New analysis pipeline for high-throughput domain–peptide affinity experiments improves SH2 interaction data

Tom Ronan¹, Roman Garnett* ©, and Kristen M. Naegle ¹*©
From the Departments of ¹Biomedical Engineering and ²Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri, USA

Edited by Karen G. Fleming

© 2020 Ronan et al. Published under exclusive license by The American Society for Biochemistry and Molecular Biology, Inc.

Protein domain interactions with short linear peptides, such as those of the Src homology 2 (SH2) domain with phosphorytrosine-containing peptide motifs (pTyr), are ubiquitous and important to many biochemical processes of the cell. The desire to map and quantify these interactions has resulted in the development of high-throughput (HTP) quantitative measurement techniques, such as microarray or fluorescence polarization assays. For example, in the last 15 years, experiments have progressed from measuring single interactions to covering 500,000 of the 5.5 million possible SH2–pTyr interactions in the human proteome. However, high variability in affinity measurements and disagreements about positive interactions between published data sets led us here to reevaluate the analysis methods and raw data of published SH2–pTyr HTP experiments. We identified several opportunities for improving the identification of positive and negative interactions and the accuracy of affinity measurements. We implemented model-fitting techniques that are more statistically appropriate for the nonlinear SH2–pTyr interaction data. We also developed a method to account for protein concentration errors due to impurities and degradation or protein inactivity and aggregation. Our revised analysis increases the reported affinity accuracy, reduces the false-negative rate, and increases the amount of usable data by adding reliable true-negative results. We demonstrate improvement in classification of binding versus nonbinding when using machine-learning techniques, suggesting improved coherence in the reanalyzed data sets. We present revised SH2–pTyr affinity results and propose a new analysis pipeline for future HTP measurements of domain–peptide interactions.

Protein domain interactions with short linear peptides are found in many biochemical processes of the cell and play a central role in cell physiology and communication. For example, SH2 domains are central to pTyr-signaling networks, which control cell development, migration, and apoptosis (1). The 120 human SH2 domains are considered “readers” because they read the presence of tyrosine phosphorylation by binding specifically to certain phosphorylated amino acid sequences. Approximately half of the binding energy of the SH2–pTyr sequence interaction is due to an invariant arginine, which creates a salt bridge with the ligand pTyr. The remainder of the binding energy results from interactions between the SH2 domain–binding pocket and the residues flanking central pTyr residues (2–4), resulting in specificity of SH2 domain interactions critical to pTyr-mediated signaling (5). Measurements of all SH2-binding affinities for target peptides would greatly aid in the decryption of domain specificity and advance understanding of cell-signaling networks that control human physiology. However, the total potential number of interactions is immense—the 46,000 tyrosines currently known to be phosphorylated in the human proteome (6) have the potential to interact with 120 human SH2 domains, resulting in over 5.5 million possible SH2–pTyr interactions.

Recent developments have expanded the measurement coverage of human SH2–pTyr interactions. Eight high-throughput (HTP) studies have been performed to measure SH2 domain interactions with specific phosphopeptide sequences (7–14) (Table 1) using either microarrays or fluorescence polarization (FP). The six studies that quantitatively measured affinity represent ~90,000 pairs of domain–peptide interactions, but these measurements cover less than 2% of possible interactions. In response, computational approaches have attempted to predict as-of-yet unmeasured interactions using the published interaction data. These methods span the range from thermodynamic models that predict interaction strength using existing structure and binding measurements (15–17) to supervised machine-learning models using patterns in peptide sequences and quantitative binding data to predict binding (14, 18). However, no computational method has used the available affinity data in their entirety. We therefore wished to leverage all available binding affinity data in a supervised learning approach to expand our knowledge of SH2–pTyr interaction space.

Unfortunately, in the process of reviewing published HTP data, we found surprising disagreement between publications about which domain–peptide pairs interacted. For the limited number of interactions for which they agreed, they reported vastly different affinities for interacting pairs. We identified two issues common to all of the data sets that could be responsible for the discrepancies in results: errors affecting protein concentration and improper use of statistical methods affecting modeling results.

First, we found potential sources of errors in protein concentration that could affect reported affinity values. Protein was minimally purified (via nickel chromatography), and protein concentration was measured by absorbance. No study used

This article contains supporting information.

*For correspondence: Kristen M. Naegle, kmn4mj@virginia.edu.
Present address for Kristen M. Naegle: Dept. of Biomedical Engineering and the Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA.

This is an Open Access article under the CC BY license.
positive controls to determine the degree of protein functionality before measuring affinity. Thus, protein of varying degrees of purity, functionality, and nonmonomeric content was used for affinity measurements. Impure or degraded protein causes overestimation of protein concentration when compared with the amount of active protein in a sample. These protein concentration errors can propagate directly to errors in affinity, because affinity is derived from concentration and activity.

Second, we found errors in model fitting and statistical methods used to evaluate model fitting, which could have significant impact on the reported affinities. All of the affinity studies used the receptor occupancy model and the coefficient of determination ($r^2$) as a determination of how well the model fit the data. For linear models, one can interpret the values of $r^2$ between 0 and 1 as the total percentage of variance explained by the fit. However, when applied to nonlinear models (like the receptor occupancy model used in each of these studies to derive affinity), the $r^2$ value cannot be interpreted as the percent of variance and has been conclusively shown to be a poor indicator of fitness (19). Although this fact has long been established in the statistical literature (20–26), $r^2$ is still commonly used to evaluate nonlinear models in pharmaceutical and biomedical publications despite being an ineffective and misleading metric. In these publications, the use of $r^2$ effectively resulted in a bias for true positive interactions at the expense of making many false-negative calls and removed many replicate measurements from incorporation into the reported affinity values.

Therefore, due to both inaccuracies in quantitative results and the significant potential for large numbers of false-negative results, we had serious concerns about using the published affinities in machine learning. To overcome these issues, we decided to retrieve and reanalyze available raw data to systematically improve classification and affinity accuracy for SH2–pTyr interactions. To accomplish this, we 1) refined model fitting techniques, 2) implemented fitting multiple models to each measurement, 3) used a statistically accurate method for model selection, 4) developed methods to identify and remove nonfunctional protein from the results, and 5) introduced a novel method to address the effects of protein concentration errors on reported affinity.

Our revised analysis improves affinity accuracy, improves specificity by reducing the false-negative rate, and results in a dramatic increase in useful data due to the addition of thousands of true-negative results. Evaluation of the revised affinities shows improved learning accuracy within an active learning model, suggesting that there is improved coherency in the features of the revised data set.

### Results

#### Evaluation of published affinity data and acquisition of raw data

In the process of evaluating published high-throughput data, we found significant disagreement between data sets. We evaluated all publications using HTP methods to measure SH2 domain interactions with specific peptide sequences. The publications containing SH2 affinity data can be grouped into three distinct data groups (Table 1). The first data group consists of the group of studies published by the MacBeath laboratory from 2006 to 2009 (7, 9, 27), which contain a body of predominantly nonoverlapping protein microarray (PM) experiments. The second data group consists of a large study published by the MacBeath laboratory in 2013 (10) with a set of new PM measurements using the protocol published in 2010 (28). The third data group consists of two nonoverlapping sets of FP experiments published in 2012 and 2014 by the Jones laboratory (13, 14). Because the other experiments (11, 12) only measured interaction and not affinity, they were not considered for this analysis.

To determine agreement between data sets, we examined both qualitative and quantitative results. First, we examined the correlation between domain–peptide affinity measurements that overlapped between any two data groups (Fig. 1, top row). We found surprisingly low correlation between affinity measurements (with a maximum correlation of $r = 0.367$). Next we asked whether the different data groups identified the same positive interactions between domain–peptide pairs, even if they did not agree on the affinity measurements. Here, we found significant disagreement over which domain–peptide pairs were found to interact (Fig. 1, bottom row). Of 347 positive domain–peptide interactions identified by one or more groups, fewer than 16% (55 of 347) were found to interact in all three data groups. No two experiments were able to agree on more than 29% of the positive interactions. The differences in interaction identification were spread randomly among SH2 domains and peptides, with no single SH2 domain, peptide, or peptide family being overrepresented (Fig. S1). These findings demonstrate significant quantitative and

| Data Group | Type | Publication | Peptides | SH2 Domains | Affinity | Results Available | Raw Data Available | Models |
|------------|------|-------------|----------|-------------|----------|------------------|------------------|--------|
| PM         | Jones et al. (7) | 61 | 159 | Yes | Yes | No | SH2Peplnt (18) | Wunderlich/Mirny (17) |
| PM         | Kaul et al. (27) | 50 | 133 | Yes | Yes | No | SH2Peplnt (18) | |
| PM         | Gordon et al. (9) | 46 | 96 | Yes | Yes | Pos | |
| PM         | Kortiger et al. (10) | 729 | 70 | Yes | Yes | No | MSM/D (16), FoldX (15) | |
| FP         | Hauser et al. (13) | 89 | 93 | Yes | Yes | PC | PEBL (14) | |
| FP         | Leung et al. (14) | 85 | 93 | Yes | Yes | PC | PEBL (14) | |
| NA         | Liu et al. (11) | 192 | 50 | No | No (fg) | No | |
| NA         | Tinti et al. (12) | 6202 | 70 | No | No | No | |

New analysis pipeline improves SH2 affinity data

Eight high-throughput experiments have been published since 2006 using experimental techniques such as protein microarrays (PM), peptide arrays (PA), and fluorescence polarization (FP). Of the published studies, only two studies have raw data available, by personal communication. Even the published data from several studies are no longer available. pos, raw data only published for positive interactions; PC, data available only by personal communication; fig, published as a figure only, numerical results available by private communication; *, original results were stored in PepspotDB but not published in the journal or supplement. PepspotDB is no longer available. NA, not applicable.
New analysis pipeline improves SH2 affinity data

Figure 1. Comparison of published affinity data. Correlation of affinity from three data groups is evaluated using scatter plots (top row). See Table 1 for group definitions. With perfect agreement, data points would fall along the dashed gray line. Surprisingly, there is very low correlation between affinities from different data groups. Even results from the same laboratory published at different times show only mild correlation ($r = 0.367$, MacBeath 2006–2009 (7, 9, 27) versus MacBeath 2013 (10)). The two PM experiments (MacBeath 2006–2009 (7, 9, 27)) identified 28.0% (97 of 347) of the positive interactions in common, but a similar number of positive interactions can be seen (25.9%, 90 of 347) when comparing one PM experiment with the FP experiment (MacBeath 2013 (10), Jones 2012–2014 (13, 14)). The highest correlation was also between experiments of different techniques ($r = 0.367$; MacBeath 2006–2009 (7, 9, 27) versus MacBeath 2013 (10)).

Raw SH2 interaction data and revised analysis

We began by examining the raw data from the Jones 2012–2014 data group to evaluate the quality and completeness of the data and to review the methods used to process the raw data into its published form. Although some raw data were missing compared with the original publication, by limiting our revised analysis to interactions of single SH2 domains with phosphopeptides from the ErbB family (EGFR, ERBB2, ERBB3, and ERBB4), as well as KIT, MET, and GAB1, the available raw data covered ~99.6% of the reported measurements.

Evaluation of the implementation of the receptor occupancy model—The raw data for each measured interaction consisted of fluorescence polarization measurements of an SH2 domain in solution with a phosphopeptide at equilibrium at 12 concentrations. In the original publication, the raw data were then used to derive an equilibrium dissociation rate constant ($K_d$) by fitting the receptor occupancy model (developed by Clark in 1926 using the law of mass action (29)). As applied to the fluorescence polarization data, the model takes the following form,

$$F_{obs} = \frac{[\text{SH2 domain}]F_{max}}{K_d + [\text{SH2 domain}]}$$  \hspace{1cm} (Eq. 1)

where $F_{obs}$ is the observed FP signal at each assayed protein concentration of the SH2 domain (measured in millipolarization units (mP)), and $F_{max}$ represents the FP at saturation (see also Fig. S2). The affinity ($K_d$) and saturation limit ($F_{max}$) are fitted parameters of the model. It is important to note that this model depends on several critical assumptions: that the qualitative differences between published data from different laboratories and even disagreement between publications from the same laboratory

Although there are significant differences between the techniques of PM and FP experimental methods, the differences between positive interactors did not group by technique (Fig. 1), and different techniques had similar numbers of common positive interactions. Although very low, the best correlation ($r = 0.367$) was between a PM experiment and an FP experiment from the same laboratory. All three data groups used similar protein and peptide production and purification methods, absorbance for determination of protein concentration, the receptor occupancy model for determining affinity, and similar methods of evaluating model fits based on $r^2$.

We concluded that we would need to look further into the methods and raw data to evaluate the differences between published data sets or even to evaluate the quality of any single set of published data. Acquisition of raw data from published studies was surprisingly difficult. No publication included raw data, only supplemental tables with post-processed values for affinity, which are insufficient for replication of published results. Furthermore, we discovered that most raw data underlying the published analysis has been lost by the original authors and is no longer available from any party (Table 1). Fortunately, we were able to retrieve raw data from the Jones 2012–2014 data group.²

² R. Jones, R. Hause, and K. Leung, personal communication.
reaction is reversible, that the ligand only exists in a bound and unbound form, that all receptor molecules are equivalent, that the biological response is proportional to occupied receptors, and that the system is at equilibrium.

We hypothesized that the specific methods used to implement the receptor occupancy model in the original publications might have affected the accuracy of the originally published results. We examined three aspects of the implementation of this model. First, we reviewed the method of subtraction of background fluorescence and found that it introduces systematic random errors in affinity results. Second, we evaluated whether the receptor occupancy model could reliably fit a non-binding sample. When we found that it could not, we implemented an alternate model and a model selection procedure to more reliably identify negative interactions. Finally, we examined the effect of dropping outlier measurements on model fitting results and implemented an alternative method to determine model fitness: signal/noise ratio (SNR).

Background subtraction causes errors in model fits and is replaced by fitting an offset—In the original analysis, the authors used a plate-wise background subtraction method, where the median baseline control value was recorded from plate measurements and subtracted from the polarization signal observed at each data point (13). When plates had excessive variation in baseline control values, the authors excluded these results from further analysis. However, in examining many measurements by eye, we found that the background values seemed uncorrelated with the signal values (Fig. S3).

A critical feature of the receptor occupancy model is that the saturation curve passes through the origin (because the point of zero signal is also the point of zero concentration). Thus, background subtraction forces the zero signal to a point other than that of zero concentration, resulting in higher residual error, which introduces errors in derived affinity (Fig. S4). These errors increase or decrease affinity (based on whether the background is high or low) and are nonlinear by affinity. Because the relationship of the background was seemingly random, and the error factors are nonlinear, background subtraction injected random error in affinity calculations. More than 54% of the replicate measurements exhibited problematic background levels (Fig. S4, bottom row). Thus, we rejected the background subtraction method in favor of fitting the model along with an offset value ($F_{bg}$).

\[ F_{obs} = \frac{[SH2 \text{ domain}]F_{max}}{K_d + [SH2 \text{ domain}]} + F_{bg} \]  
(Eq. 2)

Example fits using the receptor occupancy model with offset can be seen in Fig. S5.

The receptor occupancy model fails to accurately identify nonbinding measurements; introducing a linear model—Although the receptor occupancy model is theoretically capable of fitting a typical binding saturation curve as well as a “flat” curve representative of nonbinding interactions, we found that it fails to accurately fit nonbinding interactions in practice (Fig. S6), resulting in artificial model fits with unreasonable affinity and saturation values.

New analysis pipeline improves SH2 affinity data

Because negative interactions resemble low-slope or zero-slope lines with superimposed random noise, we hypothesized that a linear model would more reliably fit these “nonbinders” and resolve fit artifacts. The linear model is as follows,

\[ F_{obs} = m[SH2 \text{ domain}] + F_{bg} \]  
(Eq. 3)

where $F_0$ represents an offset value, and $m$ is a constant representing the slope of the fitted line (Fig. S6, red fits). Of 37,378 replicate measurements, 31,861 were best fit by the linear model. Of these, 29,778 were initially classified as nonbinders.

We also found a group of replicate measurements (~6%), which were best fit by a linear model but with steep positive slope. Linearly increasing fluorescent signal with no indication of saturation violates the assumptions of a receptor occupancy model and is more likely to represent a form of protein aggregation, peptide aggregation, or some other form of nonspecific binding. Thus, to preserve the quality of the nonbinding calls, a conservative slope cutoff of 5 mP/µM was implemented, above which replicates were identified as aggregators and removed from further consideration. Of the 31,861 replicates best fit by the linear model, 2,083 were initially classified as aggregators.

Fitting multiple models requires a model selection process—When more than one model can be used to fit the data, a method of model selection must be implemented to determine which model most accurately represents the data while balancing against adding additional parameters, which can lead to overfitting. To determine whether a measurement is best described by a receptor occupancy model or a linear model, we used the Akaike information criterion (AIC). In contrast to $R^2$, AIC is a model selection metric that is appropriate for use with nonlinear models (19, 30), is robust even with high noise data, and employs a regularization technique to avoid overfitting by penalizing models with more parameters (the receptor occupancy has three parameters; the linear model has only two parameters). In our implementation, we used a bias-corrected form of the metric, AICc, to account for only having 12 data points per saturation curve. A lower AICc score indicates a better fit. Examples of model fitting can be seen in Fig. S7.

Evaluation of model fitness—To determine how well the data were represented by a model, we used the SNR as a metric of model fitness. The SNR metric represents the magnitude of residual errors of fit to the model (a form of noise) and weighs this sum by the overall size of the fluorescent signal measured. It is calculated as follows,

\[ SNR = \frac{\max(F_{obs}) - \min(F_{obs})}{\sum_{i=0}^{n} |R_i|} \]  
(Eq. 4)

where $n$ is the number of data points, $R_i$ is the residual value of the $i$th data point, and $F_{obs}$ is the observed fluorescence (in mP).

At an SNR $\geq 1$, the measured signal is larger than the sum of all errors to the fit and represents a good quality fit in practice. We chose a ratio of 1 as the limit of a good fit based on extensive visual inspection of the fits (see Figs. S8 and S9). Replicates with SNR $< 1$ made up 5.2% of fits (1,948/37,378). These low-
New analysis pipeline improves SH2 affinity data

SNR fits fell into three classes: zero signal measurements (76.0%, 1,480/1,948), measurements where noise swamped the signal (21.9%, 427/1,948), and good measurements with large single-point outliers (2.1%, 41/1,948). Although the metric excludes some viable measurements that would be kept when reviewing by eye (e.g. 41 single-point outlier replicate measurements), this represents only one-tenth of 1% of all replicate measurements (0.11%, 14/37,378). A consistent standard is difficult to implement without such a metric, and any objective metric would likely exclude some viable measurements (see “Discussion” for thoughts on alternate metrics.)

Outlier removal biases model fitting and selection—In the original publications, the authors utilized an iterative outlier removal process. For each set of 12 data points in a replicate measurement, individual points were identified as outliers using a statistical model. The outlier was removed, and the fit was reevaluated. Up to three points were removed per replicate measurement. For measurements where more than three data points were identified as outliers, the replicate measurement was removed from further consideration.

For an ideal binding saturation experiment attempting to identify $K_d$, the concentrations tested should span either side of $K_d$ and the highest and lowest measured concentrations should establish the plateaus seen on semi-log saturation plots (see Fig. S5). Based on the concentrations selected for this experiment, the ideal range for quantification is affinity ($K_d$) in the range of 0.05–0.5 μM. For interactions with a $K_d > 1.0$ μM, the upper plateau of the semi-log saturation curve no longer has any coverage (Fig. S5, row 2). Interactions with $K_d > 5$ μM have no data points at all above $K_d$ (Fig. S5, rows 3 and 4). This suggests that every data point is critical for accuracy, particularly points above $K_d$; thus, we chose to use all data points to avoid introducing additional error and to allow the SNR metric to gauge the quality of fit.

Summary of revised analysis method for replicate measurements—Following a systematic review of each decision made in evaluating a measurement in HTP affinity studies, we developed an improved analysis pipeline (Fig. 2). For each replicate measurement, we fit two models: a receptor occupancy model with offset (Equation 2) and a linear model with offset (Equation 3). Fits were evaluated with AICc; the model with the lower quality of fit. Replicates that were fit best by the linear model and had a slope of ≤5 mP/C20 μM were classified as negative interactions, or “nonbinders.” Linear fits with a slope greater than 5 mP/μM were classified as aggregators and removed from consideration. A replicate that was fit best by the receptor occupancy model was then evaluated for SNR. If the SNR was >1, the replicate was classified as a positive interaction or “binder.” Of 37,378 replicate measurements, we identified 7.4% (2,753) as binders, 79.7% (29,778) as nonbinders, 7.4% (2,764) as low-SNR fits, and 5.6% (2,083) as aggregators (Fig. 3, left side).

High variation at the replicate level is likely caused by protein concentration errors

The original publication reported a single affinity ($K_d$) value for each domain–peptide pair, which was the average of multiple replicate domain–peptide measurements. However, we found patterns of high variation in affinity between replicates that suggested a significant problem with the either experimental design or experimental method. We hypothesized that a single variable—errors in protein concentration—could be responsible for the high variance.

In reviewing data quality, we identified a large number of domain–peptide interactions demonstrating high variance in affinity among replicates (e.g. $K_d$ values ranging from <0.5 to >20 μM for replicates from a single domain–peptide interaction). To determine the source and character of the variance, we inspected replicates as a group for individual domain–peptide interactions (for a representative example, see Fig. S10). Despite high variance between replicates, each replicate measurement had high-quality fits and low residual error, as expected from meeting an SNR > 1. (For a representative example of all measurements from one such replicate group, see Fig. S10).

To explore this further, we visualized and quantified variance (Fig. 4) for all domain–peptide interactions. Although variance tends to increase as $K_d$ increases, variance greater than 10 μM is found across a large fraction of all measurements, independent of affinity. How could high-quality individual replicate measurements result in such varied affinities for a single domain–peptide pair? We hypothesized that protein concentration error (arising from differences in protein preparations, such as impurities, degradation, and inactivity) could directly propagate to errors in modeled affinity values while still producing high-quality individual replicate saturation curves.

To test this hypothesis, we first examined the theoretical effects of protein concentration error on affinity. We demonstrate that concentration errors directly manifest as errors in affinity and that errors from impurity or degradation systematically manifest as artificially high $K_d$ (lower affinity). Next, we examined the methods and data for sources of purification errors, partial degradation, and complete protein inactivity and identified evidence of all three. Finally, we developed a method to control for these sources of protein concentration error and produce affinities with higher accuracy using the existing raw data.

Protein concentration errors propagate directly as errors in derived $K_d$ values—Although binding affinity is a molecular property—affinity is the strength of interaction between a single protein molecule and a single peptide—accurate derivation and calculation of affinity by most methods depends on the accuracy of concentration measurements for the tested protein. In the case of the receptor occupancy model used here, affinity is a derived function of concentration and FP response. Because impurities or degraded protein represent an error between the assumed concentration and the active concentration of a protein, we hypothesized that this would propagate to errors in affinity.

We examined the theoretical effect of concentration errors on measured affinity (Fig. 5). Errors in protein concentration due to impurities or degradation cause an overestimation of the true concentration of active protein. Overestimation errors in protein concentration cause errors in $K_d$, always resulting in a
higher $K_d$ (lower affinity) than the true value. This error is linearly proportional to the error in concentration.

Thus, protein concentration errors due to batch impurities or degradation can manifest as a range of $K_d$ values in replicate measurements made from different batches of protein, all of which would be equal to or higher than the true $K_{dB}$ while simultaneously coming from high-quality, low-noise replicate fits. This exact phenomenon has also been demonstrated experimentally (31).

Evidence for protein concentration errors due to protein degradation or impurity—The original publications used His6-tagged recombinant SH2 domain protein production methods and used nickel chromatography as the sole protein purification method. In theory, these methods can provide purities of up to 95% (32). However, in practice, the results can vary significantly and can be affected by the amino acid content, nonspecific binding, purification conditions, and the type of affinity matrix used (32). Our experience in the laboratory performing...
New analysis pipeline improves SH2 affinity data

Before NFPI	After NFPI
Binder(2753)	Binder(2753)
Aggregator(2083)	Aggregator(1748)
Non-Binder(29778)	Non-Binder(16859)
Non-Functional(14070)
Low-SNR(2764)	Low-SNR(1948)

Figure 3. Initial replicate-level results and the results of nonfunctional protein identification (NFPI). The categorization results of individual domain–peptide measurements are shown (Before NFPI). Of the 37,378 measurements, 7.4% (2,753) were initially identified as positive interactions (binders), 7.4% (2,083) as interactions showing aggregation, 5.6% (2,764) as low SNR, and 79.7% (29,778) as nonbinders. The subsequent identification and removal of individual domain–peptide measurements made on nonfunctional protein had a significant effect on the categorization of nonpositive replicate-level measurements. Of the 29,778 measurements initially categorized as nonbinders, 56.6% (16,859) were identified as likely to contain nonfunctional protein and were removed from further consideration.

these purifications suggests that differences in purity between different protein preparations are likely to be present. Because the method used to determine protein concentration was absorption at 280 nm, only total protein content is measured, independent of purity or activity.

If the variance in affinity was from a random (nonsystemic) source, we would expect to find no patterns of variance in time. In contrast, if variance was from batch-related protein degradation or impurities, we might see alternating patterns in affinity over time as different batches are used. For example, if a high-purity protein sample were used on run 1 and a low-purity protein sample were used on run 2, we would expect consistently higher affinities on run 1 and consistently lower affinities on run 2. Or if a partially degraded protein sample was exhausted mid-run and replaced with a fresh sample, we could see a sudden surge of higher-affinity results in the middle of a run, when compared with other runs. Similar patterns could arise from batch to batch variations in purity affecting the accuracy of expected concentration.

We examined the data for evidence of these patterns. Because we do not have true information at the batch level or activity of each protein sample, these patterns must be inferred from the data. Although these patterns are difficult to spot due to the nature of the experimental design, we find examples of nonrandom run-dependent variations in affinity in the data (Fig. S11). These patterns are not compatible with a random source of variance and are compatible with either degradation or protein impurity causing errors in protein concentration.

Evidence for complete nonfunctionality of protein domains—Because we found patterns consistent with partial degradation, we examined the data for patterns of complete protein degradation. Complete degradation, or completely nonfunctional protein, would be indistinguishable from a nonbinding measurement for a single replicate, potentially resulting in a false negative. A control experiment to determine protein functionality would normally be required to delineate these two cases. However, we hypothesized that nonfunctional protein would manifest within the data as long runs of nonbinding results across many replicates but would demonstrate contradictory evidence of binding on other runs when the protein was not degraded. We found patterns consistent with nonfunctional protein (Fig. S12). Nonfunctional protein domains were identified and removed from consideration (see “Experimental procedures” and Figs. S13 and Fig. S14).

By removing replicates where there is evidence that the protein was nonfunctional, we avoid the potential for false negatives from these ambiguous data and greatly improve the pool of true-negative calls. Removal of nonfunctional protein has a significant impact on the numbers of measurements at the replicate level. Nonfunctional replicates made up 37.6% of all replicates (Fig. 3, right side). The large number of runs showing patterns of completely nonfunctional protein contributes to the overall evidence that protein degradation is present and is a source of variance in the data.

Method for handling replicates with high variance due to protein concentration errors: Reporting the minimum instead of the mean

Two key issues arise when considering how to handle replicate measurements when impurities or degradation are suspected to be a primary source of variance. First, without knowing the exact amount of protein concentration error in any one sample, how can this error be controlled for? Second, what is the correct procedure for handling replicates when variation is primarily due to concentration errors and not random sample variation? We propose a simple but novel solution to both questions: reporting the minimum rather than the mean of the replicate measurements results in the most accurate reported measurement.

Impurities and degradation can be partially controlled for by reporting the minimum replicate $K_d$. Given some unknown amount of protein concentration error due to degradation or impurities, the active concentration of protein will always be equal to or lower than the measured concentration. And as we demonstrated above, this means that the true affinity of the protein will always be equal to or greater than the measured affinity. Put in terms of $K_d$, the true $K_d$ will always be equal to or lower than the minimum measured $K_d$. Thus, the minimum $K_d$ reflects the closest measured value to the true affinity.

Furthermore, reporting the minimum measured $K_d$ also addresses the variance problem. If the measurements were true
replicates, reflecting random noise and experimental error, taking the mean of multiple replicates would be the appropriate procedure because the sample mean would represent the highest likelihood of the true population value of affinity. However, if the variation is caused by protein concentration errors, taking the mean of multiple measurements would not reflect the true

Figure 4. Replicate measurements exhibit high variance. Variance in affinity for domain–peptide interactions was visualized by distributed dot plots (top row) for examples of higher-affinity ($K_d \leq 1.0 \mu M$, top left) and lower-affinity ($K_d > 5.0 \mu M$, top right) interactions. Each row displays all replicate measurements for a single domain–peptide pair, and the x axis position of each individual replicate reflects the $K_d$ value of that measurement. Thus, variance can be visualized as the width of spread of points along each line. Domain–peptide interactions are sorted by minimum replicate $K_d$. The relationship between variance and affinity was also visualized for all domain–peptide interactions (bottom left, note the y axis log scale). The minimum replicate variance generally increases as $K_d$ increases (trend indicated by the red dashed line), but worst-case variance is independent of $K_d$ (blue dashed line), and high variance (e.g. $K_d > 10$) is present at all $K_d$ ranges. Variance was also quantified for different minimum $K_d$ ranges against different variance ranges (bottom right). In extremely low-variance cases (e.g. variance $< 0.01$), low $K_d$ measurements (blue bar) dominate. In moderate to high variance ranges (e.g. $K_d > 1$), the distributions are more similar. These two trends support the reasonable inference that higher $K_d$ fits have higher variance in general but also demonstrate that the presence of high-variance replicates is independent of affinity.
affinity. Rather, it would inadvertently increase the reported $K_d$ value by some unpredictable amount, which depends on the number of samples and the magnitude of their degradation. In addition, because the mean is particularly affected by outliers, even one severely degraded sample would significantly increase the mean reported $K_d$ value, resulting in a reported affinity with

Figure 5. Degradation causes a decrease in measured affinity. Simulated measurements for an ideal binding saturation experiment are shown for a theoretical protein with a $K_d$ of 1 $\mu$M (row 1). (Measurements in the second column are the same data as the first column but plotted as semi-log plots with a logarithmic concentration axis.) In rows 2 and 3, the results of 50% and 75% degradation are shown. To simulate the effect of degraded protein, we plotted true activity from the ideal curve (row 1) against the erroneous assumed concentration due to degradation (rows 2 and 3). For example, to simulate 50% degradation (row 2), the true FP response for 5 $\mu$M (from row 1) is plotted at the 10 $\mu$M position (on row 2). This procedure is repeated at each concentration. When affinity is derived from these degraded protein measurements, the result is an inaccurate $K_d$ higher than the true value. Although the concentration error from degraded protein causes a nonlinear change in FP, the error in $K_d$ is linear and proportional to the concentration error. For example, if the true active protein concentration is one-half of the assumed concentration (as in row 2), the measured affinity is one-half of the correct value (meaning the $K_d$ is 2 times the true value). If the true active protein concentration is one-fourth of the assumed concentration (as in row 3), the measured affinity is one-fourth of the correct value (the $K_d$ is 4 times the true value). Therefore, errors from overestimation of protein concentration always result in higher measured $K_d$ than the true value.

New analysis pipeline improves SH2 affinity data
high error. Therefore, odd though it may seem from a statistical perspective, taking the minimum $K_d$ is the most accurate way to handle variation in replicates where errors in protein concentration overestimation represent the primary source of variation.

**Revised affinity results and comparison with the original published results**

In the results from our revised analysis, 1518 positive (binding) interactions were identified, along with 7038 negative (nonbinding) interactions. These ~7000 true-negative results represent a significant increase in information from the original publication in which no true-negative interactions were reported. For 3200 interactions, inconclusive or problematic data were present, and no conclusions about their affinity could be drawn. Of those, 2753 potential do-
tions were reported. For 3200 interactions, inconclusive or
a summary of our revised results and the originally pub-
lished results are available in the supporting information, a s

We wanted to evaluate our revised analysis compared with
original results. In a case such as this, it is difficult to evaluate
because original samples are no longer available. However, one
way to evaluate the data is to use machine-learning methods to
ascertain whether the revised data have better internal consis-
tency or predictive power than the original data set. Lacking a
biological reference, it seemed fitting to evaluate these data
using machine learning, as we originally wished to harness SH2
domain–binding measurements in machine-learning frame-
works to extrapolate from the relatively small number of avail-
able measurements.

To do this, we implemented active search, a machine-learn-
ing approach that is highly amenable to biochemistry problems
such as this. Active learning (also known as optimal experimen-
tal design or active data acquisition) is a machine-learning par-
digm where we use available data to select the next best
experiments to maximize a specific objective. Active search is a
realization of this framework where the objective is to recover
as many members of a rare, valuable class as possible. In this
case, where only 13.9% of the original data set represents posi-
tive interactions between an SH2 domain and a phosphopep-
tide (or 18.2% in the revised data set), the objective of the search
algorithm was to prioritize each sequential selected interaction
to maximize the total number of positive interactions discov-
ered. We implemented the efficient nonmyopic search (ENS)
algorithm (33) with the goal of optimizing the total positive
experiments identified in an allocated search of 100 queries.
The algorithm was seeded randomly with one example positive
before the search progressed and was repeated 50 times.

ENS showed improved average performance and higher con-
sistency with our revised data set. First, ENS worked effectively
on both the original and revised data sets, identifying positives
that far exceed the expected number by random chance by the
100th query (Fig. 9). This suggests that phosphopeptide se-
quences do encode information about whether an SH2 domain
will recognize them in a binding interaction. Second, ENS per-
formance in the revised data set was higher than in the original
data set on average, finding 45.3 positives versus 33.3 positives ($p$
value of $4 \times 10^{-12}$). Third, ENS performance is significantly more
variable on the original data set than on the final data set (ranging
between 9 and 62 positives in 50 trials (with an average of 33.3),
compared with a range of 38–67 (with an average of 45.3 posi-
tives) for the revised data set. In the worst of the 50 trials, search
in the original data set underperformed by 50% compared with
what is expected by random chance, whereas the worst random
trial within the final data set still outperformed random chance
by 2-fold. Thus, the improved average performance and lower
variability in our revised results suggest improved coherency
in our revised analysis over the original published results.

**Discussion**

Here, we performed a revised analysis of raw data from
SH2 domain affinity experiments. We presented a new analy-
isis framework, which improved on the model fitting and eval-
uation methods of previous work. We report high-confidence
true positive interactions, add thousands of true-negative interactions, and remove false-negative results due to inactive protein. We also report the minimum replicate measurement instead of the mean—an appropriate approach when protein concentration overestimation errors are the largest source of variance in the data.

Figure 6. Revised analysis final results. A heat map shows the final results of the revised analysis. A significant fraction of measurements demonstrated patterns consistent with nonfunctional protein and were removed from the analysis. Comparison with the original published results can be seen in Figs. S14 and S15.
Although the numbers of positive interactions are similar in our revised analysis, the identities of those interactions have changed significantly. The changes in calls are visualized in the Sankey map above. Of the original 1,519 positive interactions found by the original authors, 166 (10.9%) were found to be nonbinders in our analysis. Of the 10,330 rejected interactions from the original publications, 273 (2.6%) positive interactions were recovered in our analysis.
New analysis pipeline improves SH2 affinity data

Although raw data from only two experiments were available for detailed analysis, we were fortunate that it consisted of a large quantity of measurements from FP, a well-established solution-based experimental system commonly used for analytical biochemical assays. All in vitro experimental methods have limitations when attempting to understand behavior in vivo, but early high-throughput experiments used arrays that had limitations and biases for higher-affinity interactions (13). Those experiments had either the peptide (11, 12) or the protein (7–10) mounted on a surface and are less preferable to a method where both molecules were measured in solution. So despite limited availability of raw

Figure 8. Correlation between original publication and revised analysis. Affinity values were compared for the common set of positive interactions (n = 1,245, top left panel) as well as at lower-affinity thresholds (other panels, as indicated). Our revised affinity values correlate only moderately with the original publication (Pearson r = 0.635), which might be surprising, considering that the analysis is on the same raw data. Our revised results correlate best when considering all measurements under 20 μM affinity (Pearson r = 0.734). Despite choosing the minimum measured value for $K_d$, our revised data often report higher $K_d$ results than the original publication (i.e. results below the diagonal). This is due to different categorization and filtering procedures, which result in significant additions and removal of individual measurements in each set of replicates for a domain–peptide pair. It is interesting to note that correlation does not improve at higher affinity (lower $K_d$), despite the fact that the chosen raw measurement range is tailored for highest accuracy for $K_d < 1.0 \mu M$. This suggests that the differences between our revised results are independent of the accuracy of the original measurements and more likely due to the need to correctly handle variation due to protein concentration errors.
data, the data available are likely to be the best type for further analysis.

We saw very high variance in affinity within replicate measurements in these data. On its face, such high variation suggests a significant problem with either experimental design or experimental method. At a minimum, it suggests that another (uncontrolled for) variable is being measured instead of the desired variable being tested. In the worst case, the remedy requires identifying and controlling for the source of variation and redoing the experimental measurements. Even the authors of the original publication argued that the “greatest source of variability in the FP assay...is batch-specific differences in protein functionality” (13). However, we have shown that the patterns found in the data are consistent with protein concentration errors and that the likely sources of error (purification and degradation) result in overestimation of protein concentration. Because these types of errors all result in unknown amounts of active protein concentration overestimation, reporting the minimum replicate $K_D$ for each domain peptide pair represents the value closest to the true activity of the protein.

In this analysis, we implemented a simple metric of quality, SNR, which weights the total fit error by the size of the signal. The SNR metric was effective at eliminating suspect fits while rejecting very few high-quality measurements (Fig. S9). Nevertheless, this metric may not be appropriate for all types of data, particularly data with a large prevalence of single-point outliers. We extensively explored using alternate metrics, including confidence intervals (CIs). Bootstrapped CIs, established by parametric bootstrapping via residual resampling, can add more information than a single-fit result because they provide a range of certainty for a given measurement. However, we found that this method had significant limitations on these data and performed worse than the SNR metric. In these data, bootstrapped CIs have even greater vulnerability to errors from outliers, are limited by small sample sizes (only 12 residuals per measurement), and suffer from heteroscedasticity of residuals (causing the high variability in low-concentration data points), ultimately resulting in unrealistic intervals for affinity.

Several analysis methods implemented in the original publications served as sources of randomizing error and may suggest a reason for the failure to agree with other published SH2 interaction experiments. First, background subtraction caused an unpredictable increase or decrease of affinity due to forced errors in model fitting. The magnitude of the error depended on whether the published background was higher or lower than optimum and on the affinity of the interaction being measured. Even small deviations could result in significant errors. Another seemingly innocuous choice—averaging multiple replicates containing degraded protein—is likely to be a significant source of error in the originally published results from this experiment. Taking the mean of multiple replicates is a standard practice, but it serves to randomize reported values when protein concentration overestimation is the primary source of variation.

Other high-throughput SH2 domain–peptide experiments share many critical methods with the data reviewed here. In all published experiments measuring affinity, protein was minimally filtered after production. The limited purification is likely to result in errors in protein concentration measurements due to inactive protein contaminants. Furthermore, in none of the experiments was protein assessed for activity before being measured. This has two critical consequences: the inability to separate nonbinding results from negative interactions due to nonfunctional protein and additional errors in active protein concentration with respect to the measured protein concentration. Even if protein concentration errors were solely due to purification, they could be the cause of the significant discrepancies between published numerical results. Furthermore, incorrect use of statistical methods to evaluate models was common to all published work—particularly the improper use of $r^2$ to determine the quality of fit of a nonlinear model and using only a single model to fit data. These choices result in a high false-negative rate and mask the high variance in replicates that our revised analysis revealed. Our results suggest that, if the raw data were available, some of these issues could be corrected in other experiments. However, due to the lack of correlation between any published high-throughput SH2 domain data and the likelihood that similar issues plague all similar data sets, we would recommend against the use of these previously published data sets in future research or models of SH2 domain behavior. We further recommend that all derivative work should be carefully reviewed for accuracy.

We want to address the best uses of the revised affinity results we present as well as the limits of the current analysis. The negative interactions we report represent a significant improvement over theoretical methods of simulating negative interactions (18), as they are based on real measurements rather than statistical assumptions. Furthermore, the negative interactions we report are controlled for false-negative results from nonfunctional protein—something no other SH2 domain data can claim. Thus, our revised results have significant potential to improve the quality of models built on categorical (binary) binding data. The limitation of the quantitative data we report is that the highest affinity measured value may not be...
New analysis pipeline improves SH2 affinity data

the true affinity if a fully functional protein was never measured. Nevertheless, the highest measured affinity still represents the measured value closest to the true value. However, not all variation in the data is consistent with our hypothesis of protein concentration error, and some variation may represent other unknown sources of variation that we have not controlled for. For example, one key assumption of the receptor occupancy model requires measuring the reaction at equilibrium. Because no data are provided to prove that the 20-min incubation time given to all samples was sufficient to bring all reactions to equilibrium, it is possible that some variation is due to measurements made in nonequilibrium conditions.

It is concerning that an entire body of published work has developed from this class of problematic results. These experiments have had a wide-reaching effect in many areas of SH2 domain research: the data have been used to draw specific conclusions about SH2 domain biology, such as identification of EGFR recruitment targets (34), to explain quantitative differences in RTK signaling (9), and as evidence to understand the promiscuity of EGFR tail binding (35). In addition, this work has been used to guide experimental design by filtering potential binding proteins by affinity (36), to reconcile confusing experimental results (37), and to guide new experimental hypotheses testing (38). It has played a role in cancer research as context to understand kinase dependences in cancer (39) and as evidence of HER3 and P13K connections as relevant to PTEN loss in cancer (40). It has influenced evolutionary analysis (41), been used to design mechanistic EGF models (42, 43), and has been used in algorithms for domain–binding predictions (14–18, 44).

Finally, we would like to discuss best practices for future data gathering and reporting. HTP studies have great value and provide a vast quantity of often never before measured data. These methods have been useful to a wide variety of domain–motif interactions (e.g. SH3–polyproline interactions (45, 46), PDZ domains interacting with C-terminal tails (47–49), and major histocompatibility complex (MHC) interactions with peptides (50, 51)). However, just as quickly, errors in these studies propagate rapidly and thereby into research results of other investigators. This suggests that an even higher than normal standard of care is necessary when evaluating such publications. A set of best practices for HTP methods should be established in the community. We recommend that all raw data from high-throughput experiments be published, along with all code used to process that data. This would make the initial data far more valuable for future research, much like the raw arrays stored in the Gene Expression Omnibus or the raw experimental measurements stored along with the protein structure in the Protein Data Bank. To this end, we have provided the original raw data and our full revised data (including intermediate steps) on Figshare and have provided the code for the analysis pipeline on Figshare and GitHub so that future evaluation can be more easily accomplished by other researchers (see “Data availability”). Although portions of our code are highly specific to the format of these data sets, the code is written in a modular fashion that can be easily repurposed in other studies. We also recommend that methods for quantifying activity should be a best practice in studies quantitatively measuring protein. Alternatively, methods that do not depend so heavily on accurate protein concentration should be preferred. One such concentration-independent method of measuring interaction affinity was recently developed by the Stormo laboratory (52). In that method, a two-color competitive fluorescence anisotropy assay measures the relative affinity of two interactions in solution. By measuring interaction against two peptides at once from the same pool of proteins, the concentration of the protein and the proportion of active protein is the same in both interactions. When the ratios are calculated, the concentration and activity drop from the calculation of affinity. Although this method only provides relative affinity, if one could carefully establish absolutely affinity for a single peptide (or panel of peptides), absolute affinity could be extended to all interactions. Another recent experiment also uses competitive fluorescence anisotropy but measures a competitive titration curve in a single well with an agarose gradient (53). Diffusion forms a spatiotemporal gradient for the interaction, so one can produce a full titration curve in each well in a multiwell plate, measuring both affinity and active protein concentration simultaneously. Regardless of the specific method, it should be a best practice to account for or control for the concentration of active protein within the measurement of total protein concentration.

Experimental procedures

Raw data

Upon receipt of the Jones 2012–2014 raw data, we examined the data for consistency and completeness. We found that the data did not cover all interactions described in the original publication. However, by limiting our revised analysis to interactions of single SH2 domains with phosphopeptides from the ErbB family, as well as KIT, MET, and GAB1, we were able to limit the effect of missing raw data. Within this scope, only a handful of individual replicate interactions were then missing (~138 replicate-level measurements of over 37,000 measurements) and were limited to three domain–peptide pairs. Fortunately, two of the domain–peptide pairs were represented by other replicate measurements. The data we examined for this revised analysis cover the interactions of 84 SH2 domains with 184 phosphopeptides. The peptides came from receptor proteins from the four ErbB domains (EGFR/ErbB1, HER2/ErbB2, ErbB3, and ErbB4) as well as KIT, MET, and GAB1. Of SH2 proteins containing a single SH2 domain, 66 domains were measured: ABL1, ABL2, BCAR3, BLK, BLNK, BMX, BTK, CRK, CRKL, DAPPI, FER, FES, FGR, GRAP2, GRB2, GRB7, GRB10, GRB14, HCK, HSH2D, INPP1, ITK, LCK, LCP2, LYN, MATK, NCK1, NCK2, PTK6, SH2B1, SH2B2, SH2B3, SH2D1A, SH2D1B, SH2D2A, SH2D3A, SH2D3C, SH3BP2, SHB, SHC1, SHC2, SHC3, SHC4, SHD, SHE, SHF, SLA, SLA2, SOCS1, SOCS2, SOCS3, SOCS5, SRC, STAP1, SUPT6H, TEC, TENC1, TNS1, TNS3, TNS4, TXK, VAV1, VAV2, VAV3, and YES1. From SH2 proteins with double domains, C-terminal and N-terminal domains were individually measured from 10 proteins: PIK3R1, PIK3R2, PIK3R3, PLCG1, PTPN11, RASA1, SYK, ZAP70, PLCG2 (N terminus only), and PTPN6 (C terminus only). One peptide had no measurements in the raw data (EGFR pY944). Within this revised
Model fitting, model selection, and replicate-level calls

For each replicate measurement, we fit two models: the linear model (Equation 3) and the receptor occupancy model (Equation 2). Model fits were evaluated with the AICc, and the model with the lower AICc score was selected (19). The AIC as a quality metric, was calculated as follows,\[ AIC = 2p - 2\ln(L) \] (Eq. 5)
where \( p \) is the number of parameters in the model, and \( \ln(L) \) is the maximum log-likelihood of the model. In a nonlinear fit, with normally distributed errors, \( \ln(L) \) is calculated as follows,
\[ \ln(L) = -0.5N \left( \ln(2\pi) + 1 - \ln(N) + \ln \left( \sum_{i=1}^{n} x_i^2 \right) \right) \] (Eq. 6)
where \( x_1, \ldots, x_n \) are the residuals from the nonlinear least squares fit, and \( N \) is the number of residuals. The bias-corrected form of AIC, referred to as AICc, is a variant that corrects for small sample sizes (e.g. when one has fewer than 30 data points). AICc is calculated as follows,
\[ AICc = AIC + \frac{2p(p+1)}{n-p-1} \] (Eq. 7)
where \( n \) is the sample size, and \( p \) is the number of parameters in the model (19). Each replicate had a sample size of 12. The receptor occupancy model had three parameters (affinity (\( K_d \)), saturation level (\( F_{max} \)), and offset (\( F_0 \)), whereas the linear model had two parameters (slope (\( m \)) and background offset (\( F_0 \)).

Replicates that were fit best by the linear model with a slope of \( \leq 5 \) mP/\( \mu \)m were categorized as negative interactions, or “nonbinders.” Linear fits with a slope of \( \geq 5 \) mP/\( \mu \)m were categorized as aggregators. Replicates that were fit best by the receptor occupancy model were subsequently evaluated for SNR (Equation 4). If the SNR was \( \geq 1 \), the replicate was categorized as a positive interaction or “binder”; otherwise, it was rejected as a low-SNR fit and removed from consideration.

Identifying nonfunctional protein

Once all individual fits were complete, runs were examined for nonfunctional protein. If an entire run lacked even one positive binding interaction and those same interactions measured positive on another run, the nonbinder, aggregator, and low-SNR calls on that run were changed to nonfunctional protein and removed from consideration.

Replicate handling for domain–peptide measurements

For each domain–peptide pair, only replicates that were marked as binders with sufficiently high SNR were considered. For a given domain–peptide pair, the minimum numeric value of \( K_d \) (representing the strongest affinity) was reported as the final \( K_d \) for that domain peptide pair.

Active search

The probability model (33) used a simple \( k \)-nearest neighbor (\( k = 20 \)), where distance is defined by average Euclidean distance of corresponding divided physicochemical property score (DPPS) features of the amino acids (54) comprising the peptide, as in the following,
\[ d_{mn}(x,x') = \frac{1}{n} \sum_{i=1}^{n} d_e(dpps(x_i),dpps(x_i')) \] (Eq. 8)
where \( d_{mn} \) is the distance used to define nearest neighbors, \( d_e \) is the Euclidean distance, \( n \) is the number of amino acids in the peptide (here \( n = 9 \)), and \( dpps(x_i) \) is the DPPS feature vector of the \( i \)th amino acid in peptide \( x \).

Data availability

A summary of our revised results and the originally published results are available in the supporting information, as an Excel file. The complete raw data and data from each stage of the revised results are available on Figshare (doi: 10.6084/m9.figshare.11482686.v1). The code for the pipeline used to analyze the data can be found on GitHub and is archived on Figshare (https://github.com/NaegleLab/SH2fp, doi: 10.6084/m9.figshare.12326609.v1).

Acknowledgments—We thank Richard Jones, Ron Hause, and Ken Leung for providing the raw data required for this analysis.

Author contributions—T. R. and K. M. N. conceptualization; T. R. resources; T. R. data curation; T. R. software; T. R., R. G., and K. M. N. formal analysis; T. R. validation; T. R. investigation; T. R. and K. M. N. visualization; T. R. and R. G. methodology; T. R. and K. M. N. writing–original draft; T. R. and K. M. N. writing–review and editing; K. M. N. supervision; K. M. N. funding acquisition; K. M. N. project administration.

Funding and additional information—The work was supported in part by the Center for Biological Systems Engineering (to T. R).

Conflict of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: SH2, Src homology 2; SH3, Src homology 3; HTP, high-throughput; pTyr, phosphorysine-containing peptide motif(s); FP, fluorescence polarization; PM, protein microarray; mP, millipolarization units; SNR, signal/noise ratio; AIC, Akaike’s information criteria; AICc, corrected form of AIC; M, millipolarization units; SNR, signal/noise ratio; dpp, distance between peptide and protein.
New analysis pipeline improves SH2 affinity data

References

1. Yarden, Y., and Sliwkowski, M. X. (2001) Untangling the ErbB signalling network. Nat. Rev. Mol. Cell Biol. 2, 127–137 CrossRef Medline

2. Machida, K., and Mayer, B. J. (2005) The SH2 domain: versatile signaling module and pharmaceutical target. Biochim. Biophys. Acta 1747, 1–25 CrossRef Medline

3. Pawson, T. (2004) Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. Cell 116, 191–203 CrossRef Medline

4. Zhou, S., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Hayer, W. G., King, F., Roberts, T., Ratnofsky, S., Lechleider, R. J., Neel, B. G., Birge, R. B., Fajardo, J. E., Chou, M. M., Hanafusa, H., et al. (1993) SH2 domains recognize specific phosphopeptide sequences. Cell 72, 767–778 CrossRef Medline

5. Pawson, T., and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300, 445–452 CrossRef Medline

6. Matlock, M. K., Holehouse, A. S., and Naegle, K. M. (2015) ProteomeScout: a repository and analysis resource for post-translational modifications and proteins. Nucleic Acids Res. 43, D521–D530 CrossRef Medline

7. Jones, R. B., Gordus, A., Krall, J. A., and MacBeath, G. (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature 439, 168–174 CrossRef Medline

8. Kaushansky, A., Gordus, A., Budnik, B. A., Lane, W. S., Rush, J., and MacBeath, G. (2008) System-wide investigation of ErbB4 reveals 19 sites of Tyr phosphorylation that are unusually selective in their recruitment properties. Chem. Biol. 15, 808–817 CrossRef Medline

9. Gordus, A., Krall, J. A., Beyer, E. M., Kaushansky, A., Wolf-Yadlin, A., Sevka, M., Chung, B. H., Rush, J., and MacBeath, G. (2009) Linear combinations of docking affinities explain quantitative differences in RTK signaling. Mol. Syst. Biol. 5, 235 CrossRef Medline

10. Koytiger, G., Kaushansky, A., Gordus, A., Rush, J., Sorger, P. K., and MacBeath, G. (2012) Comprehensive binary interaction mapping of SH2 domains via fluorescence polarization reveals novel functional diversifications of docking affinities explain quantitative differences in RTK signaling. Mol. Syst. Biol. 5, 235 CrossRef Medline

11. Liu, B. A., Jablonowski, K., Shah, E. E., Engelmann, B. W., Jones, R. B., and Nash, P. D. (2010) SH2 domains recognize contextual peptide sequence information to determine selectivity. Mol. Cell. Proteomics 9, 2391–2404 CrossRef Medline

12. Tinti, M., Kiemer, L., Costa, S., Miller, M. L., Sacco, F., Olsen, J. V., Carducci, M., Paoluzzi, S., Langone, F., Workman, C. T., Blom, N., Machida, K., Thompson, C. M., Schutkowski, M., Brunak, S., et al. (2013) The SH2 domain interaction landscape. Cell Rep. 3, 1293–1305 CrossRef Medline

13. Hause, R. J., Leung, K. K., Barking, J. L., Ciaccio, M. F., Pin Chuu, C., and Jones, R. B. (2012) Comprehensive binary interaction mapping of SH2 domains via fluorescence polarization reveals novel functional diversification of ErbB receptors. PLoS ONE 7, e44741 CrossRef Medline

14. Leung, K. K., Hause, R. J., Barking, J. L., Ciaccio, M. F., Chuu, C.-P., and Jones, R. B. (2014) Enhanced prediction of Src homology 2 (SH2) domain binding potentials using a fluorescence polarization-derived c-Met, c-KIT, ErbB, and androgen receptor interactome. Mol. Cell. Proteomics 13, 1705–1723 CrossRef Medline

15. Sánchez, I. E., Beltrao, P., Stricher, F., Schymkowitz, J., Feringhoff-Borg, J., Rousseau, F., and Serrano, L. (2008) Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. PLoS Comput. Biol. 4, e1000052 CrossRef Medline

16. AlQuraishi, M., Koytiger, G., Jenney, A., MacBeath, G., and Sorger, P. K. (2014) A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. Nat. Genet. 46, 1363–1371 CrossRef Medline

17. Wunderlich, Z., and Mirny, L. A. (2009) Using genome-wide measurements for computational prediction of SH2-peptide interactions. Nucleic Acids Res. 37, 4629–4641 CrossRef Medline

18. Kondou, K., Cost, F., Huber, M., Reth, M., and Backoffen, R. (2013) Semi-supervised prediction of SH2-peptide interactions from imbalanced high-throughput data. PLoS ONE 8, e62732 CrossRef Medline

19. Spiess, A.-N., and Neumeyer, N. (2010) An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC Pharmacol. 10, 6 CrossRef Medline

20. Kvalseth, T. O. (1985) Cautionary note about R-squared. Am. Stat. 39, 279–285 CrossRef

21. Juliano, S. A., and Williams, F. M. (1987) A comparison of methods for estimating the functional response parameters of the random predator equation. J. Anim. Ecol. 56, 641–653 CrossRef

22. Magee, L. (1990) R2 measures based on Wald and likelihood ratio joint significance tests. Am. Stat. 44, 250–253 CrossRef

23. Nagelkerke, N. J. D. (1991) A note on a general definition of the coefficient of determination. Biometrika 78, 691–692 CrossRef

24. Anderson-Sprecher, R. (1994) Model comparisons and R2. Am. Stat. 48, 113–117 CrossRef

25. Willett, J. B., and Singer, J. D. (1988) Another cautionary note about R2: its use in weighted least-squares regression analysis. Am. Stat. 42, 236–238 CrossRef

26. Miao, S.-P., Lu, A., and Lum, H. (2007) Pitfalls of using R2 to evaluate goodness of fit of accident prediction models. Transp. Res. Rec. J. Transp. Res. Board 1542, 6–13 CrossRef

27. Kaushansky, A., Gordus, A., Chang, B., Rush, J., and Macbeath, G. (2008) A quantitative study of the recruitment potential of all intracellular tyrosine residues on EGFR, FGFR1 and IGF1R. Mol. Biosyst. 4, 643–653 CrossRef Medline

28. Kaushansky, A., Allen, J. E., Gordus, A., Stiffler, M. A., Karp, E. S., Chang, B. H., and Macbeath, G. (2010) Quantifying protein-protein interactions in high throughput using protein domain microarrays. Nat. Protoc. 5, 773–790 CrossRef Medline

29. Clark, A. J. (1926) The reaction between acetyl choline and muscle cells. J. Physiol. 61, 530–546 CrossRef Medline

30. Mazeroille, M. J. (2007) Appendix 1: Making Sense Out of Akaike’s Information Criterion (AIC): Its Use and Interpretation in Model Selection and Inference from Ecological Data. Ph.D. thesis, pp. 1–13, Université Laval, Quebec, Canada

31. Pol, E. (2010) The importance of correct protein concentration for kinetics and affinity determination in structure-function analysis. J. Vis. Exp. 17, 1746 CrossRef Medline

32. Bornhorst, J. A., and Falke, J. J. (2000) Purification of proteins using polyhistidine affinity tags. Methods Enzymol. 326, 245–254 CrossRef Medline

33. Jiang, S., Malkomes, G., Converse, G., Shofner, A., Moseley, B., and Gar nett, R. (2017) in Proceedings of the 34th International Conference on Machine Learning 70, 1714–1723

34. Tsai, C. J., and Nussinov, R. (2019) Emerging allosteric mechanism of EGFR activation in physiological and pathological contexts. Biophys. J. 117, 5–13 CrossRef Medline

35. Kennedy, S. P., Hastings, J. F., Han, J. Z. R., and Croucher, D. R. (2016) The under-appreciated promiscuity of the epidermal growth factor receptor family. Front. Cell Dev. Biol. 4, CrossRef Medline

36. Birtwistle, M. R. (2015) Analytical reduction of combinatorial complexity arising from multiple protein modification sites. J. R. Soc. Interface 12, 20141215 CrossRef

37. Leong, S. H., Iwin, K. M., Lee, S. S., Ng, W. H., Ng, K. M., Tan, S. Y., Ng, B. L., Carter, N. P., Tang, C., and Lian Kon, O. (2017) Chromosomal breaks at FRA18C: association with reduced DOK6 expression, altered oncogenic signaling independent of HER3. Cancer Res. 78, 3645–3658 CrossRef Medline

38. Ruiz-Saenz, A., Dreyer, C., Campbell, M. R., Steri, V., Gulizia, N., and Moasser, M. M. (2018) HER2 amplification in tumors activates PI3K/Akt signaling independent of HER3. Cancer Res. 78, 3645–3658 CrossRef Medline
New analysis pipeline improves SH2 affinity data

39. Campbell, J., Ryan, C. J., Brough, R., Bajrami, I., Pemberton, H. N., Chong, I. Y., Costa-Cabral, S., Frankum, J., Gulati, A., Holme, H., Miller, R., Postel-Vinay, S., Rafiq, R., Wei, W., Williamson, C. T., et al. (2016) Large-scale profiling of kinase dependencies in cancer cell lines. Cell Rep. 14, 2490–2501 CrossRef Medline

40. Stern, H. M., Gardner, H., Burzykowski, T., Elatre, W., O’Brien, C., Lackner, M. R., Pestano, G. A., Santiago, A., Villalobos, I., Eiermann, W., Pienkowski, T., Martin, M., Robert, N., Crown, J., Nuciforo, P., et al. (2015) PTEN loss is associated with worse outcome in HER2-amplified breast cancer patients but is not associated with trastuzumab resistance. Clin. Cancer Res. 21, 2065–2074 CrossRef Medline

41. Miller, M. L., Jensen, L. J., Diella, F., Jørgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhoffsky, M., Pascaulescu, A., Alexander, J., Knapp, S., Blom, N., et al. (2008) Linear motif atlas for phosphorylation-dependent signaling. Sci. Signal. 1, ra2 CrossRef Medline

42. Sites, E. C., Aziz, M., Creamer, M. S., Von Hoff, D. D., Posner, R. G., and Hlavacek, W. S. (2015) Use of mechanistic models to integrate and analyze multiple proteomic datasets. Biophys. J. 108, 1819–1829 CrossRef Medline

43. Jadwin, J. A., Curran, T. G., Lafontaine, A. T., White, F. M., and Mayer, B. J. (2018) Src homology 2 domains enhance tyrosine phosphorylation in vivo by protecting binding sites in their target proteins from dephosphorylation. J. Biol. Chem. 293, 623–637 CrossRef Medline

44. Gong, W., Zhou, D., Ren, Y., Wang, Y., Zuo, Z., Shen, Y., Xiao, F., Zhu, Q., Hong, A., Zhou, X., Gao, X., and Li, T. (2008) PepCyber: P—PEP: a database of human protein-protein interactions mediated by phosphoprotein-binding domains. Nucleic Acids Res. 36, D679–D683 CrossRef Medline

45. Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L., Schneider-Mergener, J., Volkmer-Engert, R., and Cesareni, G. (2004) Protein interaction networks by proteome peptide scanning. PLoS Biol. 2, e14 CrossRef Medline

46. Carducci, M., Perfetto, L., Briganti, L., Paoluzi, S., Costa, S., Zerweck, J., Schutkowski, M., Castagnoli, L., and Cesareni, G. (2012) The protein interaction network mediated by human SH3 domains. Biotechnol. Adv. 30, 4–15 CrossRef Medline

47. Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A., and MacBeath, G. (2008) Predicting PDZ domain-peptide interactions from primary sequences. Nat. Biotechnol. 26, 1041–1045 CrossRef Medline

48. Boisguerin, P., Leben, R., Ay, B., Radziwill, G., Moelling, K., Dong, L., and Volkmer-Engert, R. (2004) An improved method for the synthesis of cellulo-se membrane-bound peptides with free C termini is useful for PDZ domain binding studies. Chem. Biol. 11, 449–459 CrossRef Medline

49. Wiedemann, U., Boisguerin, P., Leben, R., Leitner, D., Krause, G., Moelling, K., Volkmer-Engert, R., and Oschkinat, H. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J. Mol. Biol. 343, 703–718 CrossRef Medline

50. Haj, A. K., Breitbart, M. E., Baker, D. A., Mohns, M. S., Moreno, G. K., Wilson, N. A., Lyamichev, V., Patel, J., Wei, C., Dudley, D. M., and O’Connor, D. H. (2020) High-throughput identification of MHC class I binding peptides using an ultradense peptide array. J. Immunol. 204, 1689–1696 CrossRef Medline

51. Gasteiwi, S., and Maeurer, M. J. (2009) Identification of MHC class II binding peptides: microarray and soluble MHC class II molecules. Methods Mol. Biol. 524, 417–426 CrossRef Medline

52. Zuo, Z., Roy, B., Chang, Y. K., Granas, D., and Stormo, G. D. (2017) Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. Sci. Adv. 3, eaao1799 CrossRef Medline

53. Jung, C., Schnepf, M., Bandilla, P., Unnerstall, U., and Gaul, U. (2019) High sensitivity measurement of transcription factor–DNA binding affinities by competitive titration using fluorescence microscopy. J. Vis. Exp. CrossRef Medline

54. Tian, F., Yang, L., Lv, F., Yang, Q., and Zhou, P. (2009) In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach. Amino Acids 36, 535–554 CrossRef Medline