JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

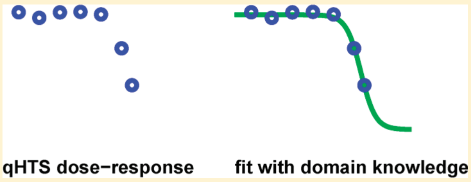
Exploiting Domain Knowledge for Improved Quantitative High-Throughput Screening Curve Fitting

Charles Bergeron,*[†] Gregory Moore,[†] Michael Krein,[‡] Curt M. Breneman,[‡] and Kristin P. Bennett[†]

[†]Department of Mathematical Sciences and [‡]Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, 110 8th Street, Troy, New York 12180, United States

Supporting Information

ABSTRACT: Least-squares fitting of the Hill equation to quantitative high-throughput screening (qHTS) assays results in frequent unsatisfactory fits. We learn and exploit prior knowledge to improve the Hill fitting in a nonlinear regression method called domain knowledge fitter (DK-fitter). This paper formulates and solves DK-fitter for 44 public qHTS data sets. This new Hill parameter estimation technique is validated using three unbiased approaches, including a novel method that involves generating simulated samples. This paper fosters the extraction of higher quality information from screens for improved potency evaluation.



1. INTRODUCTION

Screening hits are compounds that are effective at modulating a disease-specific biological pathway. They are identified by assaying a library of compounds for a chosen target using high-throughput screening (HTS) or, more recently, quantitative HTS (qHTS). Good structure−activity relationship (SAR) modeling relies on accurate screening data and proper inference from this data. Currently, the Hill equation (eq 1) is fit to each dose−response sample consisting of several activity−concentration data points by penalizing the least-squares of the residuals.[1] This algorithm, called LS-fitter in this paper, poorly estimates the Hill parameters when the highest-concentration activity measurement is less than 95% of the terminal activity.[2] The biases and variabilities associated with LS-fitter make it difficult to extract quality information from raw qHTS data and properly assess compound potency. Clearly, a reliable curve-fitting technique that can handle dose−response samples that provide only partial coverage of the activity range (AR) is needed.

This paper assembles several concepts drawn from the mathematical sciences (such as geometry, calculus, optimization, and statistics) to improve Hill equation fitting. Chiefly, we inject prior knowledge (or domain knowledge) into the curve-fitting procedure. For instance, many inhibitors have terminal activity around −100%. This known information can guide the estimation of Hill parameters, yet is not considered by LS-fitter. We exploit domain knowledge such as this to produce better-quality fits and especially on those dose−response samples, whose data points do not cover the full AR. This new technique is called domain knowledge fitter or DK-fitter. Specifically, we:

- Create a body of decisive rules for categorizing samples in terms of observed chemical behavior and data quality.
- Formulate the incorporation of domain knowledge to the Hill fitting problem.
- Propose a methodology for solving this formulation.

Improved curve fits obtained from DK-fitter are validated in three ways:

- By contrasting the fits in replicate qHTS screens.
- Through a novel validation technique involving the generation of simulated samples.
- By comparing quantitative SAR (QSAR) models generated from the output of LS- and DK-fitter.

This paper is organized as follows: Section 2 provides appropriate background, notation, and definitions. Section 3 formulates LS-fitter, identifies limitations with this technique that motivate this work, and provides an overview of DK-fitter. Sections 5−7 demonstrate that DK-fitter as the better algorithm for Hill equation curve fitting using three validation strategies. The computational complexity of both algorithms is discussed in Section 8. These are followed by a discussion and some conclusions in Section 9. Detailed methods for this paper appear at the end in Section 10. Implementation details and additional detailed results appear in Supporting Information to this paper.

2. BACKGROUND

Traditionally, a large number of compounds are assayed for activity at a single concentration: This is called HTS, a mainstay of pharmaceutical development since the development of robotic assay technologies.[3] Hit identification proceeds by selecting the compounds having activities beyond a threshold. Experimental error may cause frequent false positives and negatives.[4]

Quantitative HTS (qHTS) obtains more complete dose−response information by assaying compounds at multiple concentrations in a single experiment.[5,6] Screening large libraries of molecules by qHTS provides comprehensive information on the potency and efficacy of compounds, thereby improving the

Received: May 11, 2011
Published: October 14, 2011

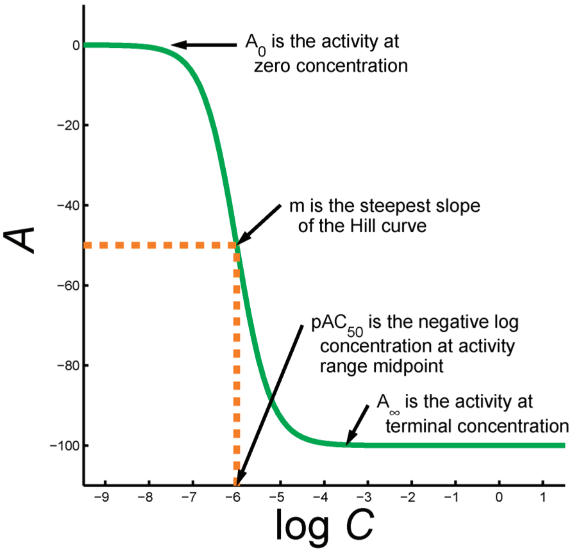


Figure 1. Interpretation for the Hill parameters is shown on this Hill curve. The naught activity $A_0 = 0$ is the activity in the presence of no compound. The terminal activity $A_\infty = -100$ is the activity attained when a sufficiently high concentration of the compound is provided. The half-maximal activity concentration AC_{50} is the concentration at which the midpoint of the activity range attained; it is usually reported using the negative logarithmic scale, so that increasing pAC_{50} corresponds to increased potency. In this example, $pAC_{50} = 6$. The Hill slope m is proportional to the Hill equation absolute steepest slope, which occurs at AC_{50}.

identification of screening hits for further analysis. To this end, qHTS has been exploited as a primary screening methodology seeking potent compounds.[7−12] Zheng et al.[13] go a step further and identify three clusters of glucocerebrosidase inhibitors from the Hill fits.

We define our notation. Denote concentration as C and activity as A. By convention, inhibitors exhibit negative activity, while agonists exhibit positive activity. Activities are in fact relative activities that are expressed as percentages, with −100% and 100% defined by controls. Concentrations are often reported on the negative logarithmic scale, denoted $pC = -\log C$, so that large values denote smaller concentrations. All logarithms in this paper are base 10.

A dose−response sample (or sample) consists of n data points $\{(C_i,A_i)\}_{i=1}^{n}$, ordered by increasing concentration. Samples having fewer than 4 data points are not considered, so $n \geq 4$. Data point (C_1,A_1) is called the starting concentration and activity, while data point (C_n,A_n) is called the ultimate concentration and activity. Interval $[C_1,C_n]$ is called the experimental concentration range (ECR). The assumption of this paper is that an appropriate ECR is chosen for each assay so as to capture relevant compound activity behavior required of screening hit identification.

The theoretical dose−response relationship is given by the Hill equation:[14]

$$A = \mathrm{Hill}(\mathrm{pC}) = A_0 + \frac{A_\infty - A_0}{1 + 10^{\mathrm{m}(\mathrm{pC} - \mathrm{pAC}_{50})}} \qquad (1)$$

All four parameters $\{\mathrm{pAC}_{50},\mathrm{m},A_0,A_\infty\}$ are interpretable (see Figure 1). The activity range (AR) of a sample is $[A_0,A_\infty]$.

To simplify analysis, we rewrite the Hill equation, introducing parameter $s = \mathrm{sign}(A_\infty)$ and making parameter change $B_\infty = sA_\infty = |A_\infty|$. Thereupon, an agonist has $s = 1$, and an inhibitor has $s = -1$. We consider transformed data points $(\log C_i,B_i)$ with

$$B_i = sA_i \qquad (2)$$

Additionally, we find it convenient to work with the logarithms of m and A_∞, and we set $A_0 = 0$ a priori, making eq 1 an increasing function that greatly simplifies the analysis of Sections 4 and 10. This Hill equation form is used in the remainder of this paper:

$$B = \mathrm{Hill}(\mathrm{pC},\mathrm{P}) = \frac{10^{\log B_\infty}}{1 + 10^{10^{\log m}(\mathrm{pC} - \mathrm{pAC}_{50})}} \qquad (3)$$

In figures, axes are clearly labeled as displaying activities A or transformed activities B. Parameters $\mathrm{P} = \{\mathrm{pAC}_{50},\log \mathrm{m},\log B_\infty\}$ are collectively known in this paper as Hill parameters.

We summarily define the following terms—they are formally defined in Section 10.2:

- Active samples present significant inhibitory or agonistic behavior. A reliable active is an active sample that contains complete dose−response information, covering its full AR. An uncertain active is an active sample that contains partial dose−response information, having a dose−response relationship which is only partly observed over the ECR. A defective active is an active sample that demonstrates some activity within the ECR but does not present the dose−response trend suggested by the Hill equation.
- Inactive samples present negligible activity across the ECR.
- Samples that are neither active nor inactive are called disrupted samples. Some samples do not fit the Hill equation: samples exhibiting nonzero constant activity across the ECR or erratic activity jumps from one concentration to the next. Experimental artifacts (aggregation, degradation, or assay disruption) and special compound behavior (quencher compounds) may be responsible for these apparently abnormal observations.

Our empirical results are based on publicly available qHTS screens. We downloaded 44 qHTS screens from PubChem (http://pubchem.ncbi.nlm.nih.gov/) between December 29, 2008 and May 20, 2010. Each is referenced by its assay identification number (AID), and the experimental protocols are found on the PubChem Web site. As a case study, the acquisition of screen AID 361 is detailed extensively.[1] Compounds are referenced by compound identification number (CID). These screens cover a wide range of interesting applications and validate our methods:

- Frequent cancer therapy target pyruvate kinase (AID 361).
- Penicillin-resistant β-lactamase enzymes (AID 584 and 585).
- Xenobiotic metabolism in the human liver by cytochrome 3A4 (AID 884).
- Better understanding of Alzheimer's disease through τ proteins (AID 1463 and 1468).
- Better understanding of Pompe's disease through α-glucosidase (AID 2100).

3. LEAST-SQUARES CURVE FITTING

Eq 3 is fit to samples having dose−response data points $(\log C_i,B_i)$ to determine Hill parameters. The state of the art for this process involves penalizing the squared residuals. This criterion

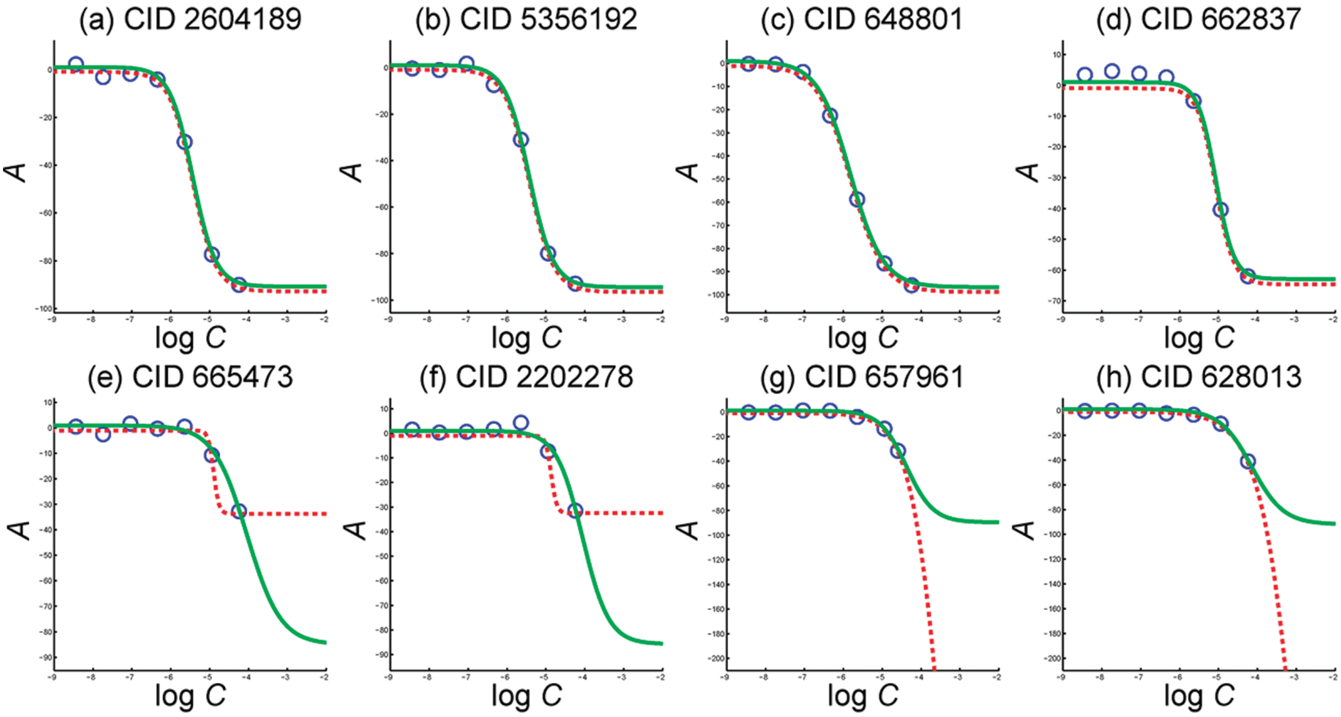


Figure 2. Data points (blue circles) are fit using LS-fitter (dashed red curve) and DK-fitter (solid green curve). Panels (a,d) show well-fit reliable actives for which LS- and DK-fitter curves are difficult to distinguish. Panels (e−h) show fits for uncertain active samples. LS-fitter curves present atypical terminal activities B_∞ and Hill slopes m. Each presents steep Hill slope m and/or atypical B_∞ values. Unsatisfactory LS-fitter performance is often observed when the true terminal activity falls outside of the ECR or when experimental error in the activity measurements suggests very gentle or steep slopes. Compounds are drawn from AID 361 and referenced by CID.

has been used since at least the early 1970s.[15] Nonlinear least-squares regression (LS-fitter) of the Hill equation to data is treated as the state-of-the-art in this paper. The following optimization problem formulates this:

$$\min_{\mathrm{P}} \quad \mathrm{LS(P)} \tag{4}$$

for P = {pAC$_{50}$,log m,log B$_\infty$} with the least-squares criterion given by the sum of squared residuals:

$$\mathrm{LS(P)} = \frac{1}{2}\sum_{i=1}^{n}(B_i - \mathrm{Hill}(\mathrm{p}C_i, \mathrm{P}))^2 \tag{5}$$

This optimization problem is solved to local optimality for each sample. Implementation details appear in Supporting Information.

Visual inspection of reliable samples and their Hill equations obtained from solving eq 5 suggest that Hill parameters for these samples are accurately determined. Results in Sections 5 and 6 support this claim. The top row of Figure 2 plots four such cases drawn from AID 361; the green solid curves returned by DK-fitter are superimposed over the red dotted curves found by LS-fitter.

Uncertain active samples have data points that partially cover their AR. The least-squares criterion is not appropriate to fit these samples as a minimum-residual fit may appear grossly incorrect to a domain expert (e.g., a chemist). Examples of poor LS-fitter curves (drawn as red dotted curves) on uncertain samples appear in the bottom row of Figure 2. In panels e,f, the least-squares criterion proposes terminal activities around −30. In panels g,h, terminal activities are below −200. With the knowledge that inhibitors present A$_\infty$ ≤ − 100, the red dotted curves in panels e−h are probably incorrect, and the green solid curves obtained by DK-fitter are probably better. It is plots such as these that motivate the need for improved qHTS

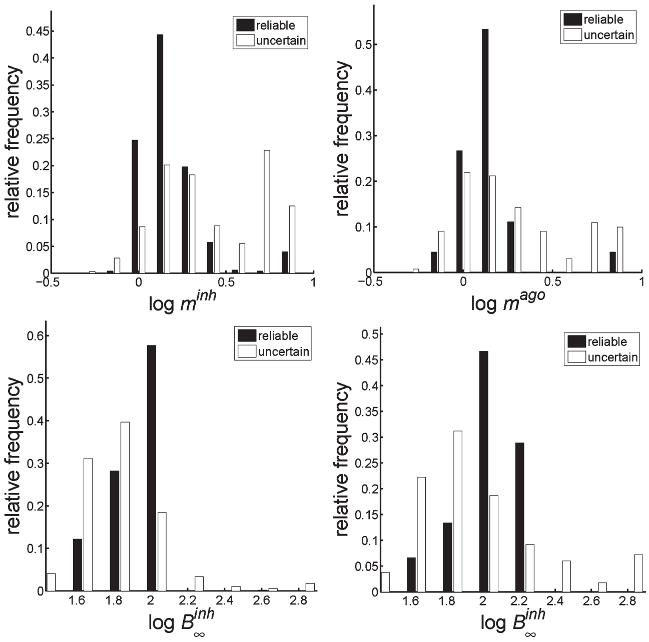


Figure 3. These histograms contrast the distribution of Hill equation parameters for reliable and uncertain (black and white bars, respectively) actives from qHTS screen AID 361.

dose−response modeling. Clearly, least-squares fits possessing zero residuals may returning grossly incorrect dose−response relations. Hence, strategies to filter out poor fits using goodness-of-fit statistics, such as Pearson correlation, cannot be successful. In the remainder of this section, we seek to understand why uncertain actives are not properly fit and what modifications are required to obtain satisfactory fits.

To this end, we investigate the distribution of Hill equation parameters m, B_∞ in reliable and uncertain actives determined from LS-fitter. Figure 3 shows these distributions for qHTS screen AID 361. Distributions for parameters associated with inhibitors and agonists are studied separately, hence the notation m^{inh}, m^{ago}, B_∞^{inh}, and B_∞^{ago}. In this figure, observed Hill parameter distributions as obtained from LS-fitter for reliable samples do not match those for uncertain samples. For example, the top panel finds steeper Hill slopes in uncertain samples and an distribution for the Hill slope in uncertain samples that is strongly bimodal. These histograms suggest a method for obtaining better quality dose−response fits: Forcing Hill equation parameters m, B_∞ for all samples to fall within reasonable ranges, as determined from LS-fitter results on reliable actives. In doing so, domain knowledge (or prior knowledge) is injected into the curve-fitting process. The next section presents a methodology for accomplishing this.

4. DOMAIN KNOWLEDGE CURVE FITTING

Denote the set of Hill parameters for an inhibitor by $P^{inh} = \{pAC_{50}, \log m^{inh}, \log B_\infty^{inh}\}$, and for an agonist, $P^{ago} = \{pAC_{50}, \log m^{ago}, \log B_\infty^{ago}\}$. The domain knowledge objective function aims to perform three tasks simultaneously:

(1) Minimize the residuals, assessed using the least-squares criterion as is done with the least-squares formulation, eq 4.

(2) Select a Hill slope that is close to the average Hill slope determined from reliable samples for that screen.

(3) Select a terminal activity that is close to the mean terminal activity determined from reliable samples for that screen.

These tasks are such that DK-fitter will deliberately trade-off the smallest possible residuals in fitting the Hill equation so as to return reasonable Hill parameters. Supporting Information deals with the situation where there is an insufficient number of reliable active samples to estimate the distributions of the Hill parameters. The resulting optimization problem is

$$\min_{P} \quad DK(P) \tag{6}$$

with

$$\begin{aligned} DK(P^{inh}) \quad &= \quad LS(P^{inh}) + \lambda_1^{inh}\frac{1}{2}\,\text{distance}(\log m^{inh})^2 \\ &+ \quad \lambda_2^{inh}\frac{1}{2}\,\text{distance}(\log B_\infty^{inh})^2 \end{aligned} \tag{7}$$

or

$$\begin{aligned} DK(P^{ago}) \quad &= \quad LS(P^{ago}) + \lambda_1^{ago}\frac{1}{2}\,\text{distance}(\log m^{ago})^2 \\ &+ \quad \lambda_2^{ago}\frac{1}{2}\,\text{distance}(\log B_\infty^{ago})^2 \end{aligned} \tag{8}$$

according to whether the sample is an inhibitor or agonist. This optimization problem is solved for each qHTS sample. Implementation details appear in Supporting Information.

The regularization term, for an arbitrary parameter Q, penalizes squared distance from the distribution mean, scaled by the variance:

$$\text{distance}(Q) = \frac{Q - \mu_Q}{\sigma_Q} \tag{9}$$

Regularization promotes the selection of realistic Hill parameter values.

The tradeoff between minimizing the curve-fitting residuals and regularizing the Hill parameters is moderated by regularization parameters $\boldsymbol{\lambda}^{inh}$ or $\boldsymbol{\lambda}^{ago}$. These hyperparameters are analogous to the tradeoff hyperparameters (often denoted C) in support vector machines. For each screen, optimal values for $\{\lambda_1^{inh}, \lambda_2^{inh}, \lambda_1^{ago}, \lambda_2^{ago}\}$ are chosen based on performance on the simulated set (see Section 6). To this end, the simulated set is evenly split into training and test sets. If there are more than 1024 simulated samples, 512 are drawn for each of training and test.

5. VALIDATION VIA REPLICATE SCREENS

The superiority of the DK-fitter approach over LS-fitter as a method for qHTS curve-fitting is demonstrated in three ways in this paper:

(1) In this section, fits to replicate qHTS screens are compared.

(2) In Section 6, results from a nonbiased validation approach are presented on 44 qHTS screens.

(3) In Section 7, the performances of QSAR models trained to predict the pAC_{50}'s obtained from both algorithms are examined.

Each of these approaches demonstrate the superiority of DK-fitter over LS-fitter.

In this section, we exploit replicate screens AID 1463 and 1468 to study the robustness of LS- and DK-fitter. For a given compound, we expect that Hill equation parameters from replicate assays will be similar. Ideally, a curve-fitting technique would robustly handle measurement errors and experimental variabilities across replicate assays.

For this analysis, we retain all samples that are classified as reliable or uncertain active in both screens. There are 163 actives that are reliable in both screens (appearing as green crosses in Figure 4), 90 actives that are reliable in AID 1463 and uncertain in AID 1468 (orange circles), 3 actives that are uncertain in 1463 but reliable in 1468 (purple diamonds), and 119 actives that are uncertain in both screens (red squares). Looking at Figure 4, we note that curve fitting of reliable actives is a very accurate process using either LS- or DK-fitter. This sustains our earlier assumption from Section 3. However, a compound that is categorized as uncertain in one or both screens will find pAC_{50}'s that are much more correlated when obtained from DK-fitter. This occurs because LS-fitter can assign unsuitable Hill equation parameters, especially when a sample's data points do not cover the full AR.

The average absolute difference $|\Delta pAC_{50}| = |pAC_{50}{}^{1463} - pAC_{50}{}^{1468}|$ across the 375 considered samples is 0.751 and 0.369 for LS- and DK-fitters, respectively. The DK-fitter error is half that of the LS-fitter, suggesting that DK-fitter calculates a more robust pAC_{50} when faced with data that includes some variability. This result demonstrates the robustness of DK-fitter in

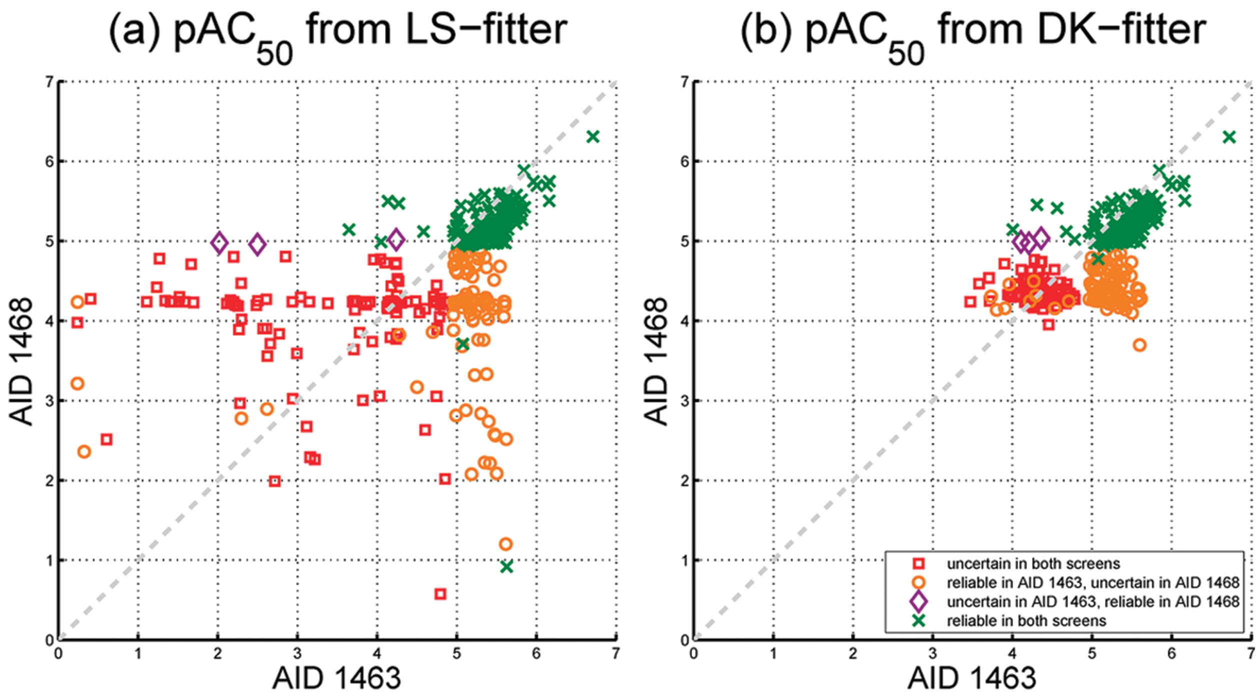


Figure 4. Figure exploits replicate screens (AID 1463 and 1468) to study the stability of both algorithms. Panel (a) is a scatter plot of LS-fitter-calculated pAC_{50} values for both screens; the pAC_{50}'s only agree when sample data are classified as reliable active. Panel (b) shows the scatter plot obtained from using the DK-fitter method: The pAC_{50}'s calculated from both screens are in much better agreement. Hence, DK-fitter is more robust and calculates more accurate Hill equation parameters.

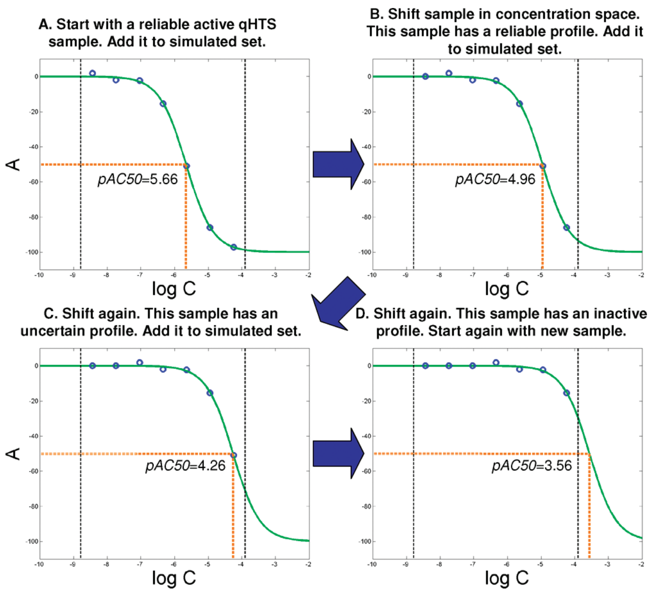


Figure 5. Repeatedly shifting a reliable sample in concentration space generates simulated samples. The Hill equation can the be fit to the data points (blue circles) of the middle panel and compared to the true fit (solid green curve) so as to validate the methods of this paper on uncertain samples.

obtaining Hill equation fits for uncertain samples and increased accuracy in calculating pAC_{50} values from qHTS samples.

We are not aware of other suitable public replicate qHTS screens to repeat this analysis. However, based on the results of this section, we believe that using DK-fitter may reduce the need for confirmatory screening in an industrial drug discovery setting.

6. VALIDATION VIA SIMULATED SAMPLES

Comparing curve-fitting techniques can be a difficult task in the absence of a priori knowledge of the true fit. Replicate screens for an analysis of the previous section are not always available. This section presents a novel approach for objectively validating DK-fitter.

As defined in this paper, reliable actives are such that any nonlinear regression technique should be satisfactory. Uncertain actives may be better fit by one method than another, but in the absence of hard knowledge of the exact Hill equation parameters for each compound, selecting the better method can be challenging. We overcome this difficulty by introducing a nonbiased validation approach to assess curve-fitting techniques objectively. This procedure requires the generation of simulated actives.

The hypotheses underlying this validation approach are:

- Hill parameters for reliable actives are those returned by LS-fitter (eq 4).
- For a given target, the distribution of Hill parameters $\{A_\infty,m\}$ is independent of pAC_{50}, e.g., the steepness and range of the Hill equation does not vary with pAC_{50}.

Under these assumptions, for a given screen, we apply LS-fitter to learn the distributions of A_∞ and m and then shift these curves in concentration space to generate new samples. These simulated actives may possess a uncertain profile when looking at their data points, but their pAC_{50} remains known. This trick of shifting reliable actives in concentration allows for a fair evaluation of LS- and DK-fitter on all active samples, especially the more difficult uncertain actives. The schematic of Figure 5 gives an overview of this procedure. A complete algorithm is provided in Supporting Information.

In passing, we mention that the hyperparameters of the DK-fitter optimization problem are determined for a screen by

Table 1. Comparison of LS- and DK-fitter AOC (eq 10) Statistics[a]

	inhibitors reliable		LS-fitter	DK-fitter	agonists reliable		LS-fitter	DK-fitter
AID	actives	testing simulated	AOC	AOC	actives	testing simulated	AOC	AOC
357	70	130	0.257	0.245	76	152	0.303	**0.166**
360	155	266	0.381	**0.070**	9	10	0.642	**0.301**
361	501	368	0.138	**0.055**	45	92	0.247	**0.113**
411	345	348	0.153	**0.075**	0	0		
444	358	359	0.348	**0.242**	63	158	**0.296**	0.376
446	56	103	0.150	0.150	59	145	0.330	**0.154**
448	54	96	0.148	**0.091**	146	252	0.415	**0.249**
450	0	0			51	149	0.541	**0.274**
530	16	31	0.151	0.162	36	38	0.817	0.817
584	12	50	0.118	**0.071**	1	0		
585	41	97	0.134	**0.064**	1194	301	0.272	**0.145**
603	38	108	0.254	**0.056**	1	4	0.676	**0.477**
875	10	19	0.102	0.117	12	21	0.203	**0.120**
881	283	273	0.207	**0.094**	402	244	0.179	**0.062**
883	512	329	0.303	**0.125**	29	57	0.177	**0.117**
884	1313	410	0.578	**0.113**	84	157	0.453	**0.316**
886	928	391	0.219	**0.101**	78	116	0.393	**0.267**
887	420	292	0.200	0.177	84	131	0.332	**0.138**
892	13	23	0.153	**0.054**	1	1	**0.000**	0.373
893	538	373	0.396	**0.331**	28	58	0.577	**0.503**
902	693	281	0.636	**0.405**	60	77	1.219	1.219
903	151	164	0.373	**0.186**	19	16	**0.077**	0.286
904	383	222	0.784	0.784	73	59	0.950	0.950
912	460	261	0.555	**0.226**	183	280	0.314	0.305
914	356	328	0.353	**0.197**	79	194	0.367	**0.341**
915	356	328	0.416	**0.204**	79	193	0.334	0.342
923	0	0			222	312	0.309	**0.157**
924	436	315	0.347	**0.111**	31	40	0.385	**0.185**
938	26	36	1.024	**0.746**	768	275	0.505	**0.232**
1379	89	222	0.307	**0.072**	0	0		
1452	203	236	0.238	**0.118**	0	0		
1454	0	0			0	0		
1458	1325	363	0.405	**0.357**	1263	333	0.703	**0.658**
1460	2852	413	0.463	**0.190**	50	59	0.256	**0.191**
1461	830	209	**0.166**	0.251	227	218	0.277	**0.116**
1463	2288	358	0.457	**0.144**	21	27	0.234	**0.102**
1466	46	28	0.221	**0.091**	40	40	0.214	**0.136**
1467	38	24	0.018	0.035	19	18	**0.158**	0.195
1468	422	289	0.314	**0.123**	6	4	0.307	0.290
1634	48	47	0.422	**0.188**	471	311	0.300	**0.132**
1688	862	255	0.255	**0.138**	258	272	0.381	**0.114**
1721	271	254	0.295	**0.096**	5	3	**0.000**	0.312
1868	13	15	0.105	**0.057**	0	0		
2100	1108	395	0.376	**0.108**	2	1	0.000	0.000

[a] Smaller AOC values are preferred and are bolded. Overwhelmingly, DK-fitter is the better method.

generating the set of simulated samples and dividing this set equally between training and test subsets. In this paper, results on simulated samples are always reported for the test subset. The number of such samples is stated in Table 1.

In the next subsection, we present the summary results across 44 screens. In the following subsection, we perform a more detailed analysis of the results for a single screen.

6.1. Quantitative Results.

Regression error characteristic (REC) curves[16] are an effective way of assessing regression quality. Figure 7 presents example REC curves that represent the ratio of samples (vertical axis) that are modeled within an error tolerance ε (horizontal axis). A desirable model is one with high accuracies across all ε. One popular measure of curve-fitting (or regression) performance obtained from the REC curve is its area over the curve (AOC):

$$\mathrm{AOC} = \int_0^\infty (1 - \mathrm{REC}(\varepsilon))\mathrm{d}\varepsilon \qquad (10)$$

The AOC is the area between a curve and the horizontal line at unit accuracy; a smaller AOC is desirable. The AOC is a reasonable estimate of the mean absolute error (MAE).[16]

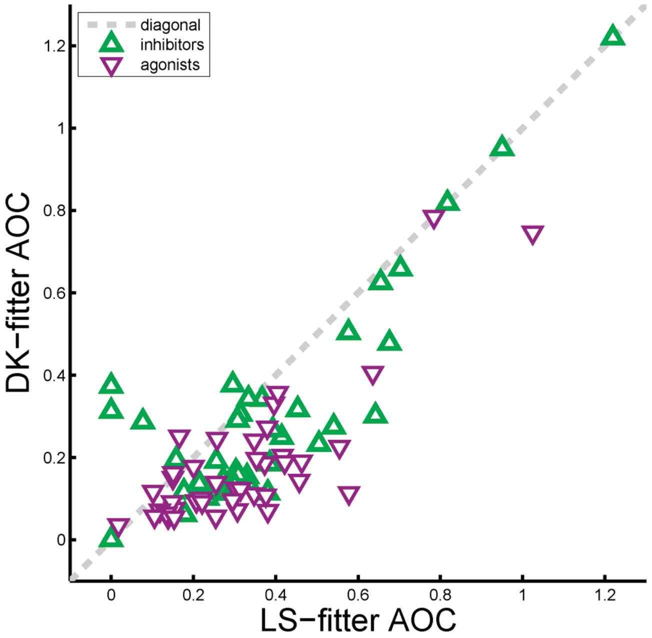


Figure 6. This scatter plot compares the AOC (eq 10) for the LS- and DK-fitter techniques for all 43 qHTS screens considered in this study. This plot shows that DK-fitter dominates, and its AOC is consistently smaller than LS-fitter.

Figure 6 contains a scatter plot for LS-fitter AOC against DK-fitter AOC. This figure shows that DK-fitter tends to be the better method, returning smaller AOC than LS-fitter.

Table 1 presents LS- and DK-fitter AOC values plotted in Figure 6. Bolding in the table indicates the better model. When the difference in the LS- and DK-fitter AOC is less than 0.025, we consider the results to be tied, and no bolding is applied.

Looking at Table 1, for each screen, there are four possible cases:

(1) DK-fitter displays smaller AOC than LS-fitter.
(2) DK-fitter is deemed to be tied with LS-fitter.
(3) DK-fitter displays larger AOC than LS-fitter.
(4) There are no reliable actives. For inhibitor curve fitting, DK-fitter is the better technique for 33 screens, both techniques perform at par for 7 screens, LS-fitter is the better technique for 1 screen (AID 1461), and 3 screens possess no reliable inhibitors. For AID 1461, overfitting in the selection of the regularization parameters appears to have occurred with DK-fitting. Three screens possess no reliable inhibitors.

For the agonists, DK-fitter dominates on 26 screens, with another 9 screens being tied between LS- and DK-fitter, and LS-fitter dominates on 5 screens (AID 444, 892, 903, 1467, and 1721, all having very small numbers of reliable agonist samples). Six screens possess no reliable agonists.

Based on these results, we conclude the superiority of DK-fitter as a dose−response modeling approach for qHTS samples.

6.2. Qualitative Results on AID 361. In this subsection, we present detailed results in graphical form for qHTS screen AID 361 in Figure 7. This screen assayed the effect of compounds

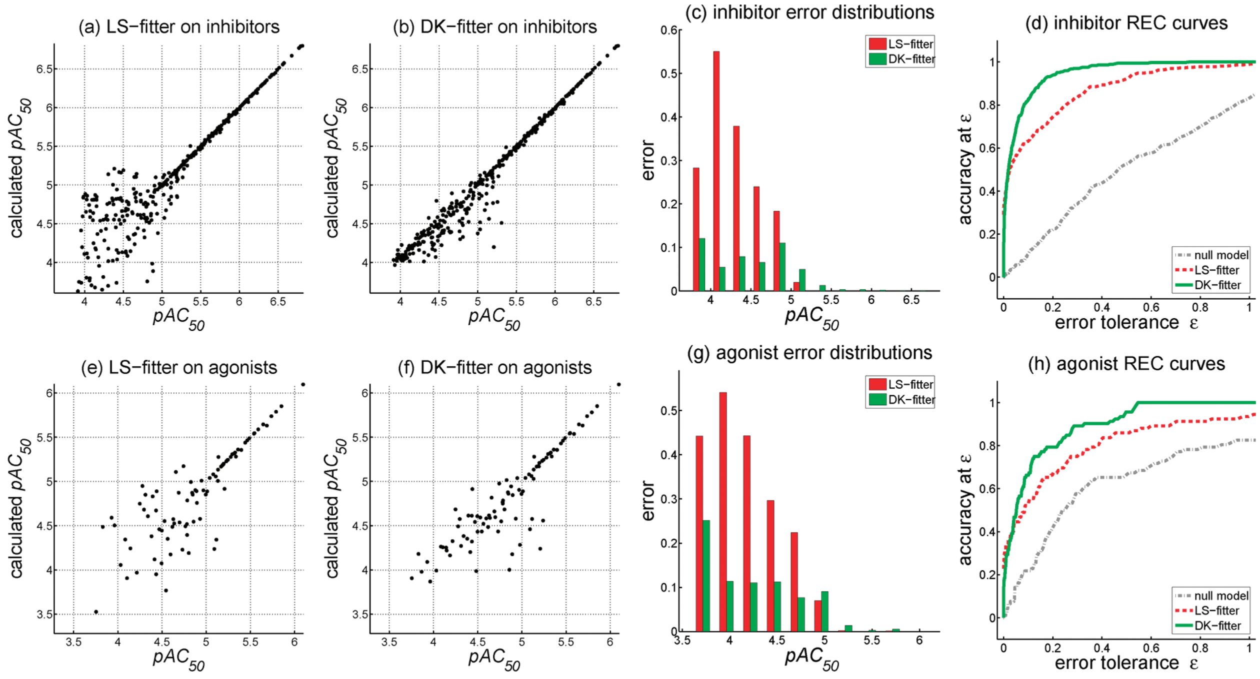


Figure 7. Results on AID 361 simulated samples. The top row presents results for inhibitors and the bottom row for agonists. Known versus calculated pAC_{50}'s are plotted using LS-fitter of panels (a,e) and DK-fitter of panels (b,f). Less variability is observed with DK-fitter. Error distributions confirm this in panels (c,g). REC curves are also presented as panels (d,h); the highest curve indicates the best method. This figure finds that DK-fitter is the better algorithm for this qHTS screen.

on pyruvate kinase, an enzyme involved in glycolysis, and reports dose−response information for 51 415 compounds. The scatter plots of panels a,b,e,f relate the pAC_{50} computed using LS- and DK-fitter, respectively, for testing samples. We observe significantly less variability using DK-fitter at smaller pAC_{50}'s. These observations are confirmed in the error distribution bar charts of panels c,g.

Examining panels c,g of Figure 7, some curve-fitting accuracy with higher pAC_{50} appears to be sacrificed so as to get a much larger gain in fitting accuracy in samples with lower pAC_{50}. This is an artifact of the method, as LS-fitter fits on reliable actives are deemed to be exact. This explains the apparent near-perfect results at the high end of pAC_{50} in panels a,e.

REC plots[16] are supplied as panels d,h. These plots clearly highlight the improved accuracy of DK-fitter. Both algorithms outperform a null model that assigns the mean pAC_{50} to all samples. We find an inhibitor DK-fitter AOC = 0.051, roughly 2.7 times smaller than LS-fitter and 4.3 times smaller than the null model.

7. VALIDATION VIA QSAR MODELING

In this final section on DK-fitter validation, we learn QSAR models to predict compound potency (as measured using pAC_{50}) from molecular descriptors. The accuracy of models using the LS- and DK-fitter responses are compared. For 8 screens (AID 357, 361, 411, 585, 884, 1463, 1468, and 2100), 851 descriptors were calculated from molecular structures: These consisted of MOE (implemented within the Molecular Operating Environment software, Chemical Computing Group, http://www.chemcomp.com/) and RECON[17] descriptors that were generated at the Rensselaer Exploratory Center for Cheminformatics Research (RECCR, http://reccr.chem.rpi.edu/). For each screen's descriptors, principal components analysis was performed for dimensionality reduction with 25 components being retained. These served as input to regression models trained using least-squares support vector machines (LS-SVM)[18] using a fast subgradient algorithm.[19−21] The tradeoff hyperparameter between structural and empirical risks was set by five-fold cross-validation. Regression accuracy is measured as the proportion of samples whose predicted pAC_{50} is within 0.5 of the response. Figure 8 presents the cross-validated results for this experiment; models trained using DK-fitter pAC_{50}'s outperform models trained using LS-fitter ones in all 8 screens.

8. COMPUTATIONAL COMPLEXITY

We briefly compare the computational complexity of both algorithms, beginning with a theoretical analysis followed by empirical execution times. In this discussion, we focus on the largest screen considered in this paper: AID 2100 consisting of 300 703 samples. Execution times are reported for a Dell Precision T5400 computer comprised of Dual Intel E5430 processors at 2.66Ghz with 8GB of RAM, running the 64-bit CentOS 5.4 operating system and Matlab 2007b.

To begin, we note that a very small number of operations are performed on an entire screen. At the start, dose−response points for all samples are loaded into memory, and basic arithmetic operations are performed to categorize each sample, such as calculating numerical first and second derivatives (subtraction and division, e.g., eq 14) and making comparisons (assessing whether they fall within prescribed intervals, e.g., eq 21). On AID 2100, sample categorization takes 30 s. From this point on, only a small percentage of samples are considered: For AID 2100, there are 1108 are reliable inhibitors and 2030 are uncertain inhibitors. (This screen possesses virtually no agonists, as seen in Table 1.)

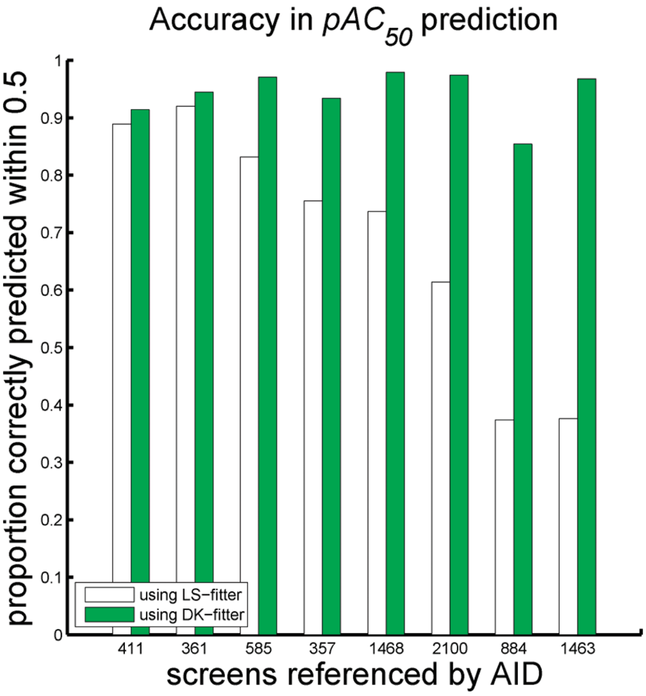


Figure 8. Results from QSAR modeling. Eight qHTS screens underwent QSAR modeling to predict their LS-fitter (white bars, left) and DK-fitter (green bars, right) pAC_{50}'s from molecular descriptors. These screens are sorted by increasing root mean squared difference between LS- and DK-fitter pAC_{50} across each screen. We see that bigger differences between both pAC_{50}'s results in predictions of LS-fitter pAC_{50} that are weaker.

The focus is on solving the optimization problems 4 and 6. All other remaining operations required fewer than 5 s. For a given sample, we found that both problems require roughly the same effort. However, fitting a reliable sample was generally faster (0.1−0.2 s) than fitting an uncertain one (0.4−0.5 s), but still within an order of magnitude of each other. Hence, we count the number of times that problems 4 and 6 are solved. Let r denote the number of reliable samples and u the number of uncertain samples. Let t denote the training set size (which is the smaller of half the simulated set or 512 samples, as seen in Section 6).

For LS-fitter, optimization problem 4 is solved $r + u$ times (once per reliable and uncertain sample), so its computational complexity is $\mathcal{O}(r + u)$.

For DK-fitter, extra work is required to set hyperparameters $\{\lambda_1,\lambda_2\}$. First, optimization problem 4 is solved r times (once per reliable sample). Second, the simulated sample set is generated, as detailed in Supporting Information. Third, up to 512 simulated samples for the training set. Fourth, a 2D grid consisting of 7 values for λ_1 and 7 values for λ_2 is generated, for a total of 49 combinations. Fourth, optimization problem 6 is solved for each hyperparameter combination and training sample, so at most $49t$ times. Fifth, the best hyperparameter combination is retained, and optimization problem 6 is solved $r + u$ times (once per reliable and uncertain sample) with the chosen hyperparameter combination. In sum, at most

$49t + 2r + u$ optimization problems are solved. Therefore, for smaller screens, the selection of hyperparameters dominates execution time. For increasingly large screens, fitting the model to each reliable and uncertain sample dominates and the algorithm scales linearly with sample size: $\mathcal{O}(2r + u)$.

Looking at empirical execution times on AID 2100, it took 3.91 min to solve 4 on the 1108 reliables (0.212 s/sample) and 22.6 min to obtain solutions on the 2030 uncertains (0.688 s/sample). Hence, LS-fitter took 26.5 min.

For DK-fitter, selecting the hyperparameters required solving 49 × 395 = 19 355 optimization problems. This was done in 84.8 min (0.263 s/problem). The chosen hyperparameters were then applied to solve 6 on the 1108 reliables taking 2.93 min (0.159 s/sample) and the 2030 uncertains taking 11.6 min (0.344 s/sample). Hence, DK-fitter took 103 min. For this screen, the cost of improved accuracy is ≈3.8 times more computational effort.

Using these results, we project that applying LS-fitter to a screen 10 times larger will take 265 min and applying DK-fitter will take 293 min. For this data set of some 3 million samples, obtaining the more accurate results provided by DK-fitter requires only 11% more effort.

9. DISCUSSION AND CONCLUSIONS

This paper developed a framework for improved curve fitting of the nonlinear Hill equation to quantitative high-throughput screening (qHTS) data. Performing curve-fitting on qHTS screens containing such a large number of samples, including outliers and erroneous activity measurements, is a challenging cheminformatics task. The result is a framework for better SAR interpretation of qHTS data.

The state-of-the-art fits each sample independently using a least-squared criterion (LS-fitter). Our work learns domain knowledge from a qHTS screen and injects this information into the curve-fitting process. Domain knowledge comes from two sources: explicitly, estimating the distributions of Hill parameters, and implicitly, selecting the regularization parameters. We also specified decisive rules for categorizing all samples in a qHTS screen.

Our new DK-fitter approach is demonstrated in this paper to outperform LS-fitter in three ways: First, we compared two replicate screens (AID 1463 and 1468). Our findings in Section 5 convincingly demonstrate that DK-fitter can better handle perturbed samples, by demonstrating substantially better agreement between DK-fitter-calculated pAC_{50}'s across both screens. We believe that our approach of encouraging the Hill slope m and terminal activity A_∞ to take values that are reasonable for that screen, where reasonable is based on observed distributions in reliable active samples, causes this stability. The quality Hill equation fits returned by DK-fitter may reduce or eliminate the need for confirmatory screens.

Second, we presented a novel validation technique that involves shifting reliable active samples (whose Hill equation parameters are assumed to be known) in concentration to simulate samples that appear to present an uncertain data profile. This technique permits the unbiased comparison of both curve-fitting algorithms on samples that are harder to fit. Using this approach, our results presented in Section 6 prove the superiority of DK-fitter in returning fits that best estimate the pAC_{50}.

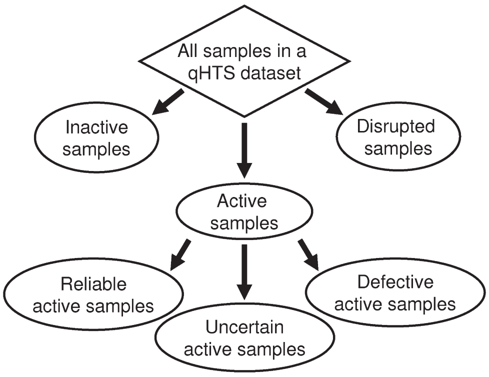


Figure 9. Each sample is categorized as active, inactive, or disrupted. Active samples are further categorized as reliable, uncertain, or defective.

Third, we used the pAC_{50}'s computed using both algorithms as the response of a QSAR model, as detailed in Section 7. In all tested cases, we found that the DK-fitter responses improved model accuracy. We believe that this is due to the DK-fitter pAC_{50}'s containing more accurate information, while the LS-fitter ones add a significant amount of noise to the model. The ability to accurately assess Hill equation parameters in a screen should increase the predictive power of models and reduce the volume of data required to produce accurate predictions.

We also showed that our DK-fitter method is linearly scalable, as is LS-fitter, and that the improved reliability afforded by DL-fitter is obtained with a decreasing additional computational effort for increasingly large screens.

The general ideas underlying this paper may extend to other nonlinear curve-fitting tasks that arise in the processing and interpretation of chemistry data or even any data set where a nonconvex function is fit to data points where reliable and uncertain cases are present.

Other research groups[5,6] have focused on methodologies that identify anomalous samples and, wherever possible, delete one or several data points from a sample. Future directions might include replicating this approach (or incorporate deleted point information provided on PubChem) so as to reduce the number of disrupted samples and hence provide a greater sample size to DK-fitter from which to learn proper fits.

We intend to use our accurate pAC_{50}'s as a measure of compound potency for virtual screening hit identification, in the vein of our preliminary studies[22] and the proof-of-concept in Section 7.

10. METHODS

This section presents detailed methods information used in this paper to implement LS-fitter, develop DK-fitter, and assess the abilities of both. In Section 10.1, we write out numerical differentiation formulas and their error terms for a dose−response sample. These formulas are used to specify conditions with respect to Hill curve shape and define dose−response sample categories based on their data points in Section 10.2.

10.1. Numerical Differentiation of Dose−Response.

Denote the true activity of a sample at concentration $\log C_i$ as $B(\log C_i)$ and the experimental qHTS measurement as B_i. Assume that the experimental error on activity measurements is normally distributed with no bias:

$$B(\log C_i) = B_i + \xi_i \tag{11}$$

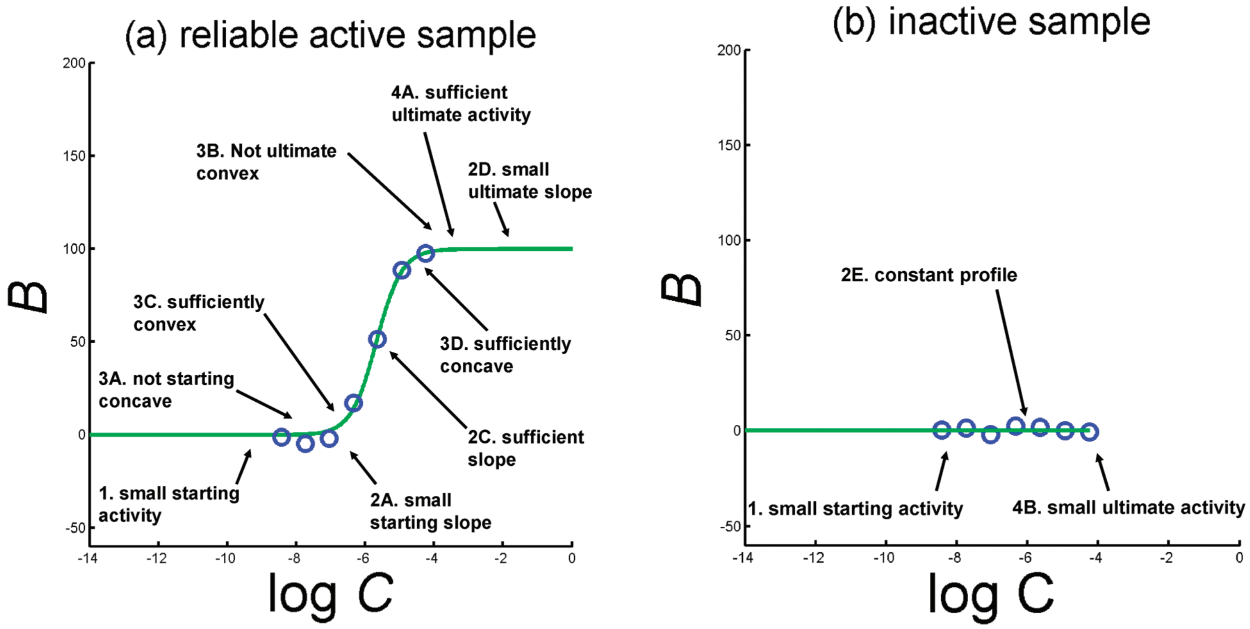


Figure 10. Qualitative characteristics of (a) reliable active and (b) inactive dose−response curves. An active sample's data points cover its full activity range. At low concentrations, the activity is zero and constant. As the concentration increases, so does the activity. The curvature is initially upward (convex) until the maximum slope is reached, after which the curvature is downward (concave). At high enough concentrations, the activity is constant with value B_∞. An inactive sample presents a zero and constant profile across the experimental concentration range (ECR). Characteristic labels (1, 2A, 2B, etc.) correspond to conditions defined in Section 10.2.

with

$$\xi_i \sim \mathcal{N}(0, \omega) \tag{12}$$

Assume that this error is independent across concentrations and samples. Parameter ω states the level of measurement error in a screen; we describe our procedure for estimating it in Supporting Information. We now differentiate eq 11 so as to quantitatively study Hill equation properties:

$$\frac{\partial B}{\partial \log C}(\log C) = B'_{i,i+1} + \xi'_{i,i+1} \tag{13}$$

where the numerical first derivative is given by

$$B'_{i,i+1} = \frac{B_{i+1} - B_i}{\log C_{i+1} - \log C_i} \tag{14}$$

This is simply the slope of the line segment connecting two successive data points. This numerical derivative estimate is valid over interval $[\log C_i, \log C_{i+1}]$. Moreover, it is exact at at least one concentration in this interval (according to the mean value theorem, since the Hill equation is smooth). In a slight abuse of notation, this interval is associated with midpoint concentration $\log C_{i+1/2}$. The error term is

$$\xi'_{i,i+1} \sim \mathcal{N}(0, \omega'_{i,i+1}) \tag{15}$$

with standard deviation

$$\omega'_{i,i+1} = \frac{\sqrt{2}\omega}{\log C_{i+1} - \log C_i} \tag{16}$$

The expression for $\omega'_{i,i+1}$ exploits the additive property of variance. Differentiating eq 13 a second time, we find

$$\frac{\partial^2 B}{\partial (\log C)^2}(\log C) = B''_{i,i+2} + \xi''_{i,i+2} \tag{17}$$

where the numerical second derivative is given by

$$B''_{i,i+1} = \frac{B'_{i+1,i+2} - B'_{i,i+1}}{\log C_{i+2} - \log C_i} \tag{18}$$

This estimate is valid over interval $[\log C_i, \log C_{i+2}]$ and is identified by its midpoint C_{i+1}. The error term is

$$\xi''_{i,i+2} \sim \mathcal{N}(0, \omega''_{i,i+2}) \tag{19}$$

with standard deviation

$$\omega''_{i,i+2} = \frac{\sqrt{(\omega'_{i+1,i+2})^2 + (\omega'_{i,i+1})^2}}{\log C_{i+2} - \log C_i} \tag{20}$$

The experimental values B_i for the activity and their numerical derivatives along with estimates of their error distributions permit the definition of criteria so as to classify samples according to each dose−response profile. This is the theme of Section 10.2.

10.2. Sample Categorization.

In Section 2, we qualitatively describe active and inactive dose−response profiles. We now undertake a quantitative analysis to characterize samples from their dose−response data points. These permit the specification of rules, or conditions, that characterize dose−response samples. We use these conditions to formally define our sample categorization at the end of this subsection.

In any qHTS screen, a large proportion of samples are inactive, meaning that small activity is recorded across the ECR. These samples might exhibit greater activity at concentrations beyond

C_n, but sample data do not suggest at what concentrations this might be observed. A much smaller number exhibit at least some active behavior. An even smaller number still present some nonzero constant activity at higher tested concentrations. The purpose of this subsection is to organize the samples into categories that are appropriate for the overall tasks of curve fitting with domain knowledge. The result is a categorization that assigns each sample as reliable active, uncertain active, defective active, inactive, or disrupted. This breakdown is summarized in Figure 9.

Figure 10 identifies qualitative characteristics that active and inactive qHTS dose−response profiles possess. Looking at panel a of Figure 10, we see that the activity is zero and constant at low concentrations. Then, as concentration increases, so does the activity. At first, the shape of the fit Hill equation is convex (upward curvature), and then it shifts to being concave (downward curvature). In doing so, the slope goes from being gentle to steeper and then gentle again. At the highest measured concentrations, the terminal constant activity is reached. These observations, quantified as conditions that are used in formal definitions. Later in this paper, it is reliable active samples that determine typical Hill slopes and terminal activities for an assay.

The remainder of this subsection formally defines these sample categories. We assume that the ECR is chosen such that samples have small activity at low concentrations. This gives the first condition.

Condition 1: Small Starting Activity. A sample satisfies Condition 1 if it has small starting activity: $-\tau_f < B_1 < \tau_f$.

Conditions 2A−E relate to the slope of a dose−response profile.

Condition 2A: Small Starting Slope. A sample satisfies Condition 2A if the starting slope is small: $-2\omega'_{1,2} < B'_{1,2} < 2\omega'_{1,2}$.

Note that, a small first derivative is within 2 standard deviations of the slope error distribution: $2\omega'$. A dose−response profile of an active sample should have activity measurements B_i that increase with concentration. Mathematically, this means that the derivative is positive for $i \in \{1, ..., n-1\}$:

$$B'_{i,i+1} > 0 \tag{21}$$

However, small decreases caused by experimental error are permitted.

Condition 2B: Increasing Profile. A sample satisfies Condition 2B if the dose−response profile is increasing (allowing for small decreases): $B'_{i,i+1} > -2\omega'_{i,i+1}$ for $i \in \{1, ..., n-1\}$.

Additionally, active samples present significant slope somewhere over the ECR.

Condition 2C: Sufficient Slope. Let

$$i^* = \underset{i \in \{1,...,n-1\}}{\arg\max} \; B'_{i,i+1} \tag{22}$$

A sample satisfies Condition 2C if maximum slope B'_{i^*,i^*+1} is sufficiently large: $B'_{i^*,i^*+1} > \tau_s$.

Finally, at ultimate concentration C_n, the slope of an active dose−response is zero.

Condition 2D: Small Ultimate Slope. A sample satisfies Condition 2D if its ultimate slope is small: $-2\omega'_{n-1,n} < B'_{n-1,n} < 2\omega'_{n-1,n}$.

Inactive samples have a constant activity profile; we define this as having small slope throughout the ECR.

Condition 2E: Constant Profile. A sample satisfies Condition 2E if the dose−response profile is constant: $-2\omega'_{i,i+1} < B'_{i,i+1} < 2\omega'_{i,i+1}$ for $I \in \{1, ..., n-1\}$.

Conditions 3A−D relate to a dose−response profile's curvature.

An active sample is convex (upward curvature, positive second derivative) at concentrations below pAC$_{50}$ and concave (downward curvature, negative second derivative) at concentrations above pAC$_{50}$. Mathematically, these are written as:

$$B''_{i,i+2} > 0, \quad C_{i+1} < \mathrm{AC}_{50} \tag{23}$$

$$B''_{i,i+2} < 0, \quad C_{i+1} > \mathrm{AC}_{50} \tag{24}$$

Once again, we must allow for experimental noise causing small deviations from these rules. Thus

$$B''_{i,i+2} > -2\omega''_{i,i+2}, \quad C_{i+1} < \mathrm{AC}_{50} \tag{25}$$

$$B''_{i,i+2} < 2\omega''_{i,i+2}, \quad C_{i+1} > \mathrm{AC}_{50} \tag{26}$$

Note that, a small second derivative is within 2 standard deviations of the slope error distribution: $2\omega''$. We do not know what AC$_{50}$ is until the Hill equation is fit, so we impose the first rule at the starting concentration and the second one at the ultimate concentration.

Condition 3A: Not Starting Concave. A sample satisfies Condition 3A if the starting curvature is not concave: $B''_{1,3} > -2\omega''_{1,3}$.

Condition 3B: Not Ultimate Convex. A sample satisfies Condition 3B if the ultimate curvature is not convex: $B''_{n-2,n} < 2\omega''_{n-2,n}$.

We also require an active sample to present sufficient convex and concave behavior over the ECR.

Condition 3C: Sufficiently Convex. Let

$$i^{\Delta} = \underset{i \in \{1,...,n-2\}}{\arg\max} \; B''_{i,i+2} \tag{27}$$

A sample satisfies Condition 3C if maximum curvature $B''_{i^{\Delta},i^{\Delta}+2}$ is sufficiently large: $B''_{i^{\Delta},i^{\Delta}+2} > \tau_c$.

Condition 3D: Sufficiently Concave. Let

$$i^{\nabla} = \underset{i \in \{1,...,n-2\}}{\arg\min} \; B''_{i,i+2} \tag{28}$$

A sample satisfies Condition 3D if minimum curvature $B''_{i^{\nabla},i^{\nabla}+2}$ is sufficiently small: $B''_{i^{\nabla},i^{\nabla}+2} < -\tau_c$.

An active sample must demonstrate sufficient activity at the ultimate concentration.

Condition 4A: Sufficient Ultimate Activity. A sample satisfies Condition 4A if $B_n > \tau_u$.

On the other hand, an inactive sample should have small activity at the high end of the ECR.

Condition 4B: Small Ultimate Activity. A sample satisfies Condition 4B if $-\tau_f < B_n < \tau_f$.

We additionally require the sufficient convex behavior occur at a lower concentration than the sufficient slope behavior and that the latter occur at a lower concentration than the sufficient concave behavior.

Condition 5A: Convex before Steep. Consider indices i^* And i^{Δ} from Conditions 2C and 3C. A sample satisfies Condition 5A if $C_{i^{\Delta}+1} < C_{i^*+1/2}$.

Condition 5B: Steep before Convex. Consider indices i and $i^{\triangledown}$ from Conditions 2C and 3D. A sample satisfies Condition 5B if $C_{\mathrm{i}^{*}+1/2} < C_{i^{\triangledown}+1}$.

Several conditions rely on thresholds that are set by the user. After inspection of samples across many qHTS screens, we find that the following choices worked well:

$$\tau_f = 15, \ \tau_s = 24, \ \tau_c = 12, \ \tau_u = 30 \tag{29}$$

Each qHTS is categorized using the following definitions.

Definition A: Active Sample. Active samples present significant inhibitory or agonistic behavior. They satisfy Conditions 1 and 4A.

Definition A1: Reliable Active Sample. A reliable active is an active sample that contains complete dose−response information, covering its full AR. These samples meet the conditions for being active, as well as Conditions 2B, 2C, 3A, 3B, 3C, 3D, 5A, and 5B. These Conditions guarantee that the AC_{50} is well within the ECR as concave behavior at a greater concentration must be observed.

Definition A2: Uncertain Active Sample. An Uncertain active is an active sample that contains partial dose−response information. Uncertain actives meet the requirements for reliables except one or several of Conditions 3B, 3D, and 5B. These conditions do not guarantee that the AC_{50} is within the ecr, but suggest that if it is not, then it it not far beyond its upper bound. This is true because convex behavior at a concentration less than AC_{50} is noted.

Definition A3: Defective Active Sample. A defective active is an active sample that demonstrates significant ultimate activity but does not display the dose−response profile of an active sample as provided by the Hill equation. These are active samples that are not reliable or uncertain samples. Some of the samples in this category have nonsmall starting concentration. Others will fail to meet one or several of the requirements to be uncertain or reliable, possibly due to an anomalous data point. Fitting the Hill equation to these samples is problematic. We believe that the framework presented in this paper improves fitting to these samples, but this is a difficult claim to prove for lack of ground truth.

Definition B: Inactive Sample. Inactive samples present negligible activity across the ECR. They satisfy Conditions 1, 2E, and 4B.

Definition C: Disrupted Sample. Disrupted samples are samples that are neither active nor inactive.

Generating a list of rules that can handle the classification of qHTS experimental samples on such large and diverse qHTS screens is a difficult task, and might require further constraining or loosening the requirements for a given category. We visually inspected the dose−response profiles for different categories in each screen to summarily confirm that the rules are functioning as intended.

We note that two conditions elaborated in this subsection are not used:

(1) Imposing Condition 2A resulted in the misclassification of many samples that would otherwise be actives with low AC_{50}.

(2) In most screens, Condition 2D was found to be too strong a requirement for a reliable active.

ASSOCIATED CONTENT

Ⓢ Supporting Information. An expanded case for the need for DK-fitter, implementation details for solving the LS- and DK-fitter optimization problems, a protocol for dealing with assays possessing insufficient reliable active samples, the formal algorithm for generating simulated samples, and detailed result analysis in the form of Figure 7 for the qHTS screens with AID 883, 884, 886, 893, 1458, 1460, 1463, and 1688. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: chbergeron@gmail.com. Telephone: +1 (518) 276-6899.

ACKNOWLEDGMENT

This work was funded by the Fonds Québécois de la Recherche sur la Nature et les Technologies doctoral fellowship program and a grant by the Office of Naval Research (number N00014-06-1-0014). The authors thank the Rensselaer Exploratory Center for Cheminformatics Research (RECCR).

REFERENCES

(1) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11473–11478.

(2) Dutta, S.; Matsumoto, Y.; Ebling, W. F. Is it possible to estimate the parameters of the sigmoid Emax model with truncated data typical of clinical studies? *J. Pharm. Sci.* **1996**, *85*, 232–239.

(3) Schnecke, V.; Bostrom, J. Computational chemistry-driven decision making in lead generation. *Drug Discovery Today* **2006**, *11*, 43–50.

(4) Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **2006**, *24*, 167–175.

(5) Auld, D. S.; Inglese, J.; Jadhav, A.; Austin, C. P.; Sittampalam, G. S.; Montrose-Rafizadeh, C.; Mcgee, J. E.; Iversen, P. W. HTS technologies to facilitate chemical genomics. *Eur. Pharm.l Rev.* **2007**, 53–63.

(6) Southall, N. T.; Jadhav, A.; Huang, R.; Nguyen, T.; Wang, Y. In *Handbook of Drug Screening (Drugs and the Pharmaceutical Sciences)*, 2nd ed.; Seethala, R., Zhang, L., Eds.; Informa Healthcare: New York, **2009**; Chapter 19, pp 442−463

(7) Auld, D. S.; Southall, N. T.; Jadhav, A.; Johnson, R. L.; Diller, D. J.; Simeonov, A.; Austin, C. P.; Inglese, J. Characterization of Chemical Libraries for Luciferase Inhibitory Activity. *J. Med. Chem.* **2008**, *51*, 2372–2386.

(8) Babaoglu, K.; Simeonov, A.; Irwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C. P.; Shoichet, B. K. Comprehensive Mechanistic Analysis of Hits from High-Throughput and Docking Screens against Beta-Lactamase. *J. Med. Chem.* **2008**, *51*, 2502–2511.

(9) Simeonov, A.; Jadhav, A.; Sayed, A. A.; Wang, Y.; Nelson, M. E.; Thomas, C. J.; Inglese, J.; Williams, D. L.; Austin, C. P. Quantitative High-Throughput Screen Identifies Inhibitors of the Schistosoma mansoni Redox Cascade. *PLoS Negl. Trop. Dis.* **2008**, *2*, e127.

(10) Auld, D. S.; Zhang, Y.-Q.; Southall, N. T.; Rai, G.; Landsman, M.; MacLure, J.; Langevin, D.; Thomas, C. J.; Austin, C. P.; Inglese, J. A Basis for Reduced Chemical Library Inhibition of Firefly Luciferase Obtained from Directed Evolution. *J. Med. Chem.* **2009**, *52*, 1450–1458.

(11) Boxer, M. B.; Jiang, J.-K.; Heiden, M. G. V.; Shen, M.; Skoumbourdis, A. P.; Southall, N.; Veith, H.; Leister, W.; Austin, C. P.; Park, H. W.; Inglese, J.; Cantley, L. C.; Auld, D. S.; Thomas, C. J. Evaluation of Substituted N,N′-Diarylsulfonamides as Activators of the Tumor Cell Specific M2 Isoform of Pyruvate Kinase. *J. Med. Chem.* **2010**, *53*, 1048–1055.

(12) Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2010**, *53*, 37–51.

(13) Zheng, W.; Padia, J.; Urban, D. J.; Jadhav, A.; Goker-Alpan, O.; Simeonov, A.; Goldin, E.; Auld, D.; LaMarca, M. E.; Inglese, J.; Austin, C. P.; Sidransky, E. Three classes of glucocerebrosidase inhibitors identified by quantitative high-throughput screening are chaperone leads for Gaucher disease. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 13192–13197.

(14) Hill, A. The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *J. Physiol.* **1910**, *40*, iv–vii.

(15) Atkins, G. L. A single digital-computer program for estimating the parameters of the Hill equation. *Eur. J. Biochem.* **1973**, *33*, 175–180.

(16) Bi, J.; Bennett, K. P. Regression Error Characteristic Curves. *Proc. Int. Conf. Mach. Learn.* **2003**, *20*, 43–50.

(17) Breneman, C.; Rhem, M. A QSPR Analysis of HPLC Column Capacity Factors for a Set of High-Energy Materials Using Electronic Van der Waals Surface Property Descriptors Computed by the Transferable Atom Equivalent Method. *J. Comput. Chem.* **1997**, *18*, 182–197.

(18) Suykens, J.; Gestel, T. V.; Brabanter, J. D.; Moor, B. D.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 2002.

(19) Joachims, T. Training Linear SVMs in Linear Time. *Proc. ACM Conf. Knowl. Discovery Data Min.* **2006**, 217–226.

(20) Teo, C. H.; Le, Q. V.; Smola, A.; Vishwanathan, S. V. N. *Proc. ACM Conf. Knowl. Discovery Data Min.* **2007**, 727–736.

(21) Bergeron, C.; Moore, G.; Zaretzki, J.; Breneman, C. M.; Bennett, K. P. Fast Bundle Algorithm for Multiple Instance Learning. *IEEE T Pattern Anal.*; in press.

(22) Moore, G.; Bergeron, C.; Bennett, K. P. Model selection for primal SVM. *Mach. Learn.* **2011**, *85*, 175–208.