ALOHA: a novel probability fusion approach for scoring multi-parameter drug-likeness during the lead optimization stage of drug discovery

Derek A. Debe · Ravindra B. Mamidipaka · Robert J. Gregg · James T. Metz · Rishi R. Gupta · Steven W. Muchmore

Received: 24 June 2013 / Accepted: 1 October 2013 / Published online: 11 October 2013

Abstract Automated lead optimization helper application (ALOHA) is a novel fitness scoring approach for small molecule lead optimization. ALOHA employs a series of generalized Bayesian models trained from public and proprietary pharmacokinetic, absorption, distribution, metabolism, and excretion, and toxicology data to determine regions of chemical space that are likely to have excellent drug-like properties. The input to ALOHA is a list of molecules, and the output is a set of individual probabilities as well as an overall probability that each of the molecules will pass a panel of user selected assays. In addition to providing a summary of how and when to apply ALOHA, this paper will discuss the validation of ALOHA's Bayesian models and probability fusion approach. Most notably, ALOHA is demonstrated to discriminate between members of the same chemical series with strong statistical significance, suggesting that ALOHA can be used effectively to select compound candidates for synthesis and progression at the lead optimization stage of drug discovery.

Keywords Domain of applicability · Model strength · Multi-parameter optimization · Model confidence · Compound fitness · Drug-likeness prediction · Lead optimization

Introduction

Drug discovery is commonly depicted as a pipeline beginning with target discovery and culminating in clinical candidate selection. It is widely appreciated that each stage of this pipeline is more resource intensive than the previous one, and therefore only a small percentage of compounds considered at any given stage are typically submitted to the next [1]. During lead optimization in particular, this attrition is accomplished not just by potency testing, but through a variety of tests to determine the so-called "drug-likeness" of each compound, including various absorption, distribution, metabolism, and excretion (ADME), multi-species pharmacokinetics (PK), and toxicology tests [2]. Depending on the envisioned therapeutic indication and dose, compounds must exhibit passing results in most or all of these tests, largely because the costs associated with failing in the clinical stages that follow are too high to risk advancing compounds that have not passed the less costly pre-clinical hurdles. With this testing fostered attrition paradigm firmly entrenched, the process of lead optimization reduces to a multi-parameter optimization [3–5] problem where the requirement of potency is balanced with a passing grade across a panel of pre-clinical fitness tests [6].

While lead optimization is inherently a multi-parameter problem, there have been many reduced parameter approaches to predicting a compound's drug likeness. For example, the most well-known example of a simple scoring metric for a compound's human oral bioavailability is

Electronic supplementary material The online version of this article (doi:10.1007/s10822-013-9679-x) contains supplementary material, which is available to authorized users.

The authors of this manuscript are employees of AbbVie. AbbVie participated in the interpretation of data, review, and approval of this article.

D. A. Debe (✉) · R. B. Mamidipaka · R. J. Gregg · J. T. Metz · R. R. Gupta · S. W. Muchmore
Platform Informatics and Knowledge Management, R&D, AbbVie, Inc., Mail Stop 10-2, 1 N. Waukegan Road, North Chicago, IL 60064, USA
e-mail: Derek.debe@abbvie.com

Lipinski's Rule of 5 [7]. Lipinski's rule of 5 is actually a set of 4 distinct compound property related thresholds (<5 H-bond donors, <10 acceptors, MW <500 Da, logP <5), and when a compound satisfies each of these criteria it is said to satisfy Lipinski's Rule of 5. Recently, there have been a variety of new "rules of thumb" in a similar spirit to Lipinski, including the 3–75 Rule [8] and the Golden Triangle [9].

While "rules of thumb" estimates of compound suitability have certainly thrived due to their ease of application and interpretability, there have also been many efforts to develop more complex algorithms and scoring functions to assess various aspects of a compound's drug-likeness. For example, the program Derek [10] (no relation to the author) is widely applied by drug discovery organizations at the recommendation of the FDA to predict potential toxicology liabilities. Simulation Plus, Inc., offers a suite of tools for ADMET model building and prediction [11]. These complex algorithmic approaches typically seek to predict the observed experimental results for a compound, but they do so at a cost of transparency and interpretability, which in turn tends to limit the scope of their deployment to domain experts. Recently, a number of methods have been developed which attempt to ease interpretation by reporting a single overall "desirability" score, notably the CNS-MPO [12] and QED [13] methods (both compared to ALOHA later in this manuscript). With or without the recent uptick in research activity, the accurately predicting the suitability of a small molecule to be a drug certainly merits consideration as one of the grand challenges of computational chemistry [14], even if today's leading methodologies typically really on empirical evidence rather than first principles.

A major limitation of most practiced approaches, whether they are a simple rule or a more complex algorithm, is that the error or confidence in the prediction is typically not reported nor well understood. For example, Lipinski's Rule of 5 does not explicitly report the likelihood of low bioavailability for a compound that satisfies the rules (that is not to say it cannot be calculated from a body of experimental data, however the rule as commonly applied does not provide this confidence information). Without such a confidence assessment, it is not possible to determine the likelihood that compound A has better bioavailability than B, even though A is a Lipinski pass, while B is a Lipinski fail. Knowing this probability conveys one's confidence in choosing to synthesize, test, or progress compound A versus compound B, highlighting a critical limitation of rules of thumb for lead optimization. The present authors [15] and others (notably Segall and co-authors [16]) have emphasized the importance of understanding prediction confidence, and view this as a critical vector for exploration in areas of predictive computational chemistry where confidence estimates have historically been absent [17].

In this paper, we present a generalized framework for scoring compounds according to their probability of passing a panel of pre-clinical tests that they would typically be expected to pass prior to advancement to clinical development. The method strives to have an interpretability and ease of application on par with a simple rule of thumb, while simultaneously considering a broad set of lead optimization related parameters. Most importantly, since the prediction error is accurately assessed, the method provides a statistically robust likelihood that any compound will pass all of the specified pre-clinical tests (as well as the likelihood that an arbitrary compound A will pass more tests than an arbitrary compound B).

Method

The ALOHA method for scoring the drug-likeness of compounds consists of three major steps:

1. Developing and training in silico models for multiple pre-clinical tests from historical data, such that each model yields the probability that a compound will pass its associated test,
2. estimating the error in each probability prediction from step (1), and
3. fusing the individual probabilities and errors from steps (1) and (2) into a single overall score representing the probability that the compound will pass all of the tests.

We begin by describing the probability fusion step (3), since it is the simplest and most general aspect of the present work. Assume that there are a set of N pre-clinical experiments that a compound must pass prior to advancement. Denote the probability of passing each of the N tests as p_1, p_2, p_3,…, p_n. Denote the standard deviation in each of these probability estimates as σ_1, σ_2, σ_3,…, σ_n. Then, the expected total number of tests passed is given by:

$$\mathrm{Pass}_{\mathrm{tot}} = p_1 + p_2 + p_3 + \cdots + p_n, \qquad (1)$$

and the error in the expected number of tests passed is given by the standard equation for the propagation of error when adding terms:

$$\sigma_{\mathrm{tot}} = [(\sigma_1)^2 + (\sigma_2)^2 + (\sigma_3)^2 + \cdots + (\sigma_n)^2]^{1/2}. \qquad (2)$$

Given $\mathrm{Pass}_{\mathrm{tot}}$ and σ_{tot}, and an assumption of a Gaussian probability distribution (we will verify this assumption later in the manuscript), the probability of passing all N tests (the ALOHA score) is determined by the proportion under the distribution curve to the right of N, for the Gaussian distribution centered at $\mathrm{Pass}_{\mathrm{tot}}$ with standard deviation σ_{tot} (practically, this calculation can be accomplished using a statistical package such as R, but since the error term is well

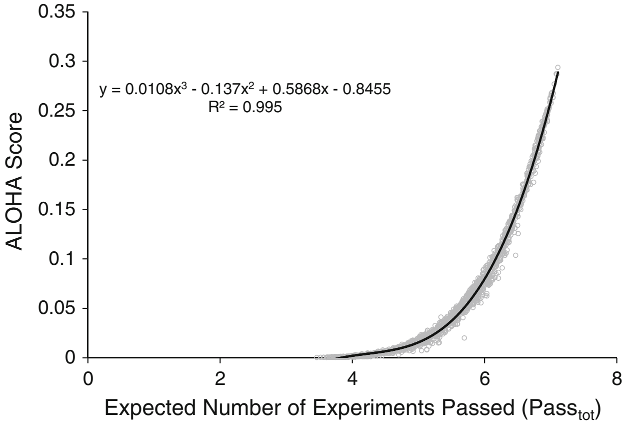


Fig. 1 Calibration of the ALOHA score using a heuristic to avoid the computational cost of determining the ALOHA score directly from Pass_{tot} and σ_{tot} assuming a Gaussian distribution. Since the error term is well behaved and decreases monotonically as Pass_{tot} increases, the ALOHA score can be estimated very accurately directly from Pass_{tot}. For N = 8, the ALOHA score is given by the equation shown on the *plot* (this equation is not valid for N $\neq$ 8, but can be redeveloped for other N as needed)

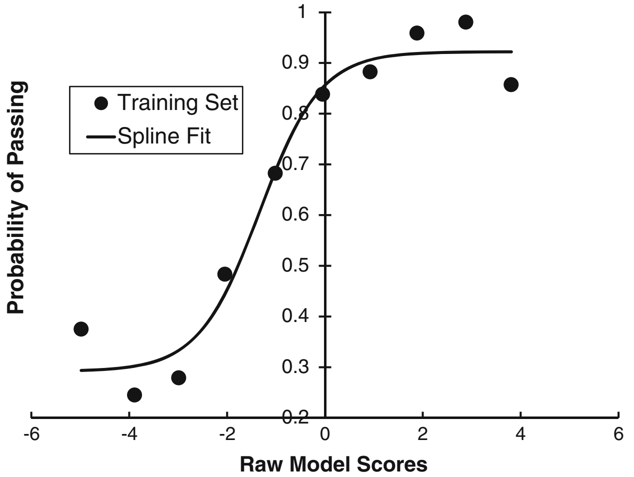


Fig. 2 The raw Bayesian modeling score can be mapped to a probability of passing by binning the raw scores for compounds in the probability calibration set and fitting with a *spline curve* (the binning algorithm used creates smaller, leftover bins at the raw score extrema, which can result in deviations from the fitted *curve* as seen in this example)

behaved, a heuristic can be developed to increase the calculation speed—Fig. 1). Note also that given Pass_{tot} and σ_{tot} for two molecules, A and B, the probability that A will pass more tests than B is determined by the area under the Gaussian distribution curve for A that is to the right of the Gaussian distribution curve for molecule B.

With step 3 achieved, our problem is reduced to determining the individual probabilities of passing each test and its associated error, steps (1) and (2) outlined above. To determine each individual test probability (step 1), we have chosen to use a widely accessible Bayesian modeling approach available in the commercial product Pipeline Pilot from Accelrys (though we certainly expect that other modeling approaches could also be adapted for this purpose). As an example, we will work through the case of developing a probability model for Dofetilide Binding, a high-throughput assay routinely run to uncover potential cardiac related liabilities. More than 10,000 compounds have been tested in this binding assay at Abbvie, so there is a large body of data to train on. The first step in our Bayesian model building is to define a passing threshold. For the Dofetilide Binding assay, we define a passing score as IC50 >10 μM (no binding). Historically, ~70 % of the compounds tested at Abbvie have passed this threshold, while ~30 % have failed. The Bayesian modeling technique we have employed is similar to that published elsewhere [18], with a few notable modifications. Most importantly, since we want our score to be easily interpretable and applicable to all potential drug-like compound classes, we use a novel two-step procedure to first train and then calibrate the model. First, we break the set of available experimental data into two separate, equally sized yet *structurally diverse and dissimilar* sets (clustered so that no compounds within or between either set share an ECFP-6 similarity >0.5). The first set serves as the *model training set*, while the other set is used to map the raw model scores to the historical probability of achieving a passing result (the *probability calibration set*). Next, a Bayesian model employing the passing threshold stated above is built using the standard Bayesian model building component in Pipeline Pilot. This Bayesian model is capable of generating a raw score for any new compound run through the model (typically some number between −10 and 10 depending on the exact parameters of the model). In order to make this raw score more meaningful and generally applicable, we score the probability calibration set using the Bayesian model and group the results into uniformly sized bins beginning with the highest resulting raw score to the lowest. Since we know the experimental data for each of these compounds, we can determine the expected probability of achieving a passing result for each bin and plot it versus the average observed model score as shown in Fig. 2.

By using an independent set of compounds to train the probabilities from the raw scores separately from the raw score training, one might assume that the model will automatically have a broad domain of applicability without bias towards the training set compounds. However, we have found that this is not necessarily the case and care must be taking when selecting parameters for the Bayesian model when a large domain of applicability (DOA) is desired. For example, we have found that including

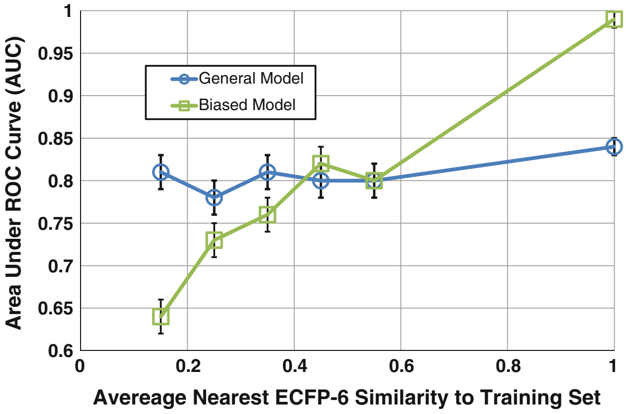


Fig. 3 Example results for DOA test #1. The *blue line* shows a general model whose performance is invariant to changes in chemical similarity to the training set. Conversely, the *green line* shows results for a biased model that performs much better on molecules near the training set. The *vertical error bars* in the AUC calculation were calculated according to the work of Nichols et al. [19]. While the *green* model clearly outperforms the *blue* model at high similarities to the training set, it actually performs significantly worse on molecules with limited or no appreciable chemical similarity to the training set. In this manner, we have verified that each of ALOHA's Bayesian models are performance invariant as the distance to the training set increases

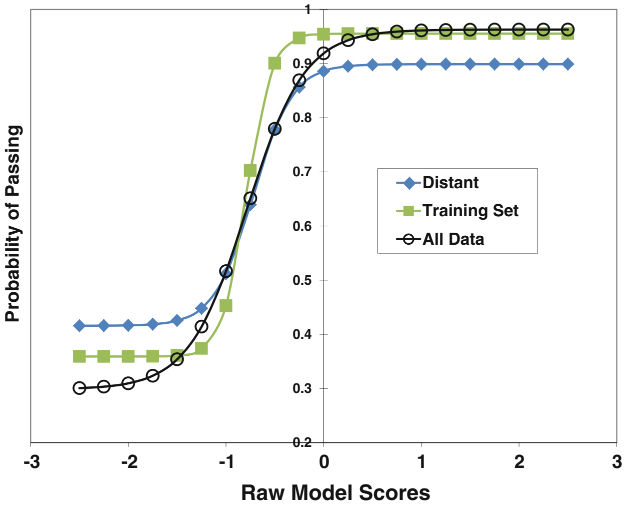


Fig. 4 Example results from DOA test #2 verifying that the raw Bayesian score to ALOHA probability score mapping is invariant as the distance to the training set increases. While the *spline curves* for this case overlap as desired, this is not necessarily the case with all types of model parameters. For example, for models using chemical fingerprint parameters, the probability score mapping spline curves for sets closer to the training set were stretched along a much greater span of raw scores than for more distant sets (having more chemical fingerprint features in common with the training molecules results in a broader raw score distribution). This stretching prevents a *single spline curve* from accurately fitting the mapping of raw scores to probabilities across all similarity domains, and so we avoided using any properties (notably fingerprints) that altered the raw score distribution width

chemical fingerprint terms in our model compromises their generality. We have developed two tests to assess and validate the domain of applicability of ALOHA's Bayesian models:

DOA Test 1 Using the area under the Receiver Operator Characteristic curve (AUC) as a proxy for model performance, we examine how the model performance changes when the model is applied to sets of compounds that are further and further from the training set (where the similarity metric is defined as the highest ECFP-6 similarity to any compound in the training set). Models with a bias toward the training set tend to perform much more poorly as the distance from the training set increases (Fig. 3).

DOA Test 2 Since a mapping from raw model scores to probability scores is performed, we also verify that the probability mapping curves are stable for sets of compounds that are further and further from the training set (Fig. 4).

To the knowledge of these authors, DOA Test 1 is a previously unreported but useful and general approach for assessing a model's domain of applicability. DOA Test 2 is specifically applied here to ensure that ALOHA's probability mapping step is stable and does not adversely affect the domain of applicability. As mentioned above and in the associated figure captions, we have found that the introduction of fingerprint terms into the Bayesian models typically narrows the domain of applicability considerably. We do not use these terms in our ALOHA models, since our objective with ALOHA is to maintain a universal small molecule applicability domain.

In addition to the Dofetilide Binding model, we have developed models for a variety of commonly employed pre-clinical tests. Table 1 summarizes the 8 tests that comprise the ALOHA results presented in this paper, the passing thresholds used, and the historic pass rates observed at Abbvie. In addition, we report a *model strength*, which we define as the dispersion (standard deviation) in the output score distribution for a diverse set of compounds. We find that this model strength term is useful for interpreting how valuable the in silico model is compared to running the actual experiment. For example, in the limit of no experimental error, the model strength for an experimental assay will approach 0.5, since running an error free experiment would yield a precise probability of either a 0 or a 1 of being above or below a certain pass threshold (if 20 molecules are tested, with 10 passing and 10 failing, then the dispersion across the scores is 0.5). This naturally represents the maximum model strength any in silico model could hope to achieve. For our in silico models, the probability scores lie between 0 and 1, and so the model strength will be something greater than 0 and

Table 1 ALOHA's individual models along with passing criteria, historic pass rates, and model strengths

Model	Passing criteria	Historic pass rate	Model strength
Dofetilide binding	IC50 ≥ 10 µM	68 %	0.22
CYP inhibition	3A4, 2D6, and 2C9 IC50 > 10 µM	57 %	0.15
Cell permeability (PAMPA)	$P_{eff} > 1 \times 10^{-5}$cm/s	65 %	0.16
Solubility	>10 µM @ pH = 7.2	62 %	0.20
Fraction absorbed (human)	>50 %	N/A (~80 % of drugs) [20]	0.12
Unbound clearance (human)	<130.0 (mL/min/kg)	N/A (~80 % of drugs) [21]	0.25
HepG2 toxicity	<50 % inhibition	60 %	0.13
Ames mutagenicity	Not "mutagenic" *or* not "mutagenic on metabolic activation"	86 %	0.07

The fraction absorbed and unbound clearance data are from the literature and hence the historic pass rates are not known. The model strength parameter (the dispersion of the model scores) was calculated using a diverse set of 500 compounds chosen from the Wombat database [22] (list with accession codes available in Supplemental material). The most discriminatory models in ALOHA are the models for unbound clearance, dofetilide binding, and solubility. Note that for the Ames test, the historic pass rate is 86 %, so the experiment itself has a model strength of 0.34 (not 0.5), which is an upper bound on the maximum possible model strength for the Bayesian model

Table 2 Data table demonstrating the declining relative error of ALOHA's probability based data fusion, even in the extreme case where $p_i = 0.5$ with a corresponding error, $\sigma_i = 0.5$ for each of the constituent models

# of experiments	Total passed	Error in total passed	Relative error
1	0.5	0.50	1.00
2	1.0	0.71	0.71
3	1.5	0.87	0.58
4	2.0	1.00	0.50
5	2.5	1.12	0.45
6	3.0	1.22	0.41
7	3.5	1.32	0.38
8	4.0	1.41	0.35
9	4.5	1.50	0.33
10	5.0	1.58	0.32
11	5.5	1.66	0.30
12	6.0	1.73	0.29
13	6.5	1.80	0.28
14	7.0	1.87	0.27
15	7.5	1.94	0.26

less than 0.5. A model strength approaching 0.5 is indicative of a model able to confidently place many compounds at the probability extremes, just as the real experiment is intended to do. The ability to report this model strength parameter is a useful byproduct of the fact that ALOHA models predict the probability (a number between 0 and 1) of exceeding a certain experimental result threshold, rather than the experimental result itself.

So, we have now accomplished step (1)—we have obtained the set of probabilities $p_1, p_2, p_3, \ldots, p_n$, that define the performance of an arbitrary new compound across our panel of tests. To apply the probability fusion method outlined by step 3, it remains to determine the standard error for each of these predictions (step 2). For a prediction, p_1, we define the standard error as:

$$\sigma_1 = ((1 - p_1) * p_1)^{1/2}. \tag{3}$$

Note that this equation corresponds to the standard error in a binomial distribution for n = 1. A few simple examples provide useful context as to the meaning of Eq. 3. For a weak prediction of $p = 0.5$ (a coin-toss as to whether the compound is a pass or fail), the associated error is $\sigma = 0.5$, and for a strong prediction of $p = 0.1$ or 0.9, the associated error reduces to $\sigma = 0.3$. The reader will note that these are relatively large error bars; the standard error bars for two compounds only become non-overlapping once there is a difference in scores greater than 0.6 e.g., once $p_A > 0.8$ ($\sigma_A < 0.4$) and $p_B < 0.2$ ($\sigma_B < 0.4$). Hence for any single test, once cannot confidently determine that molecule A is better than molecule B unless the probability scores explore the high and low probability extremes.

While the discriminatory power of any one ALOHA model may not be very high, for a series of tests, where the standard errors are governed by the propagation of error in Eq. (2), the relative error in the overall prediction decreases significantly, as outlined in Table 2. The ALOHA fusion technique is leveraging the principle that the triangulation of error has the effect of increasing prediction confidence as the number of experiments increases. ALOHA employs a philosophy wherein the results from a large number of broadly applicable models are fused together to yield an overall fitness score that is much more discriminatory than any one of the individual model scores alone.

At this point, it remains to be confirmed that the standard error estimate (Eq. 3) is accurate and the Gaussian assumption made in the data fusion step is valid. We can

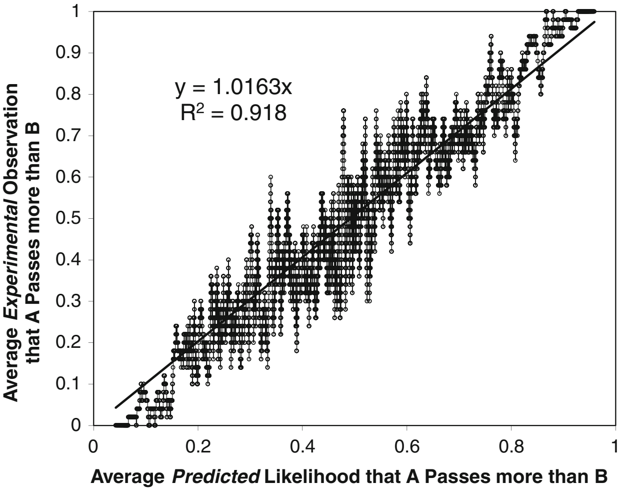


Fig. 5 The average experimentally observed probability that molecule A passes more tests than molecule B versus the average ALOHA predicted likelihood that molecule A will pass more tests than B (as noted in the text, given Pass_{tot} and σ_{tot} for two molecules, A and B, the probability that A will pass more tests than B is determined by the area under the Gaussian distribution curve for A that is to the right of the Gaussian distribution curve for molecule B). This validation test is based on 160 compounds that have all been run through six experiments. Each point on the *plot* represents a running average binning across 50 compound pairs beginning with the pairs with the highest predicted likelihood. Since the experimental results for these compounds are known, it is known for all pairs of molecules whether A is better than B, and hence for each set of 50 pairs, the actual probability that A is better than B (e.g. 30 out of 50 implies a probability of 0.6). The best fitting line through this data is indeed very close to this theoretical goal of y = x, confirming that the *error bar* approximation (Eq. 3) and the Gaussian assumption in the data fusion step are both valid

test the accuracy of the overall probability fusion predictions using compounds that have been run through multiple tests. For example, if we take a set of compounds whose experimental results are known for 6 tests (e.g. 6 passes, 5 passes, or 4 passes, and so on), then the corresponding 6-parameter probability fusion estimates for these compounds should be accurate. These accuracy validation tests are shown and explained in Figs. 5 and 6. Note that we can only use 6 tests here rather than the complete set of ALOHA 8-tests for this validation study since fraction absorbed and unbound clearance are literature data that we do not have for our corporate compound collection.

With major steps (1–3) complete and the associated confidence assessments, domain of applicability, and Gaussian approximation validated, we have described how the ALOHA determines the probability that any small molecule chemical entity will pass a pre-selected set of pre-clinical experiments. The next section presents further results and validations assessing the value of ALOHA in the context of a lead optimization campaign.

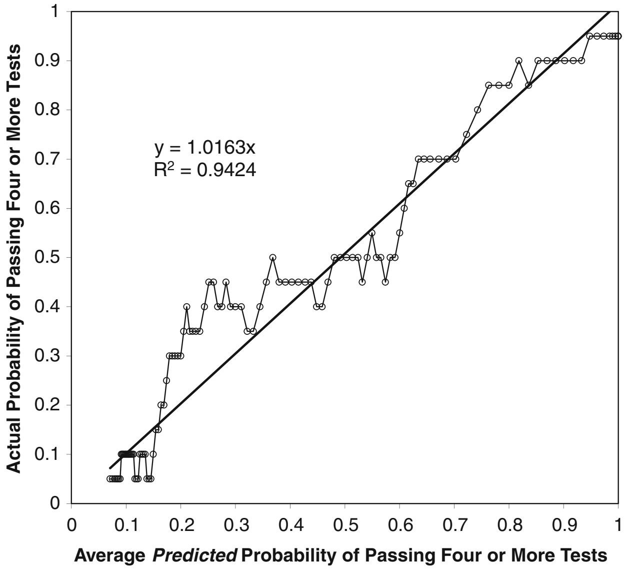


Fig. 6 Plot comparing the actual probability of passing four or more tests, versus the ALOHA predicted probability of passing four or more tests. Of the 160 molecules run through six experimental tests, 82 pass four or more tests. Each point on this plot represents a running average binning of 50 compounds, beginning with the compounds with the highest ALOHA score. Since the experimental results for these compounds are known, the actual probability that the 50 compounds in each bin pass 4 or more tests (e.g. 30 out of 50 implies a probability of 0.6). The best fitting line through this data is indeed very close to this theoretical goal of y = x, further confirming that the *error bar* approximation (Eq. 3) and Gaussian assumption in the data fusion step are both valid

Results

In this section, we present the following results:

1. Prospective Validation: What is the prediction quality for individual models on compounds synthesized and tested after the models were originally trained?
2. Comparison with well known chemical properties: What does ALOHA scoring look like on a standard logD versus Molecular Weight plot?
3. Performance compared to standard benchmarks: How well does ALOHA differentiate drugs from non-drugs compared to other common methods?
4. Score Variation versus Chemical Structure Variation: How does the ALOHA scoring vary with changes in chemical structure, and how well does it discriminate between compounds in a typical lead optimization campaign?
5. Using ALOHA: What are some practical approaches for visualizing the ALOHA output?

Prospective validation example

Since the primary subject of this paper is the ALOHA data fusion approach and not the performance of the individual

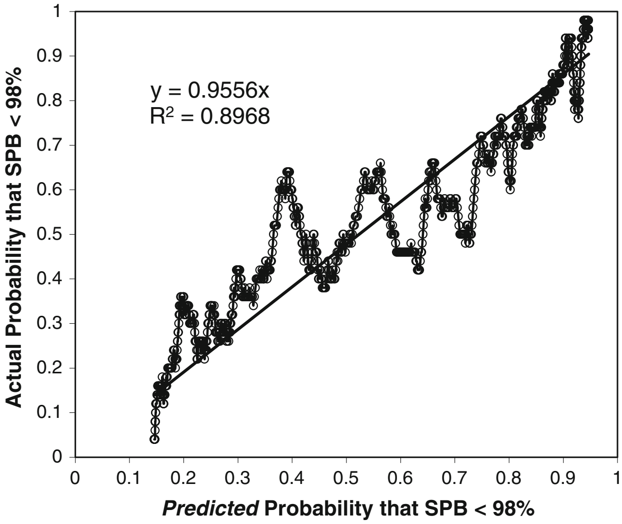


Fig. 7 ALOHA's prospective prediction results for the serum protein binding model (definition of passing in this experiment is <98 % observed binding). The model was trained on October 6, 2008, and all 1,034 compounds in this prospective study were first tested after this date and therefore not part of the training set. Each point on the *plot* represents a running average of 50 compounds, running from the highest to lowest ALOHA score. Since the experimental results for these compounds are known, the actual probability that the 50 compounds represented by each point will have <98 % observed binding (e.g. 30 out of 50 implies a probability of 0.6). The best fitting line through this data is very close to y = x, confirming that the Bayesian model probabilities are indeed accurate for prospective predictions

Bayesian models, this paper does not contain a systematic examination of the prospective performance for each and every model. However, it is important to show that the probabilities for the individual models are accurate when applied to molecules that were never tested prior to the development of the model. Figure 7 shows these prospective results for a universally applicable Serum Protein Binding Bayesian model developed using the approach described in the Methods Section. Note that Serum Protein Binding was not used in our final set of 8 ALOHA models (based on feedback from medicinal chemists that they would not typically disallow compounds based on this measure), but it was chosen for this prospective validation since a large amount of experimental data was produced using the same experimental protocol as the data used for the original model training.

Comparison with well known chemical properties

A number of recent publications have highlighted that drug-likeness is strongly correlated with calculated logD and molecular weight of the compound. The cleverly named "Golden Triangle" method maps out a triangular region in molecular weight versus logD space where compounds are much more likely to have desired PK and ADME properties. Given the emerging popularity of this metric, we have assessed how ALOHA scores compare (Fig. 8).

Performance compared to standard benchmarks

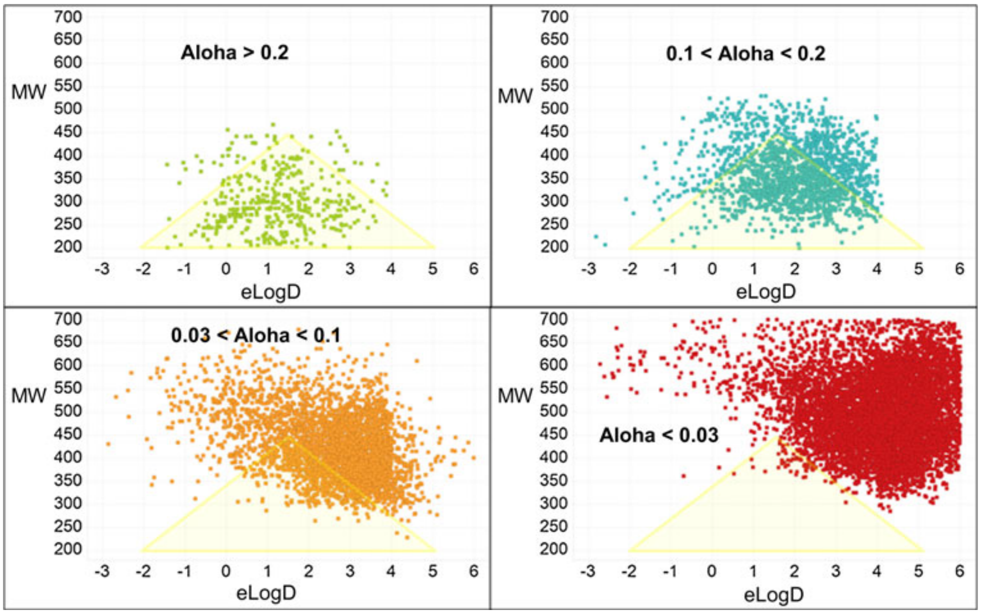
The ALOHA method, while an easily interpretable probability score, is obviously more complex to implement than a simple rule of thumb, so it certainly bears asking the question whether there is added value in using ALOHA. One approach to measuring this is to ask whether ALOHA does a better job at discriminating between drugs and non-drugs than simpler approaches. Figure 9 shows a plot of the enrichment factor obtained using ALOHA, Golden Triangle, Lipinski's Rule, CNS-MPO, and QED when asked to discriminate a set of 250 marketed neuroscience drugs from a background set of "leads" (compounds from the WOMBAT [22] database known to be micromolar or better inhibitors of some drug target, but MW <600 Da). From the figure it is clear that ALOHA significantly outperforms previously developed rules of thumb and drug suitability predictors at discriminating drugs from non-drugs.

Score variation versus chemical structure variation

As mentioned in the introduction of this paper, to be effective for the purposes of lead optimization, a scoring function used to select winners from non-winners should furnish a *statistically meaningful difference* between compounds in a typical lead optimization chemical series. Practically, this means that the scoring function must be able to differentiate between the best and worst compounds across a set of compounds in a single chemical series that a chemist would typically synthesize using a single reaction scheme and commercially available reactants. In order to explore the score variation and statistical significance of ALOHA with changes in chemical structure, we have used a test set of 71 compounds from three distinct and typical chemical series: 32 exemplified compounds from the patent for Viagra, 23 compounds from a patent on GSK-3β inhibitors, and 16 agonists of the Thyrotropin-releasing hormone receptor (molecule names and accessions supplied in the Supplementary material).

In the first study, the ALOHA score and pair-wise ECFP-6 tanimoto similarity was calculated for all 71 compounds. Figure 10 presents how the ALOHA (and QQCNS-MPO and QED) scores fluctuate as molecular similarity decreases. As the ECFP-6 similarity drops below 0.7, the variation in the ALOHA score is as significant as two compounds with no appreciable chemical similarity. Our previous work has demonstrated that an ECFP-6 similarity of 0.65 corresponds to highly similar compounds where the probability of having similar activity (within 1

Fig. 8 Molecular weight versus calculated logD (Pipeline Pilot logD) plot for 10,000 random compounds from Abbvie's compound collection, colored by the ALOHA probability score. Clearly, there is general sense of agreement between ALOHA and the Golden Triangle on the most fit compounds, though there are some notable exceptions where ALOHA scores a "non-Golden" compound very highly, and scores a "Golden" compound quite poorly (the *top right side* of the *triangle* in particular). The fittest ALOHA compounds typically have a 20-30 % probability of passing all 8 pre-clinical tests



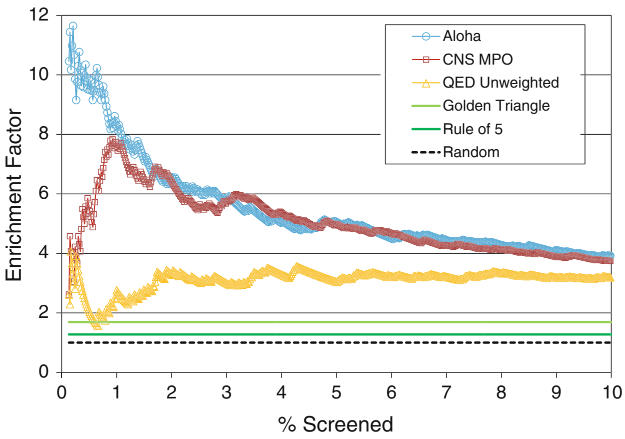


Fig. 9 An enrichment factor study comparing ALOHA to CNS-MPO, QED, Golden Triangle and Lipinski's Rule of Five at the task of separating marketed neuroscience drugs from a background set of lead compounds. Enrichment factor is a measure of how much better than random the scoring approach is performing, so when plotted versus the % of screened compounds, an enrichment factor of 8 at 1 % implies that 8 times as many drugs have been ranked in the *top* 1 % of the test set than would randomly be found. On this type of *plot*, random performance is represented by a straight line at an enrichment value of one. Since the Golden Triangle and Rule of 5 approaches provide a simple "yes" or "no" answer for each compound rather than a continuous score, they are also represented by a straight line (CNS MPO and QED produce the same score for many compounds, and the random number generator in Excel was used to provide a rank order to tie scores). The rule of thumb approaches performed less than 2 times better than random on this test set while the more complex and calculations of drug suitability were slightly better than 2 times better than random at the highest score levels (though at 1 %, CNS MPO regains a higher enrichment factor). ALOHA is nearly 12 times better than random at the highest score levels (the area under the ROC *curve* for ALOHA on this study is 0.784, *plot* not shown)

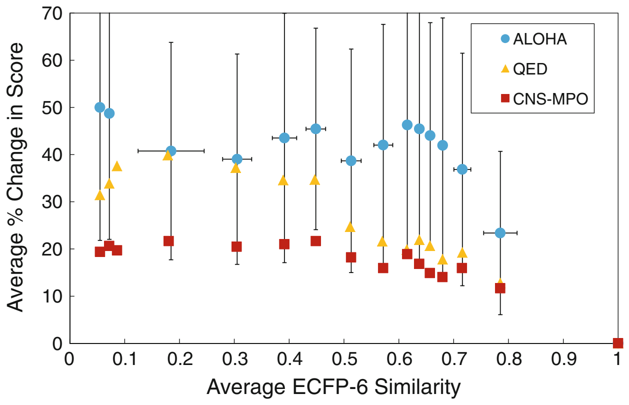


Fig. 10 A study of ALOHA's sensitivity to chemical structure change. As the ECFP-6 similarity drops below 0.7, the variation in the ALOHA score is as significant as two compounds with no appreciable chemical similarity. Hence, ALOHA is highly sensitive to small chemical structure changes within a single chemical series, allowing it be used productively during the lead optimization step in Drug Discovery where relatively small changes in chemical structure are considered. CNS-MPO also achieves its maximum dispersion in the lead optimization similarity regime; however, its maximum dispersion is significantly smaller than ALOHA. QED (unweighted) does not reach its maximum dispersion until below a tanimoto of 0.5, indicating it has a more limited ability to discriminate between members of the same chemical series. *Vertical* and *horizontal error bars* are provided for the ALOHA points

log unit) is approximately 40 % [15]. Hence, this study suggests that the ALOHA score has the potential for discriminatory power even in the regime of highly similar chemical space.

Since ALOHA provides not just a score but a score with and accurately calculated standard error, we can go further

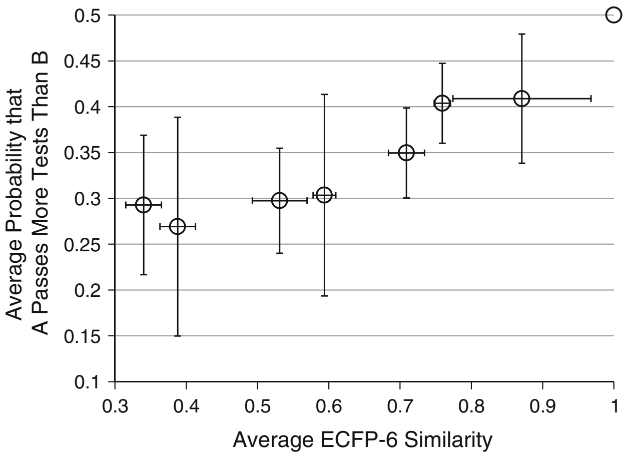


Fig. 11 A second study of ALOHA's sensitivity to chemical structure change, this time looking at the probability that compound A will pass more tests than compound B. As in Fig. 9, as the ECFP-6 similarity drops below 0.7, the variation in the ALOHA score is as statistically significant as two compounds with no appreciable chemical similarity. Hence, ALOHA scores are statistically significantly different, even for compounds within the same lead series

and evaluate the *statistical significance* of ALOHA score variation with chemical similarity (this cannot be done for methods which do not report the error). To assess this, we calculated the pair-wise ECFP-6 tanimoto similarity between the *highest ALOHA scoring* compound and constituent compounds within each class. Then, the ALOHA terms for the number of experiments passed (Eq. 1) and the error in the number of experiments passed (Eq. 2) were calculated for each molecule. Using these terms, we calculated the probability that the highest scoring molecule in each scaffold class would pass more tests than each of the other molecules in the scaffold class (by comparison of the two Gaussian distributions as was done for Fig. 5). Figure 11 shows how this probability changes as the molecular similarity from the top scoring compound decreases. As in study 1, we find that as the ECFP-6 similarity drops below 0.7, the statistical significance of the ALOHA score difference is as great as in two compounds with no appreciable chemical similarity. Furthermore, this plot conveys that in a typical lead optimization scenario where 20–30 compounds are under consideration, the top scoring compound is more than twice as likely (~ 70/30) to pass more tests than a randomly chosen compound from the set (this factor increases somewhat as a much larger set of candidates is considered). For the case of the exemplified compounds in the Viagra and GSK-3b patents (typical of compounds being considered during a lead optimization campaign), the non-averaged calculations show that the highest scoring ALOHA compound is 8 times more likely to pass more tests than the lowest scoring compound. Hence, ALOHA obtains a remarkable level of statistically significant dispersion even among compounds in the same lead series.

It is important that the reader does not conclude from these results that the ALOHA score is necessarily better than a simple rule of thumb or property categorization approaches for the purposes of HTS (high throughput screening) hit selection. The results in Fig. 10 indicate that the score landscape defined by ALOHA is much more rugged than the smooth "yes" or "no" score landscape for historic rule of thumb metrics as well as QED. If used to select from a chemically diverse set of singleton hits, the ALOHA score landscape may in fact be too rugged for optimal selection without assessing the chemical space within the vicinity of the hit more carefully. Future research will combine the ALOHA scoring approach with a chemical space exploration technique [23] to develop an approach that could assess the forward looking potential for regions of chemical space where there is little chemical precedent.

Visualizing the ALOHA output

ALOHA strives to rival the ease of use of a rule of thumb by providing a single numerical probability value representing the fitness of the compound. With that said, ALOHA also supplies information on each of the individual experiments that individuals engaged in the challenge of lead discovery will find useful. For example, ALOHA determines the experiments that are most likely to be problematic for each molecule, providing a logical road map for testing the most worrisome liabilities for a compound first. In this case, ALOHA can also supply the probability that a compound will pass the remaining $N - 1$ tests if it happens to pass the most problematic test (typically a much higher probability than the N test ALOHA score). However, as stated above, the error bars for any two individual prediction results will overlap unless there is a large separation in scores. In practice, it may be more useful to simply choose areas of chemical space that have very high ALOHA model scores for experiments that may typically be performed late in the pre-clinical testing process. That could reduce the likelihood of hitting a late-stage "dead end" after significant previous investment has already been made. Figure 12 uses TIBCO Spotfire to summarize of some of the useful ways to visualize the ALOHA output, using the three chemical series used above to study the score variation with chemical similarity.

Discussion

The most significant finding of this work is that the ALOHA method for combining the predictions for 8 pre-clinical tests (each with model strengths at or below 0.25) produces an overall fitness score that can discriminate

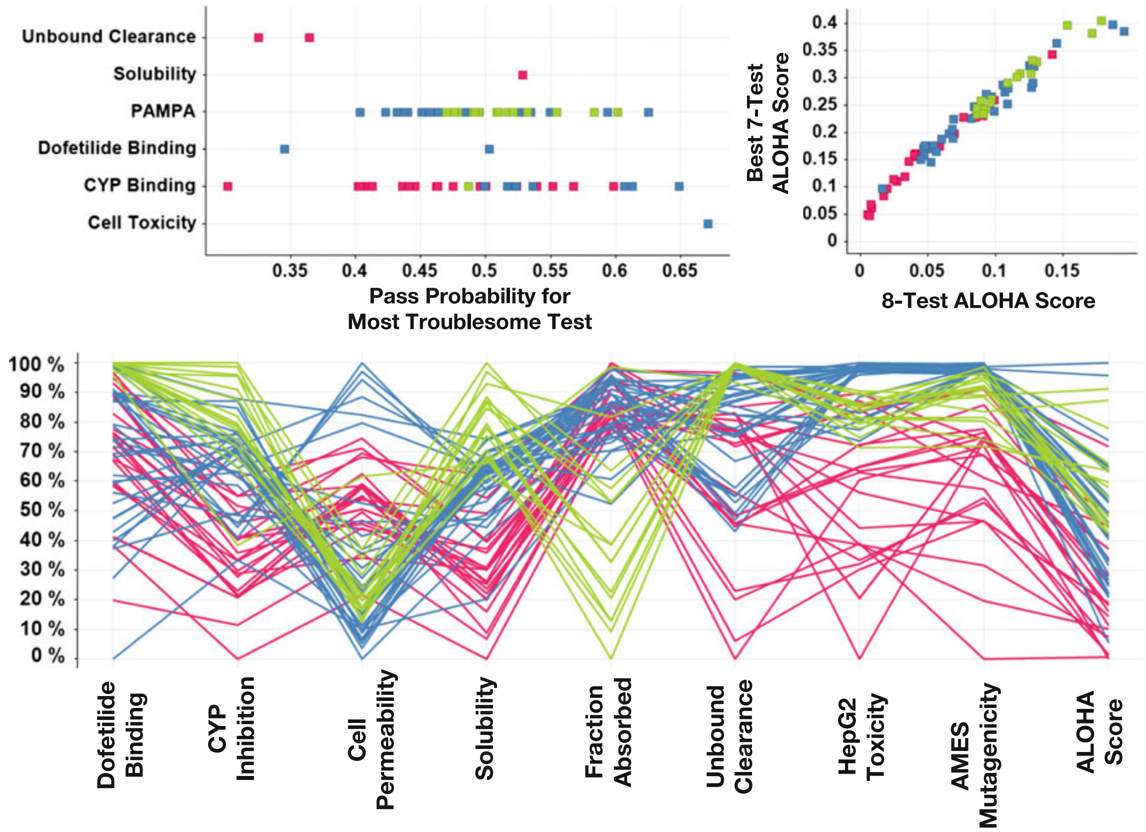


Fig. 12 A sample of useful visualization styles for visualizing ALOHA output for three distinct chemical series (GSK-3β series in magenta, Viagra series in *blue*, and the Thyrotropin-releasing hormone receptor series in *green*). In the *top left panel*, the test expected to be least likely for the compound to pass is shown along with the probability that the compounds will pass that test. The *top* compounds in these sets actually have more than a 60 % of passing each of the tests. In a lead optimization campaign, it is preferable to select compounds that are unlikely to fail tests that are costly or performed late in the Discovery process (Pharamacokinetic and toxicology tests). In the *top right panel*, the least likely test has been removed to calculate the probability that each compound will pass its 7-highest probability tests. The *top* compounds in the sets have an 8-test ALOHA probability around 0.2, but actually have a 40 % chance of passing 7 of the tests. In the *bottom plot*, a parallel coordinate plot is shown, where the y-axis shows where each compounds ranks compared to its peers for each of the tests and the overall ALOHA score. This view is useful for visualizing trends and exceptions within a chemical series

between typical compounds belonging to the same chemical series, yet at the same time perform well at the task of distinguishing drugs from leads. This interesting result is achieved via a series of interesting methodology steps that have not been previously demonstrated in the literature.

First and foremost, we have outlined the value of making predictions in probability space rather than explicit value space:

1. It is possible to fuse the probabilities together into a single overall score (Eq. 1).
2. It is possible to correctly account for the error in probability based predictions (Eq. 3) and propagate the error when fusing probabilities, resulting in a multi-parameter probability prediction with robust error bar (Eq. 2).
3. It is possible to report a model strength that estimates the discrimination power of the model compared to the original experiment. This is a very useful term that can be used to quantitatively assess whether more data is needed to build a model of desired discrimination power, or whether a training data cleaning exercise has resulted in a more predictive model.

Secondly, this work emphasizes ALOHA's philosophical bias towards models with a universal domain of applicability and presents a novel approach for assessing the domain of applicability of a prediction model. By employing a two step training process, where we first derive a raw model score and then convert the score to a success probability using a different training set, we have avoided the modeler's temptation to increase performance at a cost of domain applicability. Adding new parameters to our Bayesian models may improve the model strength, but for a multi-parameter assessment tool like ALOHA, we would advocate doing so only if the domain of applicability is not sacrificed. When training data is sparse or poor, a Bayesian model with low model strength will typically result, but with the two-step training method, it will still produce accurate probability values that do not mistakenly overpromise (the weakest possible model with zero model strength simply defaults to the background probability of passing the test for all molecules).

The ALOHA framework is straight forward to extend. While we have used the set of 8 experiments in Table 1 as the basis for this publication, we expect that in practice a variety of interesting parameters related to the fitness of a compound could be encoded into a probability. For example, we have previously published a method to determine the probability that a compound will be active for a target of interest [15]. Furthermore, we envision that methods utilizing the corpus of compounds from the patent literature could be used to code the "patentability" of a compound as a probability. At its most general, the ALOHA method provides a flexible framework for incorporating a wide variety of parameters into a single overall fitness score.

While we have used a set of Bayesian predictive models to demonstrate the ALOHA approach, the ALOHA framework can also be extended to include experimental data when it exists. When experimental data exists, the individual probability associated with the experiment becomes a 1 or 0 based on whether the result is considered a pass or fail (or even some value something in between—if error bars for the experiment are accurately known and straddle the pass-fail threshold). In this manner, the ALOHA approach offers the tantalizing prospect of calculating a fitness score for every compound in a corporate collection where experimental data is used, and modeling data is used to supplement gaps in the experimental record. In this manner, the ALOHA score could be thought of as a proxy for "Progressibility", where the progressibility score is a combination of experimental results and in silico predictions. Finally, to accommodate the variation in desired therapeutic properties for lead optimization campaigns in different disease indications, future work will focus on developing a scoring framework and user interface that allows the user to dynamically change the passing thresholds, experimental data utilization, and overall score contribution (weighting) for each individual parameter.

Due to the decade long legacy of merger activity in the pharmaceutical industry, we estimate that there may only be 15–20 drug discovery organizations worldwide that have a body of historical testing data large enough to train and validate ALOHA in a manner commensurate with that described in this paper. The ALOHA framework provides a methodology for these organizations to derive significant value from their testing legacy. And yet, there remain clinical pharmacokinetic and toxicology studies where no single organization in the world may have enough training data (at least 200 compounds for the Bayesian approaches outlined here). We expect that the results demonstrated in this manuscript could be improved significantly, both by increasing the number of models and the *model strength* of each model.

Given this potential for better results in the future, we hope that our work and the work of others on this Grand Challenge will help serve as a catalyst for pharmaceutical companies to collaborate in a so-called pre-competitive manner to share additional animal and human testing data to improve the industry's ability to discriminate safe and efficacious candidates from leads in silico. Naturally, it is difficult to inspire such collaboration without clear metrics for success, but this manuscript outlines the tools required to concisely measure the value gained by such collaboration:

1. The *model strength* term provides a useful metric for understanding whether additional experimental data has resulted in a more powerful and discriminating model.
2. The domain of applicability (DOA) tests outlined and presented in Figs. 3 and 4 provide a useful approach for understanding when model parameters limiting the applicability of the model to chemical space near to the training set.
3. The enrichment factor plotting technique (Fig. 9) provides a useful metric for understanding whether adding new experiments to the testing panel results in an increased ability to discriminate drug candidates from leads.

Given the measured but (hopefully) inevitable progress of in silico methods to provide increasingly accurate assessments of drug-likeness, our industry should carefully consider whether greater gains in global pharmaceutical small molecule drug research productivity could be achieved by improved model building based on broader collaboration and data sharing. Models could be built and quality assessed by an independent party to preserve data confidentiality (furthermore, the chemical structures could be remain confidential by reducing to an agreed standard set of parameters), and such an effort could potentially be funded by grants and commercial sales to smaller drug discovery organizations and academic institutions that lack a significant historical pre-clinical testing legacy, but would seek access once the value of the model was clearly established and communicated.

Acknowledgments The financial support for this work was provided by AbbVie

References

1. Keserü GM, Makara GM (2009) The influence of lead discovery strategies on the properties of drug candidates. Nat Rev Drug Discov 8:203–212

2. Van de Waterbeemd H, Gifford E (2003) ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2:192–204
3. Ekins S, Boulanger B, Swaan PW, Hupcey MA (2002) Towards a new age of virtual ADME/TOX and multidimensional drug discovery. J Comput Aided Mol Des 16:381–401
4. Wager TT, Hou X, Verhoest PR, Villalobos A (2010) Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. ACS Chem Neurosci 1:435–449
5. Segall MD (2012) Multi-parameter optimization: identifying high quality compounds with a balance of properties. Curr Pharm Des 18:1292–1310
6. Ekins S, Boulanger B, Swaan PW, Hupcey MA (2002) Towards a new age of virtual ADME/TOX and multidimensional drug discovery. J Comput Aided Mol Des 16:381–401
7. Lipinski CA (2009) Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods 44:235–249
8. Hughes JD, Blagg J, Price DA, Bailey S, DeCrescenzo GA, Devraj RV, Ellsworth E, Fobian YM, Gibbs ME, Gilles RW, Greene N, Huang E, Krieger-Burke T, Loesel J, Wager T, Whiteley L, Zhang Y (2008) Physiochemical drug properties associated with in vivo toxicological outcomes. Bioorg Med Chem Lett 18:4872–4875
9. Johnson TW, Dress KR, Edwards M (2009) Using the Golden Triangle to optimize clearance and oral absorption. Bioorg Med Chem Lett 19:5560–5564
10. Marchant CA, Briggs KA, Long A (2008) In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for windows, meteor, and vitic. Toxicol Mech Methods 18:177–187
11. Fraczkiewicz R, Zhuang D, Zhang J, Miller D, Woltosz WS, Bolger MB (2009) Busting the black box myth: designing out unwanted ADMET properties with machine learning approaches. CICSJ Bull 27:96–102
12. Wagner TT, Hou X, Verhoest PR, Villalobos A (2010) Moving beyond rules: the development of a central nervous system multiparamter optimization (CNS MPO) approach to enable alignment of druglike properties. ACS Chem Neurosci 1:435–449
13. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98
14. Hofer TS (2013) From molecules to electrons—grand challenges in theoretical and computational chemistry. Front Chem 1:1–4
15. Muchmore SW, Debe DA, Metz JT, Brown SP, Martin YC, Hajduk PJ (2008) Application of belief theory to similarity data fusion for use in analog searching and lead hopping. J Chem Inf Model 48:941–948
16. Chadwick A, Segall M (2010) Overcoming psychological barriers to good discovery decisions. Drug Discov Today 15:561–569
17. Stouch TR (2012) 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 26:125–134
18. Klon AE, Lowrie JF, Diller DJ (2006) Improved naïve bayesian modeling of numerical data for absorpotion, distribution, metabolism and excreation (ADME) property prediction. J Chem Inf Model 46:1945–1956
19. Nicholls A (2008) What do we know and when do we know it? J Comput Aided Mol Des 22:239–255
20. Varma MV, Obach RS, Rotter C, Miller HR, Chang G, Steyn SJ, El-Kattan A, Troutman MD (2010) Physicochemical space for optimum oral bioavailability: contribution of human intestinal absorption and first-pass elimination. J Med Chem 53:1098–1108
21. Obach RS, Lombardo F, Waters NJ (2008) Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. Drug Metab Dispos 36:1385–1405
22. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI (2004) In: Oprea TI (ed) Chemoinformatics in drug discovery, Wiley-VCH, New York
23. Stewart KD, Shiroda M, James CA (2006) Drug Guru: a computer software program for drug design using medicinal chemistry rules. Bioorg Med Chem 14:7011–7022