ELSEVIER

Healthy skepticism: assessing realistic model performance

Scott P. Brown, Steven W. Muchmore and Philip J. Hajduk

Structural Biology, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, IL 60064, USA

Although the development of computational models to aid drug discovery has become an integral part of pharmaceutical research, the application of these models often fails to produce the expected impact on productivity. One reason for this may be that the expected performance of many models is simply not supported by the underlying data, because of often neglected effects of assay and prediction errors on the reliability of the predicted outcome. Another significant challenge to realizing the full potential of computational models is their integration into prospective medicinal chemistry campaigns. This article will analyze the impact of assay and prediction error on model quality, and explore scenarios where computational models can expect to have a significant influence on drug discovery research.

Introduction

Computational modeling has become a routine scientific tool for gaining quantitative insight into potentially nonobvious relationships in relevant systems across a wide variety of disciplines [1]. When adequately developed for prospective application, computational models have the potential to produce tangible savings of time and money by systematically improving research efficiency. In industrial drug discovery, the use of computational modeling can have positive impact on research objectives in a variety of ways. Illustrative examples are the early identification and triage of problematic compounds before investing substantial research costs, or when improved decision-making translates directly into productivity gains for molecule selection during lead-optimization (LO) campaigns.

Unfortunately, one could argue that the full potential of computational modeling as a routine investigative tool in industrial drug discovery has been largely unrealized. There are several possible reasons for this, as model construction and assessment can be a challenging task with significant obstacles and pitfalls [2]. A particular challenge is defining the context within which a model can be expected to produce reliable answers. This has to do not only with the range of diversity (chemical or otherwise) within which the model can be considered valid (the 'applicability domain'), but also with the expectation or probability that the predicted results will actually influence decision-making. Thus, even when computational models have been 'validated' using conventional statistical approaches, their prospective application to naïve datasets or to address particular challenges may be inappropriate. While a range of statistical parameters can be applied to assess model performance (e.g. correlation coefficients and mean absolute errors), what is often overlooked is the impact of random error in the underlying data on which the model was built. Although this is a well-documented phenomenon [3], it is rarely addressed in the majority of papers that describe the assessment of predictive models in the life sciences. In fact, the authors are aware of only a few examples of publications that attempt to address these issues [4,5]. Ignoring these issues can lead to unrealistic expectations for model performance.

To explore the impact of errors in validation data, an analysis is presented here with the intention of addressing issues relating to realistic expectations of model performance that are consistent with the quality of the underlying data. First, historical data of measurement errors in binding-affinity assays are presented, which was acquired in the course of performing drug research at Abbott Laboratories. This allows the selection of a meaningful error estimate for investigating the effects of error as they relate to the performance of a computational method. The results of this analysis lead to some straightforward empirical rules-of-thumb for understanding the impact of error on model performance. Next,

Corresponding author: Hajduk, P.J. (philip.hajduk@sbcglobal.net)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2009.01.012

two hypothetical scenarios are constructed in which computational model is deployed to identify active molecules, and postulate a set of reasonable requirements for the method to realize systematic impact on research in drug discovery.

Assay error and correlations with experiment

One of the common approaches to assessing the quality of a model that predicts a measurable property (e.g. compound potency) is to measure the correlation of the predicted and experimental values using the Pearson product-moment correlation coefficient [6], R. Possible values of R span the interval $[-1, +1]$, and are independent of the slope of the regression. Values lying toward either end of the range, that is $|R| > 0.8$, indicate a high degree of correlation (either negative or positive) between datasets, which are typically taken to be reasonable evidence for good model performance. Absolute values less than 0.5 suggest that there is little correlation in the data. The square of the correlation coefficient, R^2, is known as the coefficient of determination, and reflects the fraction of the variance along the one axis that can be accounted for by the variance present along the other axis. There are, of course, many other statistical measures researchers have employed; however, the usefulness and interpretability of these performance measures will crucially depend on the influence of both experimental measurement error and prediction error from the model. In what follows, a strategy will be outlined for understanding the influence of error on measures of model performance and give some recommendations to aid future research.

A first step is understanding the impact of experimental assay error on model quality and predictive ability. To do this requires an estimate of the typical errors one might expect to encounter with the experimental measurement of binding affinities for industrial drug research. To produce an estimate for a probable value of such an error, we analyzed data for over 65,000 compounds from our corporate repository for which multiple measurements of activity against the same target were available, thus allowing standard deviations to be calculated. In assembling these data, the results were restricted to include only assays used to measure binding affinities or inhibitory activity of small-molecule ligands against protein targets. Shown in Fig. 1 is the resulting distribution of assay measurement standard-deviation values, which can be reasonably approximated by a Gaussian fit (red line).

The median error for this distribution is approximately 0.3 log units (corresponding to a factor of 2 in IC_{50} value), which is in good accord with the expected variability of a well-performing assay [7]. Given this estimate, it is straightforward to incorporate an expected variability in assay measurements for a given dataset solely because of the presence of experimental error. This approach is illustrated in Fig. 2 for a simple Pearson correlation. For this analysis, sets of hypothetical data were constructed with a defined number of points evenly distributed over a given number of log units. New values for the entire dataset are then sampled by adding Gaussian noise to each data point, after which the corresponding Pearson correlation is calculated and the process repeated. To illustrate the procedure, 'snapshots' of this process for three different datasets are shown in Fig. 2a, with 4 (black), 20 (green), and 50 (red) data points (all initially distributed evenly over 2 log units). When this process of sampling and resampling is simulated many times, it produces a range of Pearson correlations. Shown in

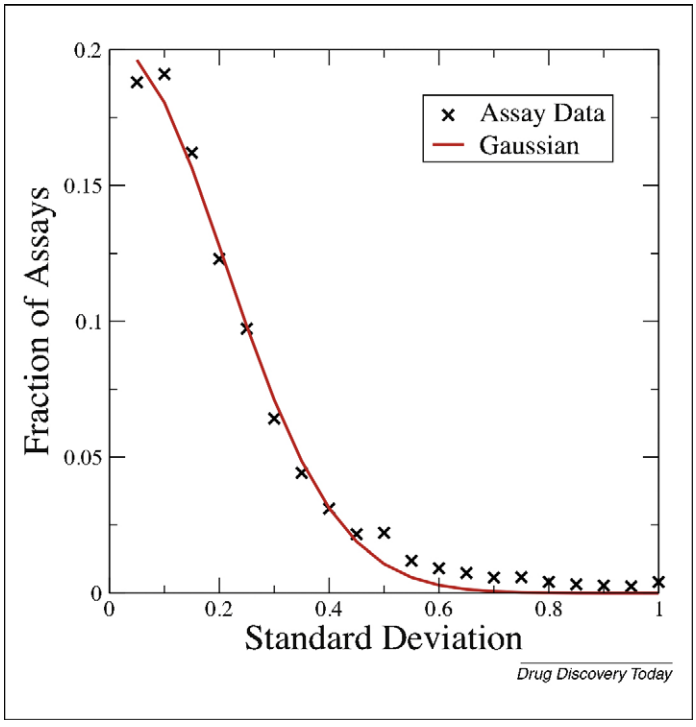

FIGURE 1

Plot of the fraction of assay measurements having a given standard deviation, in units of the negative logarithm of the measured IC_{50} value, or pIC_{50}. Results were obtained from in-house assay data on over 65,000 compounds. The distribution is reasonably approximated by a Gaussian (red line) with $\sigma = 0.20$.

Fig. 2b is the variability derived based solely on the spread induced by a 0.3 log-unit error. The median values and breadth of these distributions obviously depend on the number of data points, as well as on the number of log units over which the points are distributed.

It can be seen from Fig. 2b that the range of possible correlation coefficients for 20 data points (green line) distributed over 2 log units typically varies from 0.7 to 0.9. This has immediate utility in assessing realistic model performance. For example, consider that a model was constructed to predict the values for these 20 data points and that the observed correlation between experimental and predicted values was 0.95. On the basis of the plot in Fig. 2b, it is exceptionally improbable (0.1% chance) that a correlation coefficient between multiple experimental measurements could exceed 0.95. Therefore, this should be considered unrealistic model performance that could reflect over-fitting or insufficiently sampled data. This is based on the common sense principle that the error in predictions of ligand affinities is unlikely to be less than the resolution (e.g. error) of the experimental data.

A more comprehensive look at the behavior of R as a function of error, number of data points and potency span, is presented in the contour plots in Fig. 3. Fig. 3 shows two-dimensional (2D) contour plots for two different performance measures, organized by row. In the top row (Fig. 3a, c and e), the contoured landscapes show the 2D distribution of calculated average R-values, $\langle R \rangle$, as a function of the number of data points and spanned range of the potency values. The bottom row (Fig. 3b, d and f) shows the 2D contours for the calculated standard deviation in the R-values, σ_R, also as a function of the number of data points and spanned potency range.

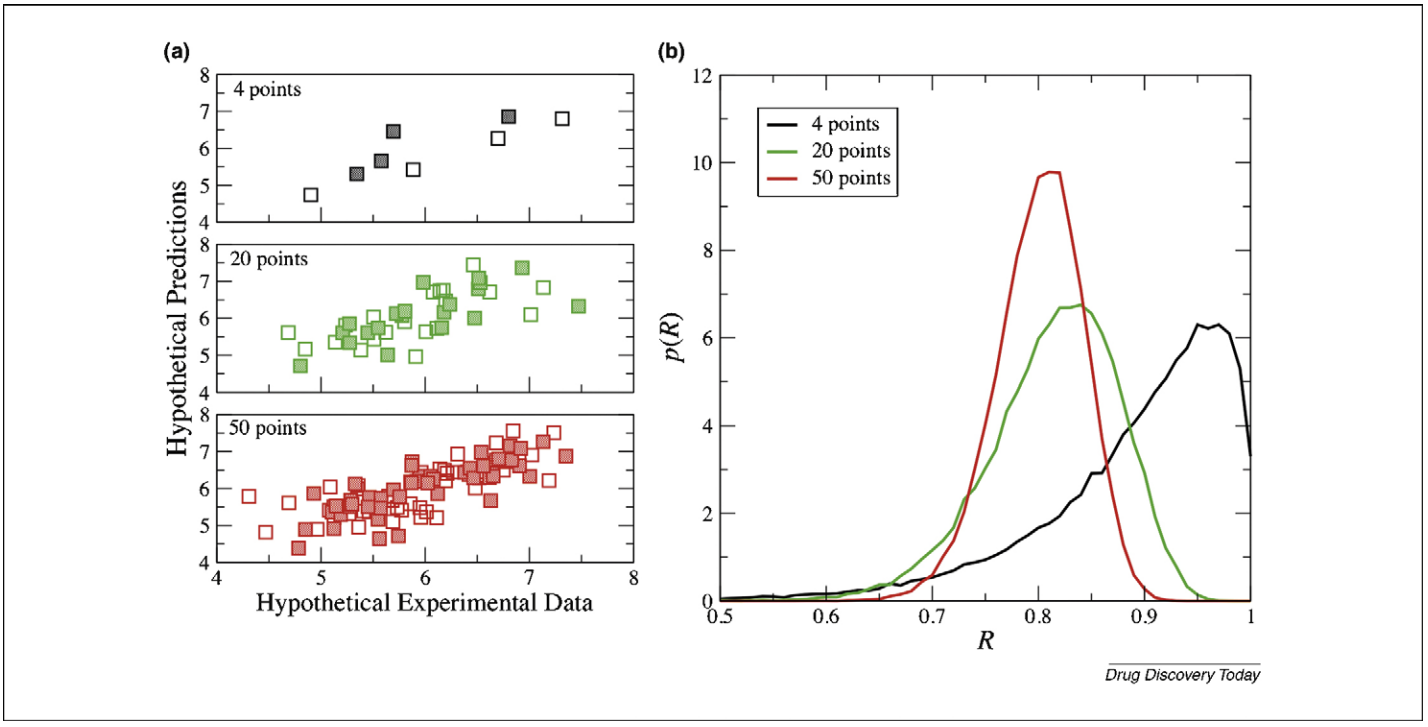

FIGURE 2

Plots illustrating the procedure for introducing sampling error into simulated data. **(a)** Two independently sampled snapshots (open and filled squares) are shown for three different numbers of points initially distributed over 2 log units. **(b)** Distribution of possible *R*-values generated by the 'snapshots' of the data in (a). 50,000 iterations were used to generate the *R*-value distributions.

For the $\langle R \rangle$ and σ_R landscapes in each row, the data in the first column (Fig. 3a and b) were produced using snapshots varied by an identical sampling error of $\sigma = 0.3$ for both experiment and prediction data. The data in the second and third columns (Fig. 3c–f) used a larger value for the sampling error in the prediction data ($\sigma_{pred} = 0.6$ and 0.9, respectively). Thus, the data in Fig. 3a and b describe an expected achievable range in *R* based on experimental variability alone, while Fig. 3c–f show the expected range in *R* based on increasing variability in predicted values. It is clear from these plots that the span of the potency range in the data has the greatest influence on the $\langle R \rangle$ value (top row), while both the span and the number of points influence the variation in σ_R (bottom row).

The plots in Fig. 3 can serve as a reality-check for model performance for any given experimental dataset. For a given number of data points and potency span, the range of probable *R*-values based on experimental and prediction error can be derived. This range should serve as a benchmark for the correlation that can reasonably be achieved with predicted values. For example, if a dataset spans a potency range of only 2 log units, it is clear from Fig. 3a that the correlation because of experimental error alone ($\sigma = 0.3$) is likely to be 0.8 – regardless of the number of data points beyond a certain threshold (minimum) value. Thus, obtaining correlation coefficients between experimental and predicted values in excess of 0.8 would be a cause for skepticism in the model, and larger prediction errors reduce this value further.

Unfortunately, this simple metric is often violated in the literature reports of model performance. In Fig. 4 we compare a representative set of literature-reported *R*-values obtained from an online search of life-science publications (the search was restricted to those articles describing predictive models with a reported *R*-value contained in the abstract and a publication date in the years 2006 and 2007). The reported *R*-values are shown as red triangles. Also shown in the figure are the ranges for calculated $\langle R \rangle$ values obtained using the reported parameters for the data (i.e. potency range and number of points) with errors of 0.3 (experimental assay error alone, blue bars) and 0.6 log units (conservative prediction error, green bars). It can be seen from this simple analysis that 8 of the 16 reported *R*-values (50%) equal or exceed that which could be expected based on experimental error alone (i.e. the reported value is within or above the blue error bars). The majority of *R*-values (11 out of 16) also exceed that which could be expected based on very conservative prediction error (i.e. the reported value is above the green error bars). Thus, it is our opinion that the majority of *R*-values obtained from this (small) literature sample are unsubstantiated given the properties of the underlying data.

Of course, a formal assessment of the impact of assay or prediction error on the examples discussed here should be explored using numerical simulation on the actual dataset used in each case (with appropriate error distributions for experimental and prediction error). In the absence of such an analysis, the data presented in Fig. 3 can serve as a reasonable guide for understanding the potential impact of such errors. In addition, these graphs allow a researcher to define an appropriate dataset (e.g. how many data points over what potency span) to achieve a desired level of performance.

Hypothetical scenarios for computational modeling

The preceding section outlined a simple approach for assessing whether the performance of a model is justified by the underlying

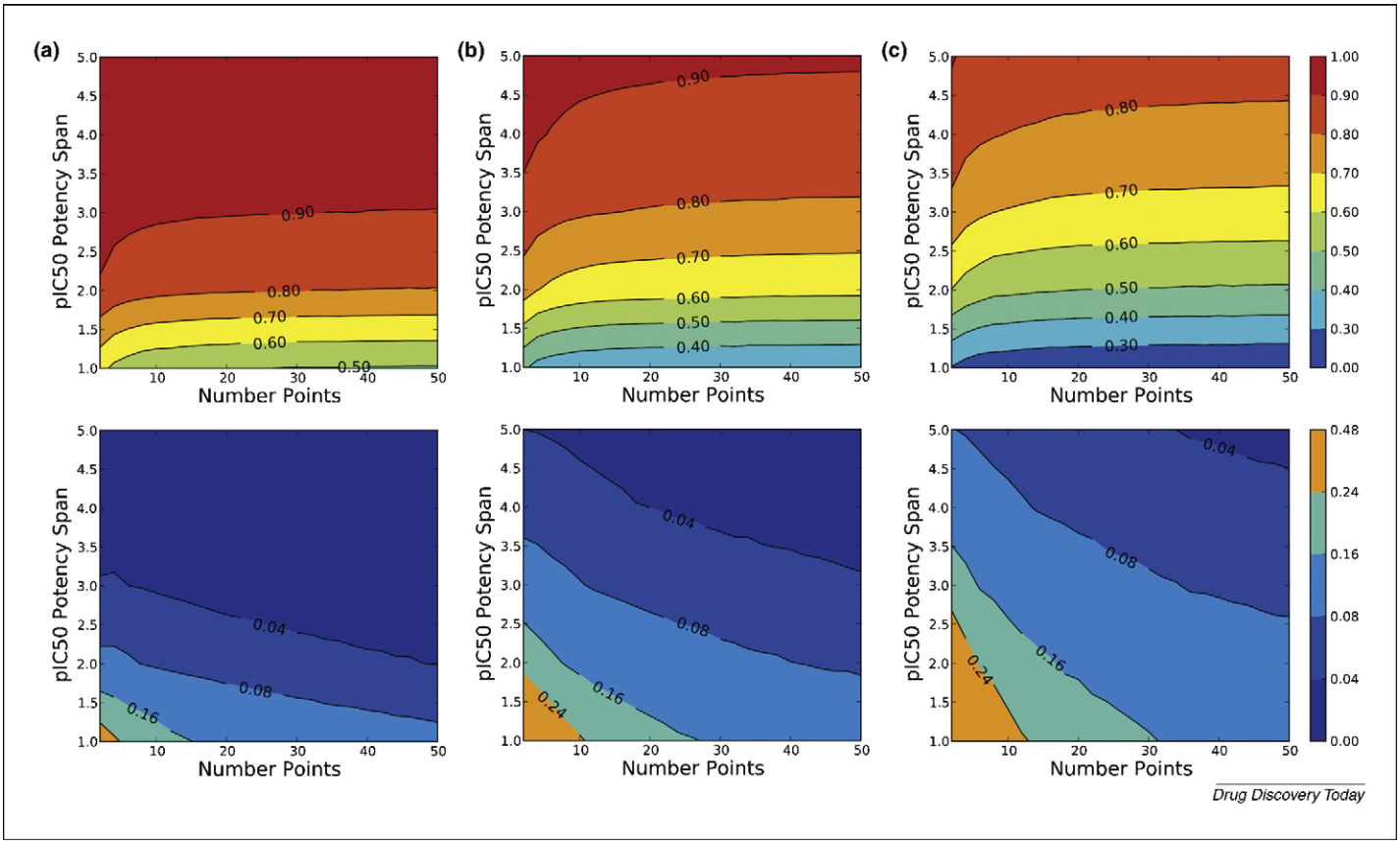

FIGURE 3

Contours showing the two-dimensional distributions of average *R*-values, $\langle R \rangle$ (top row), and the corresponding standard deviations, σR (bottom row), as a function of the number of data points and span of the potency values. Identical sampling errors are used for experimental uncertainties in each column ($\sigma_{expt} = 0.3$), whereas the sampling errors used for the prediction uncertainties are different in each column: **(a)** and **(b)**, $\sigma_{pred} = 0.3$; **(c)** and **(d)**, $\sigma_{pred} = 0.6$; and **(e)** and **(f)**, $\sigma_{pred} = 0.9$.

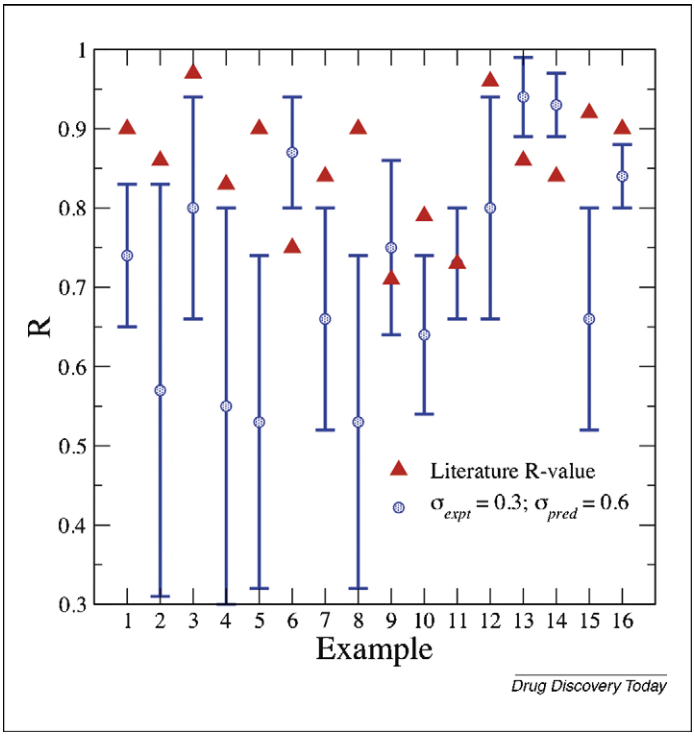

FIGURE 4

Data showing 16 examples of reported *R*-values appearing in life-science publications (*J. Mol. Model.*, *J. Am. Chem. Soc.*, *Bioorg. Med. Chem.*, *Biophys. Chem.*, *Proteins*, *J. Med. Chem.*, *J. Comput. Chem.*, *J Chem Inf. Model.*, *Biophys. J.*, *Biochem.*, *Protein Sci.*, and *Protein Eng.*) in the years 2006 and 2007 (red triangles). Also shown are the calculated $\langle R \rangle$ values assuming an error of $\sigma_{expt} = 0.3$, $\sigma_{pred} = 0.6$ (blue, with accompanying error bars to indicate the standard deviations σR in units of pIC_{50}). Note that the simulated $\langle R \rangle$ values systematically decrease as larger sources of error are incorporated.

quality of the experimental data. While some examples have been highlighted, in which reported model performance is probably unsupported by the validation data (Fig. 4), there are certainly many examples of models in the literature that are derived from sufficiently large and diverse datasets and appear accurately to capture aspects of protein–ligand recognition. The question then becomes, when can such a model be productively used in driving decision-making in drug discovery?

Interestingly, this is also a question of variability. Consider two very different scenarios: high-throughput screening (HTS) and LO. In HTS, large numbers ($\sim 10^6$) of highly diverse molecules are assessed to find a small number of active hits. Given that HTS hit rates are in the order of 0.1%, the vast majority of compounds are inactive. By contrast, during LO, modest numbers (10^2 to 10^3) of highly similar molecules are synthesized and assessed for changes in their properties. Given that highly similar molecules are likely to have similar biological activity [8,9], the variability in the potencies for a set of highly similar molecules is expected to be significantly narrower compared to HTS. How good then does a

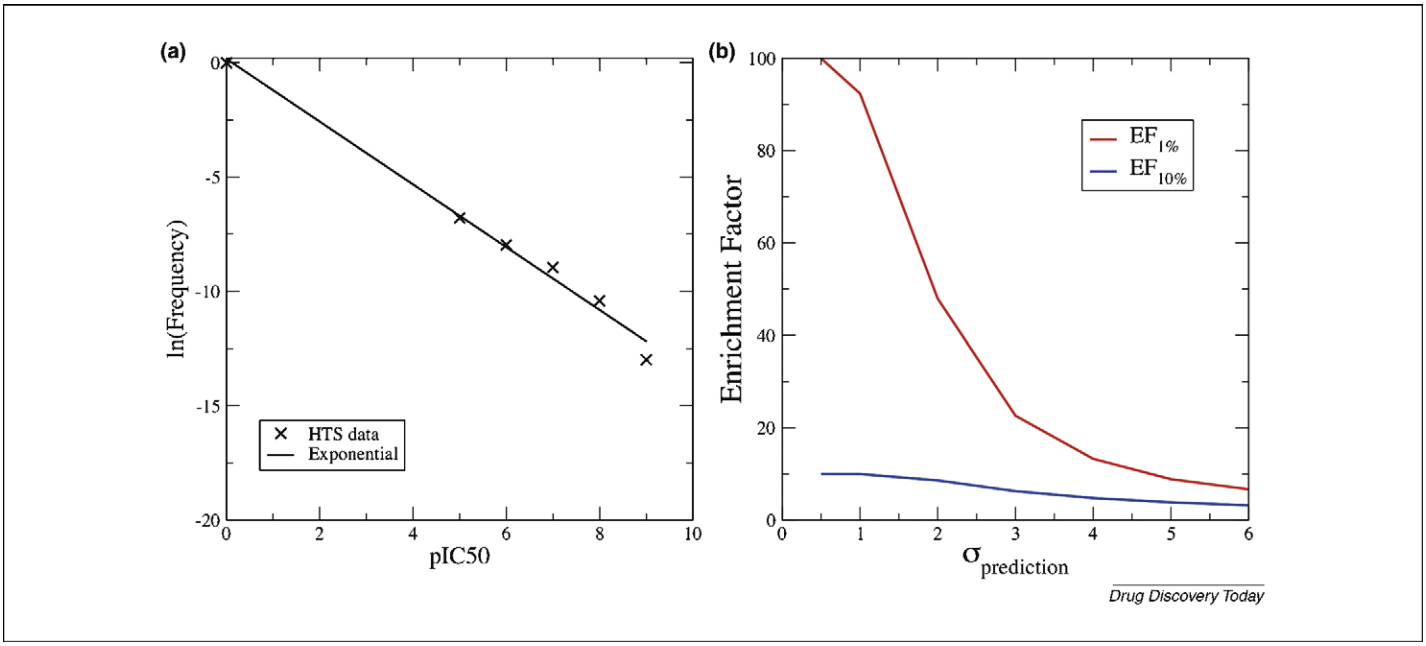

FIGURE 5

Analysis of small-molecule potency data obtained from high-throughput screening. **(a)** Binned data showing frequency of occurrence of potencies, with a linear fit to the data of $\ln(y) = 0.16 - 1.37x$. **(b)** Enrichment factors in the top 1% (red) and top 10% (blue) as a function of prediction error. The theoretical maximum enrichment factor for this dataset is 1000.

model need to be (in terms of prediction error) to actually impact these processes in real time? The relative utility of predictive models for both of these scenarios will be discussed below.

The first situation explores model impact on HTS. This is a case in which there is a large database of molecules with potencies distributed as shown in Fig. 5a. The data in Fig. 5a represent an average of observed potencies for 'active' molecules, or 'hits,' obtained from historical data over a large number of screens at Abbott Laboratories. The challenge for impacting HTS is to recognize a small number of actives in a 'sea' of inactives, which will require an ability to distinguish molecules separated by potency differences of 4 log units or more. To analyze model performance, we define an 'active' to be any molecule having a potency of at least 10 μM, that is a pIC_{50} value greater than 5. The range of potencies from $pIC_{50} = 0$ to less than 5 then constitute our set of inactives. As previously stated, the number of inactives in HTS vastly outnumbers the actives, with average hit rates of approximately 0.1%.

A reasonable quantitative measure of performance for this exercise is enrichment [10], which captures the ability of the model to 'enrich' the fraction of actives at the top of the ranked list relative to what would be obtained by randomly choosing compounds from the database. This requires us to generate ranked lists of molecules, for which the following procedure was used. 50,000 randomly sampled and stored potency values were generated according to the distribution shown in Fig. 5a. Those potencies greater than 5 were designated as 'actives,' and all others from 0 to 5 represented 'inactives.' Out of the 50,000 sampled points there were, on average, 50 actives. Each of these sampled sets was then passed to a function that adds random Gaussian noise to each pIC_{50} value to simulate the presence of scatter owing to error in experimental or predicted values. This generates a ranked list from which the enrichment factor is calculated, with final enrichments obtained by averaging over multiple samplings of the data.

Shown in Fig. 5b are the enrichment curves as a function of prediction error. It can be seen that for small errors large enrichments are achieved. For example, for $\sigma_{pred} = 1.0$ the top 1% of selected compounds have an approximately 90-fold enrichment over random, while in the top 10% of compounds there is approximately 10-fold enhancement over random. Even for an egregious prediction error of $\sigma_{pred} = 5.0$ log units, a model can still maintain a capacity to perform enrichment at close to 10-fold above random in the top 1% of the ranked list. This may seem surprising, yet it is consistent with the fact that the computational method is only required to recognize actives separated by over 4 orders of magnitude above noise (inactives). This kind of prediction error (and subsequent enrichment in virtual screening) is in fact on par with what one might expect based on a simple correlation between potency and molecular weight, as has been reported for some systems [11]. It should also be noted here that the model is not required to predict the unique identities for each active molecule. It is only required that the method distinguishes *any active* at positions in the top 1–10% of the ranked list. This scenario illustrates why the docking and scoring of molecules in virtual screening is able to provide statistical enrichment. However, ten-fold enrichments over random screening, while statistically significant, still translate to lower overall hit rates (~1% versus 0.1% for HTS). This may be at least part of the reason why a recent comprehensive study on the performance of docking and scoring across a wide range of systems documented only moderate overall performance [12].

Next, model performance in the context of LO is examined, for which the parameter to be optimized and tracked is again compound potency. Yet in this case the range of potency values is significantly narrower than for HTS. Given an active molecule, a medicinal chemist will typically envision a range of potential modifications and prioritize those that fit with established SAR

and are synthetically accessible. On the basis of the results for the first set of synthesized compounds, the next round of design will be initiated and prioritized accordingly. This process continues until not just potency but all drug-relevant properties (e.g. solubility, bioavailability and specificity) are optimized. It has long been the goal for computational models to enhance this process by expanding the numbers of compounds that a chemist can assess, prioritizing those compounds likely to be active, and deprioritizing those compounds likely to be inactive. Realization of this goal would substantially shorten the timelines to clinical candidates by reducing the numbers of compounds that actually need to be synthesized and tested. Despite huge investments into this area in both academia and industry, this goal has yet to be attained. LO can certainly be informed by computational models, but decision-making is not driven by these tools.

At least part of the reason for the modest impact of computational models in LO is the narrow range of potencies that result for groups of highly similar molecules. Figure 6a shows the average distribution of the changes in potencies that result for a set of molecules produced through the synthetic chemical transformations explored by medicinal chemists in pursuit of LO. The data in the plot were obtained by analyzing more than 2 million pairs of highly similar compounds with activity data against 50 different protein targets. This analysis is similar to that reported previously [13], but was performed on a much larger dataset that did not restrict the compound pairs to be related by a single chemical transformation. The distribution of data points in Fig. 6a is reasonably fit by a Gaussian function (red line).

The fit implies that, during the course of LO, on average ≈80% of synthetic modifications to parent molecules will result in potencies for child molecules that lie within 10-fold of the parent, and that only ≈20% of child molecules will exhibit potencies that change by greater than 10-fold. Of these child molecules with 10-fold or greater differences in potency from the parent, half (≈10%) will be 10-fold more potent than the parent, and half will be 10-fold *less* potent than the parent molecule.

The distribution in Fig. 6a can be used to quantitatively assess the ability of a computational model to impact the efficiency of LO chemistry. By sampling from the empirical fit, the observed probabilities are captured for molecule potencies generated over the course of LO. We can then statistically assess the impact of applying a computational model with varying degree of prediction accuracy and calculate a hypothetical effect on compound prioritization. To accomplish this the results are referenced to the performance of the chemist, who on average will produce one 'active' molecule (defined as having tenfold greater potency relative to parent) for every ten molecules made. This will be our 'random sampling' of actives, even though it should be stressed that the distribution shown in Fig. 6a does not result from 'random' changes to molecules, in that medicinal chemists utilize all available information in designing analogs.

Note that, as in the case of HTS, the method is not required to correctly predict the unique identities of the actives, but rather it is merely required that the presence of any active occur toward the top of the ranked lists. The formal procedure for assessing expected performance is analogous to that described for assessing model impact on HTS data as described above. 50,000 randomly sampled and stored potency changes were generated according to the distribution in Fig. 6a, with an assignment of 'inactive' to those potency changes falling between 0 and 1.0. Only half of the potency changes greater than 1.0 were designated as 'actives' (those corresponding to 1–10-fold *gain* in potency), such that

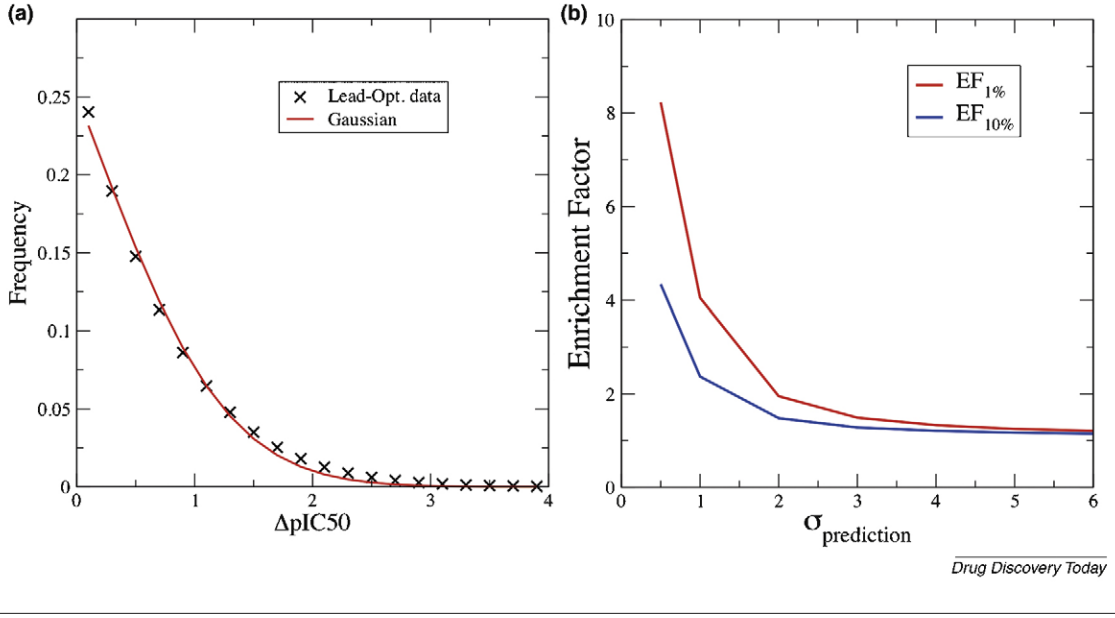

FIGURE 6

Analysis of small-molecule potency data produced during lead optimization. **(a)** Historical Abbott data showing distribution of relative potencies for over 2 million pairs of highly similar compounds against 50 different protein targets. Gaussian fit to the data (red line) with $u = -0.8$, and $\sigma = 1.1$. **(b)** Simulated impact of a computational model on the number of molecules a chemist need makes to find a single 'active' as a function of the error in the predicted values, with error bars showing possible spread in the values as the standard deviation of the mean. Superimposed in (b) are the enrichment factors for the top 1% (red line) and top 10% (blue line). The theoretical maximum enrichment factor for this dataset is 10.

on average there are roughly 5000 'actives.' Noise is then added to calculate enrichments. The cycle is repeated numerous times to obtain average expected performance. One important difference between the HTS and LO scenario is that prediction error was added twice in the case of LO, as the potency of both the parent and the child molecule needs to be predicted.

Shown in Fig. 6b are the enrichment factors as a function of prediction error for the top 1% (red line) and top 10% (blue line) of ranked compounds. From these curves, it is clear that prediction errors greater than 2.0 provide essentially no statistical enrichment for a typical LO campaign. Only when prediction errors fall below 1.0 are significant benefits expected (enrichment factors greater than 2 when the top 10% of compounds are selected). Even these potential gains must, however, be evaluated in the context of how a LO campaign actually progresses. First, the compounds produced during LO are 'active rich' (i.e. 10% of the compounds are active), such that the actual performance of any model will be significantly influenced by saturation effects [14]. The effect of saturation becomes even worse when the goal of LO is not increasing potency (which occurs ~10% of the time) but simply maintaining potency (which occurs >50% of the time) while other properties are optimized. Second, while enrichment factors at 1% are reported in Fig. 6b, it is our experience that chemists tend to make far more than 1 compound out of 100 synthetic proposals that are computationally evaluated. As shown in Fig. 6b, the actual benefit of any model decreases rapidly as larger fractions of the potential set of compounds are experimentally pursued. Thus, there are significant strategic (i.e. applying models only in situations where they are expected to produce real benefit) and cultural (i.e. prosecuting LO in such a way that maximizes the utility of computational models) barriers to realizing the full impact that even robust computational models can have on LO.

In comparing the enrichments observed for the two different scenarios in the top 1% of the ranked list, almost the entire range of prediction errors exhibit enrichments roughly tenfold greater for HTS than for LO. This is a reflection of the substantially larger tolerance of errors in HTS predictions, which seek to discriminate actives with potencies 4 log units above random noise. In the case of LO correct classification requires identifying actives in the range of 1–2 log units greater potency. While HTS has intrinsically larger tolerance for prediction error, it also requires intrinsically greater computational throughput. In LO projects it might be possible to realize impact by processing in the order of 100 molecules. This can be contrasted with HTS, for which the number of molecules is in the order of hundreds of thousands to millions. In comparing the cases of LO and HTS, it can be seen that each case requires different tradeoffs in negotiating the line between speed and accuracy.

Conclusions

In closing, the impact of experimental and prediction errors has significant effect on assessment of model quality. We have provided the results of simulations to aid researchers in evaluating the potential impact of assay and prediction error on their results. A general recommendation from these studies is that future validation datasets should (if possible) span a minimum potency range of 3 log units with a minimum of approximately 50 data points. Such datasets can be reasonably expected to have meaningful R-values in the range of 0.7–0.8, while datasets with fewer data points or a narrower potency range can be artificially inflated owing to chance correlation. Of course, this degree of quality in validation data may well be hard to achieve; however, in cases where one substantially deviates from these guidelines it is recommended that some measure of the impact of error on the assessment of model quality be incorporated in the final analysis. This was demonstrated only for the case where the use of a Pearson correlation coefficient is justified by the underlying distribution; however, it could be analogously extended to other (nonparametric) measures of correlation, such as Spearman's rho [15] or Kendall's tau [16].

Finally, provided one can construct a sufficiently validated computational model, we demonstrated hypothetical scenarios in which one might successfully employ such a computational tool and produce measurable impact on research productivity. It was found that the amount of prediction error that can be tolerated from a computational model depends on the requirements for the particular problem at hand. For example, in prioritizing compounds for HTS, there is a tolerance for prediction errors up to at least $\sigma_{pred} = 5.0$. Thus, even relatively crude models can potentially lead to productivity gains so long as they are capable of rapidly processing large numbers (10^6) of compounds. By contrast, in prioritizing compounds for LO, models with errors above $\sigma_{pred} = 1.0$ rapidly lose their ability to significantly increase discovery productivity. This is due primarily to the narrow potency range exhibited by sets of highly similar compounds and the correspondingly higher percentage of active compounds in the datasets to be examined. As a result, to fully exploit the potential of computational models during LO, one requires not only a robust prediction but also a medicinal chemistry culture that strategically and appropriately implements these models throughout the course of a synthetic program.

References

1 Silvert, W. (2001) Modelling as a discipline. *Int. J. Gen. Syst.* 30, 261–282
2 Stouch, T.R. *et al.* (2003) In silico ADME/Tox: why models fail. *J. Comput. Aid. Mol. Des.* 17, 83–92
3 Fisher, R.A. (1921) On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1, 1
4 Ozrin, V.D. *et al.* (2004) PLASS: protein–ligand affinity statistical score – a knowledge-based force-field model of interaction derived from the PDB. *J. Comput. Aid. Mol. Des.* 18, 261–270
5 Johnson, S.R. *et al.* (2007) Estimation of hERG inhibition of drug candidates using multivariate property and pharmacophore SAR. *Bioorg. Med. Chem.* 15, 6182–6192
6 Falk, R. and Well, A.D. (1997) Many faces of the correlation coefficient. *J. Stat. Educ.* 5, 1–12
7 *Assay Guidance Manual Version 5.0.* Eli Lilly and Company and NIH Chemical Genomics Center Available online at: http://www.ncgc.nih.gov/guidance/manual_toc.html (last accessed 8/17/08)
8 Martin, Y.C. *et al.* (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 45, 4350–4358
9 Muchmore, S.W. *et al.* (2008) Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* 48, 941–948

10 Stahl, M. (2000) Modifications of the scoring function in FlexX for virtual screening applications. *Perspect. Drug Discov. Des.* 20, 83–98
11 Ferrara, P. *et al.* (2004) Assessing scoring functions for protein–ligand interactions. *J. Med. Chem.* 47, 3032–3047
12 Warren, G.L. *et al.* (2006) A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49, 5912–5931
13 Hajduk, P.J. and Sauer, D.R. (2008) Statistical analysis of the effects of common chemical substituents on ligand potency. *J. Med. Chem.* 51, 553–564
14 Truchon, J.-F. and Bayly, C.I. (2007) Evaluating virtual screening methods: good and bad metrics for the ''early recognition'' problem. *J. Chem. Inf. Model.* 47, 488–508
15 Lyerly, S.B. (1952) The average spearman rank correlation coefficient. *Psychometrika* 17, 421–428
16 Kendall, M. (1938) A new measure of rank correlation. *Biometrika* 30, 81–89