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

Tom Ronan[1], Roman Garnett[2], and Kristen Naegle[1,3]

From the [1]Department of Biomedical Engineering, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130; [2]Department of Computer Science and Engineering, Washington University in St. Louis, One Brookings Drive, CB 1045 St. Louis, MO 63130; [3]Current Affiliation: Department of Biomedical Engineering and the Center for Public Health Genomics, University of Virginia, Box 800759, Charlottesville, VA 22903

Running title: New analysis pipeline improves SH2 affinity data

*To whom correspondence should be addressed: Kristen Naegle: kmn4mj@virginia.edu

Keywords: cell signaling, phosphotyrosine signaling, phosphotyrosine, Src homology 2 domain (SH2 domain), mathematical modeling, peptide interaction, epidermal growth factor receptor (EGFR), affinity, high-throughput, best practices

ABSTRACT

Protein domain interactions with short linear peptides, such as Src homology 2 (SH2) domain interactions with phosphotyrosine-containing peptide motifs (pTyr), are ubiquitous and important to many biochemical processes of the cell – both in their central importance to cell physiology, and to the sheer scale of possible interactions. The desire to map and quantify these interactions has resulted in the development of increased throughput quantitative measurement techniques, such as microarray or plate-based fluorescence polarization assays. For example, in the last 15 years, experiments have progressed from measuring single interactions to having 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 datasets led us to re-evaluate the analysis pipelines of published SH2-pTyr datasets. We identified several opportunities for improving the identification of positive and negative interactions, and the accuracy of affinity measurements. These methods account for protein aggregation and degradation, and use model fitting and evaluation that are more appropriate for the non-linear behavior of binding interaction data. In addition to improve affinity accuracy, and increased certainty in negative interactions, we find the reanalyzed data results in significantly improved classification of binding vs non-binding when using machine learning techniques, suggesting improved coherence in the reanalyzed datasets. In addition to providing the revised dataset, we propose this new analysis pipeline and necessary protein activity controls should be part of the design process of many such high-throughput biochemical measurements.

Introduction

Protein domain interactions with short linear peptides are found in many biochemical processes of the cell, and represent a vast number of potential interactions. They 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", since they read the presence of tyrosine phosphorylation by binding specifically to certain phosphorylated amino acid sequences. These domains are typically 100 amino acids long and fold into a conserved structure consisting of two α-helices and seven β-strands. At its binding core, an invariant arginine creates a salt bridge with the ligand pTyr yielding approximately half of the binding energy of the SH2-pTyr sequence interaction. Early degenerate library screens demonstrated that 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 the specificity of SH2 domain interactions within pTyr-mediated signaling (*5*). Understanding SH2 domain specificity and binding affinities with cognate ligands would greatly aid in our understanding of cell signaling networks that control human physiology. However, the total interaction space is immense – the 46,000 tyrosines currently known to be phosphorylated in the human proteome (*6*), present over 5.5 million possible SH2-pTyr sequence interactions.

Recent developments have expanded the measurement coverage of human SH2 domains with specific pTyr-containing peptides. Specifically, eight high-throughput affinity studies have been performed, using either microarrays or fluorescence polarization to measure SH2 domain interactions with specific phosphopeptide sequences (*7–14*). Although the six studies that measured affinity represent roughly 90,000 pairs of domain-peptide interactions, these measurements cover only 2% of the possible interaction space. In response, computational approaches have used published datasets to extend from the measured space into unmeasured spaces by a variety of prediction methods. These methods span the range from thermodynamic models using existing structure and binding measurements to predict interaction strength (*15–17*) to supervised machine learning models using patterns in peptide sequences and quantitative binding data to predict binding to particular domains (*14, 18*). However, no computational method has used the available affinity data in its entirety. We therefore wished to leverage all available binding affinity measurement data in a supervised learning approach to expand our knowledge of SH2-pTyr interaction space.

Unfortunately, in the process of reviewing published high-throughput data, we noticed several inconsistencies with the published results and potential problems with the methods used to produce the data sets. We found surprising disagreement between published data sets. The published data failed to agree on the identity of which domain-peptide pairs interacted, and on the small subset on which they *do* agree, they reported vastly different affinities. We hypothesized two potential causes for this variation. First, we identified potential inaccuracies in protein concentration common to all three data sets that had the potential to directly affect the published affinity values. Second, we found errors in model fitting and the statistical methods used to evaluate model fitting, which could have significant impact on the reported affinities.

In reviewing the protein preparation protocols for each experiment, we found that all of the affinity studies failed to use positive controls to determine if protein was functional before measuring affinity. Furthermore, protein was minimally purified (via nickel chromatography only), and the resulting protein concentration was measured by absorbance. Thus protein of varying degrees of purity and non-monomeric content were used for affinity measurements. Without positive protein controls, it is difficult to determine if non-interaction is due to inactive protein or true failure to interact. And testing non-monomeric protein risks violating the one-to-one assumptions of the receptor occupancy model used to calculate affinity. Errors in effective protein concentration deriving from inactive or degraded protein would result in concentration values different than the amount of active protein in the sample. These concentration errors would propagate directly to errors in the derived affinity values, as affinity values are a function of concentration.

Furthermore, all of the affinity studies used the coefficient of determination (r^2) as a determination of how well the model fits the data. In these studies, any interaction not meeting an r^2 value threshold of approximately 0.90 or 0.95 was rejected from further analysis. Unfortunately, r^2 has been conclusively shown to be a poor indicator of fitness for non-linear models (like the non-linear receptor occupancy model used in each of these studies to derive affinity) and can produce misleading results (*19*). Although this fact has long been established in the statistical literature (*20–26*) r^2 is still commonly used to evaluate non-linear models in pharmaceutical and biomedical publications despite being an ineffective and misleading metric. For *linear* data, one can interpret the values of r^2 between 0 and 1 as the total percent of variance explained by the fit. However, when applied to *non-linear* data, the r^2 value cannot be interpreted as the percent of variance and is known to fail to reflect significantly better fitting models (*19*).

Furthermore, as applied here, it effectively resulted in a bias for identification of true positive interactions at the expense of making many false negative calls.

Therefore, we had serious concerns about using the published data for use in machine learning, due to both inaccuracies in quantitative results, and the significant potential for large numbers of false negative results. Given these limitations in the published data, we endeavored to retrieve and reanalyze any raw data we could acquire in order to systematically improve SH2-phosphopeptide affinity measurement accuracy. To accomplish this: 1) we refined model fitting techniques, 2) implemented multiple models for each measurement, 3) used a statistically accurate method for model selection, 4) developed methods to identify and remove non-functional protein from the results, and 5) introduced a simple method to handle the effects of degraded protein on affinity measurements. 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 negatives. Evaluation of the revised dataset shows improved learning accuracy within an active learning model – suggesting that there is improved coherency in the features of the revised dataset. We propose this new analysis framework for improving the accuracy of high-throughput affinity domain-peptide interaction measurements, and ultimately suggest ways to improve future experiments via better experimental design.

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 high-throughput methods to measure SH2 domain interactions with specific peptide sequences, including peptide microarrays, peptide arrays, and fluorescence polarization methods. 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 lab from 2006 to 2009 (*7*, *9*, *27*) which contain a body of predominantly non-overlapping protein microarray experiments. The second data group consists of a large study published by the MacBeath lab in 2013 (*10*) with a set of new protein microarray measurements using the protocol published in 2010 (*28*). The third data group consists of two non-overlapping sets of fluorescence polarization data published in 2012 and 2014 by the Jones lab (*13*, *14*). Because the other array experiments (*11*, *12*) only measured interaction and not affinity, they were not considered for this analysis.

In order to determine how well the data groups agreed on affinity measurements, we examined the correlation between domain-peptide affinity measurements which overlapped between any two data groups (Fig. 1). We found surprisingly low correlation between affinity measurements (with a maximum correlation of $r = 0.377$). Next we asked if the different data groups identified the same positive interactions between domain-peptide pairs, even if they did not agree on the affinity measurements. We compared the identities of positive interactions measured in any of the three data groups. Here, we also found significant disagreement over which domain-peptide pairs were found to interact (Fig. 1). There were 347 positive domain-peptide interactions identified by at least one group, but less than 16% of those interactions were found to be positive 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 in the differences between any particular data group (Fig. S1).

We then considered which factors of protein preparation, peptide preparation, or experimental technology difference could have resulted in such different results. Although there are significant differences between the techniques of protein microarrays (which immobilize the SH2 proteins on the microarray and wash fluorophore-labeled peptides over the arrays) and fluorescence polarization (where both the SH2 protein domains and peptides are in solution), the differences between positive interactors did not group by technology type. The MacBeath 2013 data group (which used protein microarrays) had almost the

same size of positive interaction overlap with the MacBeath 2006-09 (also using protein microarrays) as with the Jones 2012-14 data group (which used fluorescence polarization methods) (Fig. 1). In terms of experimental and analytical methods all three data groups: 1) used recombinant SH2 domain protein production, added a His_6 tag, and used nickel chromatography as the sole protein purification method; 2) dialyzed purified protein into a buffer and added glycerol, though different buffers were used in different publications; 3) used absorbance at 280nM to determine protein concentration, though one group (*10*) measured protein concentration with a protocol (*28*) using denaturing conditions; 4) used solid phase synthesized peptides purified with reverse phase HPLC; and 5) used the receptor occupancy model, and similar methods of evaluating model fits based on the coefficient of determination (r^2). Without more detailed analysis, it would be impossible to determine which of these factors, if any, are responsible for the differences in reported results.

These findings demonstrate significant quantitative and qualitative differences between published data from different labs, and even disagreements between early results and late results published from the same lab. We concluded that we could not identify the source of differences between published data sets, or even evaluate the quality of any single set of published data, without looking further into the raw data. Acquisition of raw data from published studies was surprisingly difficult. Upon review of the affinity publications, we discovered that no publication contained raw data. Rather, publications contained 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) However, we were able to retrieve raw data from the Jones 2012-14 data group, thanks to assistance from the authors (personal communication from Richard Jones, Ron Hause, and Ken Leung).

Raw SH2 interaction data and revised analysis

We proceeded to examine the raw data from the Jones 2012-14 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 was missing in comparison to the original publication, by limiting our revised analysis to interactions of single SH2 domains with phosphopeptides from the ErbB family (EGFR, ERBB2, ERBB3, ERBB4), as well as KIT, MET, and GAB1, the available raw data covered approximately 99.6% of the reported measurements.

Evaluation of the Original 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 was then used to interpret an equilibrium dissociation rate constant (K_d) according to the receptor occupancy model, developed by Clark in 1926 and derived from the law of mass action (*29*). As applied to the fluorescence polarization data, the model takes the form:

$$F_{obs} = \frac{[SH2\ domain]F_{max}}{K_d + [SH2\ domain]} \qquad (1)$$

where F_{obs} is the observed fluorescence polarization (FP) 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 (Fmax) are fitted parameters of the model. It is important to note that this model is dependent on several critical assumptions: that the 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 fitted parameter results. We examined three aspects of the implementation of this model. First, we examined the effect of subtracting background fluorescence on model fitting and explored alternatives that introduce less bias. Second, we reviewed whether dropping outlier measurements, as used in the original publications, affected model

fitting results. Third, we asked whether the receptor occupancy model could reliably fit a non-binding sample, and examined failure modes when we found it did not.

The Effect of Background Fluorescence on Model Fitting. 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. We hypothesized that the setting of the background or "zero" polarization value (zero-signal) would affect model fitting results, because a critical assumption of the model is that the saturation curve passes through the origin (the point of zero-signal, which is also the point of zero-concentration). Because the background subtraction method results in zero-signal at a point other than zero-concentration – violating the assumptions of the model – we examined the effects of this method of background subtraction on model fitting.

In examining many measurements, we initially found that the background value was often uncorrelated with the signal values. In some cases, a strong signal with low internal noise was present, but the background value was high – far above much of the seemingly reliable signal (Fig. S3, top row). In these cases, the high quality of the data seems to contradict the limits imposed by the seemingly high background, below which measurements should be random noise. In other cases, the background level was far below the signal (Fig. S3, middle row). Subtracting a high background value would drive some FP values negative, which cannot be accommodated by the model. Subtracting a low background value forces the zero-signal point to be far below the data and causes reproducible and systematic errors in affinity parameter fitting when the curve is forced to pass through the origin. Treating the minimum measured FP value as the zero signal value can also induce unforeseen results in fit parameters, as the first data point is not always the minimum due to random noise in the data. The shortcoming of the subtraction methods is that the curve is forced through a point without using the high-quality information contained in the data forming the saturation curve.

In contrast, we chose a method in which the origin was set at a point that was extrapolated from the saturation curve data itself, instead of from the reported background values. This was accomplished by adding an offset value (F_{bg}) and fitting both the curve and the offset/origin at the same time:

$$F_{obs} = \frac{[SH2\ domain]F_{max}}{K_d + [SH2\ domain]} + F_{bg} \qquad (2)$$

(where F_{bg} represents a fluorescence background offset value). This resulted in the fewest artificial constraints on the data and high-quality fits independent of artifacts from background subtraction.

Outlier Removal Biases Model Fitting. 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 identified as outliers using a statistical model were removed iteratively and the fit was reevaluated. Up to three points were iteratively removed per measurement. For measurements where more than three data points were identified as outliers, the measurement was removed from further consideration. The approach of dropping outliers is a commonly used tool to reduce the impact of noise on a model, yet it represents a tradeoff. By using fewer data points, less of the original data is available. Furthermore, outlier identification relies on assumptions about how measured data is expected to fit a statistical model, which may not be correct. We wished to determine the impact of outlier removal on the interpretation of the raw data.

In order to determine the effect of removing a data point, we evaluated the number of data points and concentration range of those data points for suitability to the measurements attempted. The original data consisted of protein with an initial concentration of either 10µM or 5µM, and 11 serial dilutions of the protein (for a total of 12 data points), with each dilution representing a further reduction to one-half concentration. Thus the range of concentrations spanned by each measurement was either 2.4nM to 5µM or 4.9nM to 10µM. 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 (Fig. S4, second column). Given the concentrations measured, it can be seen that the experiment is designed to most accurately identify proteins with affinity (K_d) in the range of 0.05 µM to 0.5µM. However, the original publication reported values as high as 20 µM, which is more than 4-fold above the 0.5 µM limit of the highest accuracy measurement range. For interactions with a K_d of 1 µM, the upper plateau of the semi-log saturation curve no longer has any coverage from the data (Fig. S4, row 2), and interactions with K_d values higher than 5µM have few or no data points even above K_d (Fig. S4, rows 3 and 4), which significantly increases potential inaccuracies in model fitting. This suggests that every data point is critical for accuracy, particularly points above K_d. In practice, we found many cases where removal of a single data point had a large impact on the resulting fitted affinity parameter. In contrast, we found few examples where a single, obvious outlier prevented a good fit on an otherwise very high quality measurement. Based on the high sensitivity of affinity to the removal of data, we decided to use all data points (dropping no points) to avoid introducing these inaccuracies. Rather than dropping data points, we identified poor quality measurements after fitting the model by comparing the magnitude of fit error to the magnitude of the measured signal (signal-to-noise ratio, SNR).

Signal To Noise Ratio. In order to account for outlier measurements that impact fitness and to determine how well the data was represented by the model, we used a signal to noise ratio (SNR) metric. This SNR metric evaluates the magnitude of residual errors of fit to the model (a form of noise), and weights this sum by the overall size of the fluorescent signal measured. It is calculated as

$$SNR = \frac{max(F_{obs}) - min(F_{obs})}{\sum_{i=0}^{n}|R_i|} \qquad (3)$$

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 units). We chose a ratio of 1 as the limit of a good fit (see Fig. S5 and Fig. S6). At an SNR greater than one, the measured signal is larger than the sum of all errors to the fit, and represents a good quality fit in practice, with few exceptions.

Receptor Occupancy Model Failure to Fit Non-Binding Measurements. The original analysis rejected measurements below an r^2 cutoff of 0.95. Those rejected measurements were considered to be non-binders by many subsequent analysis and models. Since the use of an r^2 cutoff does not have a straightforward interpretation when evaluating a nonlinear model, we wished to understand under what conditions this approach would have produced errors in classification of binding and non-binding interactions.

Although the receptor occupancy model is theoretically capable of fitting a typical binding saturation curve as well as a 'flat' curve representative of non-binding interactions, we found that in practice it fails to identify non-binding interactions (Fig. S7, blue fits). The fitting errors follow two patterns: In the first pattern, noise in the data is over-fit. Non-binding data typically looks like a low magnitude flat line with superimposed noise. However, in practice, traditional least-squares methods will tend to over-fit noise in the data to a rapidly saturating curve, rather than fit a straight line. Ironically, this artifact results in miscategorization as a binder, with a high affinity fit. Second, when there is limited non-specific binding present, non-binding data can also present as a line with a low-to-moderate slope with superimposed noise. In these cases, the receptor occupancy model tends to fit a low-curvature arc (almost indistinguishable from a straight line). The consequences of this type of fit artifact are found in erroneous fit parameters: an astronomically high saturation value and low affinity. A saturation value of this size cannot result from the one-to-one interaction assumption of the receptor occupancy model, and clearly represents a fit artifact.

Thus, we hypothesized that a linear model would more reliably fit non-binding interactions and resolve both of these types of fit artifacts. The linear model:

$$F_{obs} = m[SH2\ domain] + F_{bg} \qquad (4)$$

(where F_{bg} represents a FP background offset value, and m is a constant representing the slope of the fitted line, Fig. S7, red fits). There are two

parameters to the linear model (slope and offset/intercept), one fewer parameter than the receptor occupancy model which has F_{max}, K_d, and offset/intercept.

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. In order to determine if a measurement is best described by a receptor occupancy model or a linear model we used the Akaike Information Criterion (AIC). In contrast to the coefficient of determination (r^2), AIC is a model selection metric which is appropriate for use with non-linear models (*19*, *30*), is robust even with high noise data, and employs a regularization technique to avoid overfitting by penalizing models with more parameters. In our implementation we used a bias corrected form of the metric, AICc, in order 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. S8. If data was best fit by the receptor occupancy model, we used SNR to identify the quality of fit of that model.

Although low-slope linear data is consistent with non-binding interactions, we also found a class of measurement which was best fit by a high-slope linear model. The high slope suggests linearly increasing fluorescent signal with concentration with no indication of saturation. This type of response is outside the scope of a receptor occupancy model, and is more likely to represent a form of either protein or peptide aggregation (or a combination of both), or a form of non-specific binding. Thus, to preserve the quality of the non-binding calls, a conservative low-slope cutoff of 5mP/µM was implemented, above which replicates were identified as aggregators, and removed from further consideration.

Summary of Revised Analysis Method for Replicate Measurements. Following a systematic review of each decision made in evaluating a measurement in high-throughput affinity studies (i.e., background subtraction, outlier removal, model fitting, and quality of fit) we developed a new analysis pipeline for each replicate measurement (Fig. 2). For each replicate measurement we fit two models: a linear model with offset (equation 4) and a receptor occupancy model with offset (equation 2). Fits were evaluated with AICc: the model with the lower score was chosen as the best fit. Replicates that were fit best by the linear model and had a slope of less than or equal to 5mP/µM were classified as negative interactions, or 'non-binders'. Linear fits with a slope greater than 5mP/µM were classified as aggregators. A replicate that was fit best by the receptor occupancy model was then evaluated for signal to noise ratio (SNR). If the SNR was greater than one, the replicate was classified as a positive interaction or 'binder'. Out of 37,378 replicate measurements, we found 2753 binders and 29,778 non-binders. There were 2764 replicates that fit best to the receptor occupancy model, but were too noisy to reliably call as binders (classified as Low-SNR fits), and approximately 2000 fits that best fit the linear model, but with high slope (classified as Aggregators, Fig. 3).

Once a fitting process is completed for each replicate, typically replicate measurements are averaged and the mean and standard deviation are reported. In the original publication, the authors averaged the affinities (K_d) derived from each replicate domain-peptide pair measurement to obtain the final published K_d value and reported standard deviations where there were three or more replicates. However, we found interesting patterns at the replicate level that made us question whether the mean was an appropriate way to handle the replicates, discussed in detail below.

High variation at the replicate level highlights protein degradation and inactivity

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 interesting patterns in the replicate-level results suggesting impaired protein functionality and problems with concentration accuracy. This led to an in-detail examination of replicate results, and made us examine the assumption of using the mean as an appropriate way to handle replicates.

In looking at replicate variance, we noticed examples of high variance in affinity among replicates (for example, K_d values ranging from 0.5µM to over 20µM for replicates from a single domain-peptide interaction). To explore this further, we visualized variance for each group of

replicates from the same peptide and domain using a distributed dot plot (Fig. 4). We found high variation in replicates across a large fraction of all measurements, independent of affinity. We then inspected the individual fits for each replicate group in order to determine the source of the variation. Most replicate fits were high quality and had low residual error to the model. By eye, the measured data looked reliable for each replicate, despite the high variation in derived affinity. (For a representative example of all measurements from one such replicate group, see Fig. S9). This pattern held true across many such examples reviewed (data not shown). How could such high-quality individual replicate measurements result in such varied affinities for a single domain-peptide pair?

On its face, such high variation in affinity between replicates 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. However, we hypothesized that a single variable – protein degradation – could be responsible for the high variance we saw in this data. 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*)

To that end, we first examined the theoretical effects of degradation on affinity. We found that degradation could produce high variance in affinity, and could also be consistent with the high-quality individual fits we saw for replicates. Next, we identified evidence of such degradation in patterns in the data. Finally, we developed a method to control for degradation in the current raw data in order to ultimately produce more accurate interaction affinities using existing raw data.

Effect of Degradation on Derived K_d. 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 function of concentration. Thus, we hypothesized that errors in protein concentration would be reflected as errors in affinity. Because a fraction of degraded or inactive protein represents an error between the assumed concentration and the active concentration of a protein, degraded or inactive protein would also propagate to errors in affinity.

The effect of 50% degradation and 75% degradation on a protein with 1µm K_d is shown in Fig. 5. For saturation binding experiments, the error in fluorescence polarization (FP) is not linear with the error in concentration – rather it is a function of the level of saturation of the protein binding the ligand. However, the error in affinity (derived from the model fit) is linearly proportional to the error in concentration.

Thus, degraded protein of varying degrees can manifest as a range of measured K_d values in replicate measurements (all of which would be equal to or higher than the true K_d), while simultaneously coming from seemingly high-quality, low-noise individual FP measurements. This exact phenomenon has also been demonstrated experimentally (*31*).

Evidence for Protein Degradation and Non-Functionality in the Raw Data. We next examined the data for evidence of protein degradation. If the variance in affinity was from random (non-systemic) sources, we would expect to find no patterns of variance in time. In contrast, if variance was from protein degradation, we might see non-random patterns in affinity over time. For example, if a fresh protein sample and a degraded protein sample were used on different runs, we might expect to see variation correlating with the day, but consistent during that run. If a protein sample was exhausted mid-run, and replaced with a fresh sample, we might see a sudden surge of increased affinity in the middle of a run. Although we don't have an exact time for each measurement, and the same peptides were measured far apart in time, we do have a pseudo-time substitute. Fortunately, on each run, the peptides were measured in approximately the same order, which allows us to see patterns of protein affinity over time and across peptides from run to run. In the first published experiment, data was primary gathered on 3 runs on 3 different days. On each run, domains were tested against hundreds of peptides providing rich data for seeing these patterns.

Fig. S10 shows this time-dependent data for interactions with three SH2 domains. For PIK3R2-N, we see that Run3 replicates consistently showed lower K_d values (higher affinities) than replicates from other days. This pattern of run to run variation suggests that the protein samples tested in Runs 1 and 2 were less active than Run 3. For RASA1-N, no single day dominated the highest affinity until plate 174, after which the highest affinity replicates all come from Run 1. This is consistent with a change to fresh, less degraded protein during Run 1. These patterns are not compatible with a random source of variance. However not all protein data shows pattern consistent with this degradation hypothesis. For SH2D2A, binding affinity tends to be weaker on Run 2, but the highest affinity (lowest K_d) experimentally alternate between Run 1 and Run 3, and significant variation appears during each run. The patterns for SH2D2A are not consistent with a simple degradation hypothesis, and may be indicative of additional sources of variation.

Because we found patterns consistent with partial degradation, we also wanted to examine the data for patterns of complete protein degradation. Complete degradation, or completely non-functional protein, would be indistinguishable from a non-binding 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 non-functional protein would manifest within the data as long runs of non-binding results across many replicates, but would demonstrate contradictory evidence of binding on other runs when the protein was not degraded.

To examine the data for patterns of non-functional protein, we plotted affinity by domain and by run for three example domains from the first publication: GRB2, BMX, and PIK3R1-C (Fig. S11). For GRB2, no positive interactions were recorded on Run 1, despite several positive interactions from Run 2 and Run 3. This is consistent with the protein on Run 1 being completely non-functional. Furthermore, because the protein on Run 1 may have been non-functional, then all non-binding interactions measured on that run are potentially false negatives. As indicated in the right panel for GRB2, by labeling all measurements from Run 1 as non-functional, we removed the replicates from consideration. For BMX, no positive interactions with any peptide were recorded on any run. While it is possible that this represents the true binding behavior, it is equally possible that the BMX protein was never functional. Since it is impossible to tell from the data, and the quality of both true positive and true negative data is of concern, the most conservative course is to consider all of these measurements as non-functional, and remove them from consideration. This is shown in the right panel for BMX. PIK3R1-C, on the other hand, shows very little evidence of degradation. However, no positive results were recorded on Run 4, which is consistent with the protein in Run 4 being non-functional.

Once all individual replicate fits were complete according to our revised protocol (Fig. 2), we added a step to the pipeline where experimental runs were examined for non-functional protein. If an entire run lacked even one positive binding interaction, but had corresponding positive interactions on another run, the entire run was marked as containing non-functional protein. By removing replicates where there is evidence that the protein was non-functional, we avoid the potential for false negatives from this ambiguous data, and greatly improve the pool of true negative calls. Non-functional protein calls for all peptides and domains can be seen in Fig. S12 and Fig. S13.

Removal of non-functional protein has a significant impact on the numbers of measurements at the replicate level. Fig. 3 shows that non-functional replicates made up 37.6% of all replicates (14070/37378). Most of the non-functional replicates were originally categorized as non-binders, but a portion came from low-SNR replicates, and replicates that demonstrated aggregation. The large number of runs showing patterns of non-functional protein contributes to the overall evidence that protein degradation is a high source of variation in the data, and the need to control for it.

Method for Handling Replicates with High Variance. Two key issues arise when considering how to handle replicate measurements in this data: both caused by the presence of degraded protein. First, we see patterns in the data strongly consistent with degraded and non-functional protein. Yet, individual measurements

from positive interactions seem to be of high-quality and relatively low noise. Without knowing the exact amount of protein degradation in any sample, how can this degradation be controlled for across replicate measurements? Second, what is the correct procedure for handling multiple replicate measurements when degraded protein is suspected in order to report a value closest to the true affinity?

We propose that protein degradation can be partially controlled for by reporting the minimum measured K_d as the affinity. Given some unknown amount of protein degradation, we demonstrated above that the *true* affinity of the protein will always be equal to or higher than the *measured* affinity for protein, because the active concentration will always be equal to or lower than the measured concentration. 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.

In addition, reporting the minimum K_d as the affinity also avoids the issues caused by averaging multiple degraded measurements. 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 mean would represent the highest likelihood of the true value of affinity. However, if the variation is known to be caused by degradation, taking the mean of multiple samples would not reflect the true affinity. Taking the mean of a varying number of samples of unknown degradation would inadvertently increase the reported K_d value by some unpredictable amount, where that amount depends on the number of samples and the magnitude of their degradation. Furthermore, since 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 high error. Therefore, odd though it may seem from a statistical perspective, taking the *minimum* K_d is the most appropriate way to handle variation in replicates where degradation represents the primary source of variation.

Revised Affinity Results and Comparison to the Original Published Results

In the results from our revised analysis, 1518 positive (binding) interactions were identified, along with 7038 negative (non-binding) interactions. These ~7000 true negative results represent a significant increase in information from the original raw data. Approximately 3200 interactions had inconclusive or problematic data and no conclusions about their affinity could be drawn. Of those, 2753 domain-peptide pairs had non-functional protein. Final affinity values were plotted for all peptide-domain interactions as a heat map (Fig. 6), and summarized by category of interaction and changes in calls (Fig. 7). Our revised results and the originally published results are available in Supporting Data as an Excel file.

Despite similar numbers of positive interactions between the original and revised results, the identities of the domain-peptide pairs comprising the positive interactions changed significantly. Changes in calls by class are visualized in Fig. 7, while the identities of the domain-peptide pairs with changed calls are visualized in Fig. S14. Results from the original publication are visualized in Fig. S15. More than 17% of the original positive interaction calls changed to either non-interactions, or rejected results due to data quality issue. In the final model, 168 interactions originally called positive in the published results are found to be true negative interactions. These changes are primarily due to using multiple models to fit the data: the added capability of the added, linear, model to identify aggregation and true non-binding interactions instead of resulting in the over-fitting artifacts or false positive results of using a single model. Similarly large changes were found in the originally published negative interactions where 273 formerly rejected interactions are classified as true positive interactions. These recovered results are primarily due to changes in baseline fitting, and using an appropriate quality metric to determine which model fits best.

Furthermore, even though 1245 domain-peptide pairs were found to bind in both the original publication and our revised analysis, the quantitative affinity of those binders changed significantly in the revised analysis (Fig. 8). Note that although the minimum of each replicate group was selected as most accurately reflecting the true

affinity, our revised affinity values are not all lower than the original publication. This is primarily due to significant changes at the replicate level – where some original replicates were removed from consideration by changes in the fitting process, and a number of new replicates were included in each replicate set.

Independent evaluation of revised analysis: measuring improved consistency via active learning

We wanted to evaluate our revised analysis compared to 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 has better internal consistency or predictive power (when compared to itself) than the original data set. Lacking a biological reference, it seemed fitting to evaluate this data using machine learning, as we originally wished to harness SH2 domain binding measurements in machine learning frameworks to extrapolate from the relatively small number of available measurements.

To do this, we implemented active search, a machine learning approach that is highly amenable to biochemistry problems such as this. Active learning (also known as optimal experimental design or active data acquisition) is a machine learning paradigm 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 dataset represents positive interactions between an SH2 domain and a phosphopeptide (or 18.2% in the revised dataset) the objective of the search algorithm was to prioritize each sequential selected interaction to maximize the total number of positive interactions discovered. We implemented the effective nonmyopic search (ENS) algorithm (*32*) 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 search progressed and was repeated 50 times.

ENS showed improved average performance and higher consistency with our revised dataset. First, ENS worked effectively on both the original and revised datasets, identifying positives that far exceed the expected number by random chance by the 100th query (Fig. 9). This suggests that phosphopeptide sequences do encode information about whether an SH2 domain will recognize them in a binding interaction. Second, ENS performance in the revised dataset was higher than the original dataset on average, finding 45.3 positives vs. 33.3 positives (p-value of 4e-12). Third, ENS performance is significantly more variable on the original dataset than on the final dataset (ranging between 9 and 62 positives in 50 trials (with an average of 33.3), compared to a range of 38 to 67 (with an average of 45.3 positives) for the revised dataset. In the worst of the 50 trials, search in the original dataset underperformed by 50% compared to what is expected by random chance), whereas the worst random trial within the final dataset still outperformed random chance by two-fold. Thus, the improved average performance and lower variability in our revised results suggests improved coherency in our revised analysis over the original published results.

Discussion

Here, we present a revised analysis of raw data from SH2 domain affinity experiments. We presented an analysis framework which improved on the model fitting and evaluation methods of previous work. We used improved methods to identify high-quality true positive interactions, and we added thousands of true negative interactions, while filtering out results from potentially inactive protein.

Although raw data from only two experiments was available for detailed analysis, we were fortunate that raw data combined a large quantity of measurements with a well-established, solution-based experimental system – fluorescence polarization – 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 would be less preferable to a method where both molecules were measured in

solution. So despite limited availability of data, the raw data available is likely to be the best example for further analysis.

Other high-throughput experiments share many critical methods with the data reviewed here. In all published experiments measuring affinity, protein was minimally filtered after production. Authors knowingly measured non-monomeric protein. 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 non-binding results from negative interactions due to non-functional protein, and additional errors in active protein concentration with respect the measured protein concentration. Incorrect use of statistical methods to evaluate models was common to all published work – particularly the improper use of the coefficient of determination (r^2) to determine the quality of fit of a non-linear model, and using only a single model to fit data. Their choices resulted in a high false negative rate, and also masked 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 for in other experiments.

One seemingly innocuous choice – averaging multiple replicates containing degraded protein – could be a significant source of error in published results from this experiment and other published high-throughput data. Taking the mean of multiple replicates is a standard practice when replicate differences represent random error, but it has drastically different results in the presence of multiple degraded measurements. The use of the mean to reconcile degraded replicate measurements could manifest as errors effectively randomizing reported affinity measurements. Even if failure to control protein degradation was the sole common error among these experiments, it could be the cause of the discrepancies between published numerical results.

It is concerning that an entire body of published work has developed from this set of problematic results. At the very least, we have shown that affinity values from the original publications were derived from data and methods causing serious inaccuracies. This data has had a wide-reaching effect in many areas of SH2 domain research: the data has been used to draw specific conclusions about SH2 domain biology such as identification of EGFR recruitment targets (*33*), to explain quantitative differences in RTK signaling (*9*), and as evidence to understand the promiscuity of EGFR tail binding (*34*). In addition, this work has been used to guide experimental design by filtering potential binding proteins by affinity (*35*), to reconcile confusing experimental results (*36*), and to guide new experimental hypothesis testing (*37*). It has played a role in cancer research as context to understand kinase dependencies in cancer (*38*), and as evidence of HER3 and PI3K connections as relevant to PTEN loss in cancer (*39*). It has influenced evolutionary analysis (*40*), has been used to design mechanistic EFGR models (*41*, *42*), and has been used in computational algorithms for domain binding predictions (*14–18*, *43*).

Furthermore, it is likely that these issues plague most other high-throughput studies of SH2 domains due to shared methodology, and thus affect works derived from those publications as well. 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 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. These negative interactions 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 from our revised set are controlled for false negative results from non-functional 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 this method is that the highest affinity measured value may not be the true affinity, if a fully functional protein was never measured. Nevertheless, the highest measured affinity should still represent the measured value closest to the true value. It is also important to

restate: not all variation in the data is consistent with the degradation hypothesis, and some variation may represent other unknown sources of variation which we have not controlled for. For example, one key assumption of the receptor occupancy model requires measuring the reaction at equilibrium. Since no data is provided to prove that the 20-minute 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 non-equilibrium conditions.

Finally, we would like to discuss methods to improve future data gathering and reporting. High-throughput 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, for example SH3-polyproline interactions (*44*, *45*), PDZ domains interacting with C-terminal tails (*46–48*), and major histocompatibility complete (MHC) interactions with peptides (*49*, *50*). 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 high-throughput methods should be established. For example, all raw data from high throughput experiments should 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 Gene Expression Omnibus, or the raw experimental measurements are 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 on Figshare (DOI: https://doi.org/10.6084/m9.figshare.11482686.v1), and provided the code for the analysis pipeline on GitHub (https://github.com/NaegleLab/SH2fp) so that future evaluation can be more easily accomplished by other researchers. Furthermore, in methods quantitatively measuring protein activity, protein degradation will always be an issue. Methods for quantifying activity should be a best practice. Alternatively, methods which 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 lab (*51*). In that method, a 2-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. Alternatively, another recent experiment also uses competitive fluorescence anisotropy, but measures a competitive titration curve in a single well with an agarose gradient (*52*). Diffusion forms a spatiotemporal gradient for the interaction, and so one can produce a full titration curve in each well in a multi-well 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.

Methods

Raw Data. Upon receipt of the Jones 2012-14 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 (approximately 138 replicate-level measurements out of over 37,000 measurements) and were limited to 3 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, 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, DAPP1, FER, FES, FGR, GRAP2, GRB2, GRB7, GRB10, GRB14, HCK, HSH2D, INPPL1, 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, SOCS6, 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-terminal only) and PTPN6 (C-terminal only). One peptide had no measurements in the raw data (EGFR pY944). Within this revised scope, the available raw data covered approximately 99.6% of the originally available raw data.

The raw data for each measured interaction consisted of fluorescence polarization measurements of an SH2 domain in solution with a phosphopeptide at 12 concentrations. The measurements were arranged on 384 well plates: 32 different SH2 domains at each of 12 concentrations, all measured against a single peptide per plate. Protein concentrations represented 12 serial dilutions of 50% starting with either 10 µM or 5 µM protein.

Model Fitting, Model Selection, and Replicate-Level Calls. For each replicate measurement, we fit two models: the linear model (equation 4) and the receptor occupancy model (equation 2). Model fits were evaluated with the bias corrected Akaike Information Criterion (AICc), and the model with the lower AICc score was selected (*19*).

The Akaike Information Criterion (AIC) as a quality metric, was calculated by

$$AIC = 2p - 2\,ln(L) \qquad (5)$$

where p is the number of parameters in the model, and $ln(L)$ is the maximum log-likelihood of the model. In a non-linear fit, with normally distributed errors, $ln(L)$ is calculated by

$$ln(L) = -0.5N\left(ln(2\pi) + 1 - ln(N) + ln\left(\sum_{i=1}^{n} {x_i}^2\right)\right) \qquad (6)$$

where x_1, ..., 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 which corrects for small sample sizes, e.g. when one has fewer than 30 data points. AICc is calculated as follows:

$$AICc = AIC + \frac{2p(p+1)}{n-p-1} \qquad (7)$$

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 background offset (F_{bg})), while the linear model had two parameters (slope (m), and background offset (F_{bg})).

Replicates that were fit best by the linear model with a slope of less than or equal to 5mP/µM were categorized as negative interactions, or 'non-binders'. Linear fits with a slope greater than 5mP/µM were categorized as aggregators. Replicates that were fit best by the receptor occupancy model were subsequently evaluated for signal to noise ratio (SNR, equation 3). If the SNR was greater than one, the replicate was categorized as a positive interaction or 'binder', otherwise, it was rejected as a low-SNR fit and removed from consideration.

Identifying Non-Functional Protein. Once all individual fits were complete, runs were examined for non-functional protein. If an entire run lacked even one positive binding interaction, and those same interactions measured positive on another run, the non-binder, aggregator, and low-SNR calls on that run were changed to non-functional 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 signal to noise ratio (SNR) were considered. For a given domain-peptide pair, the minimum numeric value of K_d (strongest measured affinity) was reported as the final K_d for that domain peptide pair.

Active search. The probability model (*32*) used a simple k-nearest neighbor (k = 20) where distance is defined by average Euclidean distance of corresponding divided physicochemical

property scores (DPPS) features of the amino acids (*53*) comprising the peptide, i.e.:

$$d_{nn}(x, x') = \frac{1}{n}\sum_{i=1}^{n} d_e\left(dpps(x_i), dpps(x'_1)\right) \quad (8)$$

where d_{nn} 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.

Acknowledgements: We would like to thank Richard Jones, Ron Hause and Ken Leung for providing the raw data required for this analysis.

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

References

1. Y. Yarden, M. X. Sliwkowski, Untangling the ErbB signalling network. *Nat. Rev. Mol. Cell Biol.* **2**, 127–137 (2001).
2. K. Machida, B. J. Mayer, The SH2 domain: Versatile signaling module and pharmaceutical target. *Biochim. Biophys. Acta - Proteins Proteomics*. **1747**, 1–25 (2005).
3. T. Pawson, Specificity in Signal Transduction: From Phosphotyrosine-SH2 Domain Interactions to Complex Cellular Systems. *Cell*. **116**, 191–203 (2004).
4. S. Zhou, S. E. Shoelson, M. Chaudhuri, G. Gish, T. Pawson, W. G. Haser, F. King, T. Roberts, S. Ratnofsky, R. J. Lechleider, B. G. Neel, R. B. Birge, J. E. Fajardo, M. M. Chou, H. Hanafusa, B. Schaffhausen, L. C. Cantley, SH2 domains recognize specific phosphopeptide sequences. *Cell*. **72**, 767–778 (1993).
5. T. Pawson, P. Nash, Assembly of cell regulatory systems through protein interaction domains. *Science*. **300**, 445–452 (2003).
6. M. K. Matlock, A. S. Holehouse, K. M. Naegle, ProteomeScout: A repository and analysis resource for post-translational modifications and proteins. *Nucleic Acids Res.* (2015), doi:10.1093/nar/gku1154.
7. R. B. Jones, A. Gordus, J. A. Krall, G. MacBeath, A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature*. **439**, 168–174 (2006).
8. A. Kaushansky, A. Gordus, B. A. Budnik, W. S. Lane, J. Rush, G. MacBeath, System-wide Investigation of ErbB4 Reveals 19 Sites of Tyr Phosphorylation that Are Unusually Selective in Their Recruitment Properties. *Chem. Biol.* **15**, 808–817 (2008).
9. A. Gordus, J. A. Krall, E. M. Beyer, A. Kaushansky, A. Wolf-Yadlin, M. Sevecka, B. H. Chang, J. Rush, G. MacBeath, Linear combinations of docking affinities explain quantitative differences in RTK signaling. *Mol. Syst. Biol.* **5**, 235 (2009).
10. G. Koytiger, A. Kaushansky, A. Gordus, J. Rush, P. K. Sorger, G. MacBeath, Phosphotyrosine Signaling Proteins that Drive Oncogenesis Tend to be Highly Interconnected. *Mol. Cell. Proteomics*. **12**, 1204–1213 (2013).
11. B. a Liu, K. Jablonowski, E. E. Shah, B. W. Engelmann, R. B. Jones, P. D. Nash, SH2 domains recognize contextual peptide sequence information to determine selectivity. *Mol. Cell. Proteomics*. **9**, 2391–2404 (2010).
12. M. Tinti, L. Kiemer, S. Costa, M. L. Miller, F. Sacco, J. V. Olsen, M. Carducci, S. Paoluzi, F. Langone, C. T. Workman, N. Blom, K. Machida, C. M. Thompson, M. Schutkowski, S. Brunak, M. Mann, B. J. Mayer, … G. Cesareni, The SH2 Domain Interaction Landscape. *Cell Rep.* **3**, 1293–1305 (2013).
13. R. J. Hause, K. K. Leung, J. L. Barkinge, M. F. Ciaccio, C. pin Chuu, R. B. Jones, Comprehensive Binary Interaction Mapping of SH2 Domains via Fluorescence Polarization Reveals Novel Functional Diversification of ErbB Receptors. *PLoS One*. **7** (2012), doi:10.1371/journal.pone.0044471.
14. K. K. Leung, R. J. Hause, J. L. Barkinge, M. F. Ciaccio, C.-P. Chuu, R. B. Jones, 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 (2014).
15. I. E. Sánchez, P. Beltrao, F. Stricher, J. Schymkowitz, J. Ferkinghoff-Borg, F. Rousseau, L. Serrano, Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. *PLoS Comput. Biol.* **4** (2008), doi:10.1371/journal.pcbi.1000052.
16. M. AlQuraishi, G. Koytiger, A. Jenney, G. MacBeath, P. K. Sorger, A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. *Nat. Genet.* **46**, 1363–1371 (2014).
17. Z. Wunderlich, L. a. Mirny, Using genome-wide measurements for computational prediction of SH2-peptide interactions. *Nucleic Acids Res.* **37**, 4629–4641 (2009).
18. K. Kundu, F. Costa, M. Huber, M. Reth, R. Backofen, Semi-Supervised Prediction of SH2-Peptide

Interactions from Imbalanced High-Throughput Data. *PLoS One*. **8** (2013), doi:10.1371/journal.pone.0062732.

19. A.-N. Spiess, N. Neumeyer, An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacol.* **10**, 6 (2010).
20. T. O. Kvalseth, Cautionary note about R-squared. *Am. Stat.* (1985), doi:10.1080/00031305.1985.10479448.
21. S. A. Juliano, F. M. Williams, A Comparison of Methods for Estimating the Functional Response Parameters of the Random Predator Equation. *J. Anim. Ecol.* **56**, 641–653 (1987).
22. L. Magee, R2measures based on wald and likelihood ratio joint significance tests. *Am. Stat.* (1990), doi:10.1080/00031305.1990.10475731.
23. N. J. D. Nagelkerke, A note on a general definition of the coefficient of determination. *Biometrika* (1991), , doi:10.1093/biomet/78.3.691.
24. R. Anderson-Sprecher, Model comparisons and r2. *Am. Stat.* (1994), doi:10.1080/00031305.1994.10476036.
25. J. B. Willett, J. D. Singer, Another cautionary note about r2: Its use in weighted least-squares regression analysis. *Am. Stat.* (1988), doi:10.1080/00031305.1988.10475573.
26. S.-P. Miaou, A. Lu, H. Lum, Pitfalls of Using R 2 To Evaluate Goodness of Fit of Accident Prediction Models . *Transp. Res. Rec. J. Transp. Res. Board* (2007), doi:10.3141/1542-02.
27. A. Kaushansky, A. Gordus, B. Chang, J. Rush, G. Macbeath, A quantitative study of the recruitment potential of all intracellular tyrosine residues on EGFR, FGFR1 and IGF1R. *Mol. Biosyst.* **4**, 643–653 (2008).
28. A. Kaushansky, J. E. Allen, A. Gordus, M. A. Stiffler, E. S. Karp, B. H. Chang, G. Macbeath, Quantifying protein-protein interactions in high throughput using protein domain microarrays. *Nat. Protoc.* (2010), doi:10.1038/nprot.2010.36.
29. A. J. Clark, The reaction between acetyl choline and muscle cells. *J. Physiol.* **61**, 530–46 (1926).
30. M. Mazerolle, Appendix 1: Making sense out of Akaike's Information Criterion (AIC): its use and interpretation in model selection and inference from ecological data. … *en Tourbières Perturbées*, *Ph. D. thesis*, 1–13 (2004).
31. E. Pol, The importance of correct protein concentration for kinetics and affinity determination in structure-function analysis. *J. Vis. Exp.* (2010), doi:10.3791/1746.
32. S. Jiang, G. Malkomes, G. Converse, A. Shofner, B. Moseley, R. Garnett, in *34th International Conference on Machine Learning, ICML 2017* (2017).
33. C. J. Tsai, R. Nussinov, Emerging Allosteric Mechanism of EGFR Activation in Physiological and Pathological Contexts. *Biophys. J.* (2019), , doi:10.1016/j.bpj.2019.05.021.
34. S. P. Kennedy, J. F. Hastings, J. Z. R. Han, D. R. Croucher, The under-appreciated promiscuity of the epidermal growth factor receptor family. *Front. Cell Dev. Biol.* (2016), , doi:10.3389/fcell.2016.00088.
35. M. R. Birtwistle, Analytical reduction of combinatorial complexity arising from multiple protein modification sites. *J. R. Soc. Interface*. **12**, 20141215 (2015).
36. S. H. Leong, K. M. Lwin, S. S. Lee, W. H. Ng, K. M. Ng, S. Y. Tan, B. L. Ng, N. P. Carter, C. Tang, O. Lian Kon, Chromosomal breaks at FRA18C: association with reduced DOK6 expression, altered oncogenic signaling and increased gastric cancer survival. *npj Precis. Oncol.* (2017), doi:10.1038/s41698-017-0012-3.
37. A. Ruiz-Saenz, C. Dreyer, M. R. Campbell, V. Steri, N. Gulizia, M. M. Moasser, HER2 Amplification in Tumors Activates PI3K/Akt Signaling Independent of HER3. *Cancer Res.* **78**, 3645–3658 (2018).
38. J. Campbell, C. J. Ryan, R. Brough, I. Bajrami, H. N. Pemberton, I. Y. Chong, S. Costa-Cabral, J. Frankum, A. Gulati, H. Holme, R. Miller, S. Postel-Vinay, R. Rafiq, W. Wei, C. T. Williamson, D. A. Quigley, J. Tym, … C. J. Lord, Large-Scale Profiling of Kinase Dependencies in Cancer Cell Lines. *Cell Rep.* (2016), doi:10.1016/j.celrep.2016.02.023.

39. H. M. Stern, H. Gardner, T. Burzykowski, W. Elatre, C. O'Brien, M. R. Lackner, G. A. Pestano, A. Santiago, I. Villalobos, W. Eiermann, T. Pienkowski, M. Martin, N. Robert, J. Crown, P. Nuciforo, V. Bee, J. Mackey, … M. F. Press, PTEN loss is associated with worse outcome in HER2-Amplified breast cancer patients but is not associated with trastuzumab resistance. *Clin. Cancer Res.* (2015), doi:10.1158/1078-0432.CCR-14-2993.
40. M. L. Miller, L. J. Jensen, F. Diella, C. Jørgensen, M. Tinti, L. Li, M. Hsiung, S. a Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovsky, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li, … R. Linding, Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.* **1**, ra2 (2008).
41. E. C. Stites, M. Aziz, M. S. Creamer, D. D. Von Hoff, R. G. Posner, W. S. Hlavacek, Use of mechanistic models to integrate and analyze multiple proteomic datasets. *Biophys. J.* **108**, 1819–1829 (2015).
42. J. A. Jadwin, T. G. Curran, A. T. Lafontaine, F. M. White, B. J. Mayer, Src homology 2 domains enhance tyrosine phosphorylation in vivo by protecting binding sites in their target proteins from dephosphorylation. *J. Biol. Chem.* (2018), doi:10.1074/jbc.M117.794412.
43. W. Gong, D. Zhou, Y. Ren, Y. Wang, Z. Zuo, Y. Shen, F. Xiao, Q. Zhu, A. Hong, X. Zhou, X. Gao, T. Li, PepCyber:P~PEP: A database of human protein-protein interactions mediated by phosphoprotein-binding domains. *Nucleic Acids Res.* **36**, 679–683 (2008).
44. C. Landgraf, S. Panni, L. Montecchi-Palazzi, L. Castagnoli, J. Schneider-Mergener, R. Volkmer-Engert, G. Cesareni, Protein interaction networks by proteome peptide scanning. *PLoS Biol.* (2004), doi:10.1371/journal.pbio.0020014.
45. M. Carducci, L. Perfetto, L. Briganti, S. Paoluzi, S. Costa, J. Zerweck, M. Schutkowski, L. Castagnoli, G. Cesareni, The protein interaction network mediated by human SH3 domains. *Biotechnol. Adv.* (2012), doi:10.1016/j.biotechadv.2011.06.012.
46. J. R. Chen, B. H. Chang, J. E. Allen, M. a Stiffler, G. MacBeath, Predicting PDZ domain-peptide interactions from primary sequences. *Nat. Biotechnol.* **26**, 1041–1045 (2008).
47. P. Boisguerin, R. Leben, B. Ay, G. Radziwill, K. Moelling, L. Dong, R. Volkmer-Engert, An improved method for the synthesis of cellulose membrane-bound peptides with free C termini is useful for PDZ domain binding studies. *Chem. Biol.* (2004), doi:10.1016/j.chembiol.2004.03.010.
48. U. Wiedemann, P. Boisguerin, R. Leben, D. Leitner, G. Krause, K. Moelling, R. Volkmer-Engert, H. Oschkinat, Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J. Mol. Biol.* **343**, 703–718 (2004).
49. A. K. Haj, M. E. Breitbach, D. A. Baker, M. S. Mohns, G. K. Moreno, N. A. Wilson, V. Lyamichev, J. Patel, K. L. Weisgrau, D. M. Dudley, D. H. O'Connor, High-throughput identification of MHC class I binding peptides using an ultradense peptide array. *bioRxiv*, 715342 (2019).
50. S. Gaseitsiwe, M. J. Maeurer, (2009; http://link.springer.com/10.1007/978-1-59745-450-6_30), pp. 417–426.
51. Z. Zuo, B. Roy, Y. K. Chang, D. Granas, G. D. Stormo, Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. *Sci. Adv.* **3**, eaao1799 (2017).
52. C. Jung, M. Schnepf, P. Bandilla, U. Unnerstall, U. Gaul, High Sensitivity Measurement of Transcription Factor-DNA Binding Affinities by Competitive Titration Using Fluorescence Microscopy. *J. Vis. Exp.* (2019), doi:10.3791/58763.
53. F. Tian, L. Yang, F. Lv, Q. Yang, P. Zhou, In silico quantitative prediction of peptides binding affinity to human MHC molecule: An intuitive quantitative structure-activity relationship approach. *Amino Acids* (2009), doi:10.1007/s00726-008-0116-8.

Data Group	Type	Publication	Peptides	SH2 Domains	Affinity	Results Available	Raw Data Available	Models	
1	PM	Jones et al. (2006)	61	159	Yes	Yes	No	SH2PepInt	Wunderlich/ Mirny
		Kaushansky et al. (2008)	50	133	Yes	Yes	No		
		Gordus et al. (2009)	46	96	Yes	Yes	Yes (pos)		
2	PM	Koytiger et al. (2013)	729	70	Yes	Yes	No	MSM/D, FoldX	
3	FP	Hause et al. (2012)	89	93	Yes	Yes	Yes (PC)		PEBL
		Leung et al. (2014)	85	93	Yes	Yes	Yes (PC)		
n/a	PA	Liu et al. (2010)	192	50	No	No (fig)	No		
		Tinti et al. (2013)	6202	70	No	No (*)	No (*)		

Table 1: Overview of Published SH2 Data and Use in Published Models. 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 is 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 are available by private communication; (*) Original results were stored in PepspotDB, but not published in the journal or supplement. PepspotDB is no longer available.

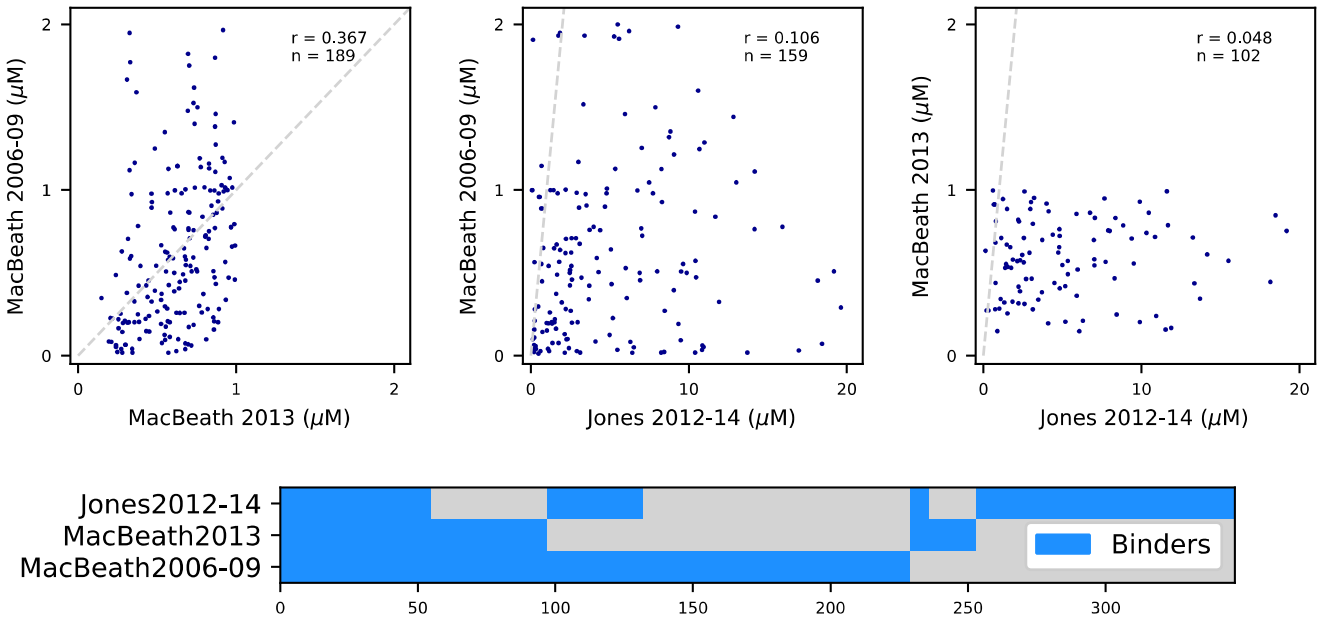

Figure 1: Comparison of Published Affinity Data. The correlation of data published by the MacBeath group between 2006 and 2009, by the MacBeath group in 2013, and by the Jones group in 2012 and 2014 is evaluated using correlation plots (top row). With perfect agreement, data points would fall along the dashed gray line. Surprisingly, there is almost no correlation between data groups. Even results from the same lab published at different times show only mild correlation (r=0.367, MacBeath 2006-09 vs MacBeath 2013). The data were also examined for agreement on positively interacting domain-peptide pairs (bottom panel). Positive interactions are identified by blue bars. Of the 347 positive domain-peptide interactions identified by at least one group, only 55 interactions were found to be positive in all three data groups (15.9%). No two data groups agreed on more than 29% of positive interactions.

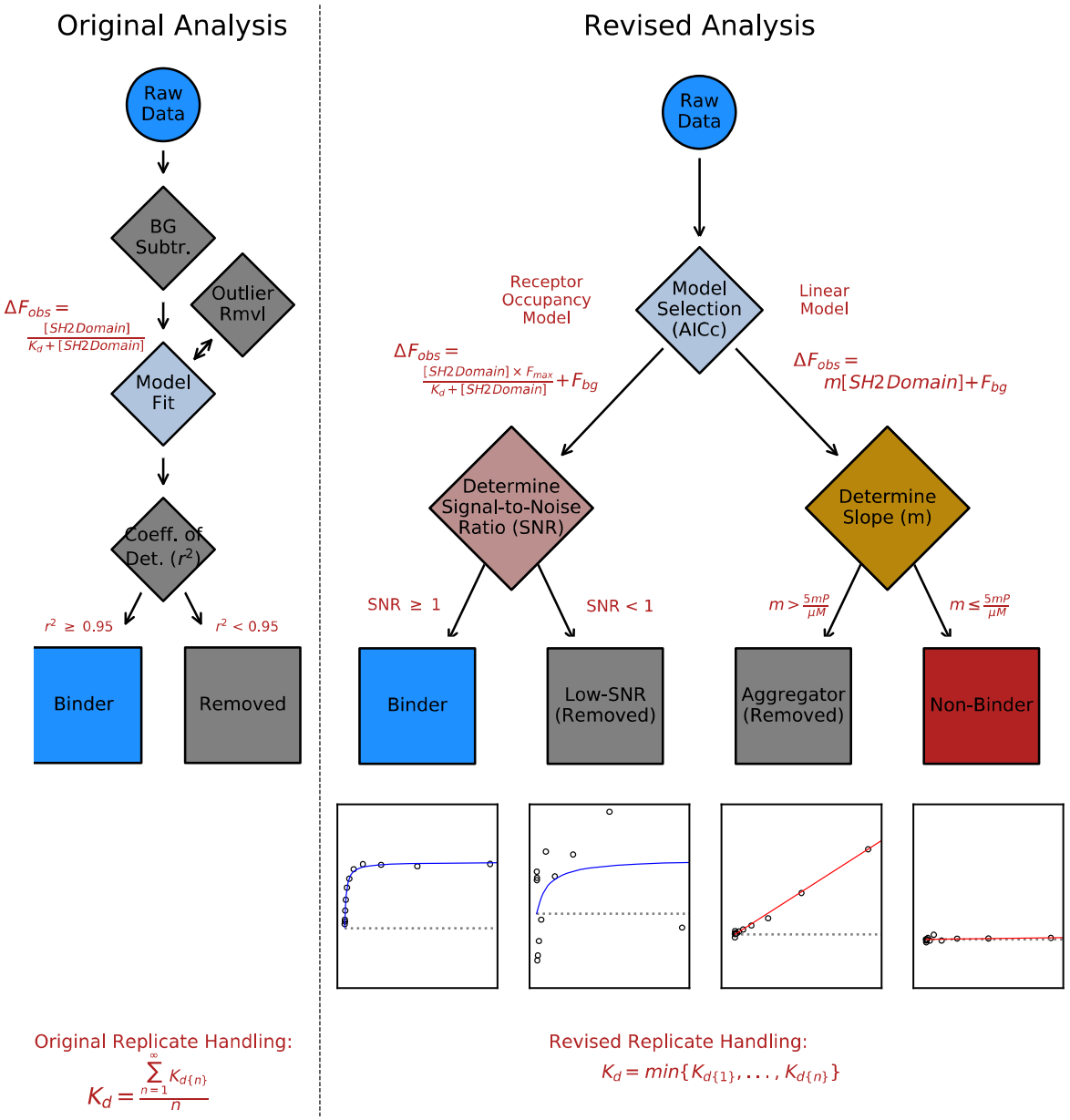

Figure 2: Flowchart of Revised Analysis Process. Comparison between the original analysis process (left panel) and our revised analysis pipeline (right panel). For our revised process, representative sample fits are shown below each of the final categorizations.

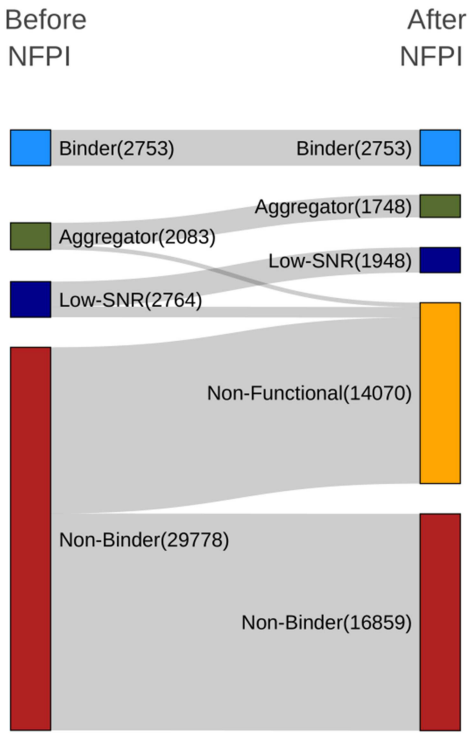

Figure 3: Initial Replicate-Level Results and the Results of Non-Functional Protein Identification (NFPI). The categorization results of individual domain-peptide measurements are shown (Before NFPI). Of the 37,378 measurements, 7.4% (2,753) were initially identified as positive interactions (binders), 7.4% (2,083) as interactions showing aggregation, 5.6% (2,764) as low signal-to-noise, and 79.7% (29,778) as non-binders. The subsequent identification and removal of individual domain-peptide measurements made on non-functional protein had a significant effect on the categorization of non-positive replicate-level measurements. Of the 29778 measurements initially categorized as non-binders, 56.6% (16,859) were identified as likely to contain non-functional protein and were removed from further consideration.

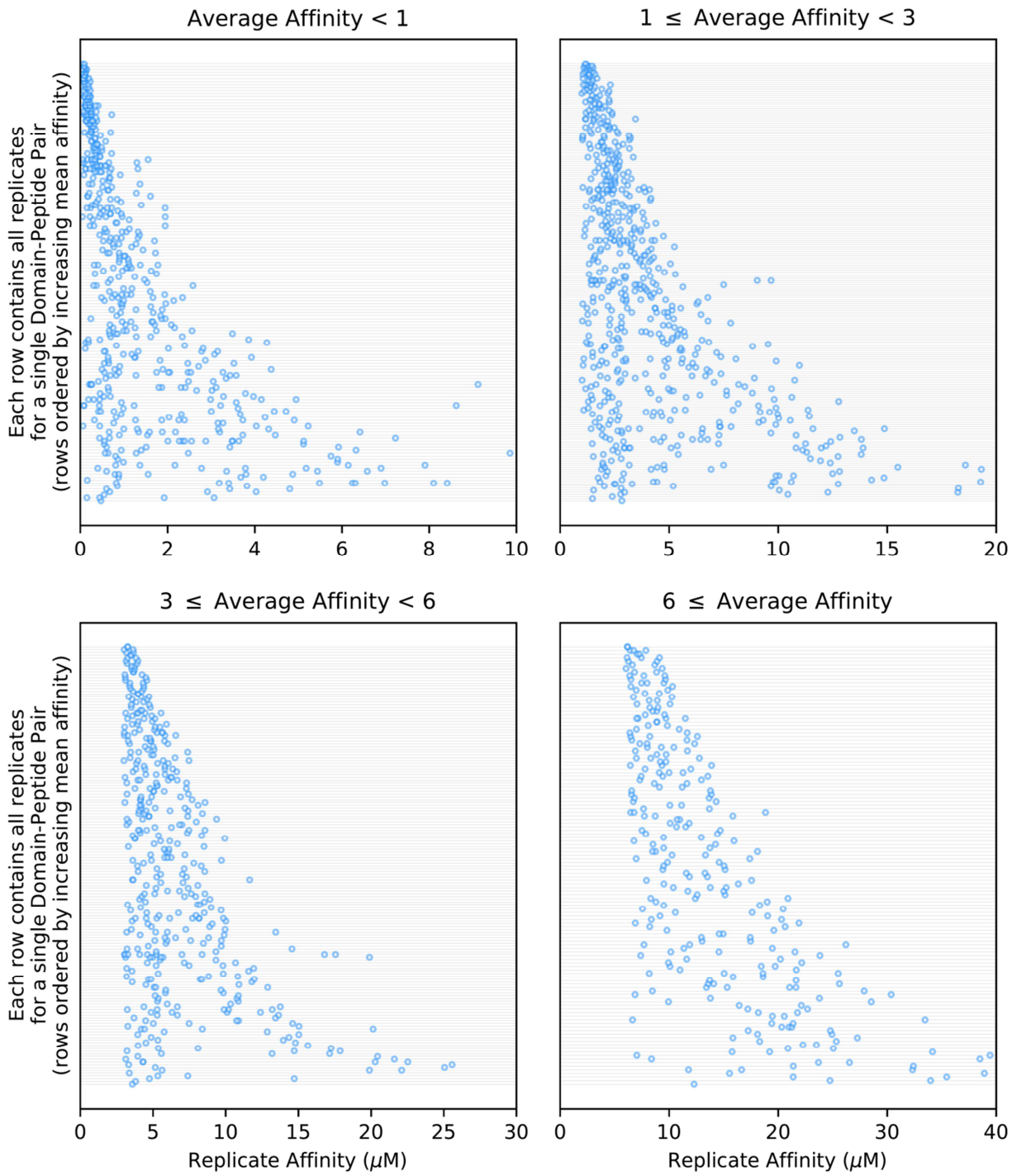

Figure 4: Replicate Measurements Exhibit High Variance. The variance in replicates for each domain-peptide interaction was visualized by distributed dot plots. Within each plot, on each row all affinity measurements for a single domain-peptide pair are plotted. Domain-peptide interactions (rows) are sorted by mean affinity, and grouped into four different affinity ranges (panels) for more detailed viewing. The plots demonstrate that high variance replicate groups can be found in domain-peptides with all ranges of mean affinity, and that very few domain peptide pairs have low variance replicates. This suggests that high replicate variance is ubiquitous, and independent of mean affinity.

Theoretical Protein
K_d = 1.0 µM

Measured Affinity at 50% Degradation

Measured Affinity at 75% Degradation

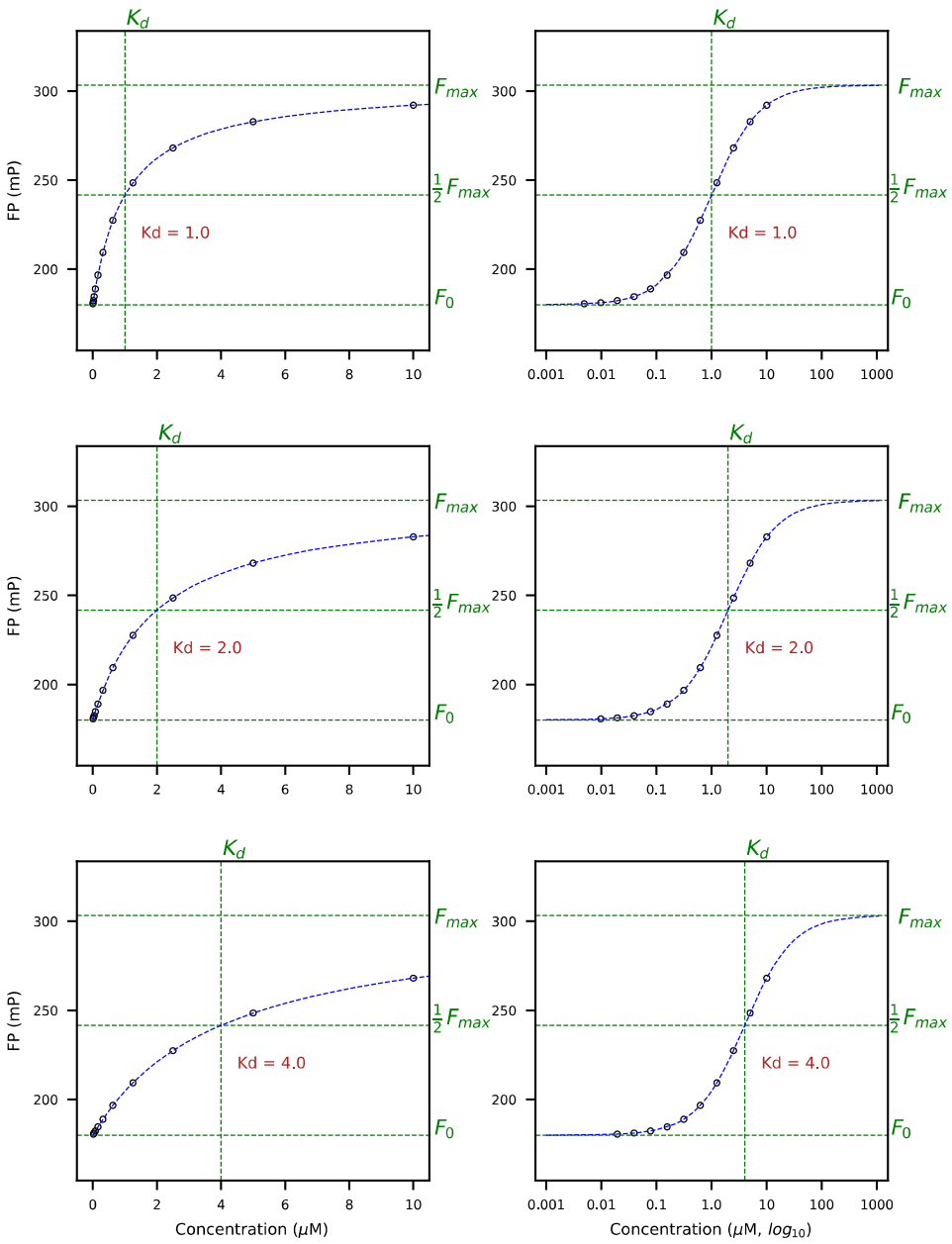

Figure 5: Degradation Effects on Measured Affinity. Simulated measurements for an ideal binding saturation experiment are shown for a theoretical protein with 1 µM affinity (row 1). (Measurements in the second column are the same data as the first column, but plotted on a logarithmic concentration axis.) In rows 2 and 3 a measurement of the same theoretical protein with 1.0 µM affinity is shown, but with the indicated fraction of degraded protein in the sample. The protein degradation is unknown by the experimenter, so the x-axis values erroneously match the full activity sample. The y-values are reduced in rows 2 and 3 because the FP signal has been adjusted to that of the active concentration. The y-values change in a non-linear fashion as governed by the fully active saturation curve (row 1). For example, in row 2 with 50% degradation, the FP measurement (y-value) at 10µM is equivalent to the FP value at 5uM in row 1. In row 3 with 75% degradation, the FP measurement (y-value) at 10µM is equivalent to the FP value at 2.5uM in row 1. When affinity is derived from these degraded protein measurements, the result is an erroneous weaker affinity as shown. Although the change in FP from degraded protein is non-linear, the error in affinity is linear and proportional to the concentration error (and inversely proportional to the K_d).

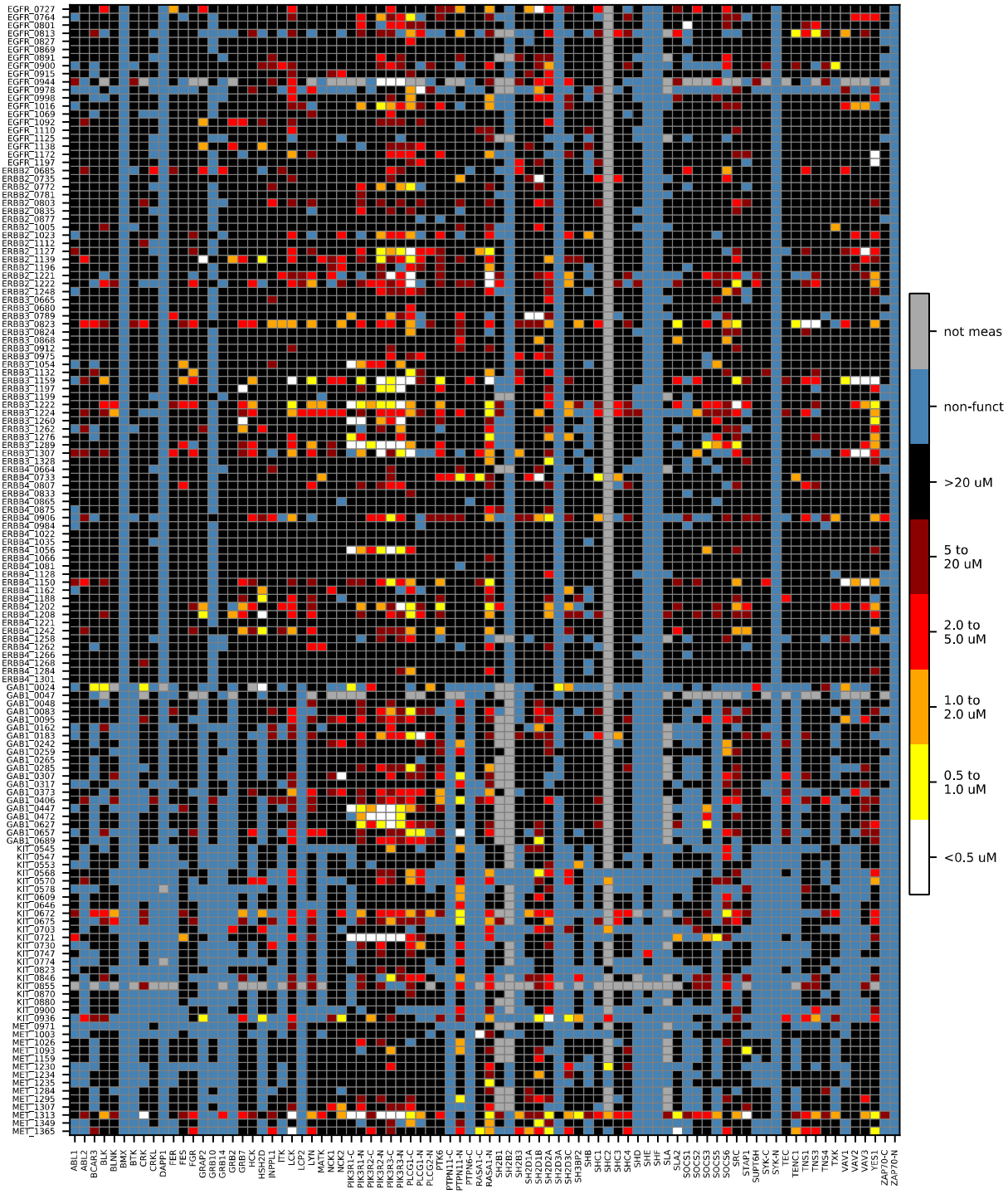

Figure 6: Revised Analysis Final Results. A heat map showing the final results of the revised analysis. A significant fraction a measurements demonstrated patterns consistent with non-functional protein and were removed from the analysis. Comparison with the original published results can be seen in Supplemental Figures 14 and 15.

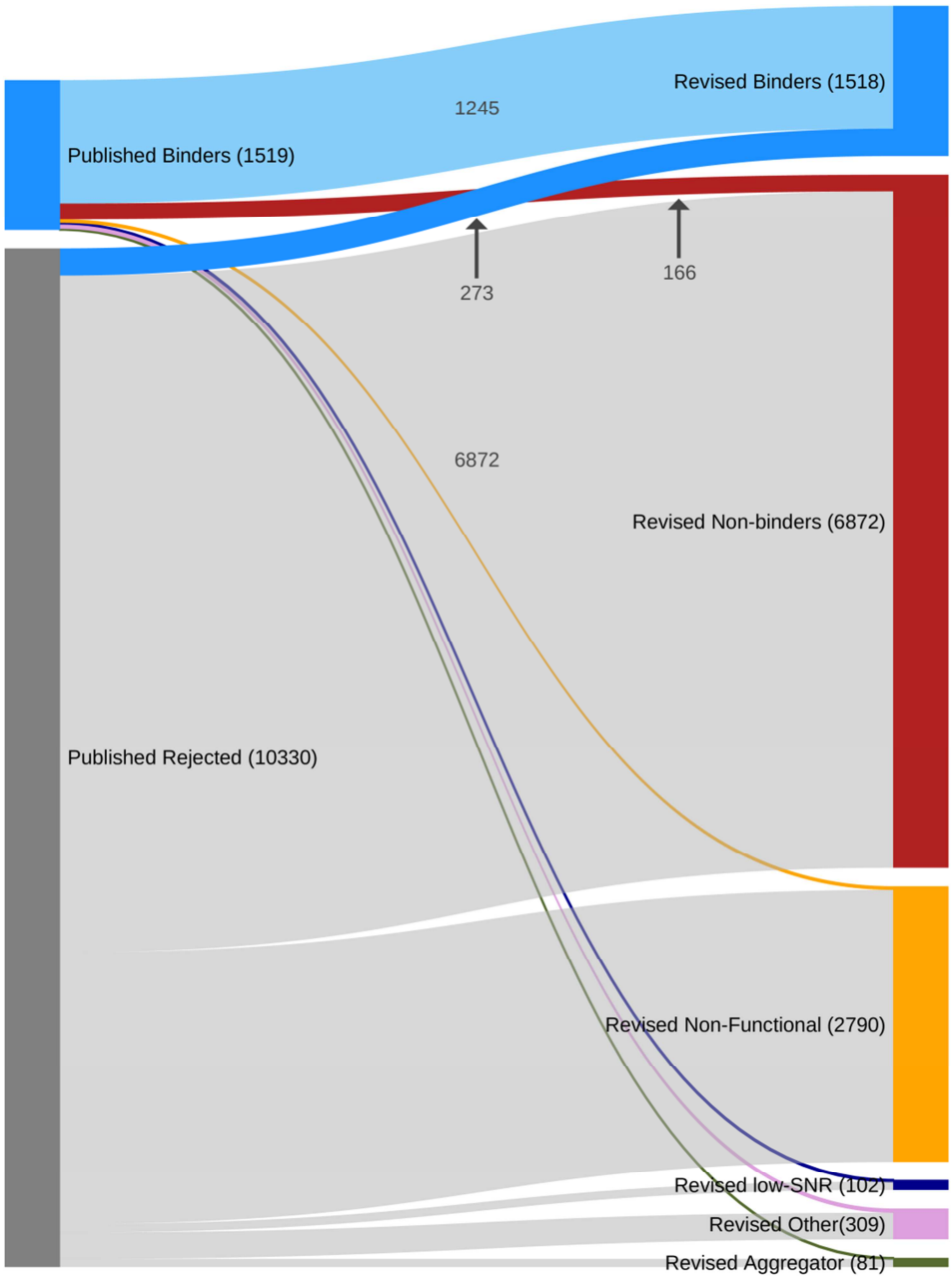

Figure 7: Changes In Calls Between Original Publication and Revised Analysis. Although the numbers of positive interactions are similar in our revised analysis, the identities of those interactions have changed significantly. The changes in calls are visualized in the Sankey map above. Of the original 1519 positive interactions found by the original authors, 166 (10.9%) were found to be non-binders in our analysis. Of the 10330 rejected interactions from the original publications, 273 (2.6%) positive interactions were recovered in our analysis.

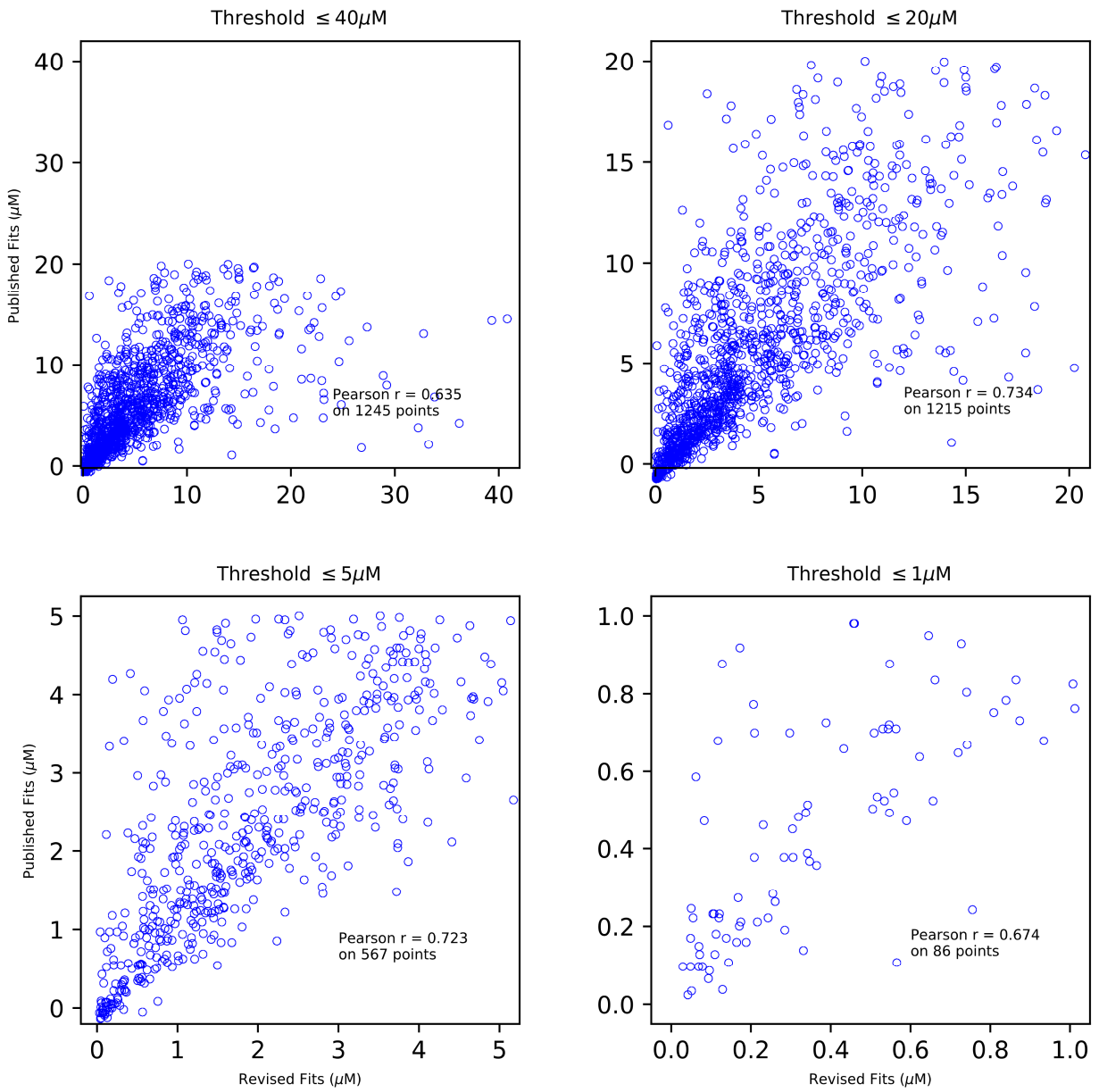

Figure 8: Correlation between Original Publication and Revised Analysis. Affinity values were compared for for the comomn set of positive interactions (n=1245, upper left panel), as well as at lower affinity thresholds (other panels, as indicated). Our revised affinity values correlate only moderately with the original publication (Pearson r=0.635), which is surprising considering the analysis is on the same raw data. Our revised results correlate best when considering all measurements under 20μM affinity (Pearson r = 0.734). Despite choosing the minimum measured value for K_d, our revised data often reports 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 before the mean or minimum are taken. It is interesting to note that correlation does not continue to 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 μM. 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 degraded protein.

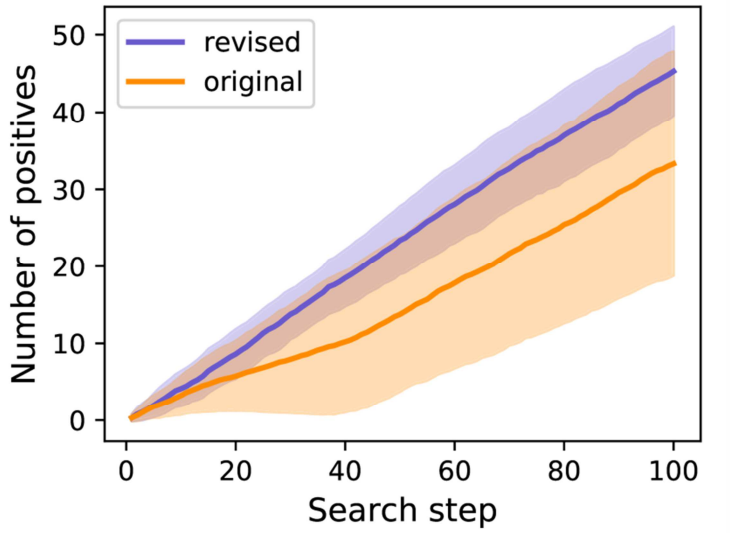

Figure 9: Enhanced Nonmyopic Active Search (ENS) Results. Performance of the active search algorithm ENS within each dataset (original or revised). The line represents mean result, with shading captures +/- standard deviation. In this context, ENS seeks to select each successive interaction such that the total number of positive interactions discovered is maximized.

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

Tom Ronan[1], Roman Garnett[2], and Kristen Naegle[3]

Supporting Information

Final Revised Affinity Data.xlsx – contains the interaction affinities between domains and phosphopeptides based on our revised analysis.

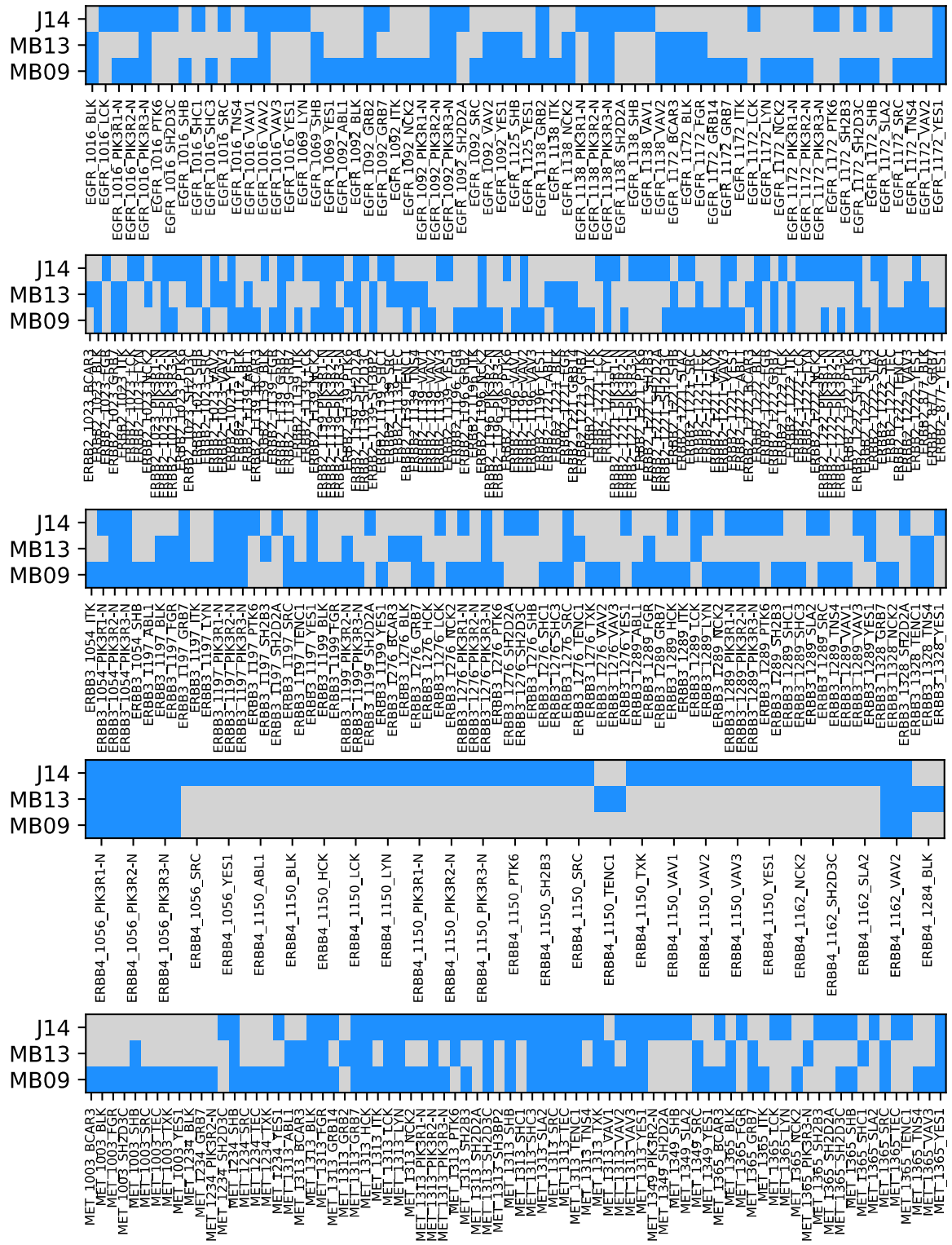

Fig. S1: Domain-Peptide-Level Comparison of Binding Between Published Results. Blue denotes positive interations (binding). J14: Jones 2012-14; MB13: MacBeath 2013; MB09: MacBeath 2006-09.

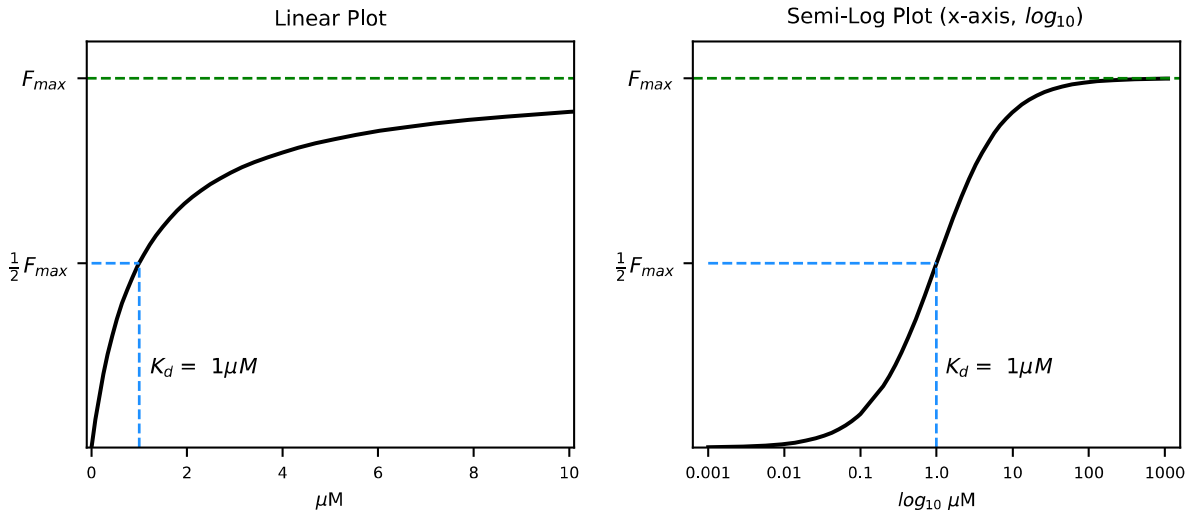

Fig. S2: The Receptor Occupation Model. Saturation plots in both linear (left) and semi-log (right) demonstrate increasing saturation with increasing fractional occupancy, with full saturation achieved at F_{max}. The affinity (K_d) can be derived by fitting a curve to the data, but can also be derived graphically as seen above. At equilibirum, the affinity is equal to the concentration when ½ of the receptor is occupied (½ F_{max}). In the semi-log curve, where the concentration axis is in $\log_{10}$ scale, K_d can be identified easily because it corresponds with the inflection point at the center of the the s-shaped curve.

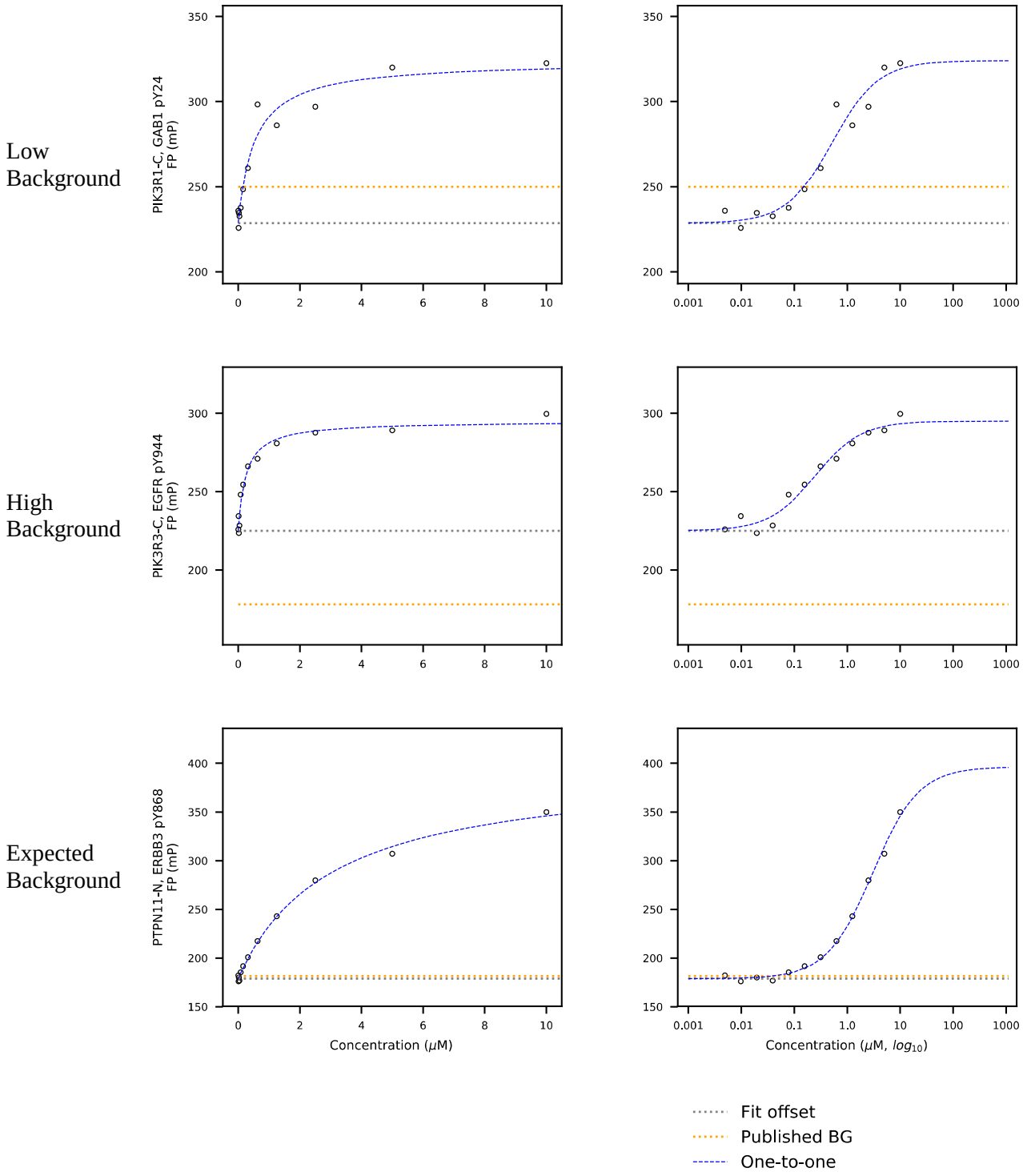

Fig. S3: Examples of Varying Background Fluorescence Polarization. The raw data exhibited highly varying levels of change in polarization (FP) for background samples containing no protein domain. In many cases, background values were higher than measurements (first row) which appeared to have high signal-to-noise ratio. In other cases, the background values were much lower (second row) than the expected value (third row). Rather than use the background levels as reported, we fit the the intercept simultaneously with fitting the model to the data.

$K_d \approx 0.1\ \mu M$

$K_d \approx 1.0\ \mu M$

$K_d \approx 5.0\ \mu M$

$K_d \approx 20.0\ \mu M$

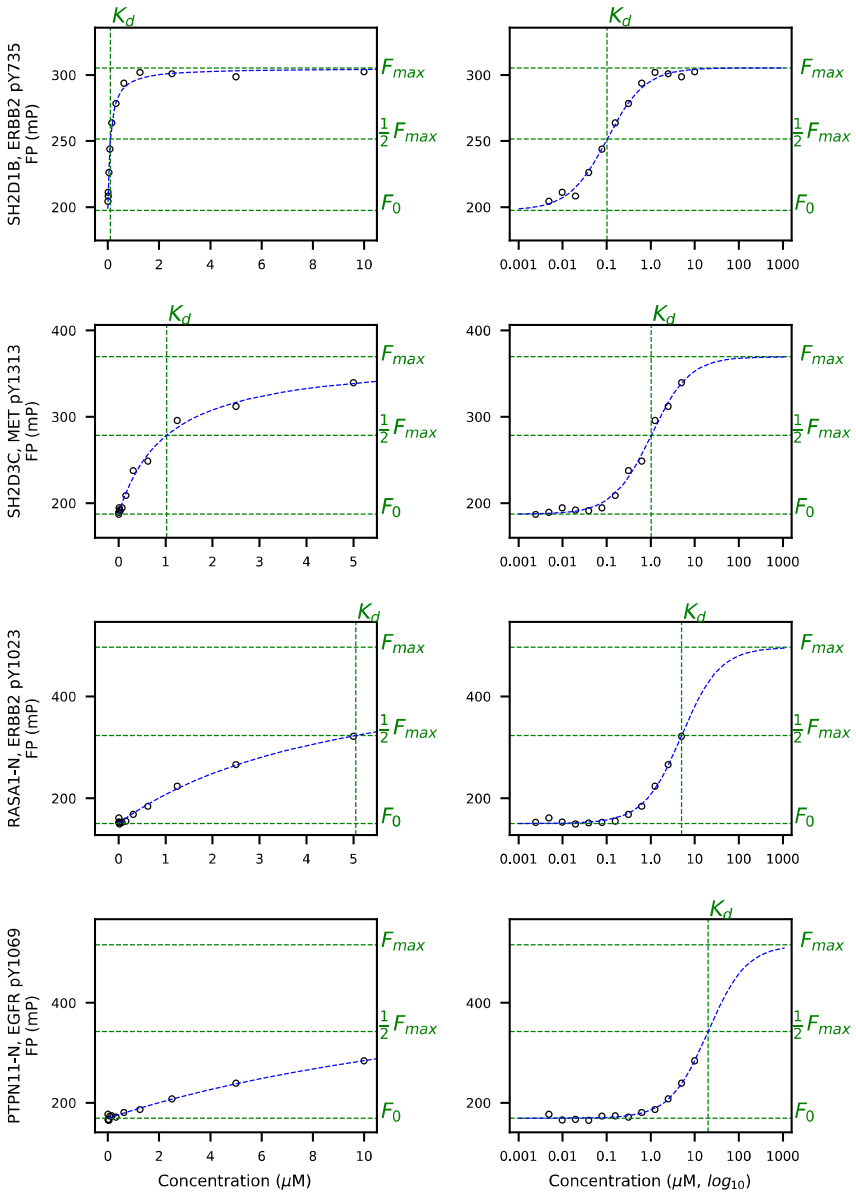

Fig. S4: Receptor Occupancy Model Fits for Various Affinity Interactions. High-quality receptor occupancy model fits showing positive interactions at varying affinities (K_d) from 0.1µM to 20µM. For interactions with 0.1 µM affinity, data points are evenly distributed on either side of the inflection point (semi-log plot), begin to establish no-signal level, and establish saturation well. As affinity decreases (K_d increases), saturation is more poorly defined, with coverage by fewer or no points. Thus, the concentration ranges chosen make this experiment best suited to identify affinity in the 0.05µM to 0.5µM range. Since data in the original publication was reported up to 20µM, results with low affinities (higher K_d values) are likely to be less accurate.

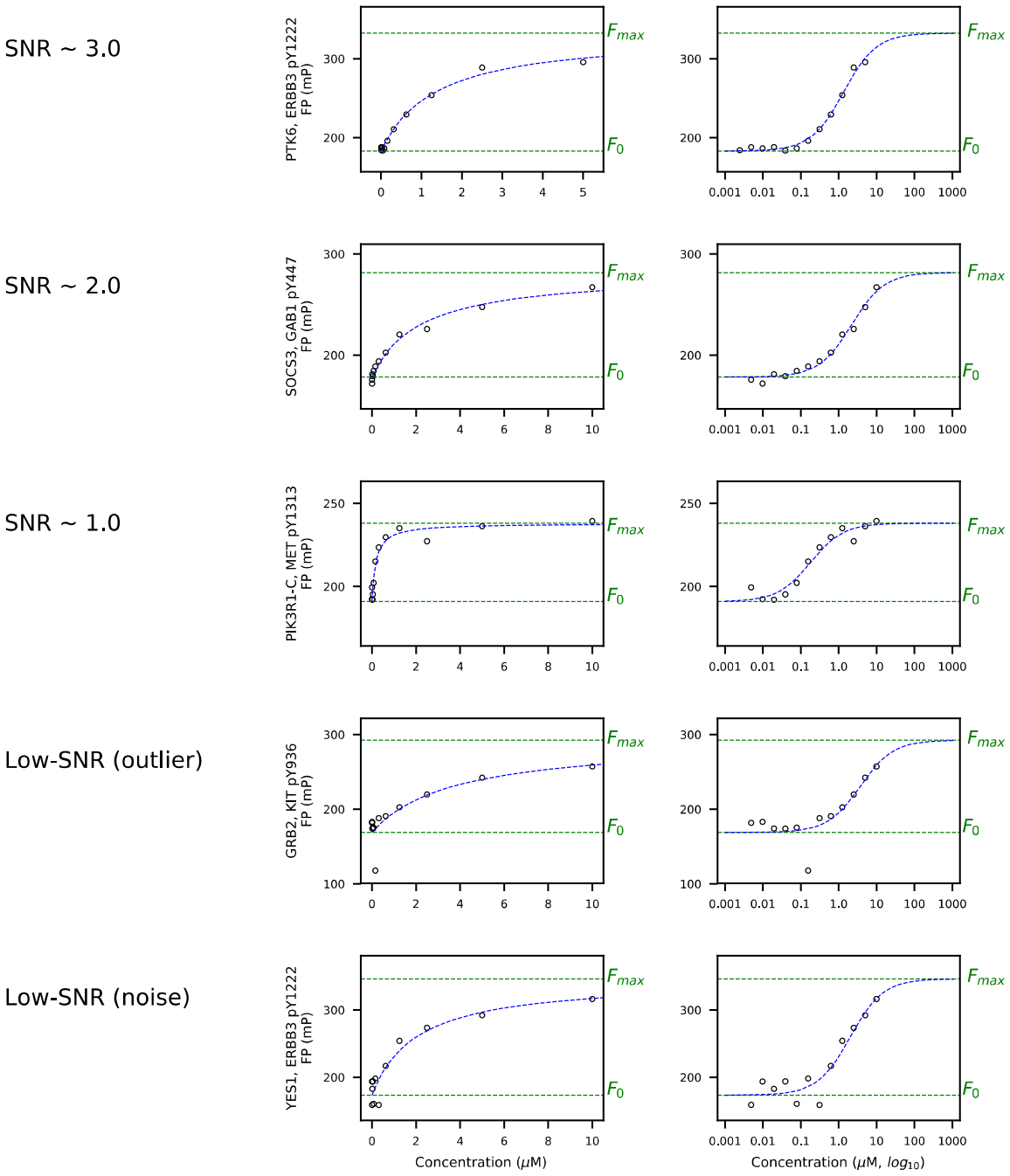

Fig. S5: Signal-to-Noise Ratio (SNR). The quality of fit metric, deemed signal-to-noise ratio (SNR), evaluates the magnitude of residual errors of fit to the model (a form of noise), and weights this sum by the overall size of the fluorescent signal measured. If the SNR is below one (such that the noise/model errror is larger than the signal of the model, the data is rejected. As can be seen from the examples above, a signal to noise ratio of 1.0 or greater represents high-quality fits to the model, with little deviation from the model fit line. An SNR below 1.0 tends to represent fits with either large numbers of deviations, or a single large outlier, and are removed from consideration. Because it is difficult to differentiate between an outlier and a noisy fit with this metric, all measurement are rejected to maintain the quality of the positive interactions.

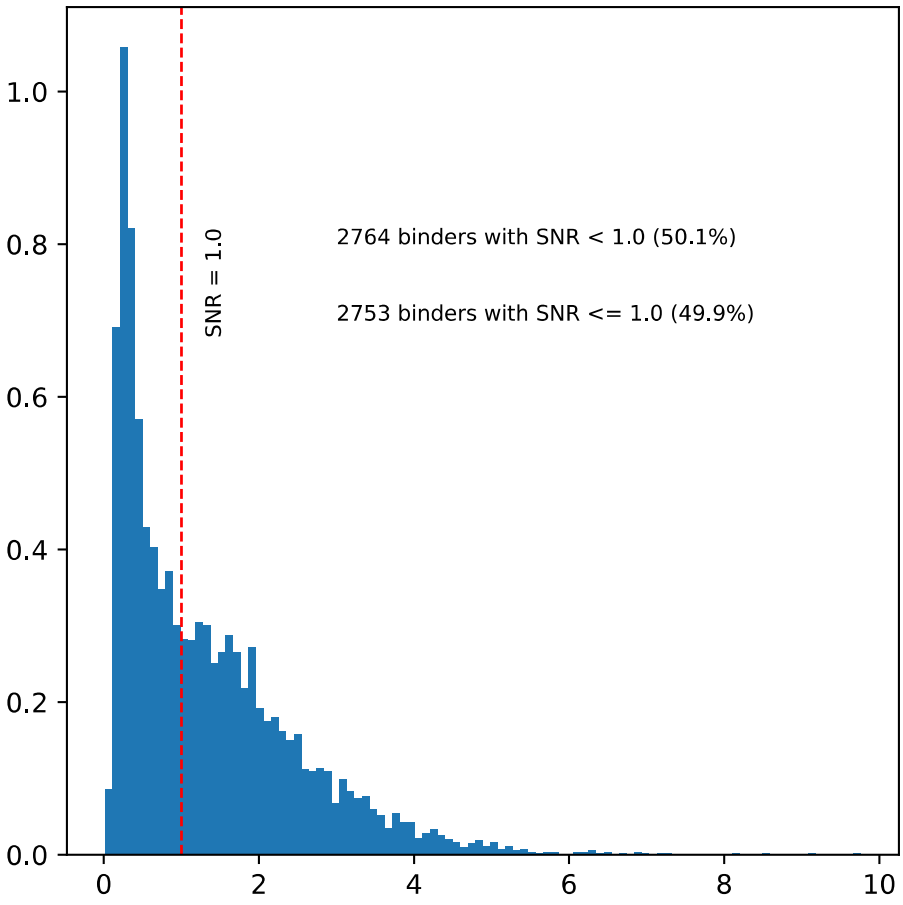

Fig. S6: Signal-to-Noise Ratio Distribution for Replicates Classified as Binders. At a signal to noise ratio (SNR) of 1.0, 50.1% of all replicate measurements which are fit best by thr receptor occupancy model represent high-quality interactions and are considered positive interactions. The remainder, which have an SNR < 1, are considered poor fits, and are removed from consideration.

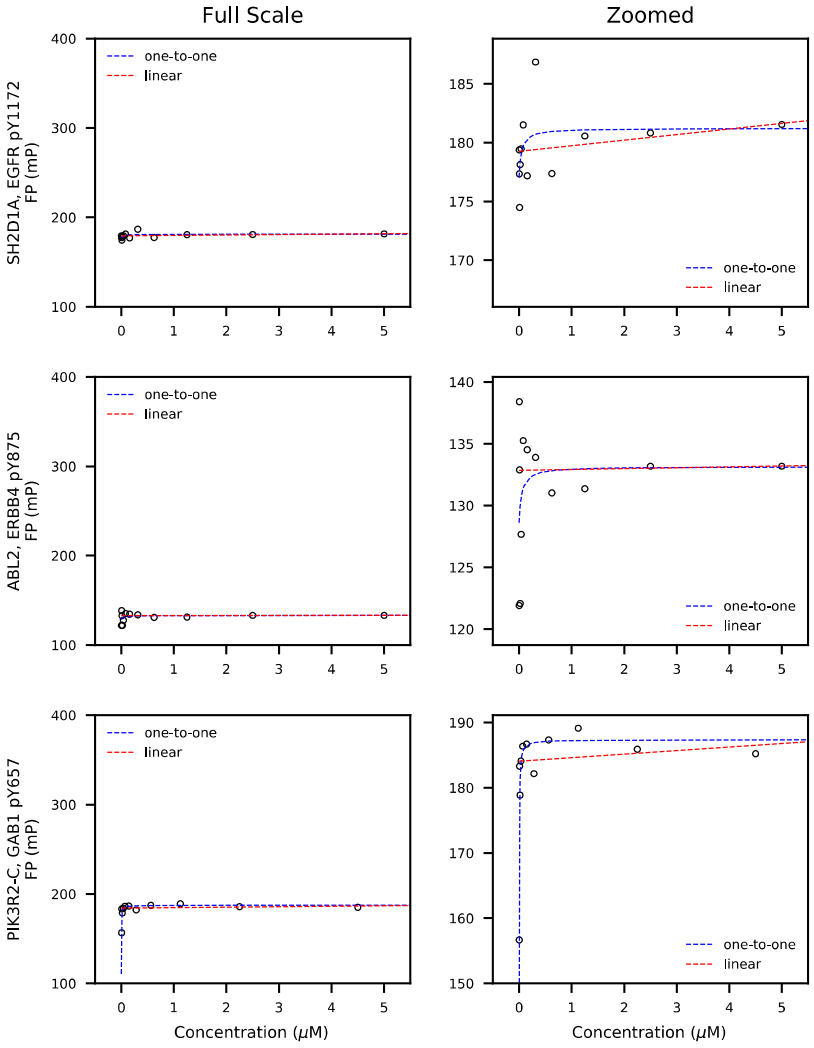

Fig. S7: Model Fitting Results for Non-binding Interactions. The receptor occupancy model fails to fit non-binding data well in practice. Non-binding data is expected to be represented by data points in a horizontal line, with some small level of superimposed noise expected (see first column for examples at full scale). Upon close examination (second column, zoomed view of the same data in the first column), noise in individual data points can be more clearly visualized. Random noise in the measurements can cause the model to force fit a 'saturation'-type curve, resulting in several fit artifacts. In one type of fit-artifact (top row), a saturation curve poorly fits the data and has a low saturation value (on the order of 5mP units). In another type of fit-artifact (bottom row), all but one data point is considered to be at saturation, while one single point sets the rest of the saturation curve, resulting in an artifically low K_d, on the order of 1nM and sometimes much lower. Most commonly, a situation between these two extremes is found (middle row). A linear model (red dashed line) has a lower AICc score when fitted to these non-binding cases than the receptor occupancy model, and can be used to reliably identify non-binding interactions while avoiding artifacts like the ones demonstrated.

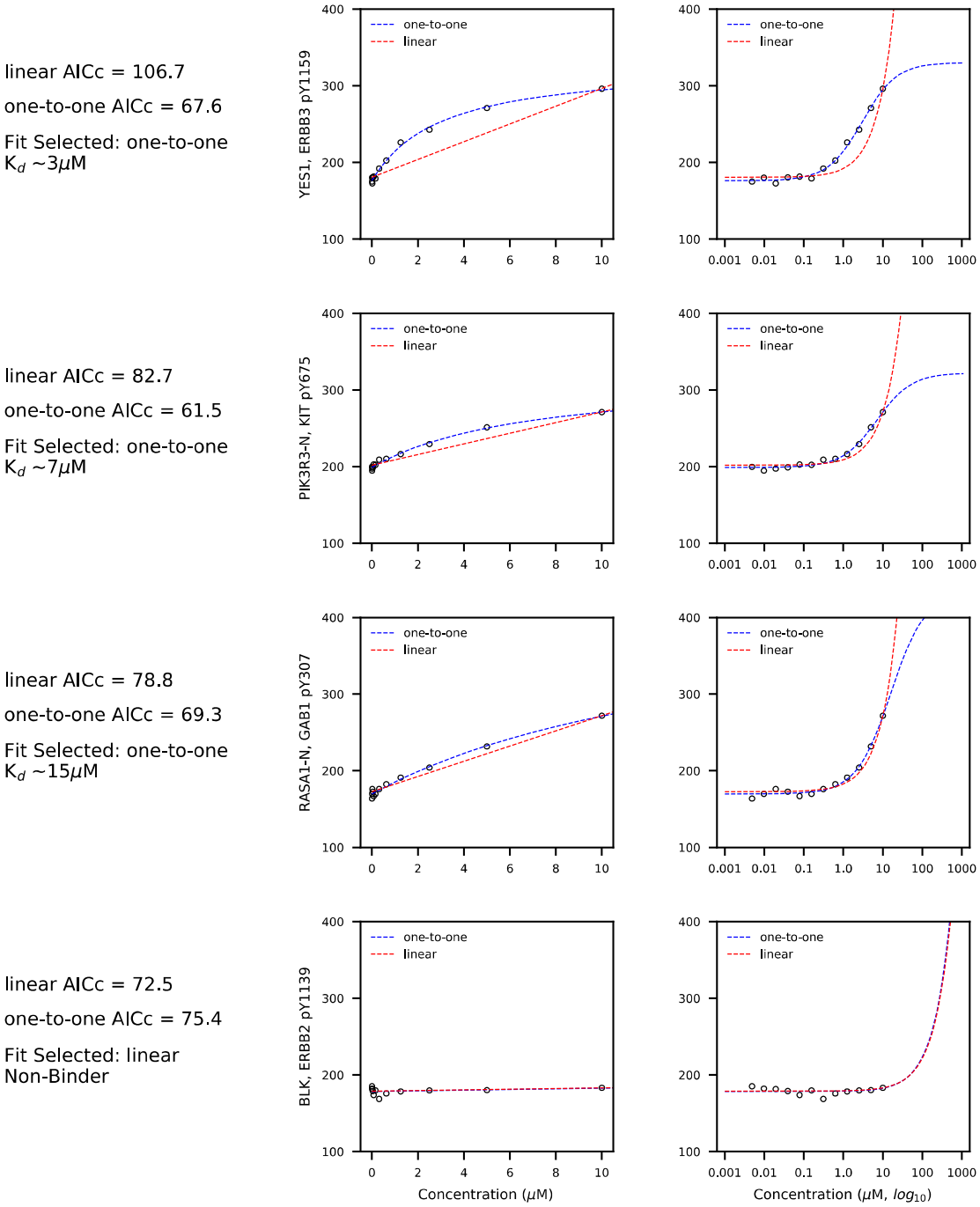

Fig. S8: Model Selection. Both linear and receptor occupancy models are fitted to the data. AICc scores are calculated and compared between models – the model with the lowest AICc score is selected as the best fit. If a linear fit is chosen, and the slope is less than 5mP/µM, the interaction is classified as a non-binding interaction.

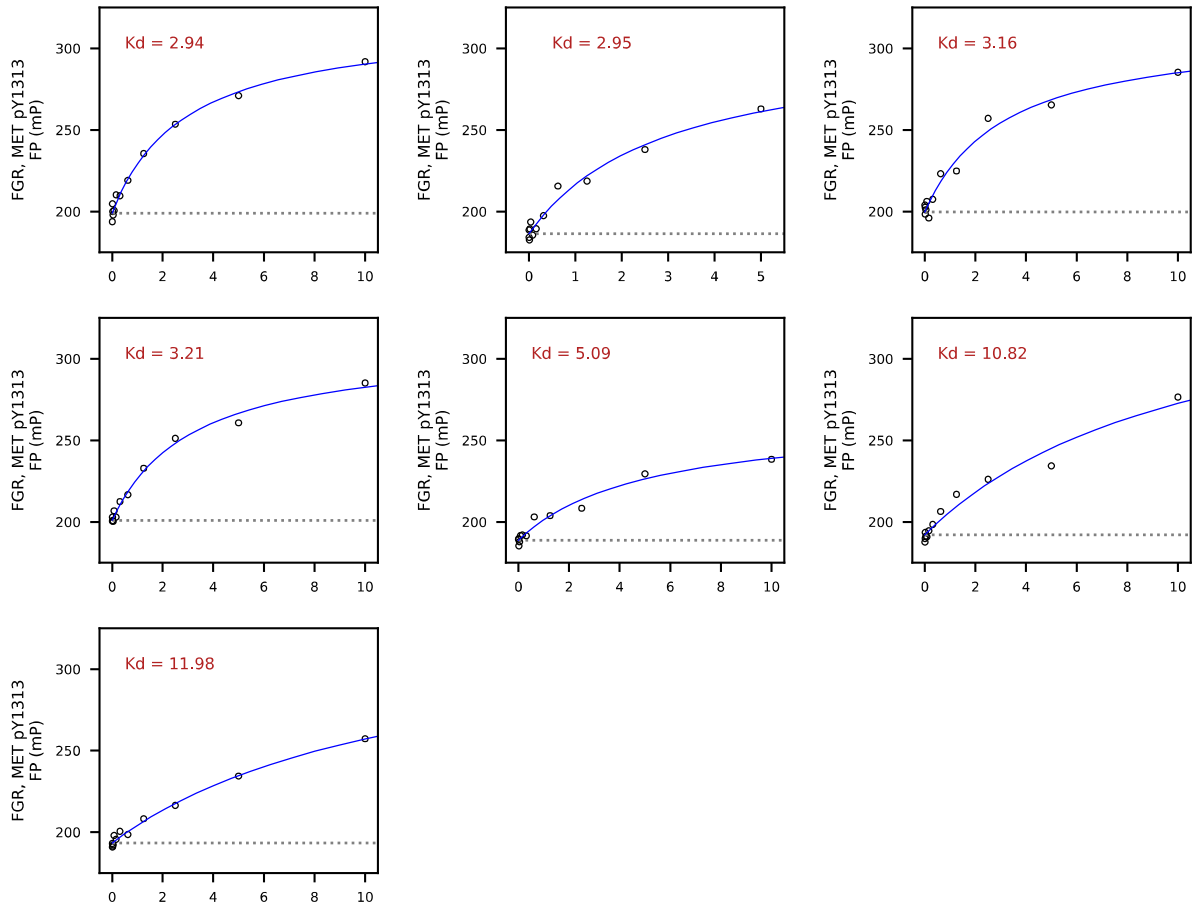

Fig. S9: Replicate Measurements for FGR Interactions with MET pY1313. An example of high-replicate variation across replicates for a single domain-peptide pair. Each individual measurement represents a high-quality fit to the receptor occupancy model, yet the resulting affinities vary from ~3µm to ~12µM. It is clear from the quality of each measurement that the variation is not due to noisy data, or fitting artifacts. Rather, each measurement seems to be a high-quality result of different affinity behavior.

Domain: PIK3R2-N

	Run 1		Run 2		Run 3	
Peptide	**Plate**	**K_d**	**Plate**	**K_d**	**Plate**	**K_d**
ERBB4 0807	129		199		222	8.59
ERBB2 1127	133	3.08	203		226	0.99
					227	1.18
ERBB4 1056	134	2.03	204	0.82	230	
ERBB4 1056	137	4.23	207		231	
ERBB3 1159	136	0.26	206	0.26	229	0.09
ERBB3 1159	139	0.70	209		233	0.23
ERBB3 1307	140	0.58	210	0.26	234	0.25
ERBB4 1150	144		214	3.66	238	3.65
ERBB2 0772	148	1.54	218		242	2.19
EGFR 0764	152	2.49	223	2.27	246	1.16
ERBB3 0823	154		225		248	4.92
EGFR 1092	160		232		254	13.55
ERBB3 0868	167		239		261	2.32
EGFR 1016	168	3.67	240	0.94	262	
ERBB2 1023	242		244		265	2.22
ERBB3 1054	243	3.14	245		266	
ERBB3 1222	250		252	0.81	273	0.84
ERBB3 1289	257	10.68	259		282	0.16
ERBB4 1202	266		268		292	1.03

Domain: SH2D2A

	Run 1		Run 2		Run 3	
Peptide	**Plate**	**K_d**	**Plate**	**K_d**	**Plate**	**K_d**
ERBB4_1202	61	3.08	186	5.53	206	4.35
ERBB4_0807	71	6.90	199		222	21.91
ERBB3_1307	82	12.28	210		234	55.47
ERBB3_0789	87	13.74	215		239	9.50
ERBB4_1262	92	15.13	220		244	9.42
ERBB4_0906	93	12.86	222		300	0.88
ERBB3_0975	101	7.56	231		253	20.46
EGFR_1092	102	5.29	232	3.23	254	13.20
EGFR_0900	105	7.50	235	18.59	257	1.96
ERBB2_1139	106	3.51	198	7.50	220	1.70
			236		221	1.44
					299	3.67
EGFR_0915	108	4.46	238	5.55	259	3.89
ERBB2_1221	213	2.80	249		270	18.26
ERBB3_1222	216	3.12	252		273	21.60
ERBB4_1188	219	9.09	255		276	7.78
ERBB3_1328	226	0.86	262		285	5.90
ERBB4_1202	233	4.36	268		292	1.23
ERBB4_1208	237	5.42	272	13.81	216	1.98

Domain: RASA1-N

	Run 1		Run 2		Run 3	
Peptide	**Plate**	**K_d**	**Plate**	**K_d**	**Plate**	**K_d**
ERBB4 1150	3	2.19	185	0.56	205	0.61
ERBB4 1202	4	15.16	186		206	4.32
ERBB4 1208	11	37.85	194	7.64	216	6.41
ERBB2 1005	19	10.87	202	1.38	225	5.14
ERBB2 1127	20	1.64	203	1.16	226	0.96
ERBB3 1159	23	0.31	206	0.44	229	0.19
ERBB3 1159	26	4.15	209	4.95	233	2.99
ERBB3 1307	27	1.77	210	1.60	234	1.93
ERBB4 1150	31	1.05	214	0.96	238	1.33
ERBB2 0772	35	6.05	218		242	9.36
ERBB4 1262	37	1.14	220	4.79	244	
ERBB4 0906	38	9.69	222	7.17	300	1.17
EGFR 0764	39	9.94	223		246	22.16
ERBB3 0823	41	2.84	225		248	2.23
ERBB3 0897	50		235	10.04	257	2.10
EGFR 1016	55	1.43	240	1.37	261	2.24
ERBB2 1023	174	5.05	244	10.55	265	9.31
ERBB4 1162	176	2.29	246	5.21	267	4.14
ERBB2 1196	178	1.62	248		269	1.71
ERBB2 1221	179	0.24	249	1.95	270	1.86
ERBB2 1222	180	0.40	250	6.89	271	2.78
ERBB3 1222	182	0.64	252	2.87	273	1.83
ERBB3 1224	183	0.75	253	2.16	274	1.25
ERBB3 1262	187	3.69	257		280	13.43
ERBB3 1289	189	3.45	259		282	4.35
EGFR 0998	190	1.30	260	7.82	283	2.82
ERBB3 1276	191	0.83	261	1.36	284	2.39
ERBB3 1328	192	1.83	262	16.03	285	9.10
EGFR 1172	193	1.31	263	5.72	286	5.37
EGFR 0727	194	3.01	264		287	6.33
ERBB4 1202	199	1.31	268	3.16	292	1.00
ERBB4 1242	204	1.53	273	1.79	297	5.90

indicates the lowest Kd (highest affinity) across the 3 runs for a domain-peptide pair.

Fig. S10: Examples of Degradation Patterns in Domain Data. In the original publication, data was primary gathered on 3 runs on 3 different days. On each run, domains were tested against hundreds of peptides providing rich data for identifying patterns. Variance in affinity from random (non-systemic) sources should manifest independent of run or sample order. In contrast, variance from protein degradation would demonstrate specific, non-random patterns in affinity. Degraded protein on a run would manifest as variance between runs, but consistently higher K_d on the degraded run across all peptides. For PIK3R2-N, we see that Run3 replicates consistently showed lower K_d values (higher affinities) than replicates from other days. This pattern of run to run variation suggests that the protein samples tested in Runs 1 and 2 were degraded. A protein sample exhausted mid-run and replaced with a fresh sample, could manifest as a surge of increased affinity in the middle of a run of lower affinity. For RASA1-N, no single day dominated the highest affinity until plate 174, after which the highest affinity replicates all come from Run 1. This is consistent with a change to fresh, active protein during Run 1. These patterns are not compatible with a random source of variance. Not all protein data shows pattern consistent with this degradation hypothesis. For SH2D2A, significant variation appears during each run. The patterns for SH2D2A are not consistent with a simple degradation hypothesis, and may be indicative of additional sources of variation.

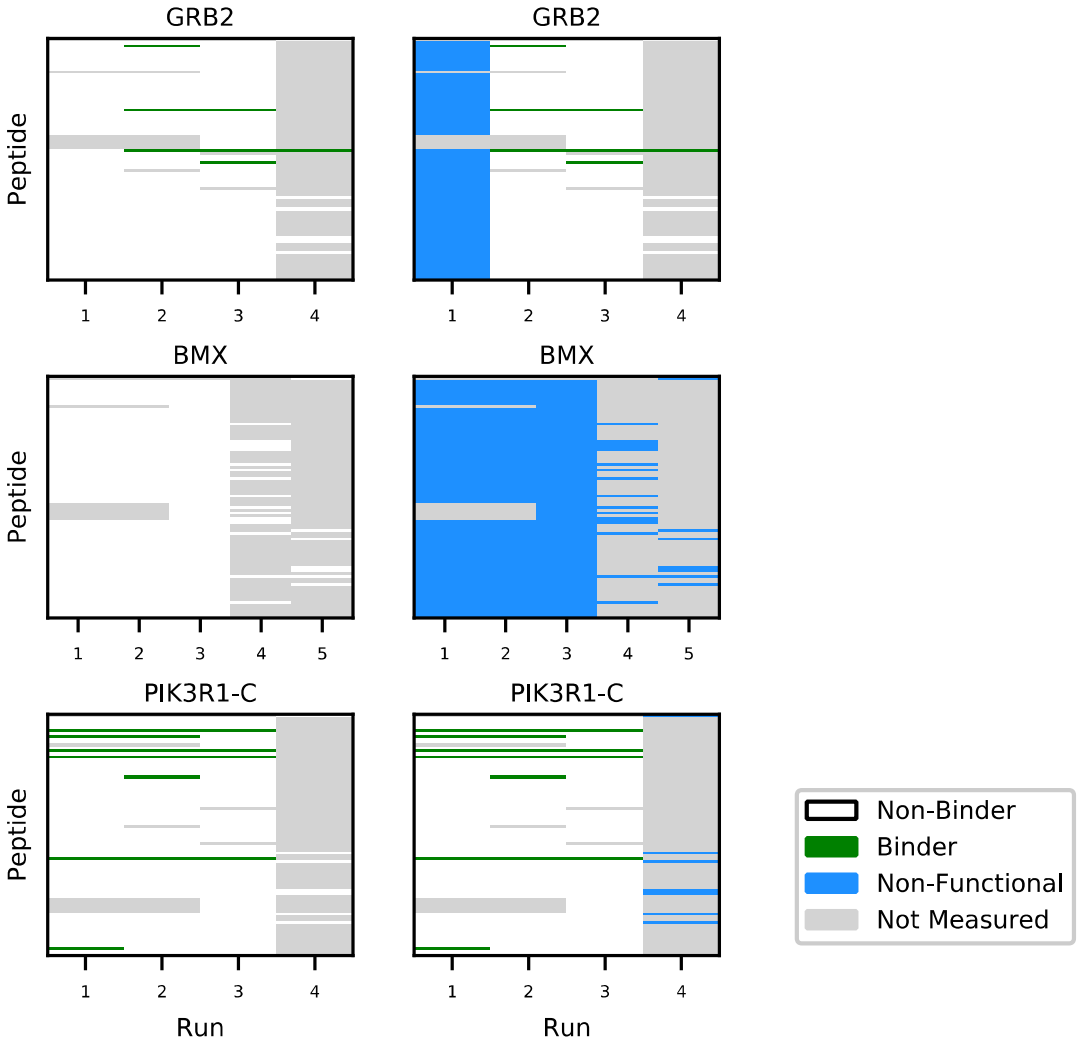

Fig. S11: Non-Functional Protein Identification – Examples. Non-functional protein identification (NFPI) can be made when plotting protein domain activity against many peptides across multiple runs. The left column figures show results before NFPI, and the right column figures show the results of NFPI on the same data. A binder is identified by a green cell, a non-binder by a white cell, non-functional protein by a blue cell, and a non-measured interaction by a gray cell. A lack of even one positive interaction on an entire run is suggestive of non-functional protein. When other runs of the same protein show positive interactions, the runs with no positive interactions are considered to be non-functional and are removed from consideration. For example, with GRB2 (row 1), runs 2 through 4 showed some positive interactions. On run 1, however, no measurements indicated positive interactions. The lack of even one positive interaction in run 1 suggests that the protein was completely degraded or non-functional, and the presence of positive interactions in the other runs acts as a positive control. Run 1 is then marked in blue in the right panel for GRB2, and removed from consideration. A less-clear case of non-functional protein can be seen with BMX (row 2). For BMX, no positive interactions were found on any run. Although it is a formal possibility that BMX simply binds none of these peptides, we simply have no information that the protein was ever active, thus we conservatively identify all runs as non-functional. For PIK3R1-C, no measurements on the fourth run were positive interactions, while other runs contain positives, thus run 4 was categorized to be non-functional.

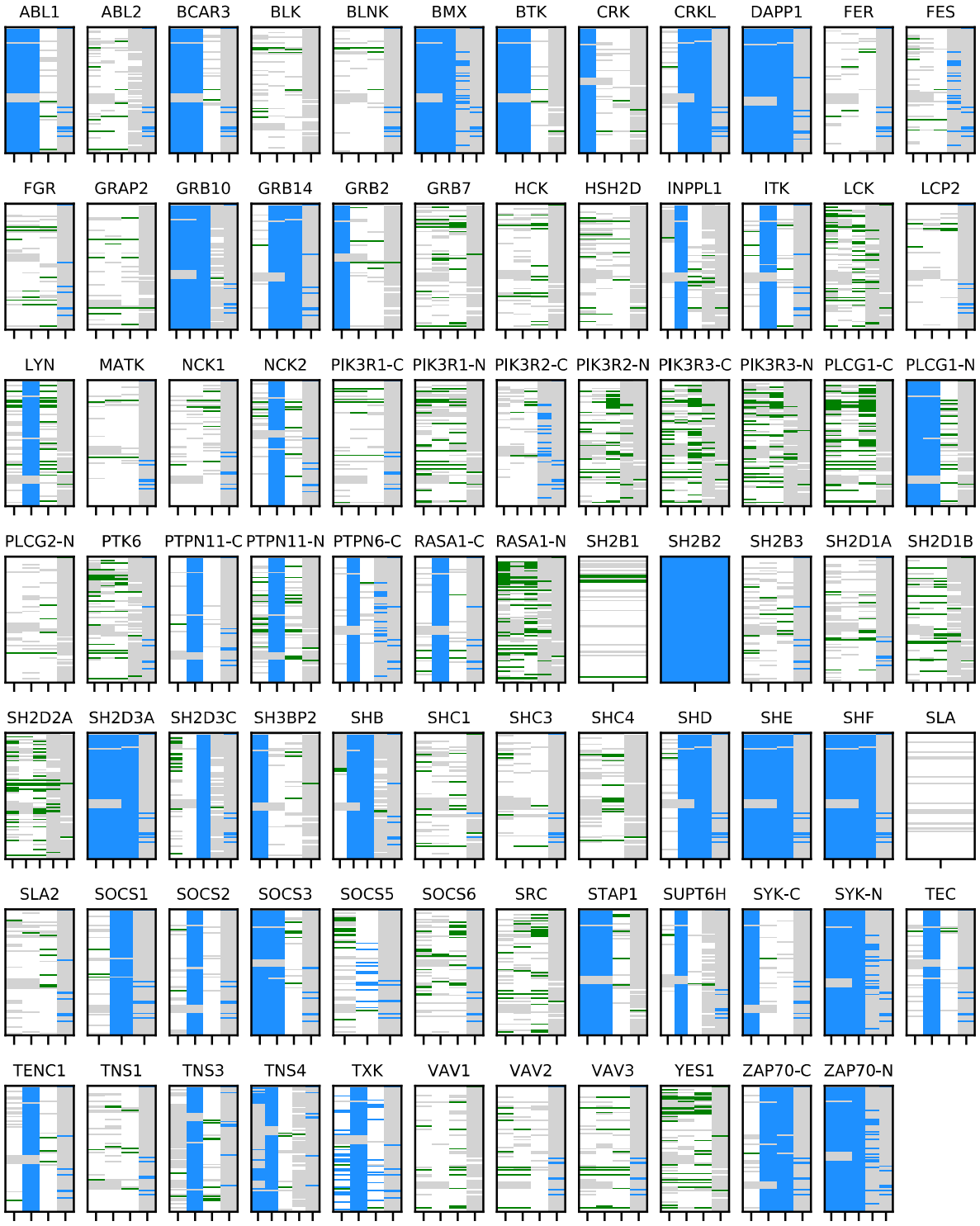

Fig. S12: Non-functional Protein in Hause, et al (2012). Non-Functional protein results for all measured interactions from the first publication, Hause, et al (2012). See legend from Fig. S11.

Fig. S13: Non-functional Protein in Leung, et al (2014). Non-Functional protein results for all measured interactions from the second publication, Leung, et al (2014). See legend from Fig. S11.

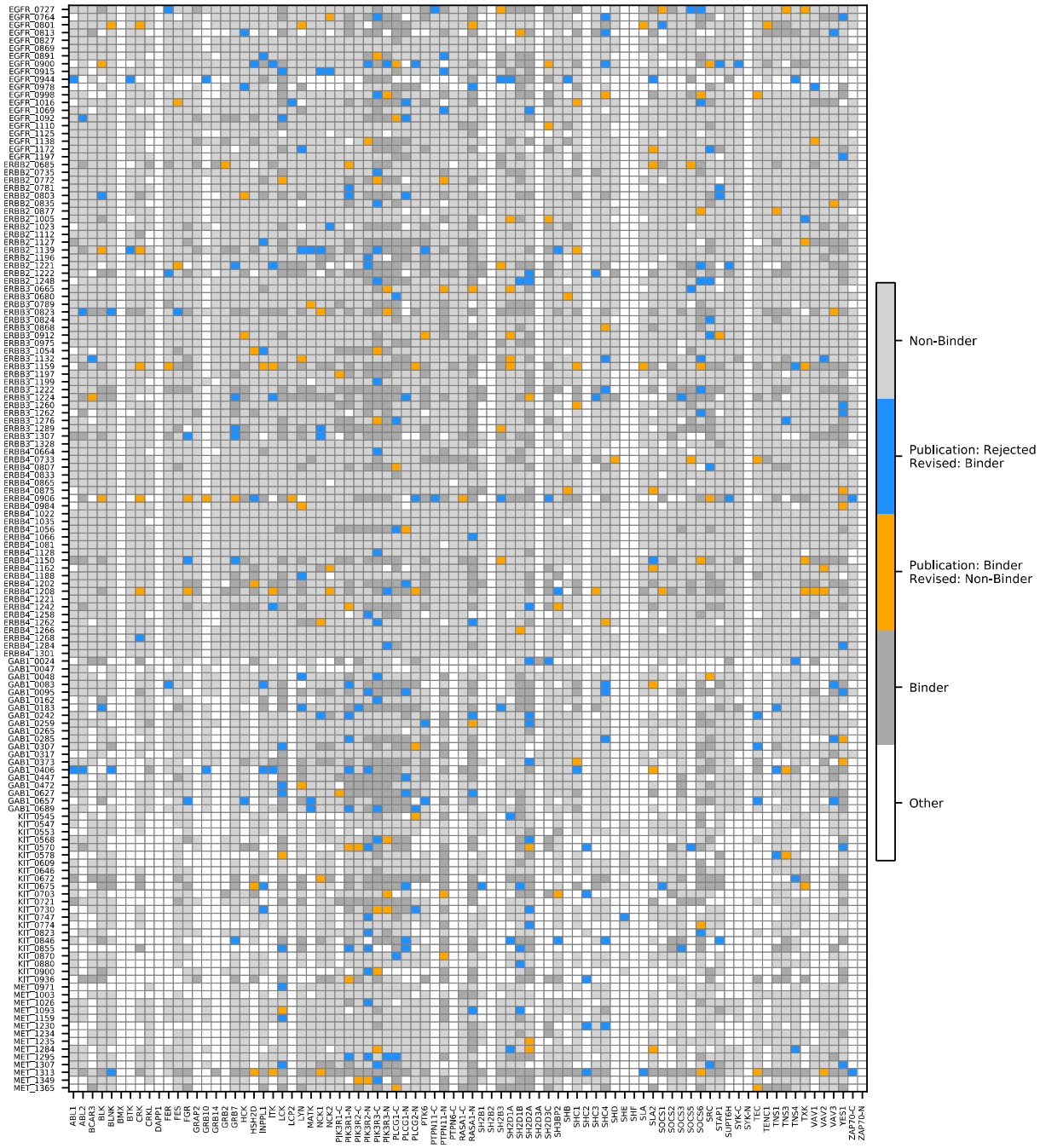

Fig. S14: Changes In Calls Between Original Publication and Revised Analysis. A heat map showing the changes in calls in our revised analysis. Differences in calls with the original publication are found across all domains and all peptides.

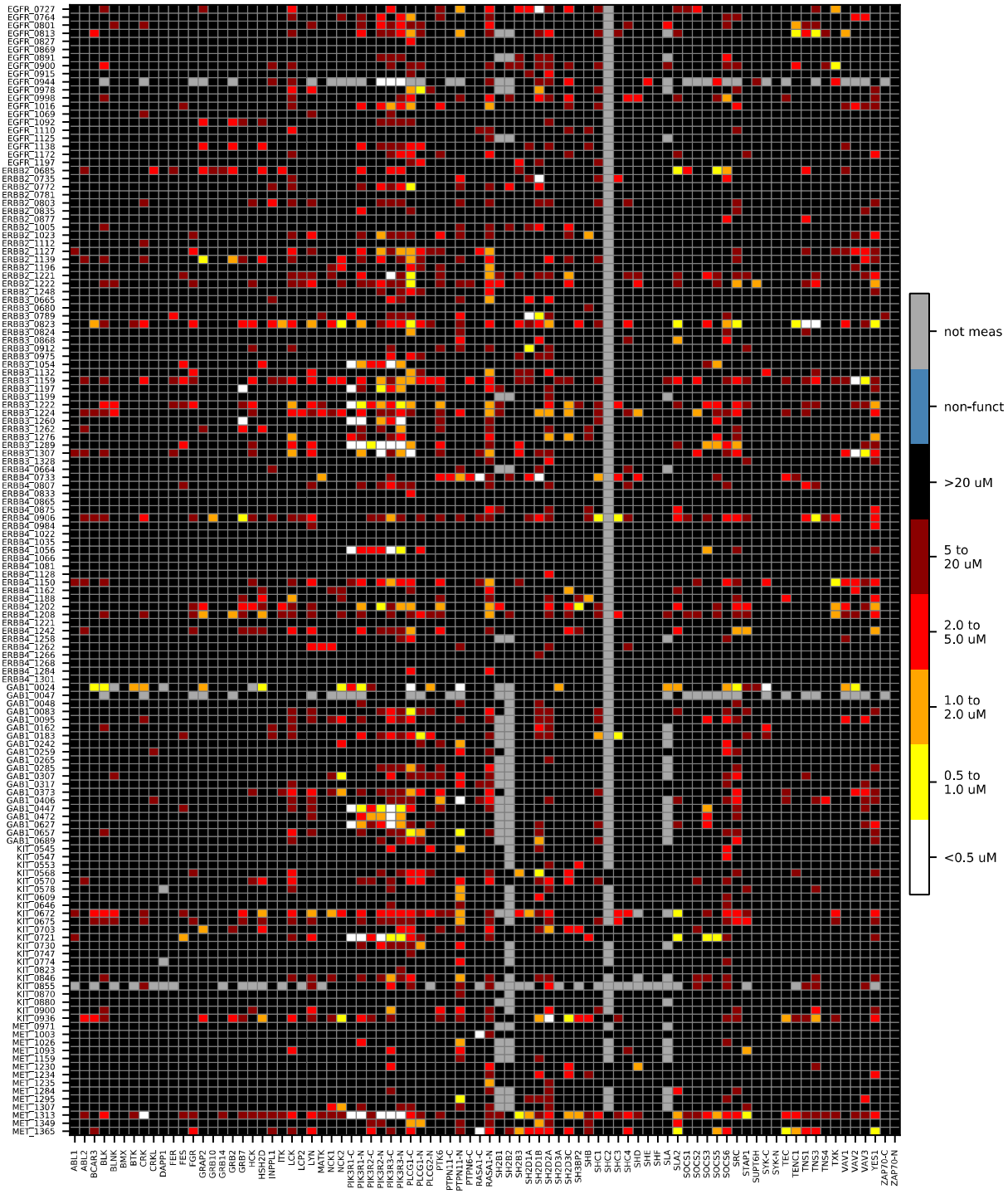

Fig. S15: Results from the Original Publication. A heat map showing the original published results in the same format, sorting order, and naming convention – for comparison with our revised analysis.