, i.e., systems with just one measurement, were removed. Measurements with CHEMBL612545 as the target ID were removed, because this is a *dummy ID* for unchecked targets.

(2) All measurements with *unclear units or values* were removed. Only the measurements with one of the units M, mM, μM, nM, pM, and fM were kept. Negative values were removed. The activity values were converted from K_i to pK_i. Measured K_i values lower than 1 fM or higher than 10 mM were removed.

(3) If a target/compound system had several *different activity values in the same publication*, only the highest pK_i value was taken into analysis to remove unclear stereoisomer annotations and reports on experimental optimization. The highest value was chosen because for stereoisomers the highest pK_i value should always correspond to the same most active stereoisomer. Some publications report assay optimization procedures, while others only report the highest activity measured. For assay optimizations, it makes sense to only include the highest activity.

(4) To *exclude citations* of previously published values, all measurements with identical reported values in different publications for the same system were removed. Values with exactly 3 or 6 pK_i unit difference were also removed to exclude citations of previously published values with *unit-transcription error*.

To *exclude rounded citations* of previously published values, the lowest pK_i values with a difference smaller than 0.02 pK_i units were removed. Upon inspection of the original literature, all randomly drawn cases from this group could be traced back to citations of previously reported values (see the Supporting Information for the lists of all randomly drawn samples).

dx.doi.org/10.1021/jm300131x | J. Med. Chem. 2012, 55, 5165−5173

(5) The list of authors from the publications was used to identify independent pairs of measurements from different laboratories. No overlap between the authors of two papers was allowed. This may unintentionally have removed some truly independent measurements due to researchers having identical names.

DATA ANALYSIS AND MEASURES OF QUALITY

The standard deviation of the measurements (σ_E), the mean unsigned error (MUE), and the median unsigned error ($M_{ed}UE$) for all pairs of measurements were used to assess the quality of the agreement between multiple measured values for the same target/compound system. For most systems, there are two or three independently measured values available. The distribution of published values per system is shown in Figure 2.

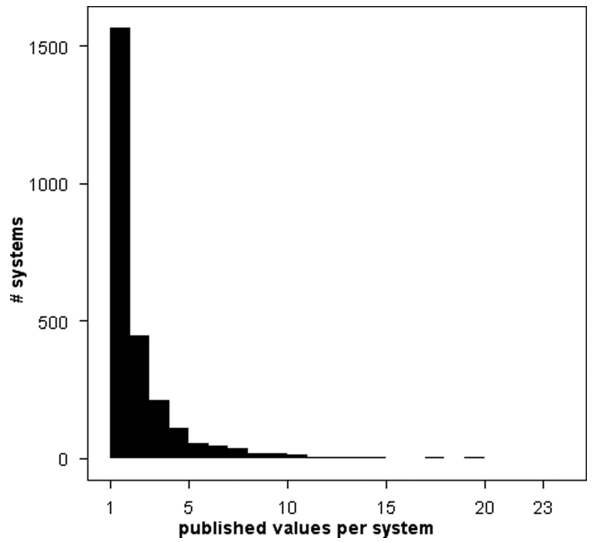


Figure 2. Published values per system. The most frequently measured system is Rimonabant (CHEMBL111) on the cannabinoid receptor 1 (CHEMBL218) with 23 independent measurements, followed by haloperidol (CHEMBL54) on the dopamine D2 receptor (CHEMBL339) with 22 measurements.

Measures of Quality from Differences between Published Values. With two or three samples only, the estimates for the average activities of individual protein−ligand systems are unreliable and therefore the standard equations for calculating the standard deviation (σ_E), the mean unsigned error (MUE), and the median unsigned error ($M_{ed}UE$) cannot be used because the total variability would be underestimated. Assuming that the experimental uncertainty (σ_E = reproducibility) is normally distributed and the same on all K_i measurements, the differences between published values can be used to calculate the standard measures of quality and the calculation of averages can be avoided. It can be shown that σ_E, MUE, and $M_{ed}UE$ of the data can be calculated from the corresponding values of the differences by division through $\sqrt{2}$. The proof is given in the Appendix.

The overall equations for calculating σ_E, MUE, and $M_{ed}UE$ from n differences between pairs of published values ($y_{pub,i,1}$ and $y_{pub,i,2}$ stand for the two measured values within a given pair i) then become

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

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

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

If more than two measurements are available for a given ligand−protein system, all possible pairs are generated for calculating the measures of quality. Another frequently used measure of quality is the squared Pearson's correlation coefficient ($R^2_{Pearson}$). If it is calculated on the pairs of measurements, it is a measure of the performance that can be achieved by an in silico model that has the same experimental uncertainty as the measurements. The order of $y_{pub,i,1}$ and $y_{pub,i,2}$ has to be scrambled in order to not bias the $R^2_{Pearson}$ calculation. Using the same notation as above for the pairs of measurements, the equation for $R^2_{Pearson}$ and the average measured value $\bar{y}$ becomes

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

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

Maximum $R^2_{Pearson}$ Achievable. To judge the quality of in silico models, it is important to know the maximum performance observable. For all biochemical pK_i data sets, be it congeneric series of ligands measured in different laboratories or ligands with varying scaffolds on different proteins as used for the evaluation of scoring functions, the maximum $R^2_{Pearson,MAX}$ that a perfect in silico model could have depends on the standard deviation of the total set of measured values and the standard deviation of the experimental uncertainty. Because published values always contain an experimental uncertainty, the maximum $R^2_{Pearson,MAX}$ observable with perfect predictions is lower than one. The equation for calculating $R^2_{Pearson,\ MAX}$ for a known experimental uncertainty with standard deviation σ_E and a given data set with an overall distribution with $\sigma(Y_{pub})$ is

$$R^2_{\mathrm{Pearson,max}} = 1 - \left(\frac{\sigma_E}{\sigma(Y_{\mathrm{pub}})} \right)^2$$

The derivation of this equation can be found in the Appendix.

All data analysis was done in R version 2.14.[24] The R script used is included in the Supporting Information.

RESULTS

Overall there were 261,746 different systems with 323,520 measured K_i values stored in ChEMBL12. Of these, 27,121 had multiple measurements assigned. The data curation procedure outlined above reduced the initial set to 2540 target/ligand systems with 7667 measurements, yielding 11,621 pairs of measurements. The data curation steps 1−4 removed more than 85% of the overall database. Another 6% of the data was removed in order ensure independence of measurements by eliminating overlap in the publications author lists (step 5).

Most pseudo multiple measurements were removed when only the most active out of multiple values per publication was selected (step 3). Inspection of randomly picked samples out of

dx.doi.org/10.1021/jm300131x | J. Med. Chem. 2012, 55, 5165−5173

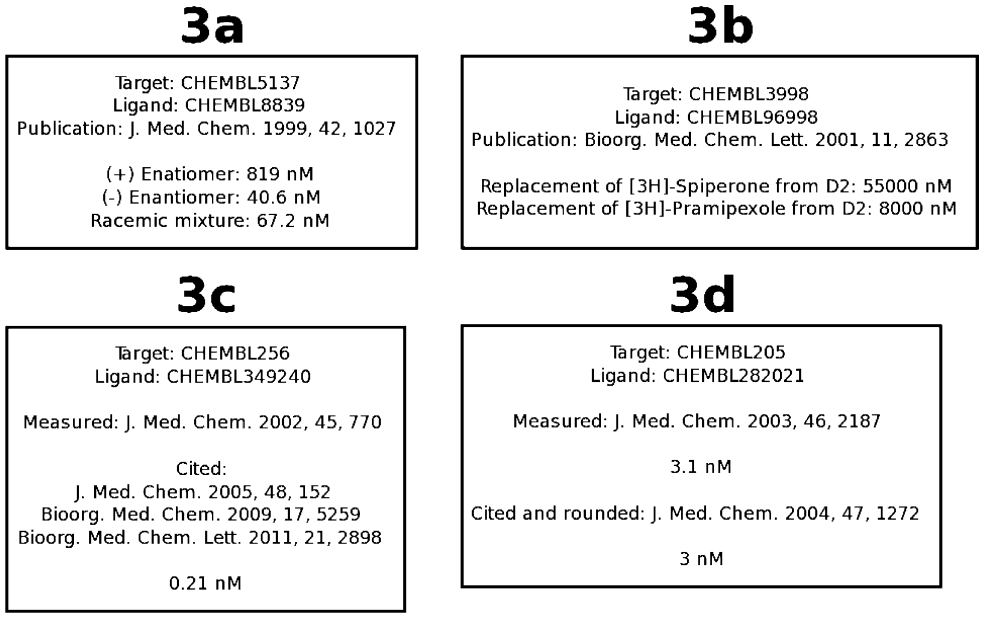


Figure 3. Examples of common issues in ChEMBL when selecting multiple measurements for a target/compound pair.

this group showed that in addition to simple typing errors, there were two major reasons why two or more K_i values for the same compound/target pair and the same publication could be found in ChEMBL. One reason was that the stereoisomers had not been properly assigned, producing three values reported for the two stereoisomers and the racemate (Figure 3a). The other common reason was that different assays were used or activities for different assay conditions were reported (Figure 3b). It is impossible to differentiate between such cases without manual inspection of the original literature. Nevertheless, as the measurements are reported in the same publication, it was assumed that these experiments were carried out in the same laboratory. The most active value was selected in order to select the most relevant (in case of assay optimization) value for the same compound (in case of undifferentiated stereoisomers), and the other values were excluded from the current analysis, as the aim of the study is reproducibility assessment of independent measurements. If all activities from publications with multiple values for one ligand−protein system were discarded, the overall results only changed by 0.01 or 0.02 units (details not shown).

In curation step 4, all pairs where a previously published value was cited were removed. In all cases of randomly picked samples from the group with identical values in different publications, in manual inspection it turned out that the later publication cited the value from the earlier one (Figure 3c). In 18 out of 20 cases of randomly picked samples where the two reported values were very close, it turned out that the later study cited the earlier one and the previously measured value was rounded (Figure 3d). Rounding differences can maximally yield a difference of 0.17 pK_i units (compare 1 nM → pK_i = 9 and 1.49 nM → pK_i = 8.83), so as a conservative filter for rounding errors a difference of 0.02 pK_i units was used for data extraction. For the analysis of measures of agreement, a threshold of 0.05 pK_i units was used. In all cases of randomly picked samples where the difference between activities was exactly 3 or 6 pK_i units, one paper was citing from the other and there was a unit transcription error in ChEMBL. In the end, a set of 7667 independent measurements for 2540 protein−ligand systems was identified (curated data set 1).

The distribution of the activities of the curated data set 1 is shown in Figure 4. It is slightly skewed to the left, indicating that

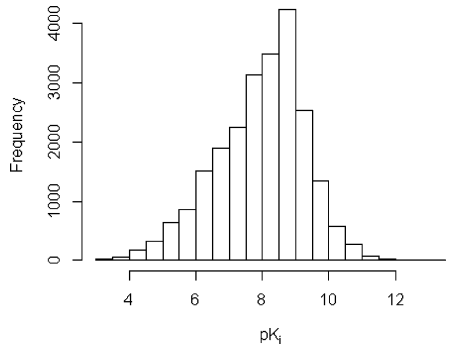


Figure 4. Distribution of ChEMBL activities after database curation.

pK_i values of highly active compounds tended to be published more often. This makes sense, given that the purpose of most medicinal chemistry programs is to generate highly active compounds.

Inspection of 10 pairs with disagreements of ∼3 pK_i units showed that 8 out of 10 pairs still had unit transcription errors. Therefore, a threshold of 2.5 pK_i units was set as maximum tolerable agreement between different measurements. Unit transcription errors of independently measured values that are detrimental to the comparison always cause disagreements of at least 1.5 pK_i units. So the threshold of 2.5 pK_i units was implemented as a better compromise between reducing unit transcription errors and keeping pairs of really independent measurements. A new curated data set without strong disagreements, from now on defined as curated data set 2, was generated by removing 404 pairs with disagreements >2.5 pK_i units. A plot of the pairs of compared ChEMBL pK_i values is shown in Figure 5. The two lines in the plot indicate the 2.5 log unit border for removing unreliable pairs of measurement.

The distribution of differences between pairs of measured values ($\Delta pK_i = |y_{i,1} - y_{i,2}|$ with $\Delta pKi < 2.5$) versus the activity

dx.doi.org/10.1021/jm300131x | J. Med. Chem. 2012, 55, 5165−5173

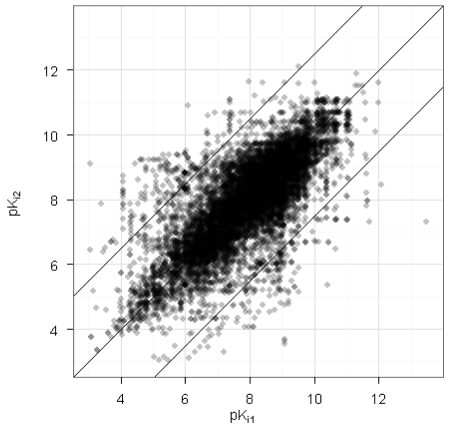


Figure 5. Plot of all pairs of measured values (curated dataset 1). Lines indicate the 2.5 log unit border for removing unreliable pairs of measurement (leaving curated dataset 2).

distribution is shown in Figure 6. One might assume that highly active compounds would get more reproducible measurements

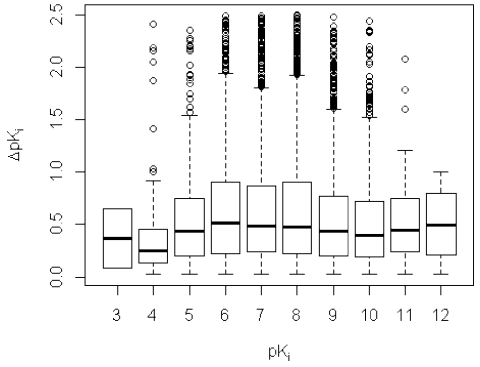


Figure 6. Distribution of differences between pairs of measurements, depending on the average activity. Mean indicated by black bar, boxes indicate 25% and 75% range, whiskers at 1.5 * interquartile range or strongest outlier if the distance to the median is smaller. Circles represent individual outliers larger than the upper whisker.

due to higher accuracy, as the complete K_i titration profile might be easier to determine. However, the data did not show a difference in reproducibility (calculated as ΔpK_i between each pairs) between high affinity and low affinity complexes, as shown in Figure 6.

To see if the physicochemical properties of ligands have an influence on reproducibility, standard chemoinformatics descriptors were calculated using an in-house tool. The distribution of differences between pairs of measured values ($\Delta pK_i = |y_{i,1} - y_{i,2}|$ with $\Delta pK_i < 2.5$) versus lipophilicity (ClogP), polar surface area (PSA), and the number of heavy atoms (nof_Atoms) for the samples where a SMILES string for the ligand was available in ChEMBL12 (11,550 pairs) is shown in Figure 7. The plots indicate that the differences between assay results become larger with very high or very low ClogP, a large polar surface area, and larger bigger ligands. However, there are only few examples for these more extreme molecules, and for the property range of standard drug molecules (0 < ClogP < 5, PSA < 250, nof_atoms < 45), there is no clear visible trend observable. Also, the few molecules with high PSA, high nof_Atoms, and very low ClogP coincide.

The statistical parameters calculated to explore the experimental uncertainty obtained from pairs of measured independent activity data generated in the two curated ChEMBL data sets are reported in Table 1. The threshold of $\Delta pK_i < 2.5$ units introduced in step 4 for the curated data set 2 did probably lead to a slight bias toward less disagreement, since a number of really independent pairs of measurements with differences between 2.5 and 3 pK_i units were also removed. As a consequence, the σ_E value for the curated ChEMBL data set 2 is slightly overoptimistic because the tail of the distribution has been cut off and thereby all compounds strongly influencing this estimate have been removed. For this case, a more robust measure like the $M_{ed}UE$ makes more sense. The $R^2_{Pearson,\ MAX}$ achievable for perfect predictions has been calculated to be 0.74 for the curated data set 1 and 0.81 for the refined curated data set 2.

The whole analysis is based on the assumption that experimental uncertainty is the same for every published value, i.e., the random error of each measurement is drawn from the same normal distribution, irrespective of the ligand chemistry and the protein target type. If the individual errors are drawn from a normal distribution, the differences between two errors are also normally distributed because adding two normal distributions yields another normal distribution. Therefore, if the differences are not normally distributed, the assumption of all experiments having the same experimental error is not true. This can be tested by fitting a Gaussian to the distribution of differences. The distribution of the calculated ΔpK_i does not fit exactly a Gaussian distribution as the calculated ΔpK_i between 1.0 and 2.5 pK_i units were always underestimated by a single fitted Gaussian. The overall distribution could be accurately fitted by using the sum of two Gaussian functions (Figure 8). This result indicates that the assumption of every published value having the same experimental uncertainty is reasonable.

DISCUSSION

In this contribution, an estimate of the experimental uncertainty of K_i variability was derived from compounds with multiple activity values reported in ChEMBL. Much care was put into the data extraction from ChEMBL to yield data which is independently measured and free from transcription errors added during the insertion of the data into the database. After this curation process, a mean unsigned error MUE = 0.44 pK_i units (corresponding to a factor of 2.8 in K_i), a standard deviation σ_E = 0.56 pK_i units, and a median error $M_{ed}UE$ = 0.34 pK_i units (corresponding to a factor of 2.2) were derived as error estimates for individual published K_i values. These values set the maximum performance achievable by any in silico models using the same data set or by extrapolation using any public K_i data from different laboratories. If the whole ChEMBL data set was used without restriction to the measurement pairs with $\Delta pK_i < 2.5$, slightly worse values were observed: $M_{ed}UE$ = 0.34, MUE = 0.48, and σ_E = 0.69. The achievable $R^2_{Pearson}$ depends on the range of the activities. If the data set used in modeling is global, normally distributed, and contains the full range of measured values from 3 to 14 pK_i, such as the PDBbind[10] or the CSARdock benchmark data set,[4] the maximum performance of a model with the same uncertainty as the experimental uncertainty can be $R^2_{Pearson,MAX}$ = 0.66. The maximum performance of a perfect model can be $R^2_{Pearson,MAX}$ = 0.81. It cannot be 1.00 because biological experiments always contain artifacts from both the measurement and the sample preparation protocol.

The observed experimental uncertainty is low compared to the usually observed uncertainty of in silico models. However, it must be clarified that the examined data was not high-throughput data and that most of the published values were averages of

dx.doi.org/10.1021/jm300131x | J. Med. Chem. 2012, 55, 5165−5173

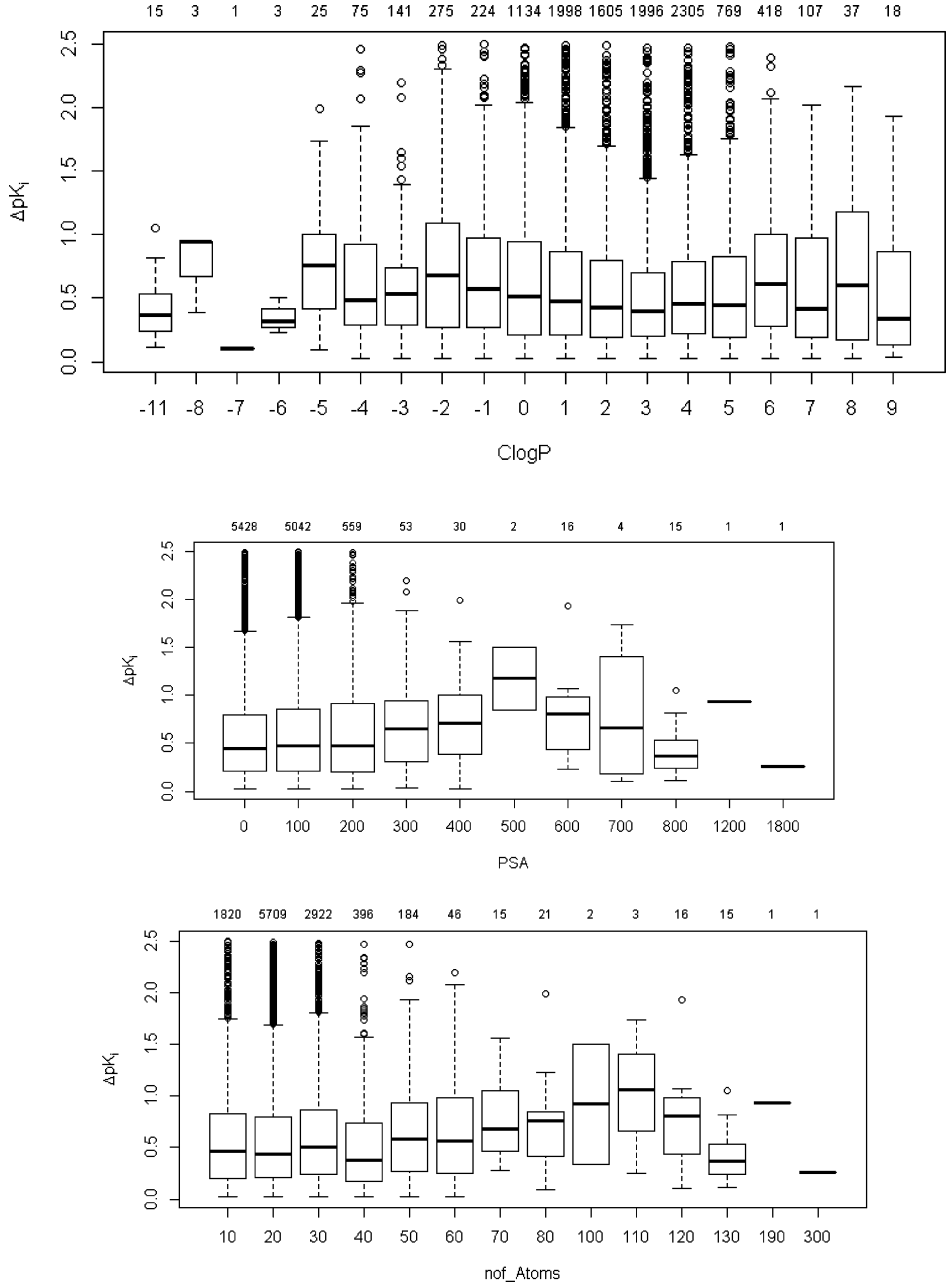


Figure 7. Distribution of differences between pairs of measurements, depending on ClogP, PSA, and the number of heavy atoms (values rounded to the corresponding values given on the *X*-axis). Mean indicated by black bar, boxes indicate 25% and 75% range, whiskers at 1.5 * interquartile range or strongest outlier if the distance to the median is smaller. Circles represent outliers larger than the upper whisker. Numbers above the plot indicate the number of pairs per bin.

Table 1. Summary of Statistical Numbers for the Experimental Uncertainty of Public pK_i Data[a]

statistics	curated data set 1	curated data set 2
MUE [pK_i]	0.48	0.44
M_{ed}UE [pK_i]	0.34	0.34
σ_E [pK_i]	0.69	0.56
$R^2_{Pearson}$	0.54	0.66
$R^2_{Pearson,\ MAX}$	0.74	0.82

[a]Curated dataset 2 is obtained by removing all pairs with $\Delta pK_i > 2.5$ from curated dataset 1.

multiple runs of the same experiment. So the variability found most probably is the interlaboratory factor, the difference between repeatability (which is usually published with each individual value) and reproducibility. In addition, it has to be kept in mind that the data set examined consisted of double measurements allowing to spot and remove strong disagreements, mostly due to unit-transcription errors.

Care must be taken when drawing conclusions on experimental variability by using multiple measurements from the original ChEMBL and probably from all other huge public data collections as well. More than 90% of the pairs of measurements that have initially been extracted were not independent because individual values were simply derived from citations of previously published data; refer to different stereoisomers/racemates or to different assays or stages of assay optimization. ChEMBL also contained a large set of compounds with unchecked biological targets that nevertheless popped up in searches

dx.doi.org/10.1021/jm300131x | J. Med. Chem. 2012, 55, 5165−5173

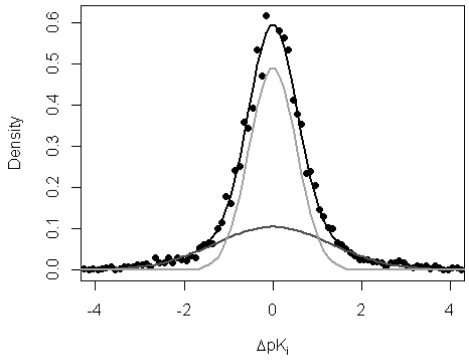


Figure 8. Two Gaussian functions (Dark gray: $\sigma = 1.3264$, weight = 0.35. Light gray: $\sigma = 0.5359$, weight = 0.66. Black: sum.) fitted to distribution of measurement differences (black dots) of the curated ChEMBL data.

(TargetID = CHEMBL612545). Nearly all pairs of measurements with a ΔpK_i greater than 3 pK_i units contained a unit transcription error, meaning that the actual agreement was much better than it seemed. To automatically identify a set of completely independent measurements from the original ChEMBL, the authors lists from the publications were taken into account and pairs of measurements with overlapping authors were removed.

The statistical values derived should be very reliable estimates because they are based on 7667 individual measurements, yielding 11,621 pairs of measurements. Although it was impossible to check all individual pairs of measurements, it can be assumed that the data remaining after the curation was clean enough to obtain a realistic estimate of the experimental uncertainty. All strong outliers with $\Delta pK_i > 2.5$ and all not independently reported values with $\Delta pK_i < 0.05$ were removed from the analysis. The thresholds used in step 4 can be defined differently; however, these subjective threshold definitions do not affect the overall results of this study. For example, varying the thresholds of step 4 between $0.02 < \Delta pK_i < 2.5$ to $0.1 < \Delta pK_i < 3.5$ changes the $M_{ed}UE$ from 0.33 to 0.39, the MUE from 0.42 to 0.51, and σ_E from 0.55 to 0.67. The magnitudes of the differences in variation are to be expected because $M_{ed}UE$ is the most robust and σ_E the least robust measure of the three. The bias of wrongly removing some very well-agreeing measurements and some strongly disagreeing measurements were assumed to be averaged out in the end. In randomly picked samples from the final curated data set 2, only rare unsystematic errors in compound and value assignments were found.

Finally, there might be some confirmation bias in the data leading to underestimation of the experimental uncertainty. If there is previous data on the same protein−ligand system available, experimentalists might tune their assay until the measured value comes close to the previously published value. The authors hope that this was not a big issue in the data, but they can neither guarantee for the absence nor exclude this bias. It is difficult to correct for this issue because it is not evident from the publication date whether or not a previously measured value was known to the experimentalist. From a strict point of view, the most accurate quantification of experimental uncertainties can only be obtained from double-blind measurements carried out in independent laboratories, as reported in refs 6−9. However, these extensive and interesting analyses have been performed on a very limited set of protein−ligand or protein−protein cases. Until a sufficiently large set of controlled independently measured pK_i values is available, the estimate obtained using the approach described here can be helpful to line a realistic estimate of experimental uncertainty in the public K_i data most commonly used by the modeling community.

CONCLUSIONS

The experimental uncertainty in K_i measurements for heterogeneous public data was estimated to be between a MUE of 0.44 and 0.48 pK_i units. This value is very important because it sets the upper limit of performance for all in silico binding affinity prediction models, be it QSAR, scoring, or MM-GBSA. If the average error of a model based on heterogeneous public data is less than 0.44 pK_i units, it is very likely that a model is overtrained.

The analysis presented further shows that more than 90% of the pairs measurements extracted from ChEMBL actually were not independent measurements. Eighty-five percent of the individual data had to be excluded, mainly because they were either citations of previously published data, unclear assigned stereoisomers, or they contained some unit transcription error. Another 6% had to be excluded because they were not measured completely independently as determined by the overlap in the authors of the corresponding publications. Therefore, a careful curation process was developed that reduces the redundancy and improves the quality of the publicly available data sets before any type of analysis.

APPENDIX

Measures of Quality from Differences between Published Values

Each published value y_{pub} is composed of the true value y_{true} and some random error ε.

$$y_{pub} = y_{true} + \varepsilon \tag{1}$$

The difference between two independently measured and published values that have been obtained for the same biological system then becomes

$$y_{pub,1} - y_{pub,2} = (y_{true} + \varepsilon_1) - (y_{true} + \varepsilon_2) = \varepsilon_1 - \varepsilon_2 \tag{2}$$

Because the overall variances for the differences of any two vectors Y_1 and Y_2 add up according to

$$\mathrm{var}(Y_1 - Y_2) = \mathrm{var}(Y_1) + \mathrm{var}(Y_2) - 2\mathrm{cov}(Y_1, Y_2) \tag{3}$$

and the covariance between two vectors of random errors $(E1,E2)$ is zero: $\mathrm{cov}(E1,E2) = 0$, the standard deviation of the differences of the measurements becomes

$$\begin{aligned}\sigma(E1 - E2) &= \sqrt{\mathrm{var}(E1 - E2)} \\ &= \sqrt{\mathrm{var}(E1) + \mathrm{var}(E2)} \\ &= \sqrt{\sigma(E1)^2 + \sigma(E2)^2}\end{aligned} \tag{4}$$

Because the standard deviation of $E1$ and $E2$ is the same $(\sigma(E1) = \sigma(E2) = \sigma_E)$, the standard deviation of the pairs of differences then becomes

$$\sigma(E1 - E2) = \sqrt{\sigma(E1)^2 + \sigma(E2)^2} = \sigma_E\sqrt{2} \tag{5}$$

Therefore, the standard deviation or experimental uncertainty of pK_i measurements can be calculated from the standard deviation of the differences by dividing by $\sqrt{2}$.

For normally distributed errors, the MUE and $M_{ed}UE$ are proportional to the standard deviation

$$\sigma_E \approx \mathrm{MUE} \approx M_{ed}UE \tag{6}$$

Therefore, the MUE and the $M_{ed}UE$ can also be derived from the MUE and the $M_{ed}UE$ of the differences by dividing by $\sqrt{2}$.

dx.doi.org/10.1021/jm300131x | J. Med. Chem. 2012, 55, 5165−5173

The overall equations for calculating σ_E, MUE, and $M_{ed}UE$ from the differences between n pairs of published values then become

$$\mathrm{MUE} = \frac{1}{n\sqrt{2}} \sum_{i=1}^{n} \left| y_{\mathrm{pub},i,1} - y_{\mathrm{pub},i,2} \right| \tag{7}$$

$$M_{ed}UE = \frac{1}{\sqrt{2}} \mathrm{median}\left\{ \left| y_{\mathrm{pub},i,1} - y_{\mathrm{pub},i,2} \right| \text{for } i \text{ in } 1...n \right\} \tag{8}$$

$$\sigma_E = \sqrt{\frac{1}{2(n-1)} \sum_{i=1}^{n} \left(y_{\mathrm{pub},i,1} - y_{\mathrm{pub},i,2} \right)^2} \tag{9}$$

Maximum $R^2_{Pearson}$ Achievable

$R^2_{Pearson}$ can be expressed as

$$R^2_{\mathrm{Pearson}} = \left(\frac{\mathrm{cov}(Y_1, Y_2)}{\sigma(Y_1) * \sigma(Y_2)} \right)^2 \tag{10}$$

with $\mathrm{cov}(Y_1,Y_2)$ being the covariance of any two vectors Y_1 and Y_2 of corresponding values and $\sigma(Y_1)$ and $\sigma(Y_2)$ being the standard deviations of the two vectors. Given that all published values Y_{pub} are composed of the true value Y_{true} plus some experimental uncertainty E

$$Y_{\mathrm{pub}} = Y_{\mathrm{true}} + E \tag{11}$$

the equation for the maximum achievable $R^2_{Pearson,MAX}$ becomes

$$R^2_{\mathrm{Pearson,max}} = \left(\frac{\mathrm{cov}\left(Y_{\mathrm{true}}, Y_{\mathrm{pub}}\right)}{\sigma(Y_{\mathrm{true}}) * \sigma\left(Y_{\mathrm{pub}}\right)} \right)^2 = \left(\frac{\mathrm{cov}(Y_{\mathrm{true}}, Y_{\mathrm{true}} + E)}{\sigma(Y_{\mathrm{true}}) * \sigma\left(Y_{\mathrm{pub}}\right)} \right)^2 \tag{12}$$

Because the covariance of a sum of normally distributed vectors can be split up according to

$$\mathrm{cov}(Y_{\mathrm{true}}, Y_{\mathrm{true}} + E) = \mathrm{cov}(Y_{\mathrm{true}}, Y_{\mathrm{true}}) + \mathrm{cov}(Y_{\mathrm{true}}, E) \tag{13}$$

the covariance between a set of random errors E and true values Y_{true} *is zero*

$$\mathrm{cov}(Y_{\mathrm{true}}, E) = 0 \tag{14}$$

the covariance between a vector and itself is the same as its variance or the squared standard deviation

$$\mathrm{cov}(Y_{\mathrm{true}}, Y_{\mathrm{true}}) = \mathrm{var}(Y_{\mathrm{true}}) = \sigma(Y_{\mathrm{true}})^2 \tag{15}$$

and the standard deviation of the true values depends on the published values and the experimental uncertainty according to (which can be derived from rearranging the variances of the vectors in (11))

$$\sigma(Y_{\mathrm{true}}) = \sqrt{\sigma\left(Y_{\mathrm{pub}}\right)^2 - \sigma_E{}^2} \tag{16}$$

the equation for the maximum achievable $R^2_{pearson,\ MAX}$ can be rewritten as

$$R^2_{\mathrm{Pearson,MAX}} = \left(\frac{\mathrm{cov}\left(Y_{\mathrm{true}}, Y_{\mathrm{pub}}\right)}{\sigma(Y_{\mathrm{true}}) * \sigma\left(Y_{\mathrm{pub}}\right)} \right)^2$$

$$= \left(\frac{\mathrm{cov}(Y_{\mathrm{true}}, Y_{\mathrm{true}}) + \mathrm{cov}(Y_{\mathrm{true}}, E)}{\sigma(Y_{\mathrm{true}}) * \sigma\left(Y_{\mathrm{pub}}\right)} \right)^2 \quad \text{(using 13)}$$

$$= \left(\frac{\mathrm{cov}(Y_{\mathrm{true}}, Y_{\mathrm{true}})}{\sigma(Y_{\mathrm{true}}) * \sigma\left(Y_{\mathrm{pub}}\right)} \right)^2 \quad \text{(using 14)}$$

$$= \left(\frac{\sigma(Y_{\mathrm{true}})^2}{\sigma(Y_{\mathrm{true}}) * \sigma\left(Y_{\mathrm{pub}}\right)} \right)^2 \quad \text{(using 15)}$$

$$= \frac{\sigma(Y_{\mathrm{true}})^2}{\sigma\left(Y_{\mathrm{pub}}\right)^2} = \frac{\sigma\left(Y_{\mathrm{pub}}\right)^2 - \sigma_E{}^2}{\sigma\left(Y_{\mathrm{pub}}\right)^2} \quad \text{(using 16)}$$

$$= 1 - \left(\frac{\sigma_E}{\sigma\left(Y_{\mathrm{pub}}\right)} \right)^2 \tag{17}$$

■ ASSOCIATED CONTENT

Supporting Information

Supporting Information available: A MySQL statement to extract all data necessary from ChEMBL12 for the analysis described, a Python script to filter the CHEMBL multiple measurement data, the final data set prepared, an *R* script to analyze the double measurements, and comparisons of randomly drawn samples from various subgroups. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*For C.K.: phone, +41 61 69 67939; E-mail, Christian.Kramer@novartis.com. For T.K.: phone, +41 61 69 66049; E-mail, Tuomo.Kalliokoski@novartis.com.

Author Contributions

†Equal contribution

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

C.K. and T.K. thank the Novartis Institutes for BioMedical Research for Presidential PostDoc Fellowships.

■ ABBREVIATIONS USED

cov, covariance; FTP, file transfer protocol; ID, identifier; IUPAC, International Union of Pure and Applied Chemistry; $M_{ed}UE$, median unsigned error; MM-GBSA, molecular mechanics Poisson−Boltzmann surface area; MUE, mean unsigned error; nof_Atoms, number of heavy atoms; PSA, polar surface area; QSAR, quantitative structure−activity relationship; SAR, structure−activity relationship; SMILES, simplified molecular input line entry specification; var, variance; σ_E, standard deviation of the measurements; $R^2_{pearson,MAX}$, maximum achievable $R^2_{pearson}$ with a given σ_E

■ REFERENCES

(1) Green, D. V. S.; Leach, A. R.; Head, M. S. Computer-aided molecular design under the SWOTlight. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 51−56.

(2) Martin, E.; Ertl, P.; Hunt, P.; Duca, J.; Lewis, R. Gazing into the crystal ball: the future of computer-aided drug design. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 77−79.
(3) Stouch, T. R. The errors of our ways: taking account of error in computer-aided drug design to build confidence intervals for our next 25 years. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 125−134.
(4) IUPAC. *Compendium of Chemical Terminology (The "Gold Book")*, 2nd ed; McNaught, A. D., Wilkinson, A., compilers; Blackwell Scientific Publications: Oxford, UK, 1997; XML on-line corrected version http://goldbook.iupac.org, 2006 created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins; ISBN 0-9678550-9-8, DOI: 10.1351/goldbook.
(5) *Community Structural−Activity Resources (CSAR)*; http://www.csardock.org (accessed December 14, 2011).
(6) Rich, R. L.; Papalia, G. A.; Flynn, P. J.; Furneisen, J.; Quinn, J.; Klein, J. S.; Katsamba, P. S.; Waddell, M. B.; Scott, M.; Thompson, J.; Berlier, J.; Corry, S.; Baltzinger, M.; Zeder-Lutz, G.; Schoenemann, A.; Clabbers, A.; Wieckowski, S.; Murphy, M. M.; Page, P.; Ryan, T. E.; Duffner, J.; Ganguly, T.; Corbin, J.; Gautam, S.; Anderluh, G.; Bavdek, A.; Reichmann, D.; Yadav, S. P.; Hommema, E.; Pol, E.; Drake, A.; Klakamp, S.; Chapman, T.; Kernaghan, D.; Miller, K.; Schuman, J.; Lindquist, K.; Herlihy, K.; Murphy, M. B.; Bohnsack, R.; Andrien, B.; Brandani, P.; Terwey, D.; Millican, R.; Darling, R. J.; Wang, L.; Carter, Q.; Dotzlaf, J.; Lopez-Sagaseta, J.; Campbell, I.; Torreri, P.; Hoos, S.; England, P.; Liu, Y.; Abdiche, Y.; Malashock, D.; Pinkerton, A.; Wong, M.; Lafer, E.; Hinck, C.; Thompson, K.; Primo, C. D.; Joyce, A.; Brooks, J.; Torta, F.; Bagge Hagel, A. B.; Krarup, J.; Pass, J.; Ferreira, M.; Shikov, S.; Mikolajczyk, M.; Abe, Y.; Barbato, G.; Giannetti, A. M.; Krishnamoorthy, G.; Beusink, B.; Satpaev, D.; Tsang, T.; Fang, E.; Partridge, J.; Brohawn, S.; Horn, J.; Pritsch, O.; Obal, G.; Nilapwar, S.; Busby, B.; Gutierrez-Sanchez, G.; Gupta, R. D.; Canepa, S.; Witte, K.; Nikolovska-Coleska, Z.; Cho, Y. H.; D'Agata, R.; Schlick, K.; Calvert, R.; Munoz, E. M.; Hernaiz, M. J.; Bravman, T.; Dines, M.; Yang, M. H. A global benchmark study using affinity-based biosensors. *Anal. Biochem.* **2009**, *386*, 194−216.
(7) Cannon, M. J.; Papalia, G. A.; Navratilova, I.; Fisher, R. J.; Roberts, L. R.; Worthy, K. M.; Stephen, A. G.; Marchesini, G. R.; Collins, E. J.; Casper, D.; Qiu, H.; Satpaev, D.; Liparoto, S. F.; Rice, D. A.; Gorshkova, I. I.; Darling, R. J.; Bennett, D. B.; Sekar, M.; Hommema, E.; Liang, A. M.; Day, E. S.; Inman, J.; Karlicek, S. M.; Ullrich, S. J.; Hodges, D.; Chu, T.; Sullivan, E.; Simpson, J.; Rafique, A.; Luginbühl, B.; Westin, S. N.; Bynum, M.; Cachia, P.; Li, Y. J.; Kao, D.; Neurauter, A.; Wong, M.; Swanson, M.; Myszka, D. G. Comparative analyses of a small molecule/enzyme interaction by multiple users of Biacore technology. *Anal. Biochem.* **2004**, *330*, 98−113.
(8) Katsamba, P. S.; Navratilova, I.; Calderon-Cacia, M.; Fan, L.; Thornton, K.; Zhu, M.; Bos, T. V.; Forte, C.; Friend, D.; Laird-Offringa, I.; Tavares, G.; Whatley, J.; Shi, E.; Widom, A.; Lindquist, K. C.; Klakamp, S.; Drake, A.; Bohmann, D.; Roell, M.; Rose, L.; Dorocke, J.; Roth, B.; Luginbühl, B.; Myszka, D. G. Kinetic analysis of a high-affinity antibody/antigen interaction performed by multiple Biacore users. *Anal. Biochem.* **2006**, *352*, 208−221.
(9) Myszka, D. G.; Abdiche, Y. N.; Arisaka, F.; Byron, O.; Eisenstein, E.; Hensley, P.; Thomson, J. A.; Lombardo, C. R.; Schwarz, F.; Stafford, W.; Doyle, M. L. The ABRF-MIRG'02 study: assembly state, thermodynamic, and kinetic analysis of an enzyme/inhibitor interaction. *J. Biomol. Technol.* **2003**, *14*, 247−269.
(10) Gedeck, P.; Rohde, B.; Bartels, C. QSAR—How good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924−1936.
(11) Stumpfe, D.; Bajorath, J. Assessing the Confidence Level of Public Domain Compound Activity Data and the Impact of Alternative Potency Measurements on SAR Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 3131−3137.
(12) Bolton, E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry*; American Chemical Society: Washington, DC, 2008; Vol. *4*, Chapter 12.
(13) Gaulton, A.; Bellis, L.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y; McGlinchey, S; Akhtar, R.; Atkinson, F.; Bento, A. P.; Al-Lazikani, B.; Michalovich, D.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database For Chemical Biology And Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(14) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(15) Wang, R.; Fang, X.; Lu, Y.; Yang, C.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(16) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 333−340.
(17) Tiikkainen, P.; Franke, L. Analysis of Commercial and Public Bioactivity Databases. *J. Chem. Inf. Model.* **2012**, *52*, 319−326.
(18) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38−52.
(19) Kramer, C.; Gedeck, P. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 1961−1969.
(20) Kramer, C.; Gedeck, P. Global Free Energy Scoring Functions Based on Distance-Dependent Atom-Type Pair Descriptors. *J. Chem. Inf. Model.* **2011**, *51*, 707−720.
(21) Kramer, C.; Gedeck, P. Three Descriptor Model Sets a High Standard for the CSAR-NRC HiQ Benchmark. *J. Chem. Inf. Model.* **2011**, *51*, 2139−2145.
(22) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. H. Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discovery Today* **2011**, *16*, 1019−1030.
(23) Williams, A. J.; Ekins, S. A quality alert and call for improved curation of public chemistry databases. *Drug Discovery Today* **2011**, *16*, 747−750.
(24) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, 2011; ISBN3-900051-07-0, http://www.R-project.org.