GRAM: a True Null Model for Relative Binding Affinity Predictions

Guanglei Cui,* Alan P. Graves, and Eric S. Manas

Computational and Modeling Science US, Platform Technology and Sciences, GlaxoSmithKline Pharmaceuticals, 1250 South Collegeville Road, Collegeville, PA 19426

E-mail: guanglei.x.cui@gsk.com

Abstract

Relative binding affinity prediction is a critical component in computer aided drug design. Significant amount of effort has been dedicated to developing rapid and reliable in silico methods. However, robust assessment of their performance is still a complicated issue, as it requires a performance measure applicable in the prospective setting and more importantly a true null model that defines the expected performance of random in an objective manner. Although many performance metrics, such as correlation coefficient (r^2), mean unsigned error (MUE), and room mean square error (RMSE), are frequently used in the literature, a true and non-trivial null model has yet been identified. To address this problem, here we introduce an interval estimate as an additional measure, namely prediction interval (PI), which can be estimated from the error distribution of the predictions. The benefits of using the interval estimate are 1) it provides the uncertainty range in the predicted activities, which is important in prospective applications; 2) a true null model with well-defined PI can be established. We provide one such example termed Gaussian Random Affinity Model (GRAM), which is based on the empirical observation that the affinity change in a typical lead

optimization effort has the tendency to distribute normally $N(0, \sigma)$. Having an analytically defined PI that only depends on the variation in the activities, GRAM should in principle allow us to compare the performance of relative binding affinity prediction methods in a standard way, ultimately critical to measuring the progress made in algorithm development.

Introduction

Ligand binding affinity prediction has been a long standing research focus in computational chemistry because of its practical applications in drug discovery. Of the many objectives of a typical lead optimization, target binding is arguably one of the few that are most amenable to explorations via first-principle approaches. The thermodynamics of *in vitro* target-ligand binding can often be probed both experimentally and theoretically, and there is often a direct correspondence between the two.

This pursuit has led to a variety of computational methods over the last few decades, ranging from physics-based approaches to pure statistical models. Among the physics-based approaches, techniques built on the foundation of statistical mechanics theories appear promising, especially those that explicitly sample the configurational space of the receptor and the ligand. Although it is still a considerable challenge to create a robust computational method for binding affinity predictions with a broad domain of applicability, rapid advances in computing power, accelerated sampling algorithms, and greater coverage of force field parameters for drug-like molecules have fueled a renaissance of molecular dynamics (MD) simulation-based approaches in recent years. For example, Schrodinger Inc. applied their free energy perturbation approach (FEP+) to a broad range of non-covalent protein targets and reported reasonable agreement with experiment (mean unsigned error or MUE around 1 kcal/mol).[1] The theoretical framework of FEP was first introduced in the early 1950s by Zwanzig for studying the thermodynamic properties of homogeneous condense-phase systems.[2] Applying the theory to heteroge-

neous molecular systems appeared much later and was pioneered by Berendsen, McCammon, Jorgensen, and Kollman in the 1980s.[3–8] In spite of the theoretical rigor of the approach, there had been relatively low adoption of FEP by the pharmaceutical industry, mainly due to time and complexity considerations. Hence, Schrodinger's commercial platform immediately generated a lot of interest, which subsequently led to independent evaluations conducted by the industry practitioners. The key question that these evaluations looked to address is whether the FEP technique in its current form can differentiate itself from other physics-based methods that are less computationally demanding, albeit often at the cost of theoretical rigor. This differentiation could be purely statistical (i.e. do two methods differ from one another in some statistically significant way?) or based on impact on decision-making (i.e. would demonstrably different decisions be reached using one method relative to another?).[9,9] This paper focuses on a question related to the former: whether any one specific method is significantly different from a reasonably chosen null model.

The importance of having a proper statistical comparison to a null model has been well recognized by the scientific community. However, there is no general consensus as to how an appropriate null model for binding affinity prediction should be constructed. In addition, methods are often compared without paying adequate attention to the sensitivity of various performance metrics (e.g. Pearson correlation coefficient (r^2), root mean square error (RMSE), and mean unsigned error (MUE)) to the distribution of activities in the test datasets.

This perspective outlines considerations over the choice of performance metrics and highlights the importance of designing an appropriate null model for binding affinity prediction. With the rationale described in detail below, prediction interval (PI) is proposed as a companion performance metric to capture the uncertainty in individual prospective predictions. It is shown that an appropriate null model for relative affinity prediction can be established with non-trivial and analytically determined PI. This is contrary to other

null models proposed in the scientific community, such as molecular weight or partition coefficient (LogP), which are more or less arbitrarily selected and are neither truly random, nor have well characterized performance. The prediction interval of our null model, termed GRAM (Gaussian Random Affinity Model), only depends on the observed variability and distribution in the measured affinities of the test set, therefore providing a robust and non-arbitrary baseline estimate of a zero-effort prediction.

Describing Error in Predictions

Ideally, a single performance metric would be used to quantify the agreement between our predictions and experimental observations. In reality, multiple performance metrics are typically employed, especially when the overall agreement of the prediction with experiment is far from perfect. In the case of binding affinity predictions, correlation coefficient (such as Pearson's r and Spearman's rank correlation coefficient), MUE, and RMSE are widely used in the literature. Among them, Pearson correlation coefficient (r^2) is a particularly poor metric for two well documented reasons. Firstly, r^2 is sensitive to sample size and sample variance or dynamic range. This is straightforward to understand in that a larger dynamic range in the dataset or a bimodal distribution in the observations tend to produce larger r^2, even though this may not have anything to do with the underlying physics or technique used. Secondly, the slope and intercept of the regression line from which r^2 is determined often have large confidence intervals when predicting binding affinities, since the correlation with experiment is often medium-to-low and can vary condiserably by targets and chemical series.

In a prospective scenario where predicted affinity or affinity changes are used to inform design, it is important to understand confident in the predictions. To a medicinal chemist, his or her key interest is the confidence with which a particular calculation predicts the outcome of the next experiment or set of experiments.

To account for the uncertainty in a prediction explicitly, it is necessary to focus on the spread of error produced over a sufficiently large sample. Formally, this is captured by prediction interval (PI), which is an empirical estimate of a range in which a percentage of all prediction errors have been observed. Mathematically, PI is defined as the following,[10]

$$PI = \pm t_{(\alpha/2,n-2)} \sqrt{MSE \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}\right)}$$

where MSE is the mean square error, $t_{(\alpha/2,n-2)}$ is the Student t-multiplier for significance level α, and n is the sample size. For a large sample, the error for which is approximately normally distributed, the prediction interval at the 95% confidence level is roughly $2 \times RMSE$ and it does not vary significantly over the entire activity range, which means a stable estimate of the upper and lower bounds of individual prospective predictions.

As an initial example, we examined the performance of FEP+ for a set of 37 Smyd3 compounds via 57 free energy perturbations, and compared the predictions with the experimental pIC50s (Figure 1). The MUE, RMSE, and Pearson correlation coefficient (r^2) are 0.8 kcal/mol, 1.0 kcal/mol, and 0.54, respectively. We chose to analyze the error in the predicted relative affinities rather than the affinities derived from a reference compound for two reasons – a) the relative affinities are the direct outcome of FEP+ calculations; b) empirically, the relative affinities for a set of congeneric compounds approximately center around 0. Figure 2 shows the histogram of the prediction error with a bin width of 1 kcal/mol, from which the prediction interval for any confidence levels can be estimated. For instance, roughly 95% of the predictions fall in the range of [-2.2, 2.2] kcal/mol or [-1.6, 1.6] pIC50 units. Therefore, the prediction interval for FEP+ at the 95% confidence level is estimated to be 2.2 kcal/mol or 1.6 log units from the Smyd3 set. This estimate appears to be robust across multiple targets used in a large scale internal evaluation of FEP+ at GSK. Over 251 compounds (8 distinct targets including Smyd3) that were tested both experimentally and with FEP+, 11 predictions (or 95.6%) were found to differ from the ex-

periment by more than 2.0 kcal/mol (Figure 3). This level of performance is qualitatively similar to what was reported by Wang et al.[1,11]

Proper Null Model for Affinity Predictions

A null model is a generative process based on either a known or a hypothetical statistical distribution of phenomenon being studied. This generative process is contructed such as it removes as best as possible any components that may contribute to the phenomenon or the mechanism behind. Null model gained popularity in ecology and biogeography in the early 1980s to measure the significance of the observed patterns in ecological and biogeographical data, such as taxonomic ratios, biodiversity measure, species co-occurrence, and community assembly rules.[12] In recent years null hypothesis testing has become more prevalent in the computational chemistry and computer-aided drug design community, for example, in the prediction of protein-ligand binding affinities, which is important for reducign the number of design-make-test cycles in lead optimization. Null model and null hypothesis are closely related, but different concepts. The latter is a general statement or default position that is either accepted or rejected during test, while the former focus on identifying any hidden bias in null hypothesis testing that may introduce the risk of accepting a false null hypothesis or vice verse. By the definition used in ecological and biogeographical context,[13] a null model for binding affinity or relative affinity is a statistical process for generating random activities from a known or hypothetical distribution. The random aspect of a null model is often not taken into account, as it is easy to misinterpret it as zero- or little-effort model, and therefore it is common that a variety of simple measures have been used as substitutes, including a simple constant,[14] the number of heavy atoms,[14] molecular weight (MW), partition coefficient (LogP), or other calculated molecular properties.[15] MW or LogP may have utilities as convenient comparator models for casual juxtapositions with non-trivial affinity predictions, but they are

not valid null models and their expected performance is not well defined – they are not suitable for being used as the standard reference for comparative studies. Their use as null models has also been criticized elsewhere.[16]

To design a proper null model, it is necessary to have the knowledge on the statistical distribution of binding affinities. Clearly, this cannot be obtained experimentally without the risk of introducing sample bias and has to be addressed theoretically. The only attempt of this kind so far was based on the energy landscape theory developed for describing protein folding problems and an analytical model for molecular interactions. Zheng and Wang[17] proposed a universal law that governs the distribution of affinities in biomolecular recognition, which was described as Gaussian near the mean and exponential near the tail. Whether this is a universal theory or not, the Gaussian-like distribution is largely consistent with our observations from medicinal chemistry campaigns at GSK – 1) the distribution of binding affinities (10^3 compounds) generated over the course of a fully executed program is roughly Gaussian around the mean; 2) the distribution of affinity evolution for compounds separated over a suitable period of time (e.g. 3-4 weeks) during an active lead optimization is approximately a Gaussian function centered around 0. For example, we plotted the histogram of pIC50s for about 386 CD73 compounds synthesized over a period of a few months (Figure 4) as well as the histogram of pIC50 evolution between newer compounds and older compounds (Figure 5). This, particularly the latter observation, may be the consequence of the shape of the SAR landscape as well as the chemistry exploration strategy adopted, which is usually conservative exploitation combined with occasional exploration. With this characteristic distribution of binding affinities and relative binding affinities, a random process for generating relative binding affinities can be proposed as such,

$$\Delta\Delta G_{GRAM} \sim N\left(0, \sigma_{GRAM}\right)$$

where $\Delta\Delta G$ denotes the relative free energy difference between a pair of compounds, $N\left(0, \sigma_{GRAM}\right)$ denotes a normal distribution function whose mean is 0 and the standard deviation is σ_{GRAM}. Assuming a normal distribution in observed relative affinities, $N\left(0, \sigma_{obs.}\right)$, the error of this random prediction, which we call Gaussian Random Affinity Model or GRAM, is also normally distributed, $N\left(0, \sqrt{\sigma_{GRAM}^2 + \sigma_{obs.}^2}\right)$. The prediction interval at a given confidence level (e.g. 95%) is then exactly determined,

$$PI_{GRAM}^{95} = 2\sqrt{\sigma_{GRAM}^2 + \sigma_{obs.}^2}$$

and the RMSE of the model is simply $\sqrt{\sigma_{GRAM}^2 + \sigma_{obs.}^2}$. For a dataset whose$\sigma_{obs.}$ is 1.76 kcal/mol, such as the Smyd3 example above, the PI^{95} and RMSE for the GRAM model that draws from the same distribution (i.e. $\sigma_{GRAM} = \sigma_{obs.}$) is then approximately 5.0 kcal/mol and 2.5 kcal/mol, respectively. The performance of GRAM is fully determined by the variance in the dataset, therefore removing any arbitrary aspect in the assessment of *in silico* affinity prediction methods.

Concluding Remarks

Robust prediction of the binding affinities of new compounds is important to the efficiency and the quality of medicinal chemistry campaigns. Despite the importance of these predictions in rational drug discovery and their strong theoretical foundations, achieving chemical accuracy of binding affinity measurements has been difficult despite of the intense research over the past decades. Part of the reason is the access to high-quality experimental data. In recent years there has been a strong push led by both the industry and the academic communities to share industry data,[18] including compound structures, binding affinities, and the structures of drug-target complexes, which would facilitate methodology development and evaluation. CSAR and D3R/SAMPL blind challenges are two most notable examples of such community-driven efforts, in which several phar-

maceutical companies including GSK have contributed structural and binding affinities measurements from their internal drug discovery programs.[19–21] Improving data availability is an important but first step toward better binding predictions. To objectively compare the quality of predictions across different protein targets and datasets, robust performance metrics and appropriate null model analysis must also be carefully considered.

A null model in its true sense needs to reflect the outcome of a random process that samples from the underlying statistical distribution of interest. For binding affinity, the true distribution is of course not known. However, a Gaussian or Gaussian-like distribution is a good approximation and this is supported by both theoretical estimate and empirical observations. In the case of relative binding affinities, the distribution mainly depends on the standard deviation (σ) as the mean is often close to 0 in our experience. For the purpose of assessing affinity prediction methods, knowing the detail of this distribution function is actually not necessary because it is usually possible to construct such a dataset so that the distribution of the selected affinities is approximately Gaussian. This then allows us to design a simple but true null model (GRAM) that generates binding affinities from $N(0, \sigma)$. Unlike other models (constant, simple molecular properties, etc.) casually employed, GRAM has an analytically determined, non-trivial PI and RMSE that depend on a single parameter σ or the variability in the affinity distribution – reducing σ leads to reduced PI and RMSE or better performance of the null model. In blind affinity prediction challenges, the variability in the contributed affinity data is often not managed purposefully. The general practice has been making sure that the dynamic range within the datasets is sufficiently large, usually 2 to 4 log units, which may not be always feasible. This makes it difficult to compare the quality of predictions (or methods) done on different datasets. By explicitly taking data variability into account, GRAM not only allows such comparisons to be made, but also relaxes the dynamic range requirement and potentially makes more industry datasets eligible.

It is also worth noting that prediction interval is a more informative performance metric for assessing the quality of affinity predictions in prospective situations. The popular performance metrics such as correlation coefficients, $RMSE$, and MUE do not provide direct guidance on how good the next prediction is. On the contrary, PI offers a range where the experimental affinity is likely to be found with a certain level of confidence. The estimated upper and lower boundaries, or the best and worst scenarios due to a proposed chemical modification, can be meaningful for chemistry planning. Based on our internal evaluation, FEP+ is evidently a more accurate and robust method for relative affinity predictions than GRAM and possibly other simpler scoring functions used in the industry. Driving down RMSE to 1.0 kcal/mol is a significant achievement for the field as a whole. However, the PI^{95} of FEP+, roughly 2 kcal/mol (or more than 25% of the dynamic range in the Smyd3 example), suggests the need for further improvements and the size of the gap to close for the goal of achieving chemical accuracy. The experimental variability in highly reproducible potency measurements can be obtained from a control compound used. For CD73, the estimated standard deviation in the measured pIC50s is 0.15 log units or 0.2 kcal/mol ($N = 87$, Figure 6). Assuming this is a constant over the entire pIC50 scale, the standard deviation in the measured relative affinities is roughly 0.3 kcal/mol. Figure 7 illustrates a simulated scenario of chemical-accuracy predictions, where we added a normally distributed random error, $Err \sim N(0, 0.3)$, to the observed relative affinities in the Smyd3 dataset. This may serve as a useful reference and an inspiration for future methodology development.

References

(1) Wang, A. L. et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* **2015**, 1–21.

(2) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954**, *22*, 1420–1426.

(3) Postma, J. P. M.; Berendsen, H. J. C.; Haak, J. R. Thermodynamics of cavity formation in water. A molecular dynamics study. *Faraday Symposia of the Chemical Society* **1982**, *17*, 55–67.

(4) Tembre, B. L.; McCammon, J. A. Ligand-receptor interactions. *Computers & Chemistry* **1984**, *8*, 281–283.

(5) Jorgensen, W. L.; Ravimohan, C. Monte Carlo simulation of differences in free energies of hydration. *The Journal of Chemical Physics* **1985**, *83*, 3050–3054.

(6) Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. Calculation of the relative change in binding free energy of a protein-inhibitor complex. *Science* **1987**, *235*, 574–576.

(7) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. Free energy calculations by computer simulation. *Science* **1987**, *236*, 564–568.

(8) Merz, K. M.; Kollman, P. A. Free energy perturbation simulations of the inhibition of thermolysin: prediction of the free energy of binding of a new inhibitor. *Journal of the American Chemical Society* **1989**, *111*, 5649–5658.

(9) Sherborne, B.; Shanmugasundaram, V.; Cheng, A. C.; Christ, C. D.; DesJarlais, R. L.; Duca, J. S.; Lewis, R. A.; Loughney, D. A.; Manas, E. S.; McGaughey, G. B.; Peishoff, C. E.; van Vlijmen, H. Collaborating to improve the use of free-energy and other quantitative methods in drug discovery. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 1139–1141.

(10) 3.3 - Prediction Interval for a New Response | STAT 501. https://onlinecourses.science.psu.edu/stat501/node/274.

(11) Abel, R.; Wang, L.; Harder, E. D.; Berne, B. J.; Friesner, R. A. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Accounts of Chemical Research* **2017**, *50*, 1625–1632.

(12) Gotelli, N. J.; Graves, G. R. *Null Models in Ecology*; The Smithsonian Institution, 1996.

(13) Gotelli N. J.,; McGill Brian J., Null Versus Neutral Models: What's The Difference? *Ecography* **2006**, *29*, 793–800.

(14) Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K. The SAMPL4 host-guest blind prediction challenge: an overview. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 305–317.

(15) Nicholls, A. A Different Conference. 2013; https://www.eyesopen.com/ants-rants/different-conference.

(16) Abel, R.; Bhat, S. In *Annual Reports in Medicinal Chemistry*; Goodnow, R. A., Ed.; Platform Technologies in Drug Discovery and Validation; Academic Press, 2017; Vol. 50; pp 237–262.

(17) Zheng, X.; Wang, J. The Universal Statistical Distributions of the Affinity, Equilibrium Constants, Kinetics and Specificity in Biomolecular Recognition. *PLoS Computational Biology* **2015**, *11*, 1–24.

(18) Warr, W. A. Blowing a breath of fresh share on data. *Journal of computer-aided molecular design* **2016**, *30*, 1143–1147.

(19) Carlson, H. A.; Smith, R. D.; Damm-Ganamet, K. L.; Stuckey, J. A.; Ahmed, A.; Convery, M. A.; Somers, D. O.; Kranz, M.; Elkins, P. A.; Cui, G.; Peishoff, C. E.; Lambert, M. H.; Dunbar, J. B. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *Journal of Chemical Information and Modeling* **2016**, *56*, 1063–1077.

(20) Carlson, H. A. Lessons Learned over Four Benchmark Exercises from the Community StructureâActivity Resource. *Journal of Chemical Information and Modeling* **2016**, *56*, 951–954.

(21) Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K. D3R grand challenge 2015: Evaluation of proteinâligand pose and affinity predictions. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 651–668.

Figures

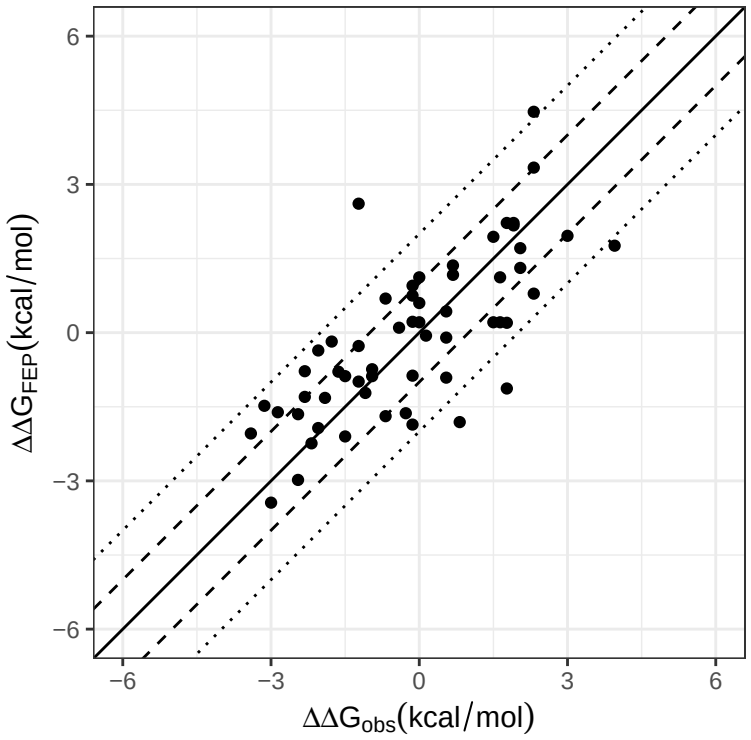

Figure 1: The comparison of the FEP+ predicted relative binding affinities of 57 pairs of Smyd3 compounds ($\Delta\Delta G_{FEP}$) and the corresponding experimental measurements ($\Delta\Delta G_{obs}$). The MUE, RMSE, and Pearson correlation coefficient (r^2) are 0.8 kcal/mol, 1.0 kcal/mol, and 0.54, respectively. A set of lines ($y = x$, solid, $y = x \pm 1$, short dash, and $y = x \pm 2$, dotted) are included for visual guidance.

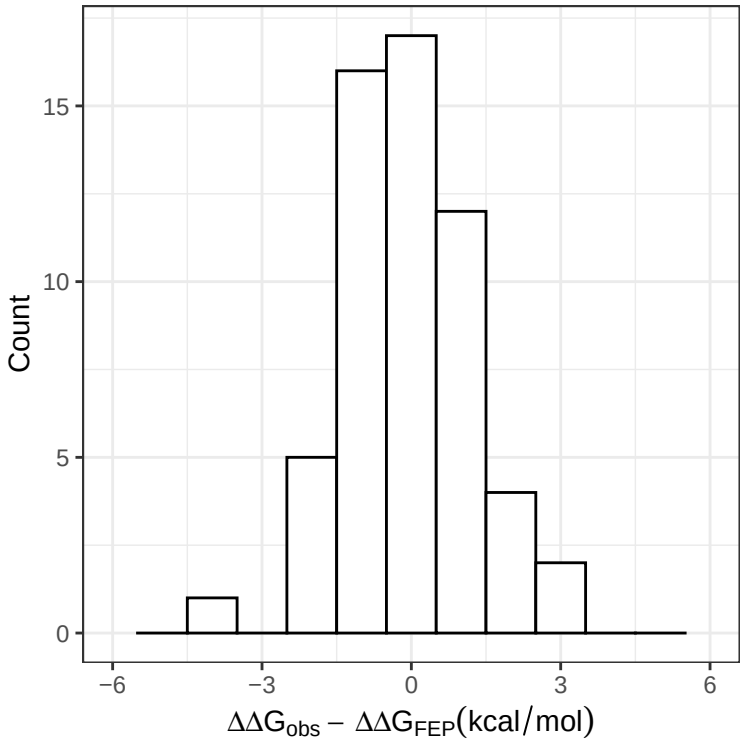

Figure 2: The histogram of prediction errors ($\Delta\Delta G_{obs} - \Delta\Delta G_{FEP}$) with a bin width of 1 kcal/mol.

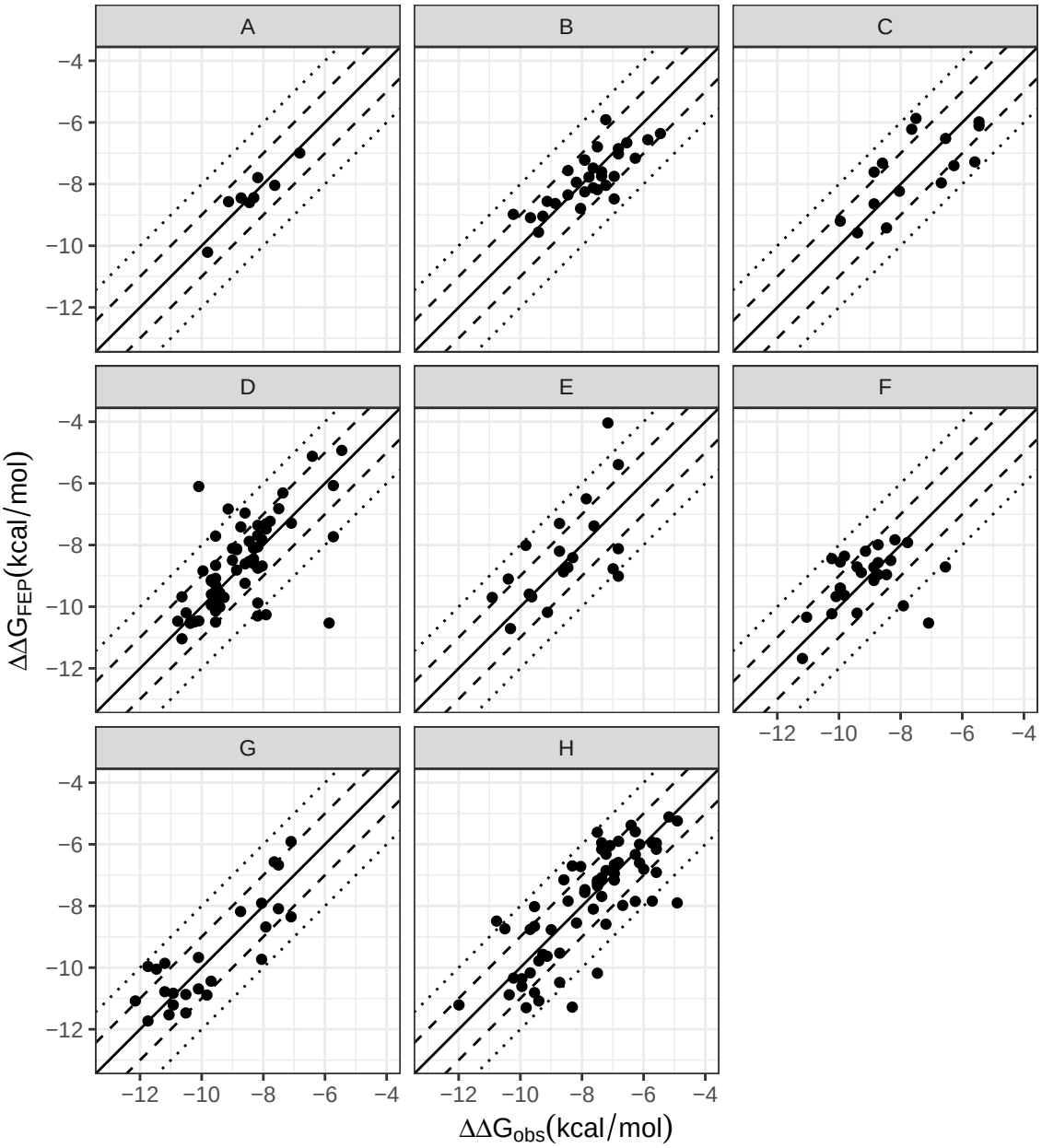

Figure 3: The comparision of the FEP+ predicted binding affinities (ΔG_{FEP}) and the experimental observations (ΔG_{obs}) over 8 protein targets.

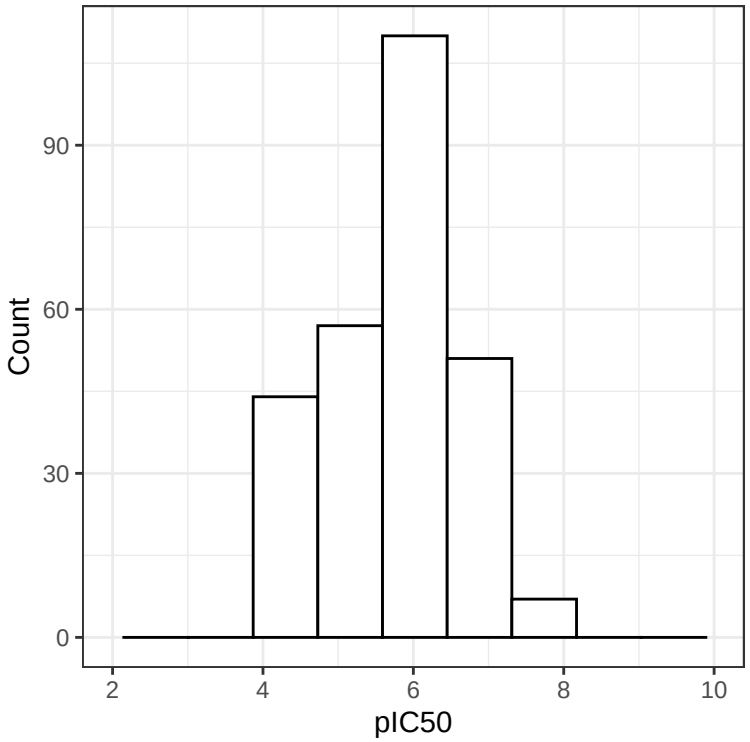

Figure 4: The histogram of the measured pIC50s (Rapid Fire Mass Spectroscopy) for 386 CD73 compounds with a bin width of 0.86 log units.

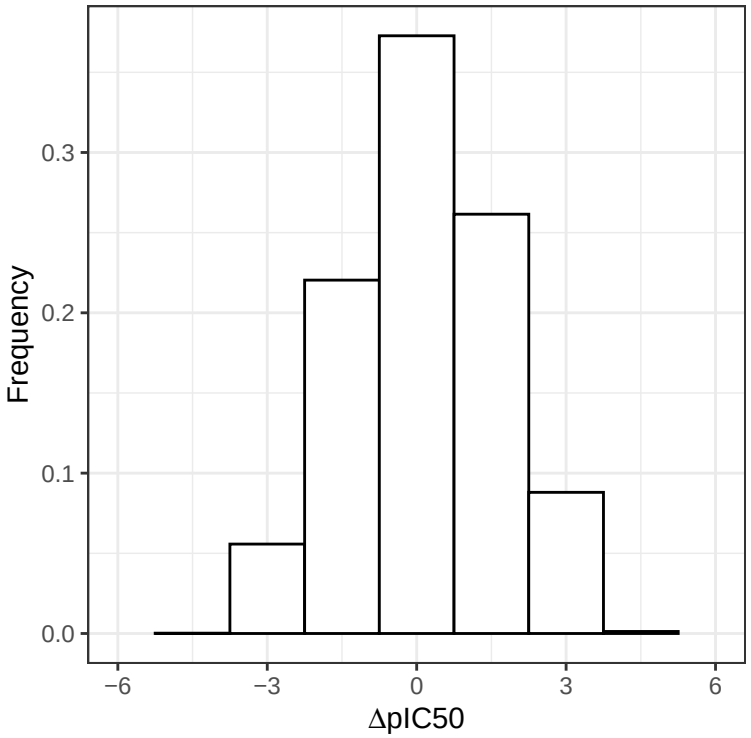

Figure 5: The histogram of potency (pIC50) evolution in 386 CD73 compounds over 30 days with a bin width of 1.5 log units.

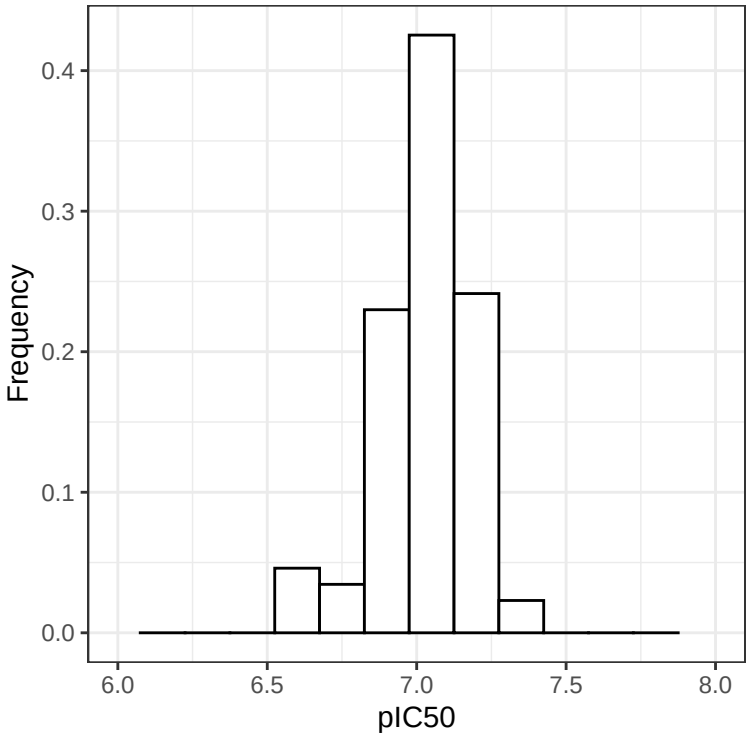

Figure 6: The histogram of measured pIC50s for a control compound used in the CD73 program ($N = 87$) with a bin width of 0.15 log units.

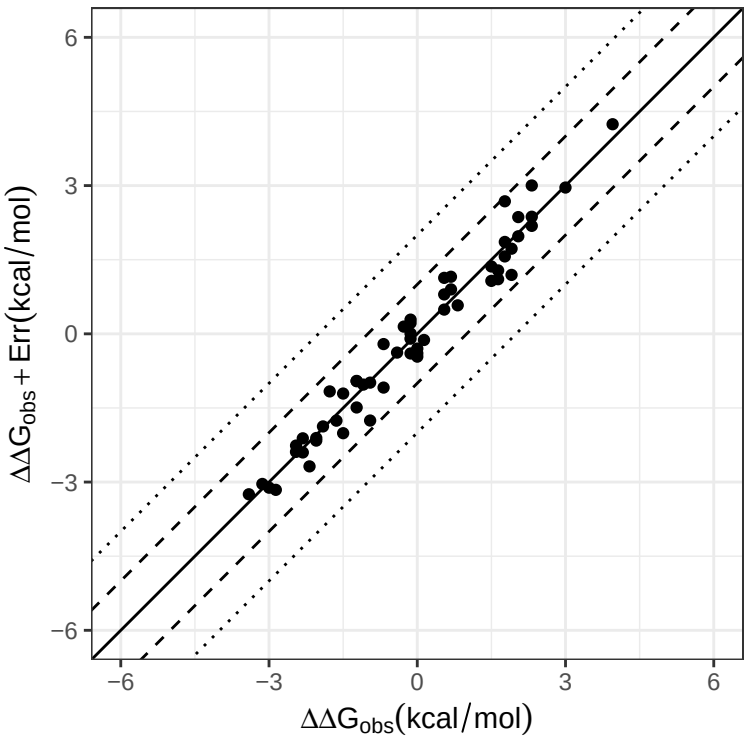

Figure 7: The comparison of the simulated prediction results obtained by adding a normally distributed random error ($Err \sim N(0, 0.3)$) to $\Delta\Delta G_{obs}$ to illustrate what the agreement may be, if the RMSE of future prediction method is as low as 0.3 kcal/mol. The Pearson correlation coefficient (r^2) is 0.96 for this particular run.