Winter 12-15-2017

Binding Affinity and Specificity of SH2 Domain Interactions in Receptor Tyrosine Kinase Signaling Networks

Tom Ronan
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering & Applied Sciences
Department of Biomedical Engineering

Dissertation Examination Committee:
Kristen Naegle, Chair
Jan Bieschke
Roman Garnett
James Havranek
Rohit Pappu

Binding Affinity and Specificity of SH2 Domain Interactions in Receptor Tyrosine Kinase Signaling Networks

by
Thomas James Ronan III

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2017
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would first like to give thanks for the welcoming and inspiring environment in the lab where much of this work was conceived. Kristen's infectious enthusiasm and constant willingness to dig into new problems is something I inspire to. I want to specifically thank Roman for being a constant sounding board for new ideas and a willing participant in vigorous discussion – some of it even about science. Thank you to Kathy for making the time in between all of the hard work as fun as can be imagined.

Thanks to great collaborators and mentors for challenging projects and hard questions: specifically Linda Pike, Rohit Pappu, James Havranek, Jan Bieschke, Roman Garnett, Mark Anastasio, and Barani Raman. I would like to express gratitude for financial support from the Center for Biological Systems Engineering (CBSE).

Thank you to my mother for constant support. Finally, a profound thanks to my wife for helping me, and caring for me, and putting up with me on this quest to expand my understanding of this amazing universe.

Tom Ronan

Washington University in St. Louis

December 2017

ABSTRACT OF THE DISSERTATION

Binding Affinity and Specificity of SH2 Domain Interactions

in Receptor Tyrosine Kinase Signaling Networks

by

Thomas James Ronan III

Doctor of Philosophy in Biomedical Engineering

Washington University in St. Louis, 2017

Kristen M. Naegle, Chair

Receptor tyrosine kinase (RTK) signaling mechanisms play a central role in intracellular signaling and control development of multicellular organisms, cell growth, cell migration, and programmed cell death. Dysregulation of these signaling mechanisms results in defects of development and diseases such as cancer. Control of this network relies on the specificity and selectivity of Src Homology 2 (SH2) domain interactions with phosphorylated target peptides. In this work, we review and identify the limitations of current quantitative understanding of SH2 domain interactions, and identify severe limitations in accuracy and availability of SH2 domain interaction data. We propose a framework to address some of these limitations and present new results which improve the quality and accuracy of currently available data. Furthermore, we supplement published results with a large body of negative interactions of high-confidence extracted from rejected data, allowing for improved modeling and prediction of SH2 interactions. We present and analyze new experimental results for the dynamic response of downstream signaling proteins in response to RTK signaling. Our data identify differences in downstream response depending on the character and dose of the receptor stimulus, which has implications

for previous studies using high-dose stimulation. We review some of the methods used in this work, focusing on pitfalls of clustering biological data, and address the high-dimensional nature of biological data from high-throughput experiments, the failure to consider more than one clustering method for a given problem, and the difficulty in determining whether clustering has produced meaningful results.

Chapter 1: Introduction, Background, and Review of the Current State of Quantitative Measurements of SH2 Domain Interactions

1.1 Introduction and Overview

Receptor tyrosine kinase (RTK) signaling mechanisms play a central role in intracellular signaling and control development of multicellular organisms, cell growth, cell migration, and programmed cell death. Dysregulation of these signaling mechanisms results in defects of development and diseases such as cancer. Control of this network relies on the specificity and selectivity of Src Homology 2 (SH2) domain interactions with phosphorylated target peptides. In this work, we identify the limitations of current quantitative understanding of SH2 domain interactions, review analysis methods, present improved methods and identify best practices, and present new data for SH2 interactions, and network behavior.

In Chapter 1, we review the current state of quantitative measurement data for SH2 domain interactions, and identify severe limitations in accuracy and availability. In Chapter 2, we propose a framework to address some of these limitations and present results which improve the quality and accuracy of currently available data. Furthermore, we supplement published results with a large body of negative interactions of high-confidence extracted from rejected data, allowing for improved modeling and prediction of SH2 interactions. In Chapter 3, we present and analyze new experimental results for the dynamic response of downstream signaling proteins in response to RTK signaling. Our data identify differences in downstream response depending on the character and dose of the receptor stimulus, which has implications for previous studies using high-dose stimulation. In Chapter 4, we review some of the methods used in this work,

focusing on pitfalls of clustering biological data, and address the high-dimensional nature of biological data from high-throughput experiments, the failure to consider more than one clustering method for a given problem, and the difficulty in determining whether clustering has produced meaningful results.

1.2 Background

1.2.1 The Reader, Writer, Eraser Paradigm of Phosphotyrosine Signaling

Src Homology 2 (SH2) protein domains are small, 100 amino acid modular domains first identified in the human SRC (sarcoma) proto-oncogene (1). SRC encodes a multi-domain protein with three domains: a Src Homology 3 (SH3) domain, an SH2 domain, and non-receptor protein tyrosine kinase domain. SH2 domains are found in a wide range of signaling proteins. Each SH2 domain interacts specifically protein domains containing a phosphorylated tyrosine residue and flanking residues which convey specificity (2).

In phosphotyrosine signaling mechanisms, like found in receptor tyrosine kinase (RTK) signaling networks, the SH2 domain serves as the 'reader' of signal in the reader, writer, eraser paradigm of signaling (3). When a tyrosine residue is phosphorylated by a kinase (the 'writer' of phosphotyrosine signaling), proteins containing SH2 domains can interact and bind that residue, subject to the compatibility of the flanking peptides. A phosphatase (the 'eraser') can then later remove the phosphate group and terminate the signal.

1.2.2 Role of SH2 Domains in Receptor Tyrosine Kinase Networks

SH2 domains are believed to have evolved at the dawn of multicellularity, and allowed a new, orthogonal, signaling mechanism to develop allowing communication between cells (3). They play a key role defining specificity within RTK networks, which control cell development, migration, and apoptosis (4). When a transmembrane cell surface receptor, such as the Epidermal Growth Factor Receptor (EGFR) binds an extracellular ligand from another cell such as Epidermal Growth Factor (EGF) a signaling cascade ensures translating the extracellular signal

into intercellular response. A ligand-stimulated receptor becomes dimerized, and makes a conformational change. This change causes activation in the intercellular kinase domain, causing cross phosphorylation of specific tyrosine residues the dimerized receptor tails. These tails are targets for SH2 domains. Proteins containing SH2 domains bind these tails bringing additional kinase and scaffolding domains causing a cascade of SH2-mediated phosphotyrosine signaling. These signals integrate through several pathways (such as PI3K, MAPK, JAK/STAT) resulting in changes in gene expression. Dysregulation of RTK signaling networks is a cause of several developmental diseases and forms of cancer (4, 5).

1.2.3 Qualitative Measurements of Binding

Early experiments attempted to classify the binding profile of SH2 domains using degenerate libraries. Resources like SMALI (6) and Scansite (7), were able to identify residues at each position of a phosphorylated peptide that contribute to binding. A limitation of this method is that influence at each position is identified independent of each other position. These results can be combined into a statistical model like a position specific scoring matrix (PSSM) under the assumption that each position acts independently.

Although useful for identifying general binding trends, these models have significant limitations. Since contributions between positions can be interdependent, some interactions cannot be captured by a model assuming independence. Conditionally dependent interactions are believed to play a role in determining selectivity and specificity of interactions. One of the most important manifestations of this phenomenon is ‘non-permissive residues’ – residues in a peptide that disrupt binding despite the presence of other residues correlating with strong binding (8). Predictions from a model based on independence would result in false positive predictions when non-permissive residues were present. A key source for identifying conditionally dependent

interactions is in data showing non-binding or negative interactions, which degenerate libraries do not effectively demonstrate. Furthermore, data like degenerate binding libraries result in qualitative models, and cannot provide quantitative measures of affinity required to predict competition and network outcomes (see Chapter 2.5.1).

1.2.4 Accurate Quantitative Measurements Are Required to Predict Network Outcomes

Interaction of SH2 domains with phosphorylated peptides is the key step in determining signaling outcomes (9). In order to successfully model signaling network outcomes, one must be able to predict the outcome of competition for phosphorylated residues. To predict which interactions occur, one must know what domains are present that have affinity for the phosphorylated site, the strength of the affinity, and the effective concentration of the domains available to interact with the phosphorylated site. Concentration and identity of interaction partners is likely to be both cell-specific and condition-specific, but affinity is likely to depend on physical characteristics and structures of the SH2 domain and target peptide. Thus accurate quantitative measurements are a critical step in order to do accurate modeling, predict outcomes of competition, and ultimately to predict network outcomes.

1.3 A Review of the Current State of Quantitative Measurements of SH2 Domain Interactions

Significant effort was made by four research groups from 2006 through 2014 to acquire high-throughput quantitative measurements of interaction between most human SH2 domains and a large number of phosphorylated targets in signaling proteins. Despite this effort, the bulk of published data is unusable for future research. Significant errors have been identified in large portions of reported data, rendering remaining data of limited value, and some experimental design choices have resulted in data that cannot be compared to results from any other study. Furthermore, some data was only published in print as a summary or as a figure, and in the intervening years, the electronic forms of that data have been lost. Although these issues could be rectified by access to the raw data, raw data from most groups has been lost. Data from only one group is available for any further analysis. Here, we summarize the experiments conducted, and discuss the experimental design choices, availability problems, and errors identified in the data which affect the suitability of this data for future work.

1.3.1 Overview of Published Data

High-throughput measurements of SH2 domain interaction with peptides have been made by four research groups and published in nine publications from 2006 to 2014. Most groups have focused on the response of phosphorylated tyrosine containing peptides in the most well studied receptor tyrosine kinase (RTK) tails: EGFR(ErbB1), ErbB2, ErbB3, and ErbB4. Later measurements expanded to additional families of RTKs, and to a larger pool of suspected phosphotyrosine containing peptides in the human proteome. Each group used different experimental techniques, resulting in various limitations in the collected data (Table 1.1).

Group	Year	Ref	Method	Peptides	SH2 Domains	Pairs Tested	Pairs Reported	Thres-hold	Data Type	Orig Data Avail.	Raw Data Avail.
MacBeath	2006	(10)	PM	61	159	9699	383	≤ 2 µM	K_d	Yes	No
	2008	(11)		50	133	6650	482			Yes	No
	2008	(12)		16	96	1536	25			Yes	No
	2009	(13)		46	96	4416	740			Yes	No*
	2013	(14)		729	70	51030	2808	≤ 1 µM		Yes	No
Nash	2010	(8)	PepA	192	50	9600	n/a	n/a	Intensity	No**	No
	2014	(15)		22	4	88	60	≤ 9 µM	K_d	Yes	
Jones	2012	(16)	FP	85	93	7905	1395	≤ 20 µM	K_d	Yes	Yes
	2014	(17)		85	93	7905	2216			Yes	Yes
Cesareni	2013	(18)	PepA	6202	70	434140	317613	n/a	Intensity	No***	No

Table 1.1: Overview of Published SH2 Domain Interaction Data.
*Raw data available for positive interactions; **Only published as figure; ***Only published as summary. No longer publically available, extracted from PepspotDB web server in 2015/2016. PM-Protein Microarray; PepA-Peptide Array; FP-Fluorescence Polarization.

The first high-throughput measurements were made between 2006 and 2009 by the MacBeath group (10, 11, 13, 19). These measurements were made using functional protein microarrays (PM), where proteins are immobilized onto a glass slide and a peptide with a covalently attached fluorophore is presented to the slide and then washed with a buffer. When a peptide and SH2 domain have sufficient interaction affinity, the peptide is not washed away and the fluorescent peptide signal can be detected. A microarray slide can have an entire panel of different proteins printed onto it, allowing testing of many proteins at once. This technique represented a significant increase in the quantity of measurements able to be made in a short period of time when compared to earlier techniques. Furthermore, they measured eight concentrations per measurement at equilibrium, and from the multiple measurements they were able to calculate the dissociation constant (K_d) at equilibrium. Later work demonstrated that protein microarrays were unable to reliably detect low affinity SH2 domain interactions with dissociation constants higher than 2µM (16). Ultimately, in 2013, the MacBeath group published

a significantly larger data set also using protein microarray technology, repeating and superseding earlier measurements and adding a large number of previously unmeasured protein domains and peptides (14).

In 2010, the Nash group published an interaction experiment using solid-phase peptide arrays. In theory, a peptide array should provide superior results to a protein microarray, while still maintaining the high-throughput capabilities of protein microarrays. By plating peptides instead of proteins, the more delicate proteins could be maintained in soluble conditions closer to their native environment and only presented to the array during an experiment. As a control for this technology, they performed lower-throughput fluorescence polarization experiments to validate their peptide array results. Fluorescence polarization can be used to measure protein-peptide interaction maintaining both the protein and the peptide in solution. Unlike the Jones group measurements, the Nash group only tested a single concentration per protein. The value they reported was proportional to fluorescence and was deemed by the authors to be 'semi-quantitative'. Other work has called into question the quantitative validity of a single measurement (20). In 2014, additional work from the Nash group demonstrated a new type of peptide array on a small number of interactions, using multiple concentrations of protein, which was both quantitative and highly reproducible (15).

In 2013, the Cesareni group presented results from the largest scale experiment to date with over a 10-fold increase in tested interactions. They used a glass-slide based peptide array and GST-tagged SH2 domain proteins. They used a single concentration value of protein per interaction, and thus reported a value proportional to fluorescence. This technique is known to limit the accuracy of quantitative measurements (20). Furthermore, we demonstrate that GST-

tagged proteins behave differently in a non-systematic way than untagged proteins, further calling into question these results (see Chapter 2.5).

In 2012 and 2014, the Jones group published two high-throughput experiments using fluorescence polarization (FP) for all measurements. For validation, they used Surface Plasmon Resonance (SPR) – a more accurate but low-throughput technique – to validate their FP measurements. The Jones group also measured multiple concentrations per interaction, and were able to report the dissociation constant (K_d) at equilibrium.

1.3.2 Published Experiments Have Significant Limitations

The usefulness of published data is limited by multiple factors: limitations of experimental techniques and experimental design choices, errors in published data, and limitations in future data availability. Some experimental techniques have been subsequently shown to have significant limitations. Experimental design choices limit the usefulness and ability to compare results between data sets. Demonstrable errors and inaccuracies in reported data limit the usefulness of some data. Finally, lack of availability of published data and also raw data limits current and future usefulness of experimental results.

Three difference experimental approaches have been used to publish high-throughput measurements of interactions between SH2 domains and phosphorylated peptides: protein microarrays (10–14), peptide arrays (8, 15, 18), and fluorescence polarization (16, 17). Of the three techniques, fluorescence polarization has the best sensitivity and reproducibility on large scale data sets. Protein microarrays can only detect interactions with higher affinity (<2μM) (16). Early implementations of peptide arrays suffer from the same issues with reproducibility and sensitivity, as well as issues with noise and high background signal (data not shown), but the

most recent implementation of peptide arrays may have overcome these issues (15), but have not yet been used for large scale measurements of SH2 domains. Fluorescence polarization techniques are able to detect lower affinity interactions (< 20 μM) and thus have increased sensitivity. Although the authors of the FP experiments identified problems with reproducibility (16) (including a low 63.3% validation rate for true interactions on a single FP run), they suggest (and we also demonstrate) that problems with protein preparation are more likely responsible for this issue than the technique.

Experimental design choices play a major role in the current and future usefulness of some of the published data. Experiments which measure only one concentration of protein per interaction, such as from the Nash and Cesareni groups (8, 18), represent potentially inaccurate measurements and fundamentally lack the controls to determine if those inaccuracies exist (20). Furthermore, a protein-peptide interaction cannot be compared directly with other data sets – it can only be compared relative to some other interaction. Understanding these limitations, the MacBeath and Jones groups (as well as the latest Nash experiment) measure interactions at multiple concentrations of the SH2 domain allowing for equilibrium measurements of the dissociation constant (K_d). This dramatically increases the usefulness of the data, and allows for comparison of results, despite difference in experimental conditions.

One data source is severely limited in usefulness due to errors in the published data. Although the PepspotDB data published by the Cesareni group is the single largest data set available, it has a significant inconsistency in 2/3 of the measurements made, drawing the accuracy of the published data into question. The data contains columns for the foreground (FG) and background (BG) fluorescence measurements, as well as the difference between the foreground and background ($FG - BG$), and the fold change ratio of the foreground to the

background ($\log_2 \frac{FG}{BG}$). For approximately 2/3 of the data (over 280,000 protein-peptide interactions) the difference and fold change columns cannot be computed from the reported foreground and background values (see Table 1.2 for an excerpt of PepspotDB data). This draws into question all other values in these rows, as they are derived from these raw measurements. These errors were not limited to a subset of protein domains, or peptides, a data range of gathered data, or any other logical subset of the measurements that we could identify. Since 1/3 of the measurements do compute for these columns, it seems that at some point, a portion of the database became scrambled, by row or by column. Unfortunately, these experimental results have been used in multiple published analyses and models. Based on these findings, the vast majority of this data should not be used in any future work and previous publications should be reevaluated.

SH Domain	Peptide	FG	BG	Reported $FG - BG$	Calculated $FG - BG$	Reported $\log_2 \frac{FG}{BG}$	Calculated $\log_2 \frac{FG}{BG}$
ABL1	TRFDDWyLWVQMY	162	146	16	16	0.15	0.15
ABL1	LKDKEGyTSFWND	166	139	27	27	0.25	0.25
ABL1	NITDPEyGYLARE	149	140	9	9	0.09	0.09
ABL1	YPREGKyGHAACF	178	140	38	38	0.36	0.36
ABL1	AFFNPKyQHEGFY	154	140	22	14	0.21	0.14
ABL1	ALVDLDyEDRPEY	142	138	3	4	0.03	0.04
ABL1	IIEEGKySLVMEY	155	142	17	13	0.14	0.13
ABL1	QFSKGVyAIFGFY	136	132	2	4	0.02	0.04
ABL1	FPFNFSySDYDMP	154	143	12	11	0.12	0.11
ABL1	AKLKDYyIFNKYL	141	142	4	-1	0.04	-0.01
ABL1	GQMKDLyHYITSY	139	132	5	7	0.05	0.07
ABL1	STPKVLyEIPDTY	177	141	43	36	0.37	0.33

Table 1.2: Excerpt of PepspotDB.
Sample excerpt from PeptspotDB demonstrating inconsistencies in published data. Green cells indicate calculated values match published values; blue cells indicate calculated values do not match published values.

Original data was never published for two data sets, and is currently very difficult to acquire. In the first Nash group publication (8), data was published as a binned heatmap, but a table containing the measured values was never published. The data is privately available (Ron Hause, personal communication), but not from the Nash group which has since disbanded. The data collected from the Cesareni group (18), was also never published. A summary of data in the form of sufficient statistics was included in the original publication, but the data that the calculations were based on was not published in a journal, and the statistics used were insufficient to completely describe the data. That data was subsequently displayed in an online database, but that database has since been taken offline and is no longer available from the Cesareni group. We were able to evaluate and identify the errors in this data as described above as we retrieved a copy of the information in the database before it became unavailable.

The state of availability of raw data underlying these experiments is dire. Since much of the useful data published is in the form of a dissociation constant, the reported data is actually based on a calculation made upon one or more raw data measurements over multiple concentrations. These raw measurements are required to evaluate the fitting methods used to calculate dissociation constants, to question the assumptions in the models used, and to determine if the measurements in the experiments were valid. Off all high-throughput data gathered since 2006, only one group has raw data available. Reanalysis of the raw data from the one available data set is found in Chapter 2.

1.3.3 Intergroup Experimental Results Correlate Poorly

A further complication in working with the interaction data is that there is practically no agreement of measured data between different groups (Figure 1.1). Although both the MacBeath and Jones groups were able to successfully validate a random selection of measurements against

a lower throughput and more accurate method of measuring interactions – the MacBeath PM data was validated successfully against low-throughput FP, and the Jones FP data validated successfully against SPR analysis – neither group's data validates well against published sets of curated low throughput data.

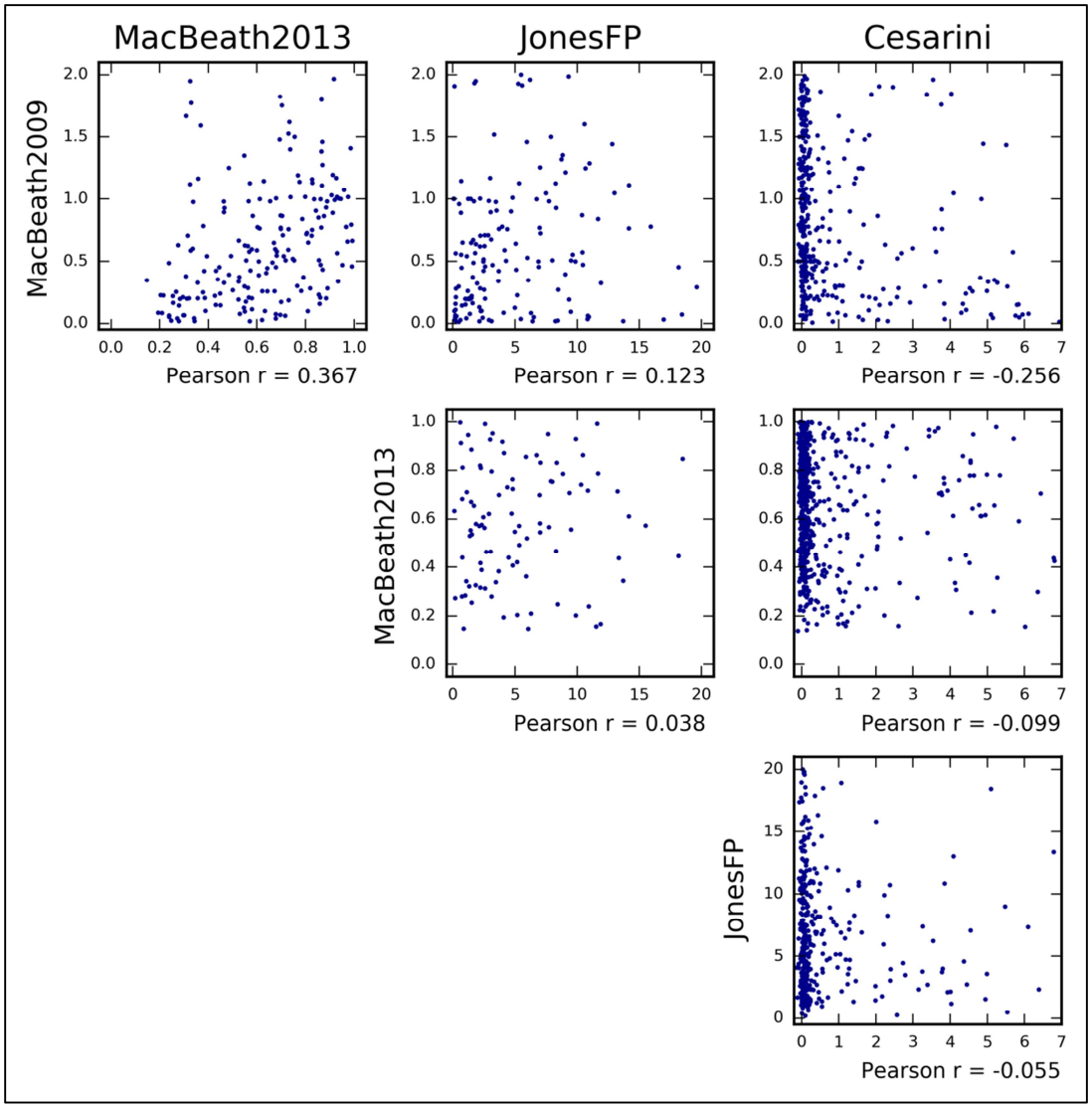

Figure 1.1: Between Group Comparisons of Published Data.
Correlation between published quantitative data from different groups is plotted as a scatter plot. Data units represent K_d values (µM), except for the Cesareni group data which is published as a z-score of an signal intensity-based scale. Pearson correlation coefficients are indicated below each plot.

The disagreement between labs on dissociation constants is troubling. The dissociation constant, measured at equilibrium, should be the same despite systematic differences in measurement techniques. If the primary cause for error was the technology used (e.g. PM vs FP), one technology should have compared unfavorably to the lower-throughput validation methods. On the other hand, if differences in protein preparation are responsible for the majority of the variance, low-throughput validation methods could also be measuring that same protein preparation. If the variance is due to the conditions and method of protein preparation in a particular laboratory, low-throughput measurements would correlate well to high-throughput techniques within the same laboratory even given significant differences between laboratories.

Based on our analysis (Chapter 2), it is very likely that a significant component of measurement variance is due to protein preparation. These differences can be magnified by some post-measurement modeling techniques and methods of handling replicate measurements. Eliminating these sources of variance might very well ameliorate the significant differences in measurements found between difference research groups.

1.3.4 Conclusion

Despite significant effort to measure and understand SH2 domain interactions with their target peptides, critical flaws or lack of access limits the usefulness of most data. Interactions measured with single measurements are known to be inaccurate, and comparison to other measurement is severely limited. Errors in reported data make that data suspect. Lack of access to published data prevents use. Lack of access to raw data hinders our ability to evaluate the analysis process that produced such data. Of all the high-throughput interaction data gathered to date, a complete set of raw data is available from only one group.

Chapter 2: Revised Analysis of SH2 Domain Interaction Data

2.1 Motivation for Reanalysis

The primary motivation for reanalysis of existing SH2 domain peptide interaction data is to identify true negative interactions (pairs of SH2 domains and peptides that have very low or no affinity for one another). All quantitative high-throughput studies-to-date made tradeoffs to produce positive interactions at the expense of valid negative interactions. Although the raw measurement techniques were essentially neutral to positive and negative interactions, choices in analysis techniques were made to focus on identifying the most likely positive interactions such as using quality metrics (16) or statistical tests (18) that favored only positive interactions. Thus, results were tuned to maximize true positive detection at the expense of making false negative calls. In these data sets, lower-certainty interactions were relegated to an 'out' group, along with potential negative interactions, poor fits, and noisy data.

Negative interactions are as important as positive interactions when building accurate models of binding (21). Supervised machine learning techniques rely on training data sets in developing a function to map new input to a category or value. In most cases, training sets leading the highest accuracy must contain both positive and negative examples. Similarly many statistical modeling techniques benefit from negative data for increased accuracy. In order to address this shortcoming, researchers using this data to create models of interactions have either used methods that do not take into account negative interactions (7), or used methods to generate synthetic negative interactions (21). Other researchers ignored the false negative issue and treated all interactions that were not positive interactions as non-binders (17, 22). Since all of

these methods will produce less accurate models than models using true negative interactions, we hoped to extract additional information about negative interactions from existing data.

Based on our previous work examining the current state and availability of SH2 domain interaction data (see Chapter 1), we identified only one data set which did not contain obvious systematic errors, and for which raw data was available: the FP data from the Jones group. Although this data was the only raw data available, the FP technology used has the highest sensitivity for weak interactions and is solution based, allowing the proteins to be closest to their native states when tested. Thus, this data set would likely have been the starting point of any reanalysis even if more data were available. We contacted the original authors and were generously provided with all raw data and files available (Richard Jones, Ron Hause, and Kin Leung, personal communication).

In reviewing the experimental details and analysis methods from the Jones group experiments, it became clear that best practices were not followed in both the design of the experiments and the analysis of the data. During experimental design and data gathering, insufficient controls were used, making it difficult to distinguish between true negative interactions and non-functional proteins or peptides. Despite identifying non-monomeric protein, limited purification was undertaken. It would be reasonable to assume that limited purification and controls were related to the tradeoff between cost and gathering more data. In addition, in modeling and analysis, several critical steps were overlooked. No assessment of the appropriateness or deviation from one-to-one models was made, no evaluation for non-specific binding was used in modeling, and an inappropriate quality metric was used, resulting in discarding of a majority of the measured data. Furthermore, protein preparation was also known

to be a major source of variation, but the variation was treated as typical sample variation in analysis, and no controls for non-functional or partially-functional protein were utilized.

These deviations from best practices suggested that alternative analysis methods might uncover useful information.

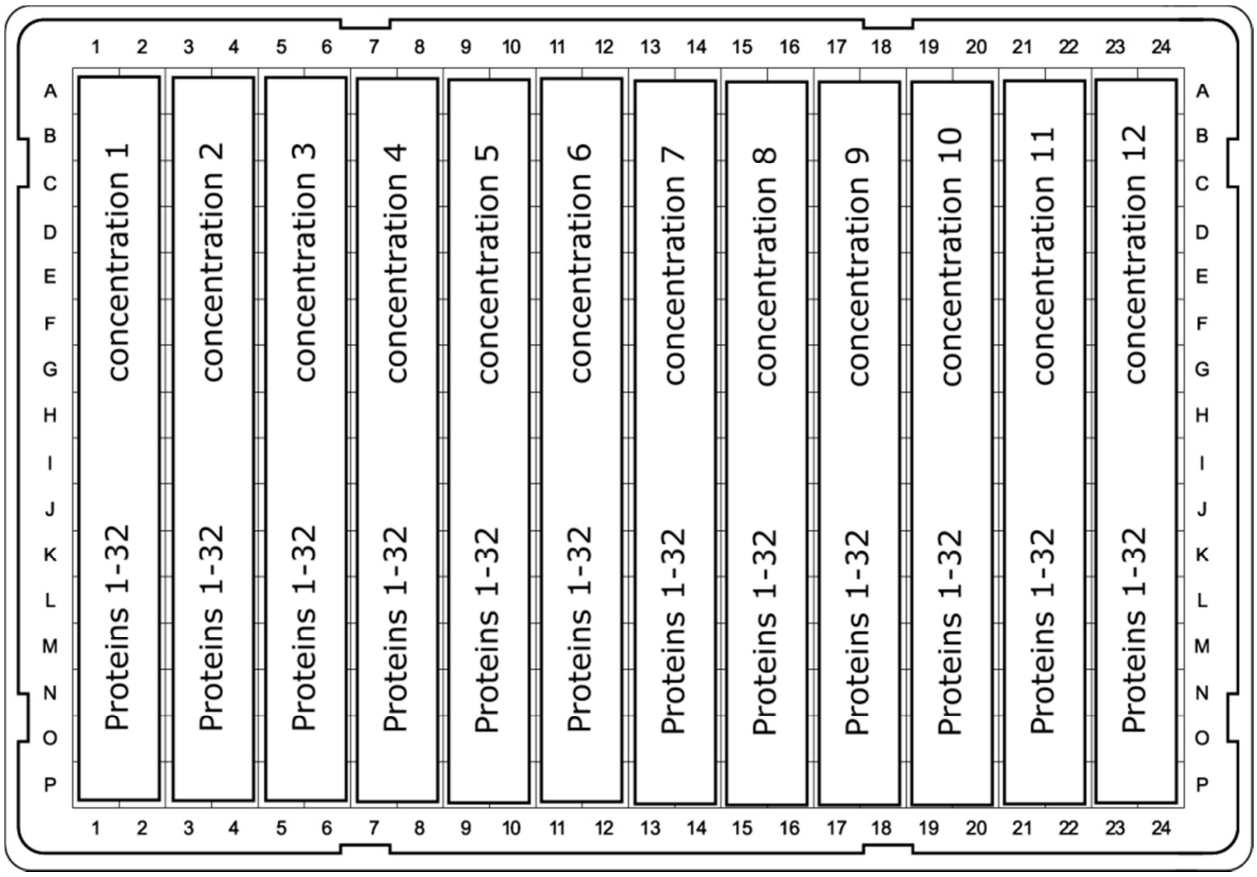

Figure 2.1: Experimental Layout of 384-well Plate.
Layout diagram for the 384-well plates used in the Jones FP experiments. Thirty-two proteins at 12 concentrations were tested per plate against a single peptide concentration. Proteins were placed on the plate in the pattern above.

2.2 Preliminary Data Analysis

2.2.1 Description of Raw Data

The Jones group interaction data consists of two separate data sets published in 2012 and 2014 (16, 17). Although the 2014 study added additional peptides, both studies used the same technology (fluorescence polarization), covered the same SH2 domains, and used the same protocols. Some aspects of the availability and formatting of raw data, and aspects of the experiment design and sample preparation, are relevant to the reanalysis of the raw data.

The published data contains interactions between 89 single-domain SH2 proteins with 165 peptides from 8 different receptor proteins (EGFR, ErbB2, ErbB3, ErbB4, GAB1, Kit, Met, and the human androgen receptor). The set of raw data provided by the Jones group did not contain all published data. Although it contained data from all 89 domains, it was limited to 142 peptides from 7 receptor proteins, missing data from the human androgen receptor, and a handful of other domain-peptide combinations. Comparisons are based on this slightly smaller set of interactions common to both the published and raw data.

The raw FP data consisted of numerical fluorescence polarization data (in mP units) in a 16x24 grid, from each 384 well plate that it was scanned from. The plate contained 32 SH2 domain proteins at each of 12 concentrations organized spatially as in Figure 2.1. The formatted raw data contained labels and concentrations for each well of the plate, as well as the identity of the fluorescently labeled peptide presented to the entire plate (see Table 2.1 for a slightly reformatted example of raw data from a plate). For each plate a 17th row in the data contained background intensity but was not used in this analysis. Data for all plates scanned in a particular run were concatenated end-to-end in a single Excel file.

10.0 μM		5.0 μM		2.5 μM		1.25 μM		0.625 μM		0.313 μM		0.156 μM		0.0781 μM		0.0391 μM		0.0195 μM		0.00977 μM		0.00488 μM	
PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN	PIK3R1-N	TXN
TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3	TEC	TNS3
PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC	PIK3R1-C	SRC
SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3	SOCS1	SOCS3
ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2	ZAP70-C	GRB2
PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N	PLCG1-NC	PLCG1-N
FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2	FGR	GRAP2
TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C	TNS1	PLCG1-C
PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC	PIK3R1-NC	PTPN11-NC
SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C	SOCS6	PTPN11-C
BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1	BLK	VAV1
SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1	SOCS2	NCK1
CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3	CRK	VAV3
PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN	PTPN11-N	LYN
VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5	VAV2	SOCS5
SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2	SYK-C	NCK2
bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	bg	
137.74	135.88	147.66	135.29	147.76	134.3	147.32	134.48	145.04	140.87	139.88	146.42	146.11	145.44	139.42	143.56	136.04	151.76	142.28	144.77	151.06	144.52	140.27	155.25
141.18	144.84	132.07	134.93	143.15	143.64	141.68	145.94	147.33	141.37	137.8	138.04	133.9	142.95	144.37	137.71	136.45	139.45	145.54	143.85	141.77	141.36	149.75	141.71
139.29	154.53	141.32	147.29	143.46	146.32	141.93	131.54	142.36	135.62	141.83	137.3	139.28	143.35	144.89	143.96	142.45	139.54	135.38	146.54	136.09	150.25	152.06	143.81
134.9	148.11	139.44	140.02	143.51	145.2	131.88	138.22	131.2	133.11	141.88	137.45	143.8	148.78	139.58	139.25	142.12	145.72	136.61	146.7	148.07	141.51	147.74	143.69
146.06	138.14	136.48	134.64	126.42	137.1	138.25	134.47	144.21	145.61	142.76	144.21	139.07	136.25	136.73	141.56	146.95	133.45	139.72	135.51	136.65	149.63	139.68	144.11
141.2	134.97	142.13	146.69	146.12	137.39	152.14	142.02	144.51	151.14	133.37	130.31	142.52	141.66	138.18	138.76	141.8	146.83	139.19	148.03	138.46	147.75	141.79	140.63
130.95	134.11	135.74	141.75	140.14	134.4	136.51	140.6	139.95	141.08	138.39	139.61	145.52	136.44	139.11	131.44	144.47	140.15	139.23	136.91	144.36	143.3	144.34	141.55
131	142.97	137.44	139.66	142.07	139.92	144.07	152.73	143.51	143.91	138.59	134.38	142.22	141.34	140.71	136.23	133.25	141.03	143.46	147.7	142.19	143.25	141	147.29
141.08	121.97	133.8	133.68	138.19	139.46	133.12	134.02	136.16	140.94	144.01	144.57	132.58	140.75	141.25	137.63	135.88	132.56	134.85	136.86	143.32	142.4	142.86	145.96
131.09	133.57	132.71	138.58	140.34	141.35	142.59	142.47	134.5	137.19	140.94	142.91	149.65	139.07	138.79	145.1	143.36	137.59	144.76	132.26	143.11	138.46	143.69	132.23
123.1	137.03	139.1	130.81	141.04	135.75	129.83	143.13	131.71	138.29	146.33	146.01	140.32	140.18	144.39	135.92	142.95	143.52	142.67	139.89	148.02	137.52	138.05	143.86
132.45	139.27	134.92	136.74	138.78	140.64	140.86	137.08	139.8	134.97	141.53	152.38	135.08	146.18	143.24	148.42	134.55	140.33	141.13	143.44	139.91	128.22	143.95	130.41
125.46	139.8	135.95	140.49	133.68	135.91	140.76	126.67	139.68	137.6	132.6	137.17	140.06	141.97	142.77	148.5	141.67	135.57	135.65	145.59	140.71	141.85	139.12	147.87
145.36	138.75	134.55	139.72	132.23	134.95	137.35	143.61	134.05	138.9	145.3	142.28	142.9	132.36	134.9	135.31	143.53	144.99	132.41	136.21	142.39	143.3	138.15	138.67
118.01	139.48	136.48	141.3	141.17	138.6	142.16	144.61	139.32	139.93	140.46	135.98	147.22	145.29	138.13	140.82	149.49	133.26	148.12	135.97	154.97	149.72	143.13	150.19
140.92	141.64	150.64	155.25	143.22	143.81	149.41	150.82	152.49	141.51	145.62	150.52	146.53	149.05	144.94	151.74	141.56	149.13	148.44	158.42	154.49	161.34	145.69	148.02
135.88	147.66	135.29	147.76	134.3	147.32	134.48	145.04	140.87	139.88	146.42	146.11	145.44	139.42	143.56	136.04	151.76	142.28	144.77	151.06	144.52	140.27	155.25	

Table 2.1: Raw Data Format

Raw data was stored as intensities, labels, and concentrations for each well of the plate. The plate number, and the identity of the fluorescently labeled peptide presented to the plate were recorded. For each plate a 17th row in the data contained background intensity information. The information from the original excel data is reproduced here but has been modified to fit the page.

The peptides in the Jones FP data are predominantly 13-mers, with a central phosphorylated tyrosine residue. In the first Jones experiment, approximately 32 peptides were initially made and tested as 18-mers. Jones and colleagues found the 18-mers to have similar binding free energies, but the 13-mer peptides generated higher maximum polarization, and thus gave better sensitivity and signal-to-noise (16). The SH2 domain proteins were primarily expressed as single-SH2-domain-containing constructs. When a native protein contained more than one SH2 domain, constructs were built with each domain independently (and referred to with a suffix of "–N" or "–C" to indicate which terminal the domain was from). Although a few proteins were also expressed as a double domain, those constructs are excluded from this revised analysis. These SH2 constructs that the Jones group selected for the experiment each met the following criteria: "1) fraction of monomeric protein observed in previous study following expression and purification >=50% by size exclusion chromatography; 2) previous evidence of functionality by PM as evidenced by interaction with one or more phosphopeptides with an apparent midpoint binding constant KD<=1 mM." (16)

2.2.2 Model Selection

The mathematical model underlying a saturation binding experiment is based on a theoretical one-to-one kinetic interaction:

$$F_{obs} = \frac{F_{max}[domain]}{K_d + [domain]} + F_0$$

where F_{obs} is the observed fluorescence (in mP units), *[domain]* is the concentration of the SH2 domain, F_0 is the baseline value (in mP units), F_{max} is the value at saturation , and K_d is the dissociation constant. This form of model was used in the original publication, as well as all

other high-throughput analyses which computed dissociation constants from the raw binding data (10, 11, 13, 14, 16, 17, 19).

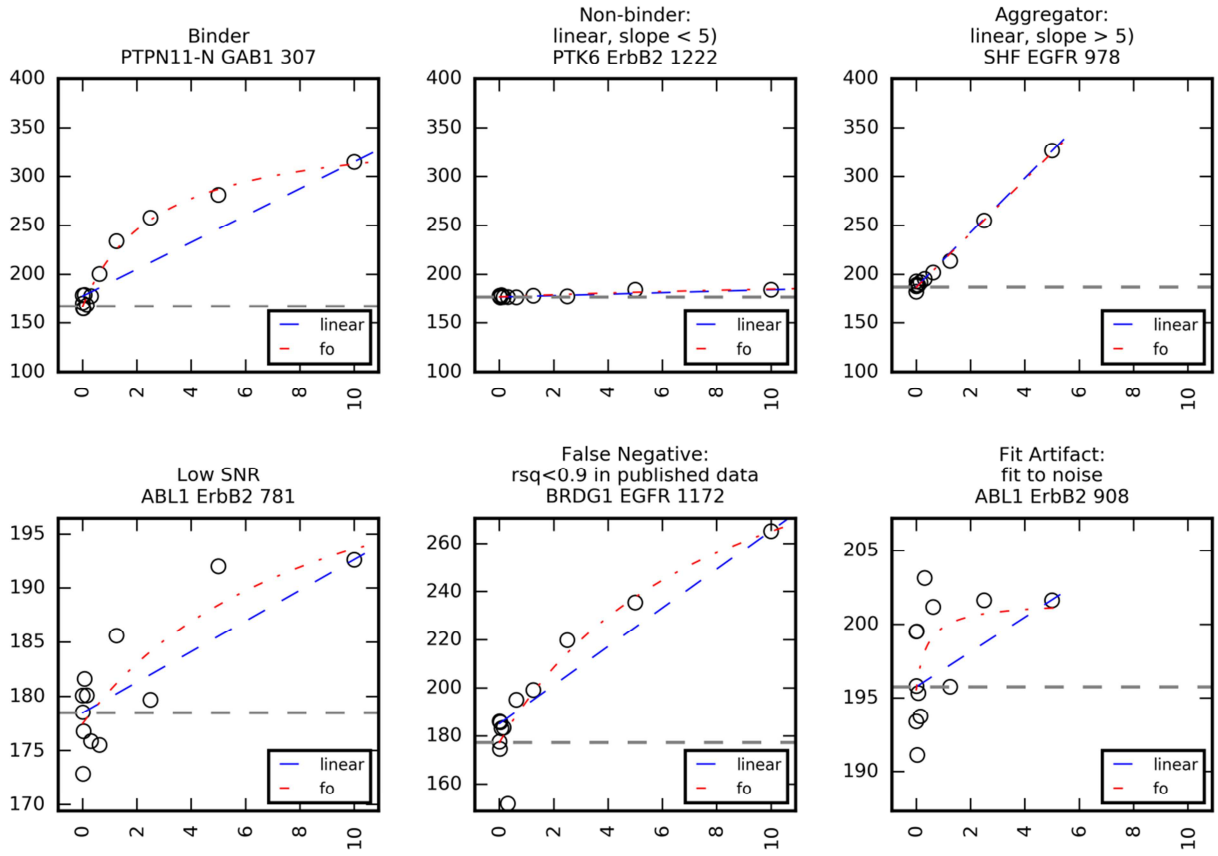

Figure 2.2: Examples of fitting results for individual replicates.
Several illustrative examples of fitting results for individual replicates are shown. (Upper Left) Binder example with moderate K_d, and signal approaching saturation. (Upper Middle) Non-binder example, with low magnitude, flat signal . (Upper Right) Aggregation example: strong linear response of high slope with no saturation evident. (Lower Left) An example of a fit for which a first order fit was best, but which had high residual error. (Lower Middle) An example of a binder called in our revised analysis which was rejected as a binder in the original publication. (Lower Right) Fit Artifact: Despite clearly being noisy data with little form, a first order fit is selected over a linear fit. These examples are also typically filtered out by signal to noise criteria, but they represent a third category of data that fits no true model, thus they are difficult to identify. Key: Dashed gray line – fitted offset. Dashed blue line – linear fit. Dashed red line – first order fit.

This seems to be the appropriate model to describe typical positive SH2 domain interaction with phosphorylated peptides (see Binder, Figure 2.2). However, preliminary analysis of the data indicated that fitting only a single one-to-one model of binding via least squares failed in some cases due to fitting artifacts, particularly on many non-binding interactions and

interactions that suggested aggregation. For example, a non-binding interaction (see Non-Binder, Figure 2.2) is typically indicated by low but constant magnitude data with random noise superimposed. When the random noise at the first point or two coincides with a positive slope, the least square solution was often a very sharply bending saturation curve with an extremely low dissociation constant (K_d on the order of 0.0001 μM), and a very low F_{max}, often approaching zero (see Fit Artifact, Figure 2.2). Thus the greedy fitting algorithm was parametrizing the fitting results based on random noise. In the cases which we have described as aggregation (see Aggregator, Figure 2.2), the florescence increases linearly with concentration and never shows signs of saturation. When fit with a one-to-one model, dissociation constants and concentration at saturation both approached infinity (and resulted in fits with K_d and F_{max} values on the order of $1x10^7$ μM or greater. In the latter case, a one-to-one model is inappropriate to model this phenomenon, resulting in the improper parameterization.

In order to overcome these fitting artifacts, we fit both a linear and one-to-one model to the interaction data. The linear model was in the form:

$$F_{obs} = m[domain] + F_0$$

where F_{obs} is the observed fluorescence (in mP units), *m* is the slope of the fit line, *[domain]* is the concentration of the SH2 domain, and F_0 is the baseline value (in mP units). The best model fit was evaluated using the Akaike Information Criterion (AIC) (23).

2.2.3 Determination of Saturation Conditions

We hypothesized that the maximum polarization measured – the polarization at saturation – should be similar across all domains and peptides. Thus, in a saturation binding experiment such as this, as protein of increasing concentrations are exposed to labeled peptide, polarization

values should monotonically increase until they plateau – despite increasing protein concentration – as all peptides become bound to an SH2 domain. This saturation behavior should follow the classic hyperbolic binding curve describing one to one binding (Figure 2.3). The labeled peptides in this experiment are all similarly sized small molecules – 13 or 18 amino acids in length, and the SH2 domains are very close to 100 amino acids in length and globular in shape. Since fluorescence polarization measurements effectively measure the volume difference between an unbound labeled molecule and a complex containing the labeled molecule (Figure 2.4), volume differences between free peptides and bound peptides should be very similar across all domains and peptides. Thus, polarization at saturation should be similar across all measured domains and peptides. However, based on the presence of non-monomeric protein reported in the original publication, we hypothesized that we might also find results that might not match perfect theoretical behavior. Jones and colleagues reported that all domains used in the experiment had a "fraction of monomeric protein observed" greater than or equal to 50% "following expression and purification by size exclusion chromatography" (16). Unfortunately, the percent of non-monomeric protein was not recorded in the published data or in the raw data. Knowing that proteins tested were not pure monomers, we hypothesized that we might see effects resembling 'larger volume' binders than expected for a monomeric one-to-one binding experiment.

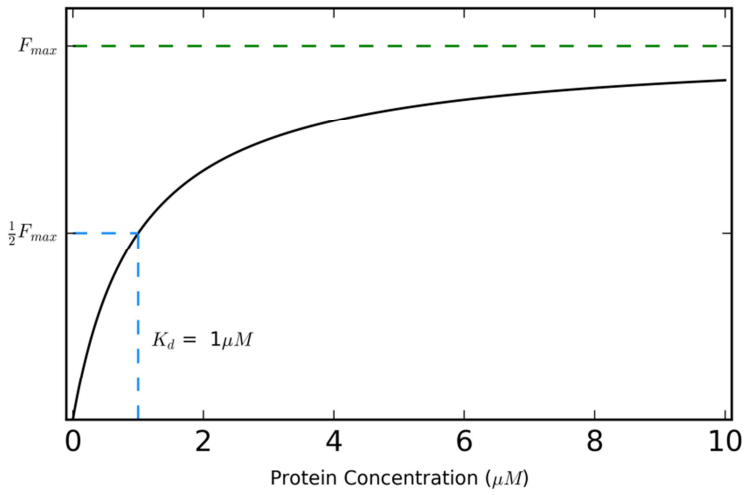

Figure 2.3: One-to-One Binding.
Example of one-to-one saturation binding curve. The F_{max} is the signal expected at the asymptote of the curve at complete saturation. The dissociation constant (K_d) at equilibrium can be computed graphically from the figure. When the fluorescence value on the curve is ½ F_{max}, that concentration is the K_d.

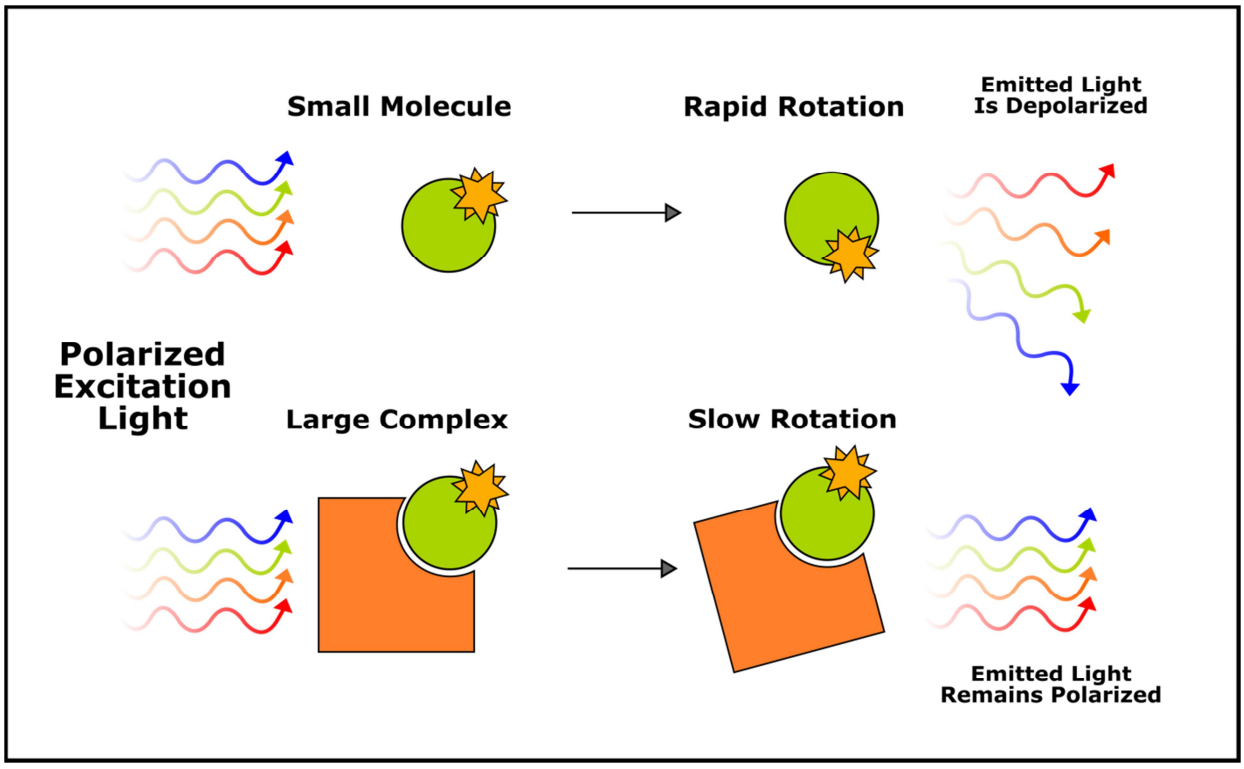

Figure 2.4: Fluorescence Polarization (FP) Measures Binding.
Polarized light is used to excite a fluorescently labeled molecule. (Top) Small molecules rotate more rapidly than larger molecules so light emitted from the fluorophore is emitted in different planes depolarizing the light. (Bottom) When bound by a larger molecule, light emitted from the small molecule remains close to the exciting plane, because large molecules have rotated less during the time between excitation and emission. Thus FP effectively measures volume differences, and is used to identify molecular interactions.

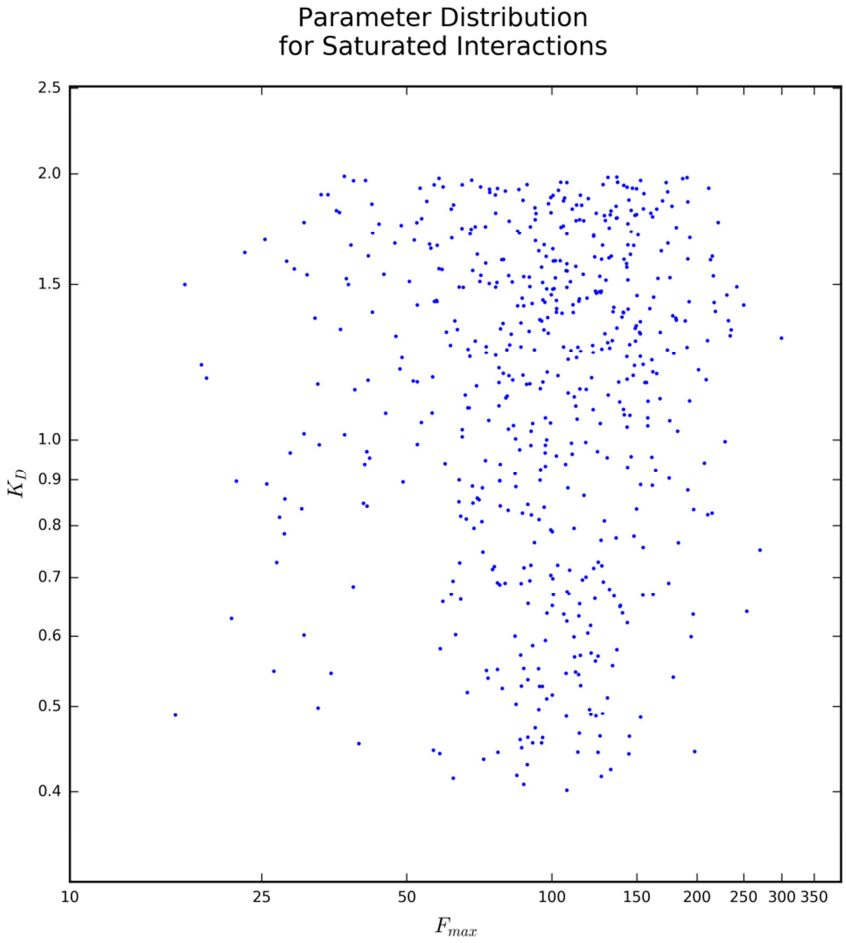

Figure 2.5: Range of F_{max} and K_d for Control Set of Binders.
The distribution of K_d vs F_{max} is plotted for a set of control interactions. Positive interactions with good fits and a K_d between 0.4µM and 2µM were chosen to represent a control set of one-to-one fits with good saturation signals. The bulk of fits had an F_{max} between 25 and 300, suggesting this is a reasonable expected range for maximum fluorescence of a one-to-one fit for a 100AA SH2 domain to a short peptide.

In order to find identify a reasonable range of values for the maximum fluorescence (F_{max}) of a true one-to-one interaction in this experiment, we looked at a subset of measurements likely to represent saturated binding. A subset of positive interactions with a K_d between 0.4µM and 2µM were chosen. This subset represented a K_d range for which the experimental concentrations were likely to detect saturation well. It's also in a reasonable range for affinity for an SH2 domain for a peptide according to previous studies. Thus, this set of interactions is likely to predominantly consist of true binding interaction of SH2 domains of moderately high affinity

and provide a valid sample of saturation conditions. A visual inspection of the fits confirmed these hypotheses. Interactions in this K_d range had F_{max} values between 25 and 316 mP units, and no maximum values in this set extended above 316mP units (Figure 2.5). Thus, we concluded that true one-to-one interactions were unlikely to have fitted F_{max} values far above this range.

However, we were surprised to find such large variation between the interactions. In order to determine the source of the variation, we examined the relationship between F_{max} range and protein domain. We chose a set of interactions for which a first order fit represented the best fit, and which had low residual error. F_{max} ranges vary significantly by domain – with some domains having relatively tight windows of F_{max} values, and others with high variance. The median F_{max} also varied significantly by domain (Figure 2.6 and Figure 2.7). Only a small fraction of interactions fell below an F_{max} of 50mP. The small differences in protein size (all close to 100AAin length) are unlikely to be the source of this variation. Protein with high average F_{max} values may have a higher content of non-monomeric protein. Proteins with high F_{max} variation may have been composed of multiple difference protein preparations with differing percentages of non-monomeric protein.

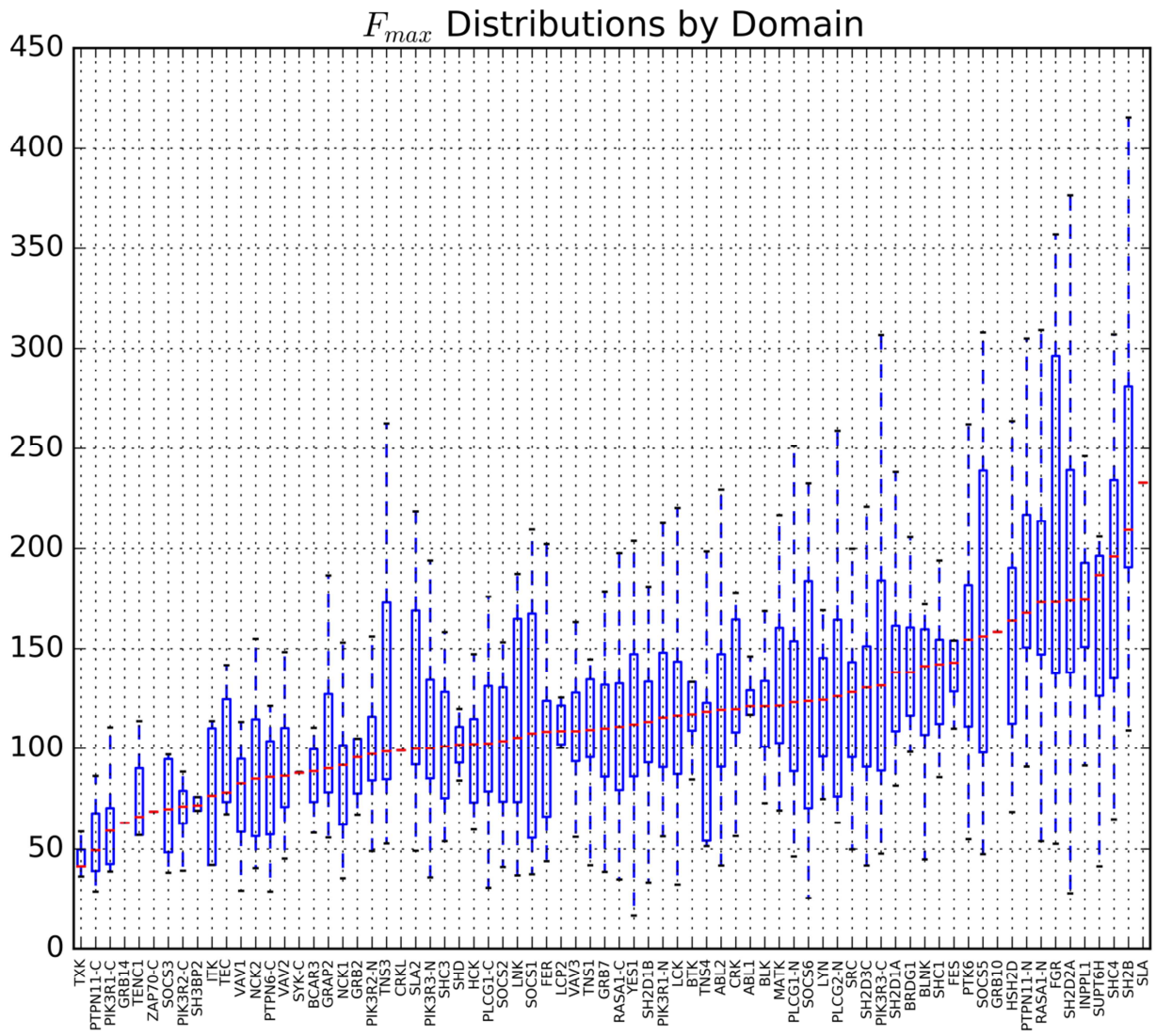

Figure 2.6: F_{max} Distributions by Domain for 2012 Experiment.
F_{max} distributions by protein domain are plotted as box plots. The box extends from the lower quartile to the upper quartile value in the data. The red line represents the median. The whiskers extend to the smallest and largest values which are not outliers (more than 1.5 times the interquartile range past the quartile). Data from (16).

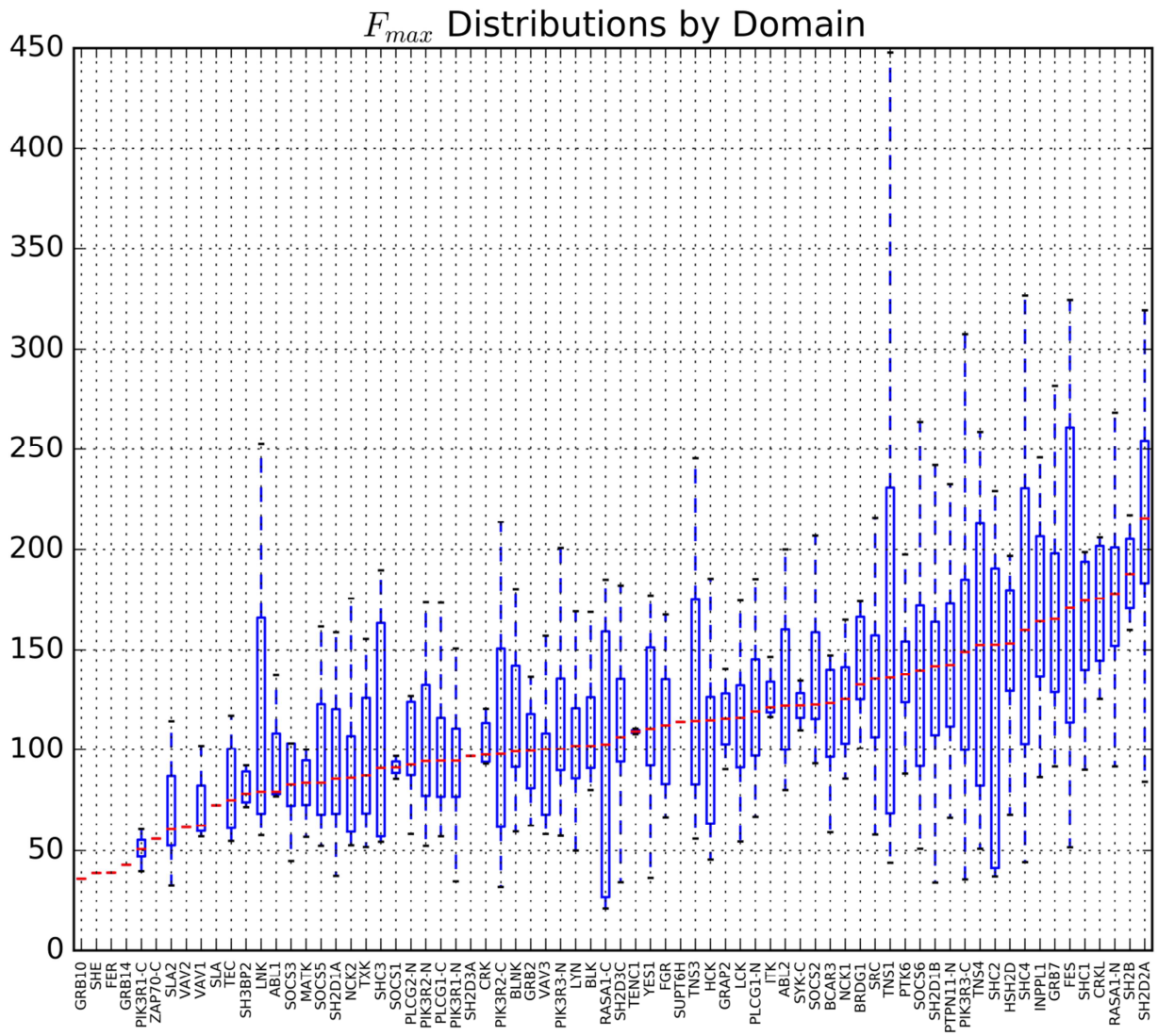

Figure 2.7: F_{max} by Domain for 2014 Experiment.
F_{max} distributions by protein domain are plotted as box plots. The box extends from the lower quartile to the upper quartile value in the data. The red line represents the median. The whiskers extend to the smallest and largest values which are not outliers (more than 1.5 times the interquartile range past the quartile). Data from (17).

2.2.4 Protein Functionality Impacts Identification of Negative Interactions

In addition to accurately identifying true positive interactions, one significant challenge when analyzing this interaction data is the task of identifying true negative interactions as distinct from negative interactions due to missing or non-functional protein, or experimental failure. The ability to identify non-functional protein greatly increases the quality and accuracy of non-binding calls and is vital to future modeling of SH2 signaling systems.

Several parameters related to expected protein behavior affect our ability to accurately identify true negative interactions. SH2 domain recognition and binding to a phosphorylated peptide seems to the key step in determining signaling specificity. Thus, in order to maintain selectivity, we hypothesized that most SH2 domains would only bind a limited set of phosphorylated peptides. If true, we would expect that a large majority of experimental interactions for each domain would measure as negative, non-binding, interactions. In addition, based on prior knowledge of their function, we expect that each SH2 protein tested should interact with at least one or more phosphorylated peptide found in the human proteome. Thus true positives are likely to be rare. However, since the experiment does not cover all possible interactions in the proteome, it is possible that some proteins tested will truly interact with no tested peptides. This circumstance highlights one difficulty: although non-interaction may represent a series of true negative results, complete non-interaction would be indistinguishable from a situation with non-functional protein unless controls were included to distinguish such a result. In addition, due to the experimental methodology chosen in the original publication, one

significant practical difficulty in this experiment lies in distinguishing proteins which have degraded or are non-functional from proteins which truly fail to bind any tested peptide.

In an ideal experiment, controls should be chosen to assure that the results from the experiment are due to the parameters and components being tested and not due to outside unintended influences. When evaluating protein interaction with a fluorescent peptide, controls should establish that the protein is folded and functional, that the peptide is phosphorylated and properly labeled, and that the protein and peptide demonstrate the expected activity. A non-functional or unfolded protein or an improperly labeled peptide would lead to a false negative interaction. In this experiment, the fluorescent polarization experiments were carried out on 384-well plates. Each plate contained 32 different proteins at 12 concentrations measured against a single peptide. The set of 12 concentrations are fit with a curve resulting in fitting parameters of K_d (equilibrium dissociation constant), F_{max} (fluorescence value at saturation), and F_0 (baseline or background fluorescence). This results in 32 different protein measurements against a common peptide. Unfortunately, explicit controls were not included in the experimental design, and thus were not present any plate tested. On its face, this is a significant limitation in the original experimental design.

Although there are no explicit controls on a single plate, some implicit patterns in the data can be used as controls. For example, baseline fluorescence for all 32 proteins at every concentration would be a pattern consistent with a non-functional peptide. Although a single plate might be useful to diagnose an issue with a peptide, a single plate does not contain a useful pattern to determine problems with protein activity or functionality. Since the behavior of any particular protein-peptide interaction is not known, any single negative interaction may represent a true negative non-binding event. Nevertheless, a combination of protein results across

replicates from different runs and across multiple peptides can demonstrate a pattern helpful in diagnosing a problem with protein functionality. We identified several such patterns that were effective at identifying varying types of protein degradation.

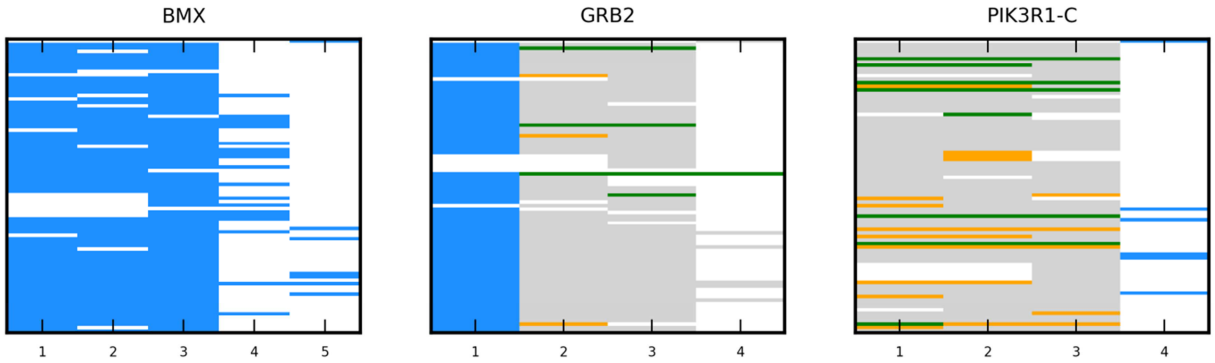

Figure 2.8: Examples of Categorical Binding Activity at the Replicate Level.
Heatmaps of activity for the SH2 domains from BMX, GRB2, and PIK3R1-C interacting with ErbB peptides (16). Rows represent a peptide tested, and columns represent different runs. Green: positive, binding interactions. Gray: negative, non-binding interactions. White: not-tested. Orange: Aggregation. Blue: Non-functional protein.

We first looked for gross patterns of protein non-functionality. Binding and non-binding results for each protein were plotted as a heatmap with each row representing a different peptide and each column representing replicates from a different run or different day. We found both proteins that displayed patterns consistent with normal biological variation, and patterns consistent with degraded protein. For example, in PIK3R1-C interactions with ErbB peptides, eight peptides show positive interactions Figure 2.8. Five of those peptides show consistent results across 3 test runs. One peptide showed consistent positive results across two runs, but a negative interaction in the third run. One peptide showed a positive result on run 1, but aggregation behavior on the remaining two runs. (One other peptide with positive interactions was only tested on one run.) For PIK3R1-C, there are no systemic patterns that suggest protein degradation. Contrast this result with GRB2 and BMX interactions with ErbB peptides (Figure 2.8). For GRB2, four peptides show positive interactions. However none of the positive results

are replicated on run 1. This pattern is strongly suggestive of protein non-functionality of GRB2 on run 1. For BMX, no positive interactions are recorded during any interactions with ErbB peptides. Although it is possible that those interactions are all true negatives, it is also possible that the protein was never functional at all. Heatmaps of this form for each protein can be seen for the 2012 experiment (16) in Figure 2.9 and for the 2014 experiment (17) in Figure 2.10.

2.2.5 Partial Protein Functionality Can Affect Affinity Measurements

The protein functionality analysis indicated that some proteins were likely to be non-functional during the experiment. Since proteins that displayed positive binding interactions at one point in time could seemingly degrade sufficiently to result negative interactions during other runs, we hypothesized that proteins might also partially degrade resulting in effects on the quantitative measurements of K_d. Fitting and parameter evaluation depends on the assumption that the reported concentration of protein is all functional: the K_d parameter is equal to the concentration at which half-saturation is achieved (Figure 2.3). If the effective or functional protein concentration was less than the reported concentration, it would affect K_d values. For example, suppose 25% of the protein was degraded. Then the effective concentration – the concentration of protein available to bind a peptide – would be 75% of the reported value. A saturation binding experiment and parameter fit that reported a K_d of 1μM, would actually be measuring a reaction with a K_d of 0.75 μM. Protein consisting of a mixture of functional and non-functional molecules would result in measurements will artificially higher K_d values (which are lower affinity reactions). Thus partially degraded protein would represent a systematic bias towards weaker interactions and higher K_d values.

Figure 2.9: Categorical Plots of Binding Activity at the Replicate Level from the 2012 Experiment.
Heatmaps of activity for the SH2 domains interacting with ErbB peptides (16). Rows represent a peptide tested, and columns represent different runs. Green: positive, binding interactions. Gray: negative, non-binding interactions. White: not-tested. Orange: Aggregation. Blue: Non-functional protein.

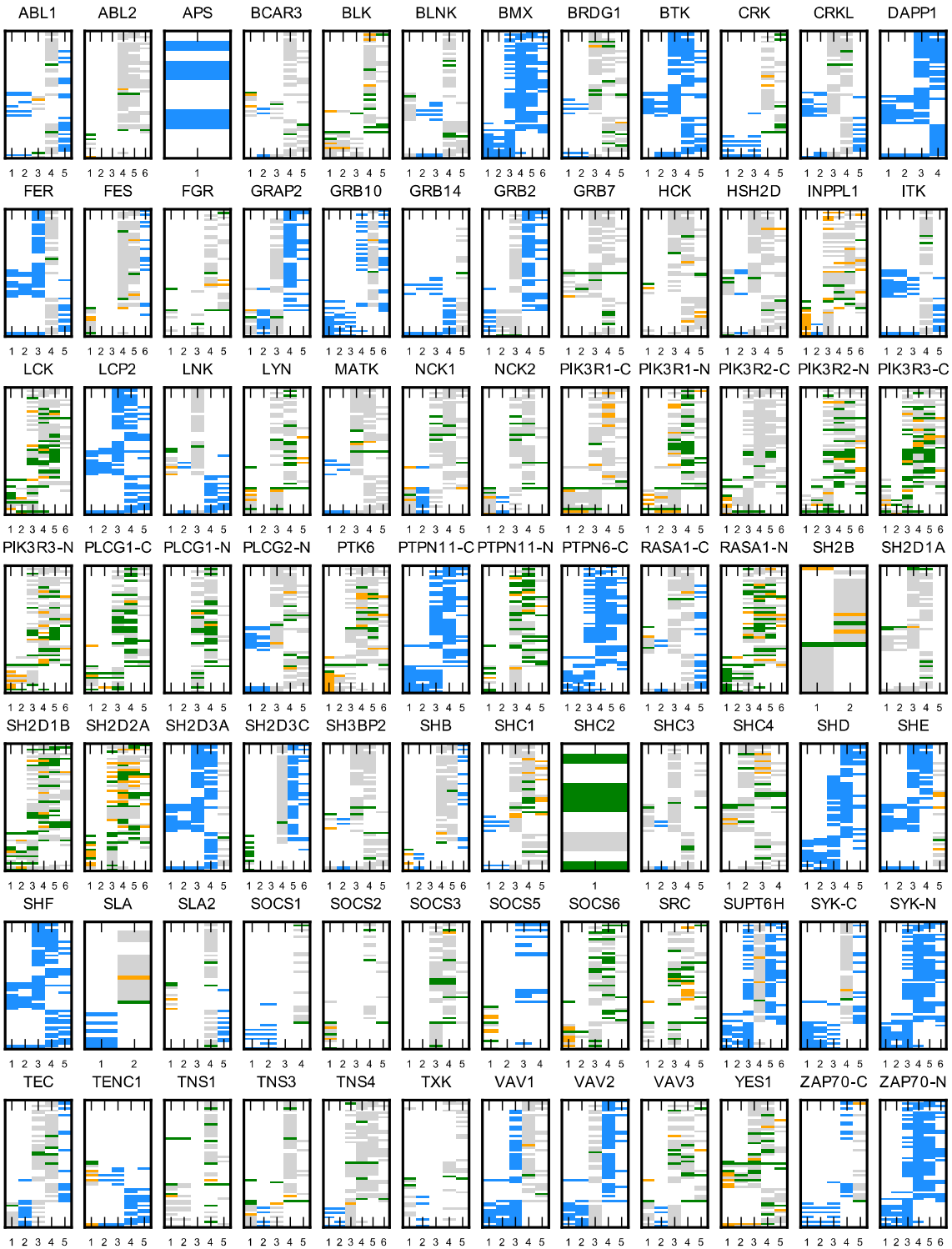

Figure 2.10: Categorical Plots of Binding Activity at the Replicate Level from the 2014 Experiment. Heatmaps of activity for the SH2 domains from non-ErbB peptides (17). Rows represent a peptide tested, and columns represent different runs. Green: positive, binding interactions. Gray: negative, non-binding interactions. White: not-tested. Orange: Aggregation. Blue: Non-functional protein.

We hypothesized that this partially degraded protein would manifest in K_d patterns in the experimental results. If a run contained degraded protein, but the protein was not so degraded as to prevent binding, binders should have a systematic bias towards higher K_d values (lower affinity) than a run with fresh, functional protein. Since each plate compared a panel of 32 proteins against a single peptide, this pattern would require data gathered across sequential plates containing the same protein mix. The pattern would be easier to identify if data were acquired in approximately the same order on different runs. Luckily, the order of tested peptides was predominantly preserved from one run to another in the 2012 experiment (16) enabling exactly this analysis.

In order to test for partial degradation, fit K_d values were arrayed in a table (Table 2.2). Data for each run is sorted in order of data acquisition, and each row represents a peptide. Thus, the columns can be viewed as a short time scale (on the order of minutes between measurements) and different runs representing a long time scale (with hours or days between measurements). In the data for PIK3R2-N, one can see that run 3 contains almost all of the highest affinity (lowest K_d) replicates and run 1 contains the lowest affinity (higher K_d) replicates. Only one peptide in run 1 shows the strongest binding. This pattern suggests that protein in runs 1 and 2 was partially degraded, and protein in run 3 was the least degraded. A related pattern can be seen in the binding data for SH2D2A. In this case, there is no strongest run (with strongest binder distributed randomly between runs 1 and 3), but run 2 shows consistently weaker binding.

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**

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	**12.3**	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	###	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	###	216	**1.98**

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

Table 2.2: Patterns in Kd Values Demonstrate Protein Degradation Effects.
Green highlighted values represent the highest affinity (lowest K_d) measurement for each peptide. K_d values are in µM.

There is no record in the original data for when protein sample changes were made (i.e. if a supply of protein was exhausted and replaced with another supply). We hypothesized that one might find runs of moderate or weak affinity followed suddenly by a very strong affinity. This pattern, also very unlikely due to random chance, would be consistent with an older partially degraded protein sample being replaced by a new, fresh sample of protein. In the data for RASA1-N, one sees exactly this pattern. For the earlier tested peptides in each run, there is no clear stronger sample. However, at plate 174 in run 1, the highest affinity measurements all come from run 1 for the rest of the protein data.

Although these patterns are difficult to detect and are not completely consistent, they are unlikely to be due to systematic errors in data acquisition. Any systematic bias in FP data (such as higher or lower absolute readouts on a particular run) would not result in a similar bias for the calculated K_d. Shifting a saturation curve up or down does not change the calculation of the K_d parameter because Kd is based on the horizontal axis value at half-saturation. Similarly such a bias would not affect the F_{max} parameter which is based on the difference between the maximum value at saturation and the baseline value – both of which would be shifted and thus the difference would remain unaffected. This is exactly the value and rationale in using a parameter derived from multiple measurements at equilibrium.

Unfortunately, there is not enough compatible data to produce control patterns for all proteins, and not enough to be used to infer quantitative protein activity for each measurement. Nevertheless, the patterns strongly suggest that partial protein functionality biases the acquired parameters towards weaker (higher K_d) interactions. As this seems to be a primary source of variation, and it is not random, taking the mean of multiple replicates would result in an inaccurate estimator of the population activity. Instead, the minimum K_d value – the representing

the most functional protein tested – should be used. Although this value still may not represent the true K_d (as the best protein tested may still itself be only partially functional) this value would still be closer to the true value than any other replicate.

2.2.6 Identification of Potential Protein Aggregation

We also hypothesized that non-monomeric protein might manifest with patterns outside those expected of a one-to-one interaction. We found two major trends in interaction data outside of classic saturation binding curves. The first trend was a relatively flat linear response independent of concentration with slopes from 0 to 40mP/uM (see Non-Binder, Figure 2.2). These concentration-independent results with low magnitude signal are exactly what would be expected of true negative interaction for a non-binder. The second trend was a linear response with a high slope, and strong, protein concentration dependent signal (see Aggregator, Figure 2.2). Since signal failed to saturate in these experiments, one-to-one assumptions produced binding saturation curves with extremely high F_{max} values, on the order of $1x10^7$ mP units. Interactions that do not saturate are unlikely to represent true one-to-one binding, and are more likely to represent some type of aggregation phenomena of multiple proteins binding one or more labeled peptides resulting in high volume change and high signal with increasing concentration.

After identifying this phenomenon in individual replicates, we looked for patterns across proteins. Binding, non-binding, non-functional, and aggregation interaction results for each protein were plotted as a heatmap with each row representing a different peptide and each column representing a different run or different day (Figure 2.9 and Figure 2.10, aggregation in orange). Proteins showed mixed results with aggregation and non-binding (e.g. ZAP70-N), or a mixture of aggregation, binding, and non-binding results (e.g. LCK). We did notice a significant decrease in aggregation from the 2012 data to the 2014 data, though no obvious change in

pattern (such as limiting to a time range or run or subset of protein domains). This suggests that this phenomenon might be related to protein preparation, purification, or handling, and that it could be minimized.

2.2.7 Final Process for Selecting Fits

Our fitting procedure uses a classic one-to-one kinetic model to identify positive interactions, as well as alternate models to aid in identification of non-binders and aggregators. Additional criteria, including a measure of signal-to-noise, aid in identifying low-quality measurements. Individual replicate measurements for a particular domain-peptide pair are evaluated categorically and quantitatively. If replicate results indicate positive interaction (binding), the minimum K_d reported is chosen, as explained in the analysis of partial protein functionality. This method results in higher confidence calls for both positive interactions (binders) and negative interactions (non-binders) than earlier methods which only accepted a small fraction of the data and only focused on positive interactions. This is accomplished with limited loss of data into indeterminate categories. An overview of the fitting method used on each replicate can be seen in Figure 2.11.

For domain-peptide pairs where a one-to-one model is the best fit, results were categorized as ‘potential binders’. Potential binders are then tested for signal-to-noise. If the sum of the absolute value of the residuals from the fit (noise) is lower than the difference between maximum signal and the baseline signal (signal), the pair was categorized as a ‘binder’ and the fit parameters (K_d, F_{max}, and F_0) were recorded. Otherwise, if the signal was less than the noise, the measurement was categorized as ‘low signal-to-noise’ and the interaction was considered inconclusive. When a linear model is the best fit, zero or low-slope (under 5mP/μM) indicates a non-binding interaction. A linear fit with a higher slope indicates aggregation, and the replicate is

set aside as inconclusive. A portion of potential binders best fit a one-to-one model according to the AIC, but in essence represent a straight-line no-slope fit, just like a non-binder. Results where the $K_d \geq 1000$ μM (straight line), the $F_{max} \leq 1$ mP (low slope), or the $F_0 \leq 100$ mP (a fitting artifact due to noise on a low-slope straight line) all represented cases where the fit was for all purposes a linear fit, and were categorized as such.

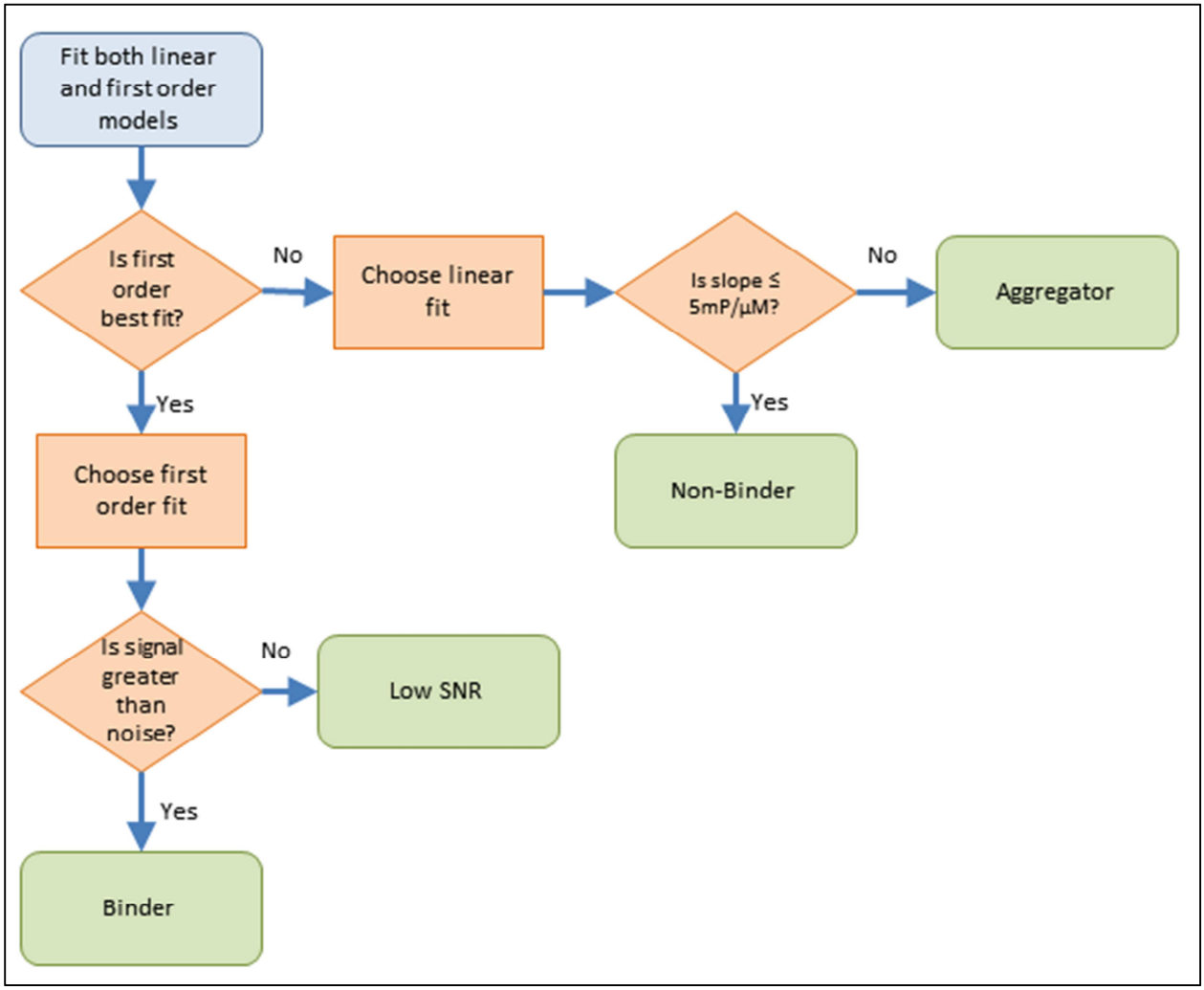

Figure 2.11: Flowchart Describing Fitting Process for Individual Replicate Measurements.

A series of one or more measurements were made for each domain-peptide pair tested. Each measurement has been assigned a category (binder, non-binder, aggregator, or low-signal-

to-noise) as part of the analysis. In order to resolve a call for a domain-peptide pair for the experiment, the calls for each measurement must be considered, and inconsistencies in categorization must first be reconciled. Those conflicts are resolved as in Table 2.3. If the replicate group contains one or more binders, the domain-peptide pair is categorized as a 'binder'. We chose to believe evidence of a binding reaction over inconclusive or potential false negatives because evidence of binding (even in a single replicate) suggests an interaction is possible, despite failure to observe the reaction in other replicates. If a replicate group contains one or more 'non-binders' and zero 'binders', it was categorized as a 'non-binder'. The process used to assign a non-binding result already excludes multiple potentially inconclusive situations. Thus when a group has both non-binding and inconclusive evidence, but no evidence of binding, the group is treated as a confirmed negative interaction, and assigned a non-binding category. If a replicate group contained no 'non-binders', and no 'binders', and only considered of one or more 'aggregators', or one or more 'low signal-to-noise' measurements, it was categorized as inconclusive, and removed from further analysis.

Rule for Handling Replicate Measurements	**Outcome**
Contains at least one binder	binder
Contains no binders AND contains at least one non-binder	non-binder
Contains no binders AND no non-binders AND contains at least one aggregator, 'low signal-to-noise', or 'non-functional' measurement	inconclusive

Table 2.3: Rules for Making Calls on Groups of Replicate Measurements.

When a group contains more than one binder, it has multiple replicates for the fit parameters of K_d, F_{max}, and F_0. Multiple replicates are typically averaged, as they were in the original publications (16, 17), as a sample mean is typically a more reliable indication of a true mean than any single measurement. However, due to systemic evidence of protein degradation (see Results), variances between measurements are likely due to difference in protein

functionality. In the case where a portion of the protein is degraded, the effective concentration available for interactions is always less than the expected concentration. Thus the most accurate measurement would be from the protein with the highest functionality (as the functional concentration would be closest to the expected concentration). Of course, even the highest measurement might not represent a fully functional protein – but it would be closer to the true value than any other single measurement. For this reason, the parameters from the sample with the strongest binding (lowest K_d) were selected as the value for the domain-peptide pair. While this method is a significant departure from convention, the statistical convention is based on assumptions of the source of noise which seem to be violated in these experimental results, requiring a new method to reconcile replicates.

2.3 Methods

2.3.1 Data Description

For each SH2 domain and peptide pair, a Fluorescence Polarization (FP) Saturation Binding Assay was originally performed by Hause and colleagues as described in (16). Data for each protein-peptide pair consists of twelve domain concentrations, ranging from 0.002–10μM, measured interacting with a fixed peptide concentration (20nM), after 20 minutes to obtain equilibrium. Experimental interaction magnitude was recorded in millipolarization (mP) units.

2.3.2 Model Fitting and Selection

Data for each replicate for each pair were fit to two models – a linear function (equation 1) and a function representing a non-linear one-to-one equilibrium interaction (equation 2)

$$F_{obs} = m[domain] + F_0$$

$$F_{obs} = \frac{F_{max}[domain]}{K_d + [domain]} + F_0$$

where F_{obs} is the observed fluorescence (in mP units), *m* is the slope of the fit line, *[domain]* is the concentration of the SH2 domain, F_0 is the baseline value (in mP units), F_{max} is the value at saturation , and K_d is the dissociation constant. Data for each replicate for each pair were fit independently and replicates were handled as described below.

The linear model fit parameters for slope and baseline value, while the non-linear model fit parameters for K_d, F_{max} and baseline value. Baseline values were fit along with the other parameters because background values recorded in the original data were often incongruous with respect to the measurements. Both linear and non-linear fits were performed with least squares regression using the Trust Region Reflective algorithm as encoded by the *optimize.least squares*

module from the SciPy package in the Python programming language. Non-linear fits used the standard least-squares loss function (the default loss function), while linear fits used a modified soft loss method to limit influence of outliers (options “loss='soft l1' f scale=0.1”). Model selection was performed using the Akaike Information Criterion (AIC) (24), with the assumption of Gaussian-distributed errors as described in (23).

For domain-peptide pairs where a linear model represented the best fit (as determined by the AIC score), results were separated into two categories based on the slope of the linear fit. Fits with slopes below 5mP/μM were categorized as ‘non-binders’, otherwise fits with higher slopes were set-aside from further analysis and marked as ‘aggregators’ (see Results section on Determination of Aggregation, elsewhere). For domain-peptide pairs where a one-to-one model represented the best fit, results were categorized as ‘potential binders’. Potential binders were then further categorized into binders, marked as having low signal-to-noise ratio, or flagged as artifacts representing linear fits.

A portion of potential binders best fit a one-to-one model according to the AIC, but essentially represented a straight-line fit. Results where the $K_d \geq 1000$ μM, the $F_{max} \leq 1$ mP, or the $F_0 \leq 100$ mP all represented cases where the fit was for all purposes a linear fit, and were treated as such.

Potential binders were also tested for signal-to-noise (equation 3):

$$\sum_{[domain]_i} |R_i| \geq max(F_{obs}) - F_0$$

where $[domain]_i$ is the i^{th} concentration of the SH2 domain, F_{obs} is the observed fluorescence (in mP units), F_0 is the baseline value (in mP units), and F_{max} is the value at

saturation. If the sum of the absolute value of the residuals from the fit was greater than the difference between the difference between maximum signal and the baseline signal, the pair was categorized as ‘low signal-to-noise’. Otherwise, the pair was categorized as a ‘binder’ and K_d and F_{max} were recorded.

2.3.3 Replicate Analysis

For each domain-peptide pair, based on the number of replicates measured, one or more fits were obtained from the raw data. To assign a final category for the pair taking into account all replicates, the type of fit for each replicate was considered as described in Table 2.3. If the replicate group contained one or more binders, the domain-peptide pair was categorized as a ‘binder’. If the replicate group contained one or more binders, the domain-peptide pair was categorized as a ‘binder’. If a replicate group contained one or more ‘non-binders’ and no ‘binders’, it was categorized as a ‘non-binder’. If a replicate group contained no ‘non-binders’, and no ‘binders’, and one or more ‘aggregators’, it was categorized as an ‘aggregator’ and removed from further analysis. If the replicate group contained only ‘low signal-to-noise’ measurements, it was categorized as ‘low signal-to-noise’ and removed from further analysis.

For remaining domain-peptide pairs categorized as ‘binders’, the K_d, F_{max}, and associated confidence interval of the replicate with the minimum K_d was selected for the final value of the domain-peptide pair.

2.4 Results

2.4.1 Description of Results

Of 12147 interactions tested, 1506 interactions (12.5%) resulted in a positive interaction, and 10628 interactions (87.5%) resulted in negative interactions. The remaining 1117 interactions (9.2%) were indeterminate. Of the indeterminate interactions, 382 (3.1%) were due to non-functional protein, 312 (2.6%) were due to low signal-to-noise issues, and 277 (2.3%) were due to protein aggregation. The results are plotted as a heatmap in Figure 2.12.

We hypothesized that domains would bind a relatively small proportion of peptides, on the order of 20%, as the SH2 domain interaction with phosphorylated peptides is believed to be at the core of specificity determination. Although the total binding fraction was even lower – closer to 12.5% – individual domains actually bind widely varying percentages of tested peptides: from less than 1% to 58.5% of tested peptides (Figure 2.13). Different domains have a wide range of selectivity and specificity for their targets, suggesting that more promiscuous domains have a purpose outside of determination of specificity.

In order to make calls on the 12147 interactions tested, a total of 37488 replicate measurements were analyzed. Fitting results show a substantial portion of replicates (over 30%) resulted in indeterminate results (e.g. Low SNR, Non-Functional, and Aggregator categories, see Table 2.4). Yet, final determinations of domain-peptide interaction results (the calls made on all replicates for a domain-peptide pair) only resulted in 9.2% indeterminate results. This suggests that many results which had one or more indeterminate results were ‘rescued’ by either a true positive result or a high-confidence negative result.

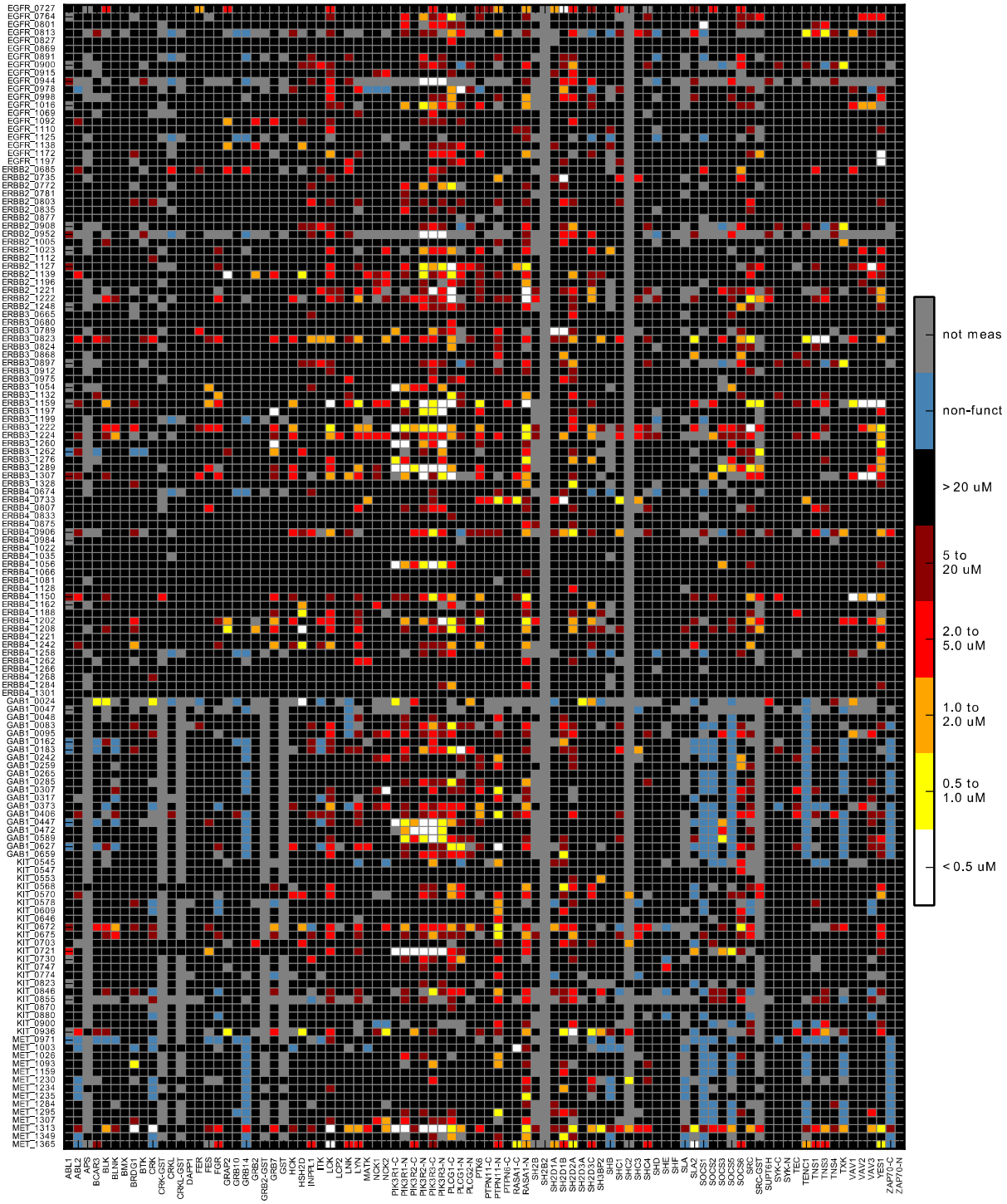

Figure 2.12: Results of the Revised Analysis of Jones Group FP Data.
Heatmap plot of the revised analysis. Peptides (y-axis) are plotted against domains (x-axis) with K_d values represented by a heatmap. The color scale is based on a temperature scale, with lower affinity signals displaying as dark squares, and higher affinity (lower K_d) signals rising through red to orange, yellow, and white. Gray squares represent interactions not measured. Blue squares represent protein identified as non-functional, and thus represent indeterminate results.

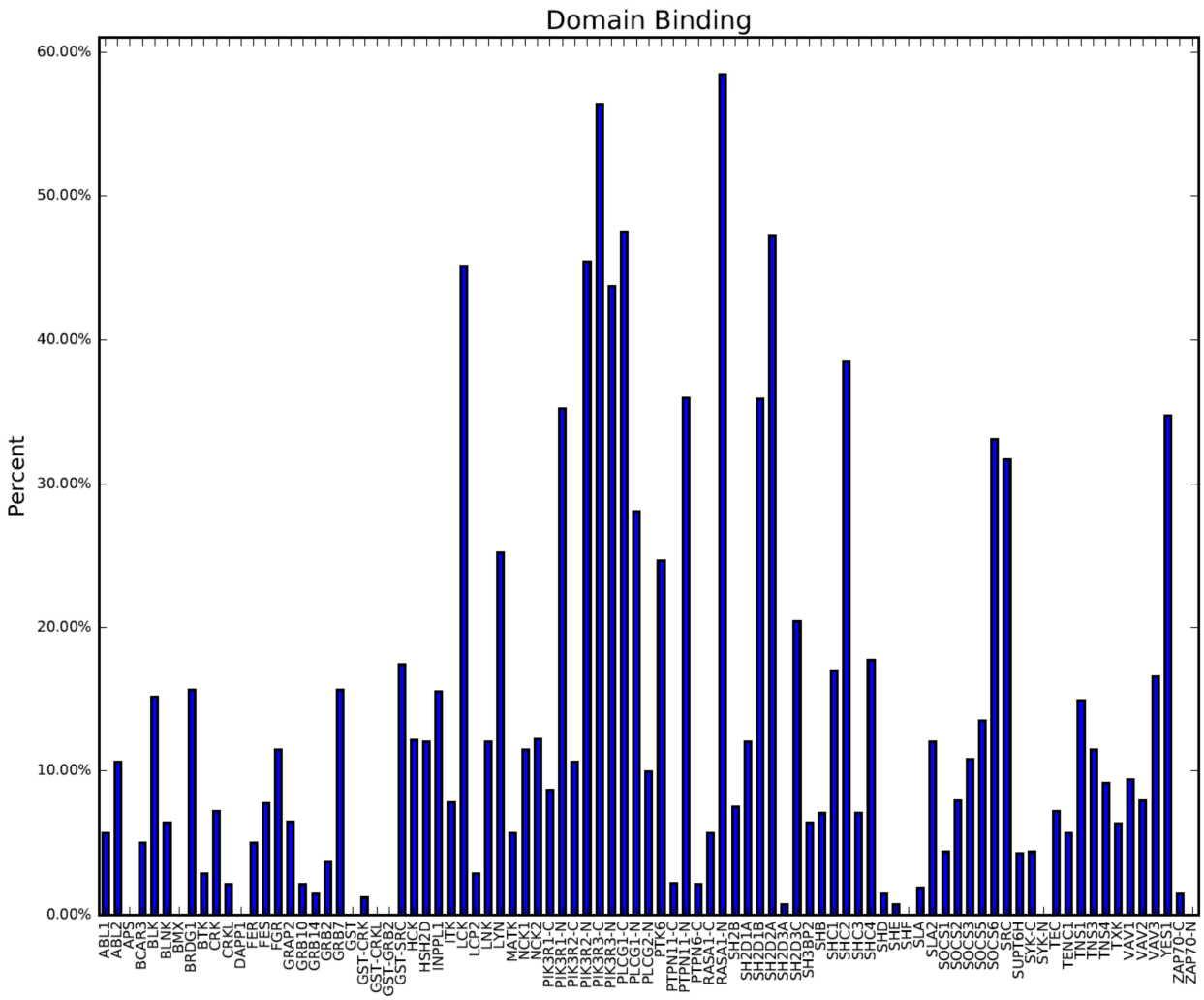

Figure 2.13: Fraction of Peptides Bound by Each SH2 Domain.
A plot of the fraction of peptides tested resulting in a positive interaction (binder) for each domain.

Categorical Binding Call	Number of Replicates	Percent
Binder	2880	7.7%
Non-Binder	23208	61.9%
Low SNR	2555	6.8%
Non-Functional	6976	18.6%
Aggregator	1867	5.0%

Table 2.4: Summary of Fit Results at the Individual Replicate Measurement Level.

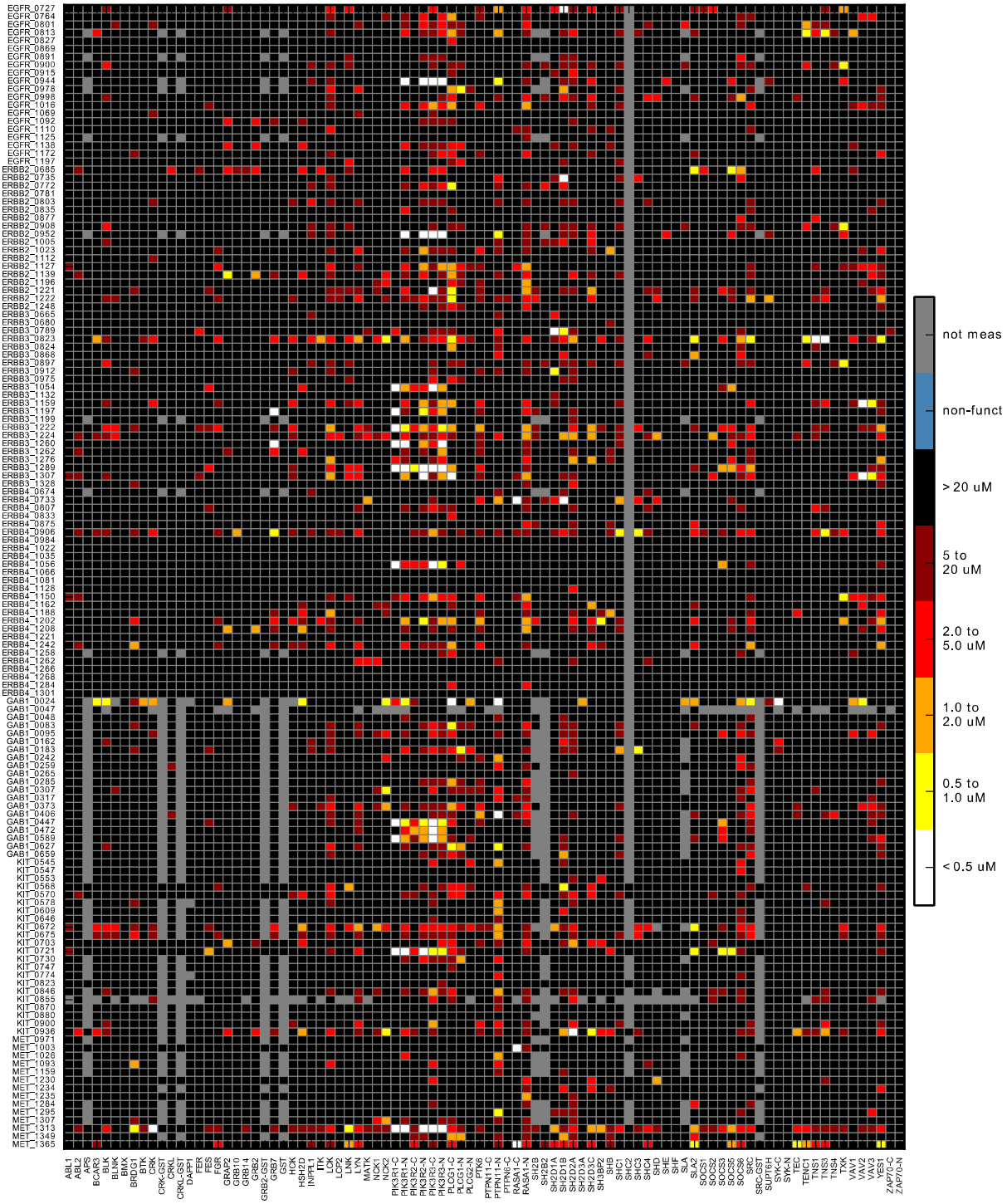

Figure 2.14: Published Results From The Jones Group FP Data.
Heatmap plot of the original published results – revised and replotted here to match the format of results from the reanalysis. Peptides (y-axis) are plotted against domains (x-axis) with Kd values represented by a heatmap. The color scale is based on a temperature scale, with lower affinity signals displaying as dark squares, and higher affinity (lower Kd) signals rising through red to orange, yellow, and white. Gray squares represent interactions not measured.

This aspect of replicate-level analysis represents a significant improvement in confidence of quantitative results for positive interactions, and certainty over validity of negative interactions. The failure to remove questionable replicates would significantly impact any metric based on average replicate values.

		Revised Analysis Calls					
		Binder	**Non-Binder**	**Aggregator**	**Low-SNR**	**Not Functional**	**n/a**
Published Calls	**Binder**	1225	198	47	27	10	12
	Dropped	281	9326	230	285	372	134

Table 2.5: Comparison of Qualitative Binding Results.

2.4.2 Comparisons with Published Results

The original published fitting results are also plotted as a heatmap (Figure 2.14). Results from this reanalysis differ significantly from the published results on the same underlying data. Both quantitative and categorical differences can be seen.

The original data identified 1519 binders, out of 12147 interactions (12.5%). The remaining 10628 measurements (87.5%) were discarded in an outgroup containing the poor fits, noisy data, non-binders, etc. This new analysis produces approximately the same number of binders (1506 vs 1519), but the identity of these binders has shifted dramatically. In addition, the analysis recovers 10268 high-confidence non-binding interactions representing approximately 87.5% of the original data. With respect to positive categorical interactions, out new analysis agrees on approximately 71% of the positive calls. Of the positive calls, 1225 overlap between the two sets, but we disagree on 479 interactions. Our new analysis recovers 281 binders from

the ‘dropped’ pool of the published data, but 294 interactions called binders by the original data are classified as non-binders or indeterminate by this work (Table 2.5).

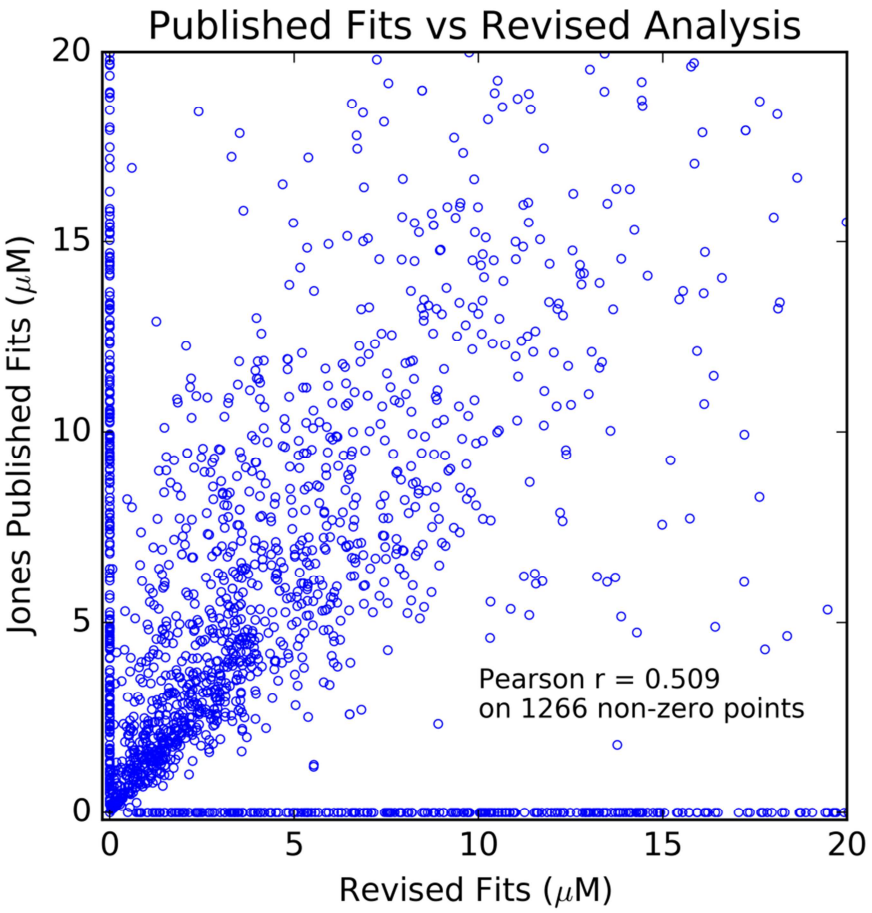

Figure 2.15: Scatter Plot Comparing Published Results With the Reanalysis.
Published fits are compared with our revised fits. Points falling along the x- or y-axes represent data in one set that was identified as a binder but not by the other. Pearson correlation was calculated only on the positive interactions common to both sets.

Quantitative results differ even more drastically (Figure 2.15). Since we select minimum K_d to represent a group of replicates instead of mean K_d, we expected that a significant fraction of our calls would result in lower K_d (higher affinity) calls. We found that of the 1226 interactions, we reported a lower K_d on 830 (67.7%). Nevertheless, we also matched the originally published K_d calls (±10%) on 201 interactions (16.4%), and reported higher K_d values on 235 interactions (19.2%). The fraction of higher K_d values is not surprising, since the

changes we proposed in the fitting process resulted in the addition of new positive interactions (previously abandoned in the published work) as well as exclusion of negative and inconclusive interactions (previously considered to be positive interactions in the published work). Nevertheless, we were surprised to see that our new results do not correlate well with the originally published results (giving a Pearson correlation coefficient (*r*) of only 0.509). Although K_d calls between 0μM and 5μM have slightly stronger correlation, most interactions (especially those above 5μM) are essentially randomized when compared to the original published results.

2.4.3 Validation with Known Interactions

In order to determine if these new fitting results were consistent with the known response we examined a well-studied biological system: the Epidermal Growth Factor Receptor (25–27). Some examples of expected downstream responses to EGF stimulus are shown in Figure 2.16. Multiple SH2 domains show affinity for each of these receptors, so the predicted outcome of competition should match the known response of the system.

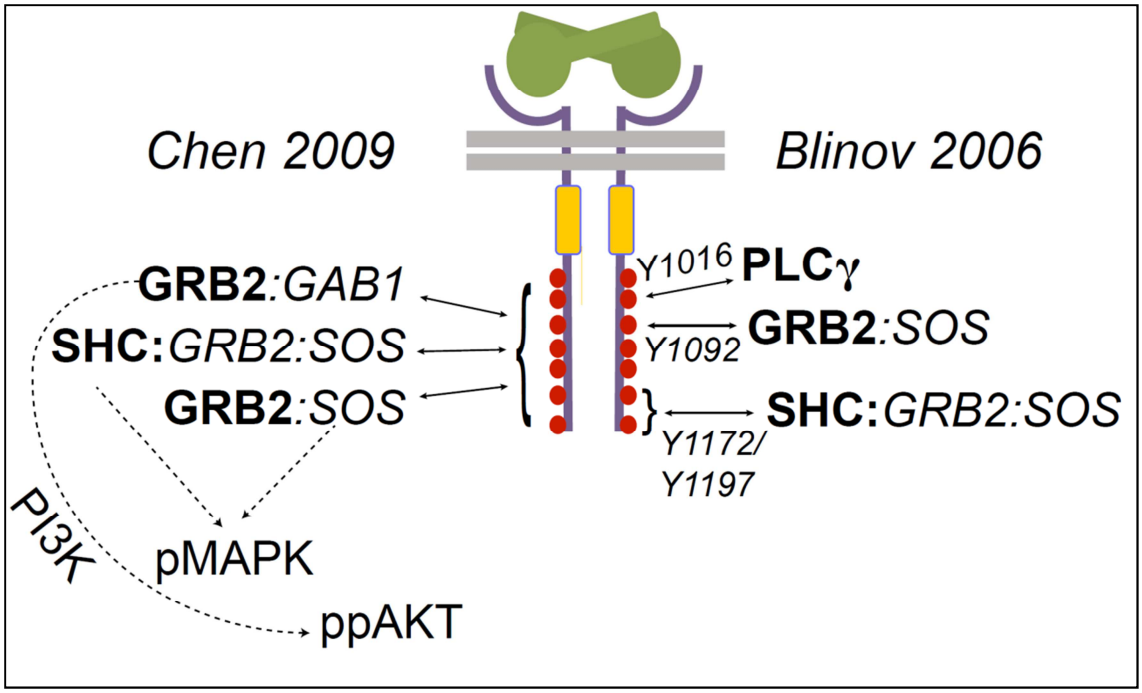

Figure 2.16: Validated SH2 Domain Interactions on the Intracellular Tail of EFGR.
A diagram of expected interactions on the intracellular tail of EGFR.

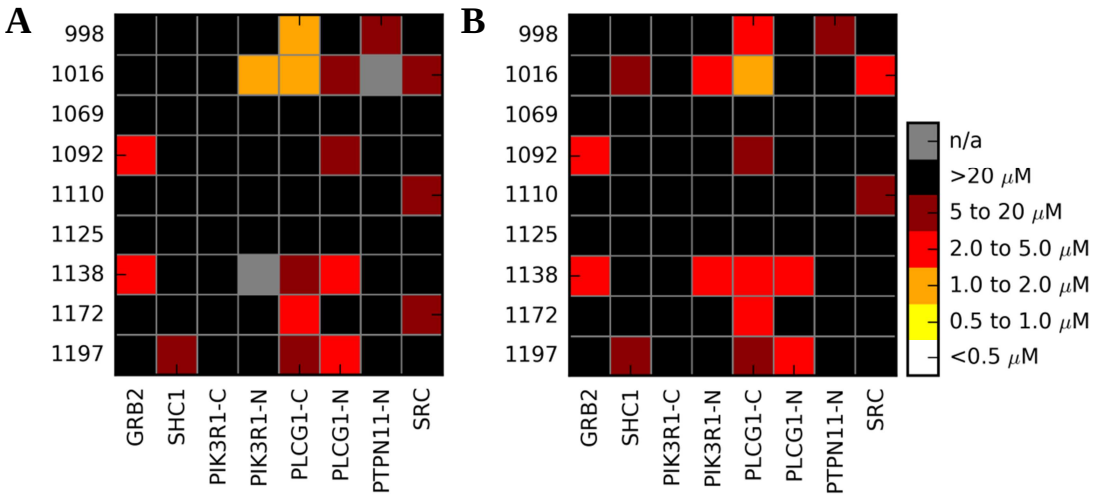

Figure 2.17: Heatmap of EGFR Tail Interactions from the Reanalysis.
(A) Results from our revised analysis. (B) Published results from the Jones FP data (16). Peptides (y-axis) referenced by position of the phosphotyrosine are plotted against domains (x-axis) with K_d values represented by a heatmap.

At tyrosine 1016, PLCγ is predicted to be the strongest interaction. Our reanalysis find that either the N-terminal domain from PIK3R1 or the C-terminal domain from PLCγ should have the strongest interactions, consistent with the predicted outcome. The published data (16) is also consistent, predicting the C-terminal domain from PLCγ. Similarly, for tyrosine 1092 both our reanalysis and the published data suggest that GRB2 has the highest affinity (lowest K_d). Although both our data and published data predict PLCγ as the strongest interaction for tyrosine 1172 and 1197, the models are merged for those residues and predict SHC as the strongest binder. This model from Blinov, et al.(25) has some limited usefulness for validation purposes, as it was primarily attempting to connect MAPK and PLCγ signaling, so no predictions were made for PI3K, for example. Nevertheless, these results suggest that on a well-known system, our reanalysis has not resulted in a divergence from expected interactions, and that the best understanding similarly accurate on this system.

2.5 Implications of Revised Results

This revised analysis, despite using the same raw data, presents significantly different results. Qualitatively, we only have approximately 70% agreement with the published results on positive domain-peptide interactions. Quantitatively, there is almost no agreement with the published work. We wanted to determine if these quantitative results changed the conclusions one could draw from the underlying behavior of SH2 domain data.

First, we explore the accuracy of binding models that assume independence between positions (e.g. PSSM/PFM-based models). Here, we examine the predictive power of Scansite binding motif models and compare to published work and to the revised analysis. We also evaluate the accuracy of new protein binding models derived from our analysis. Second, we explore the conclusion drawn from earlier binding results that domains which are more closely related interact with similar sets of peptides. Finally, we examine a surprising conclusion from this analysis about the behavior of GST-labeled proteins, which has implications for interpreting previous high-throughput studies, and impacts the work that is derived from them.

2.5.1 Evaluation of Scansite Protein Binding Models

Scansite protein binding models are derived from degenerate library experiments (28). These experiments were designed to identify residues at each position of a phosphorylated peptide that contribute to binding. Each finding at each position was identified independently of each other position. Scansite models are matrices representing the frequencies of each amino acid at each position and are in the form of a matrix like a position specific scoring matrix (PSSM) or position frequency matrix (PFM). This matrix can be used to score any peptide, with the highest scores representing true binders, given the assumption that each position acts

independently of every other position. The primary limitation of this model is that amino acid contributions to binding can be interdependent. 'Non-permissive' residues – residues in a peptide that disrupt binding despite the presence of other residues correlating with strong binding (8) – were identified as one such example. We would expect that a model based on independence would result in false positive predictions when non-permissive residues were present.

First, we evaluated whether Scansite scores correlated with affinity published affinity data. Since we have already established that the Jones fluorescence polarization (FP) data (16, 17) and MacBeath protein microarray (PM) data (14) do not correlate with one another, we evaluated both data sets against the Scansite scores. Scansite models are available for the following domains: ABL1, CRK, FGR, GRB2, ITK, LCK, NCK1, PLCG1-C, PLCG1-N, SHC1, SRC, PIK3R1-C, PIK3R1-N. Scansite scores do not substantively correlate with binding affinity for either data set (Figure 2.18).

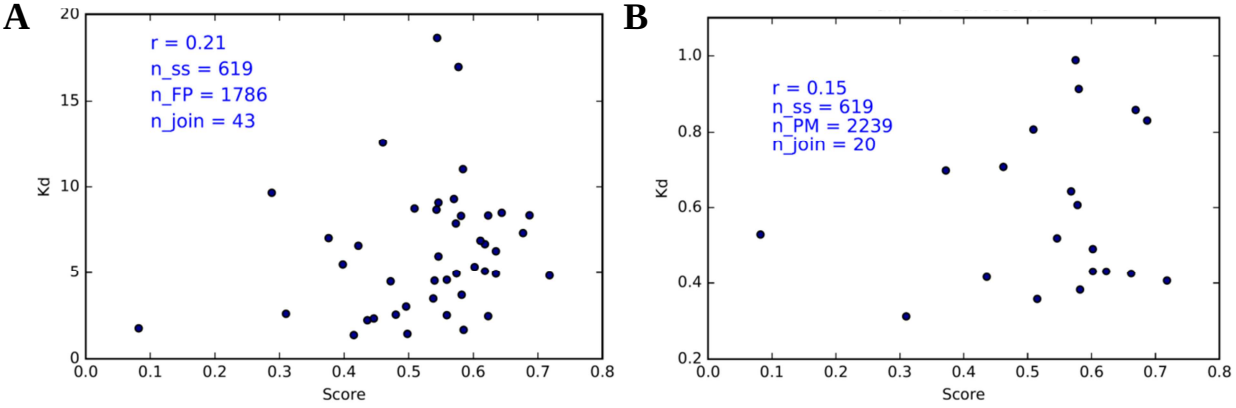

Figure 2.18: Comparison of Published Data with Scansite Scores.
A comparison of Scansite scores to published quantitative data from two groups was performed. (A) Jones group FP data (16, 17). (B) MacBeath group PM data (14). K_d values are in µM.

We next compared to our revised results. Although correlation analysis can identify similar quantitative values, and also might identify similar ranking, it would not clearly identify a case where true binders are enriched near the top ranks of Scansite scores. In order to evaluate

the Scansite models, we plotted receiver operator characteristic (ROC) curves for each model, and calculated the area under the ROC (AUROC) (Figure 2.19). Most Scansite models had poor predictive power with AUROC scores between 0.520 and 0.650. Two models performed worse than chance (CRK and ITK), but since most positive binders were ranked near the bottom of the list, CRK would perform much better if the scoring ranking were reversed. Although three domains scored well via Scansite (NCK1, AUROC 0.718; PIK3R1-C, AUROC 0.835; and GRB2, AUROC 0.955) overall performance was very poor.

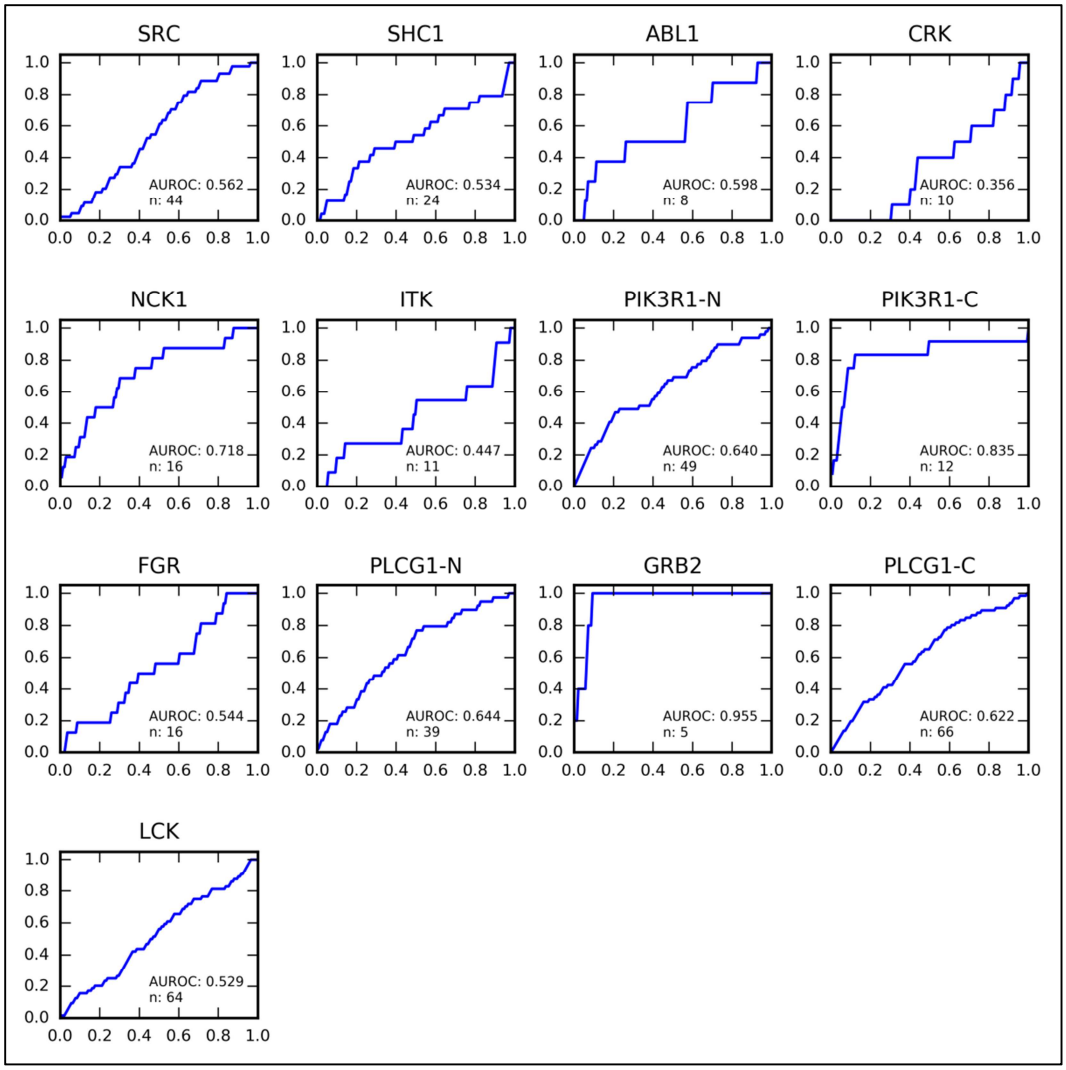

Figure 2.19: Evaluation of Scansite Binding Motifs.
Scansite scores were calculated for each peptide from several domains. Scansite scores were then ranked, and compared to binding results from our revised analysis. Results were plotted as ROC curves, and the area under the curve (AUROC) was reported.

2.5.2 Evaluation of Newly Constructed Protein Binding Models

We were interested in exploring whether PFM-based protein binding motif models would be more successful at predicting binding when built from the data in the revised analysis. In order to evaluate this, we built a PFM-based model from the 'binder' peptides for a domain, and then scored all peptides for that domain with the model. Although this method is not as rigorous as a cross-validated test set with holdouts, it was sufficient to identify general trends in quality for the models.

When comparing qualitative calls for binders and non-binders to the Jones FP data, it was reasonable to match the same 20μM cutoff as used in the original publication. Here, however, a call of 'binder' vs 'non-binder' is arbitrary, and calls at different K_d values might lead to different results. In order to examine the effect of making a categorical call about a continuous affinity phenomenon, we built PFM-based models for 'binder' calls at different K_d thresholds. Increasing the K_d threshold for binder calls tends to increase the number of peptides incorporated in the model, and since peptides are unique, this increases peptide diversity.

Results for several examples are plotted as ROC curves in Figure 2.20 and a summary of all results by domain can be found in Figure 2.21. Several trends were obvious from this analysis. First, all models performed well, as would be expected given than they were tested on the same data they were trained on. Also, all models performed better than Scansite models. However, all models performed progressively worse as the K_d threshold for binding was increased. This trend is reasonable to expect, because as the K_d threshold increases, more peptides are included. If those peptides are significantly different than peptides already included, the model will become increasingly degenerate.

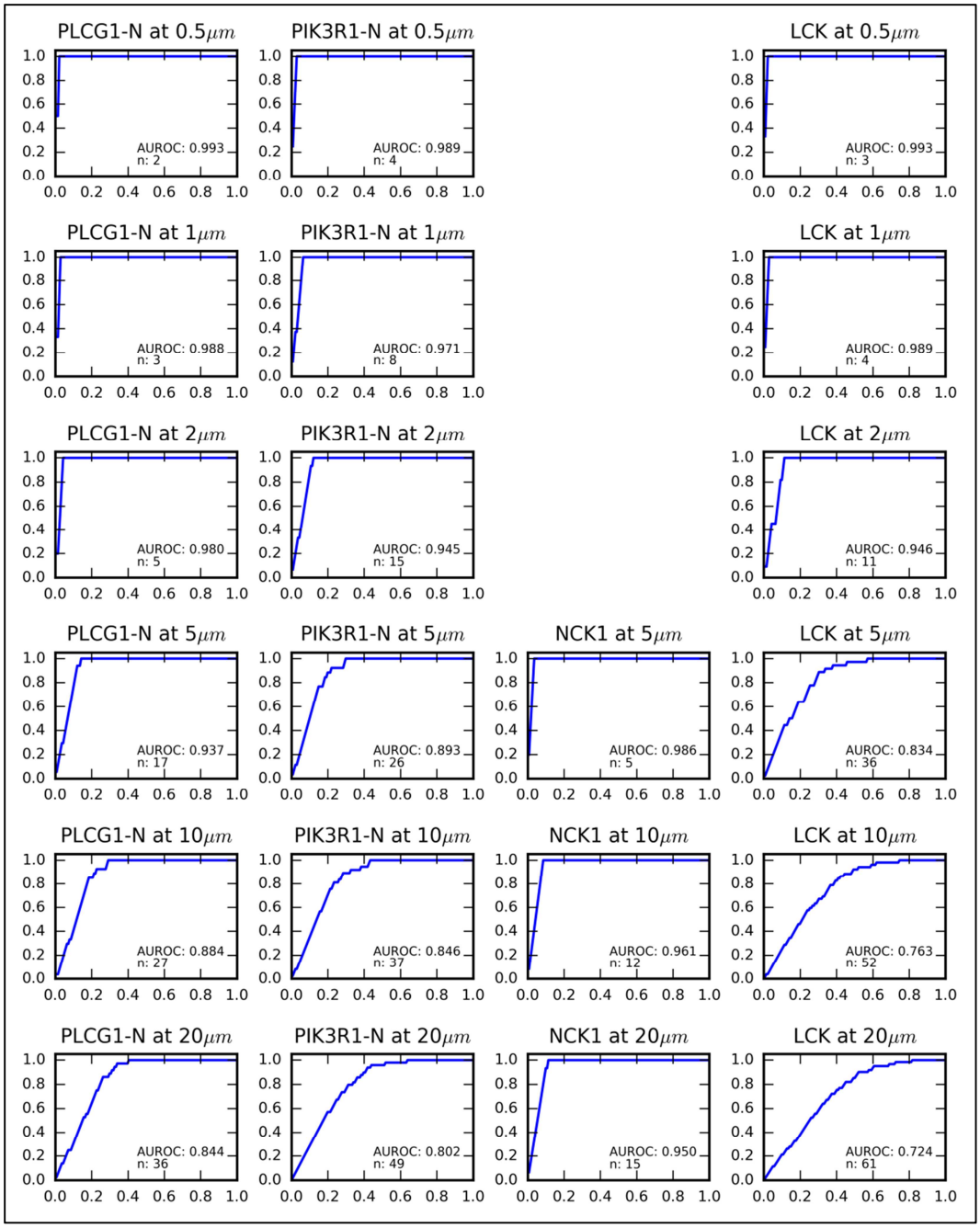

Figure 2.20: Evaluation of Protein Binding Motifs Created at Different K_d Thresholds.
New motifs for each domain were created from positive interactions in our revised analysis. A different motif was created for positive results at each K_d threshold. Peptides were scored and then ranked, and compared to binding results. Results were plotted as ROC curves, and the area under the curve (AUROC) was reported.

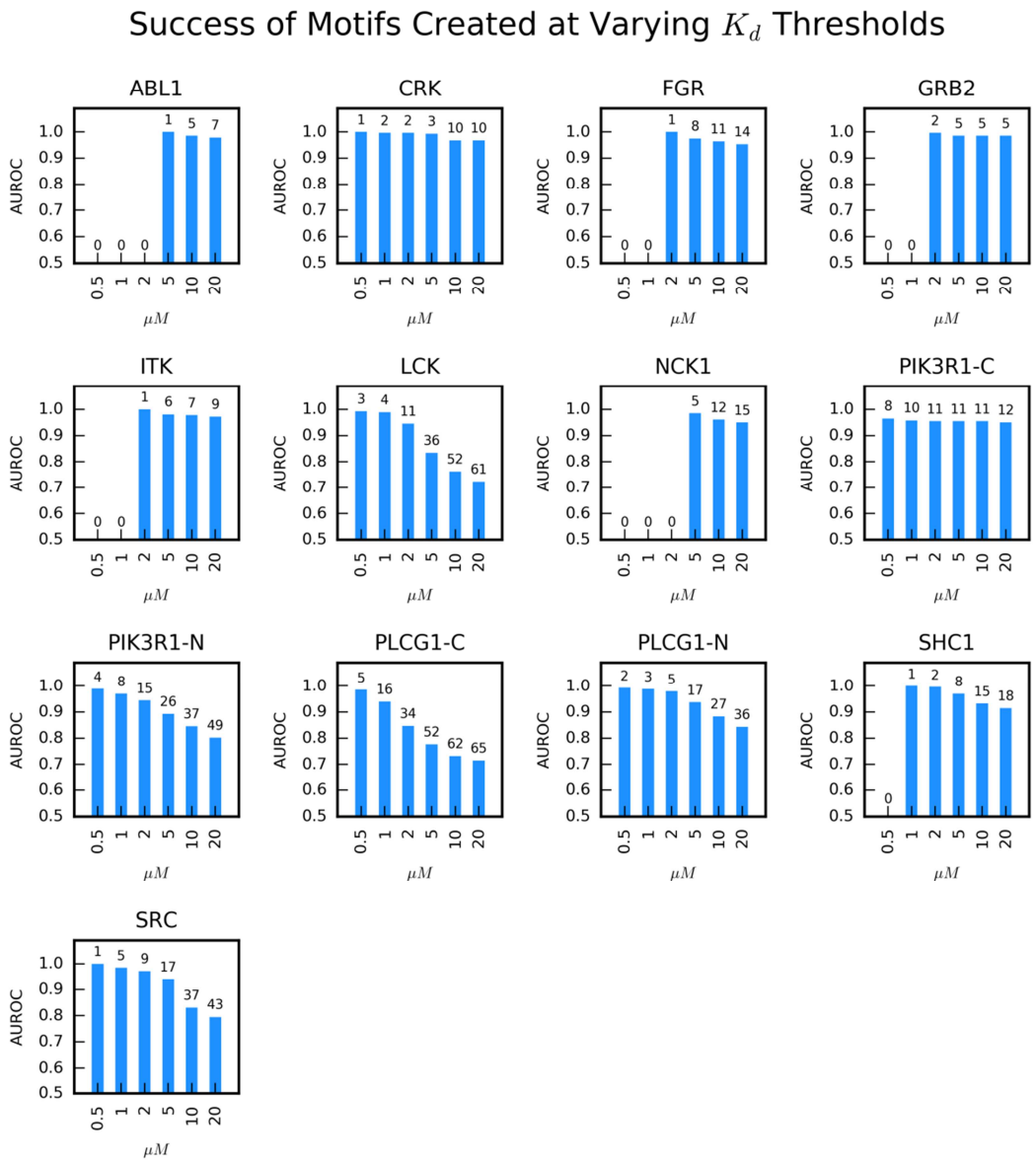

Figure 2.21: Summary of Protein Binding Motif Evaluation.
Summary plot showing predictive power of protein binding motifs from revised analysis results. AUROC scores from each domain at each threshold concentration (see Figure 2.20) were plotted. The value above the bar represents the number of positive results used to build each motif at that K_d threshold. The relationship between number of positive results and AUROC score are plotted in Figure 2.22.

However, trends in the relationship between number of peptides included and AUROC suggested that we examine that relationship. We plotted AUROC vs K_d threshold for each domain and threshold as a scatter plot (Figure 2.22). The results indicate a very strong negative correlation (Pearson's r: -0.991) between AUROC and K_d. This suggests that PFM-based models

might have built in boundaries on predictive power with this type of diverse peptide data, and is worthy of further exploration in future work.

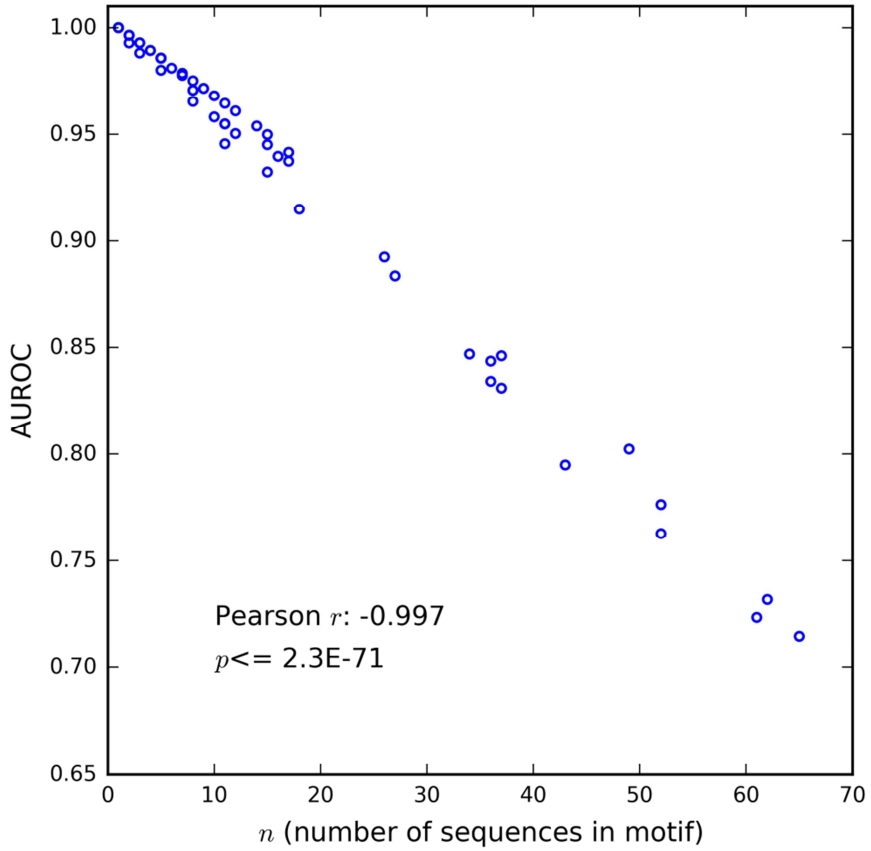

Figure 2.22: Relationship of AUROC with Number of Sequences in Protein Binding Motif.
The number of positive interactions (binder) sequences used to create each motif is plotted against the predictive power of the motif (as measured by AUROC score). The relationship is strongly negatively correlating (Pearson's r: -0.997).

2.5.3 Many Closely Related Proteins Do Not Have Similar Binding Profiles

Analysis of binding results from some SH2 and SH3 domain interaction studies have suggested that closely related domains (domains from same evolutionary family) bind similar peptide profiles. The Nash group reported finding multiple domains which had similar binding profiles as other members of the same family (8). Other groups report similar findings for SH2 and SH3 domains (29, 30). This finding has been used in algorithms to predict peptide interactions (31). Although closely related protein domains have less sequence divergence, and exhibit more similar fold structure, closely related protein domains arise from duplication events. A duplication event can free one copy of a protein from selection pressure, potentially allowing for rapid functional divergence. Plus, even a single amino acid mutation in a protein domain could radically affect protein function, and could enhance or destroy binding, or change specificity. We examined the data from our revised analysis to determine protein binding behavior by domain family.

Using the Jaccard similarity coefficient,

$$J(A,B) = \frac{|A \cup B|}{|A \cap B|}$$

(which compares the membership of the intersection of two sets with the union of the sets), we compared the binding profiles all domains. Binding was 'binarized' at a 20μM threshold. Protein families were defined by Ensemble gene trees. As domains with limited number of positive binding examples may skew results, we only considered domains with 10 or more binders. In Figure 2.23, we compare binding profiles for protein domains within closely related families.

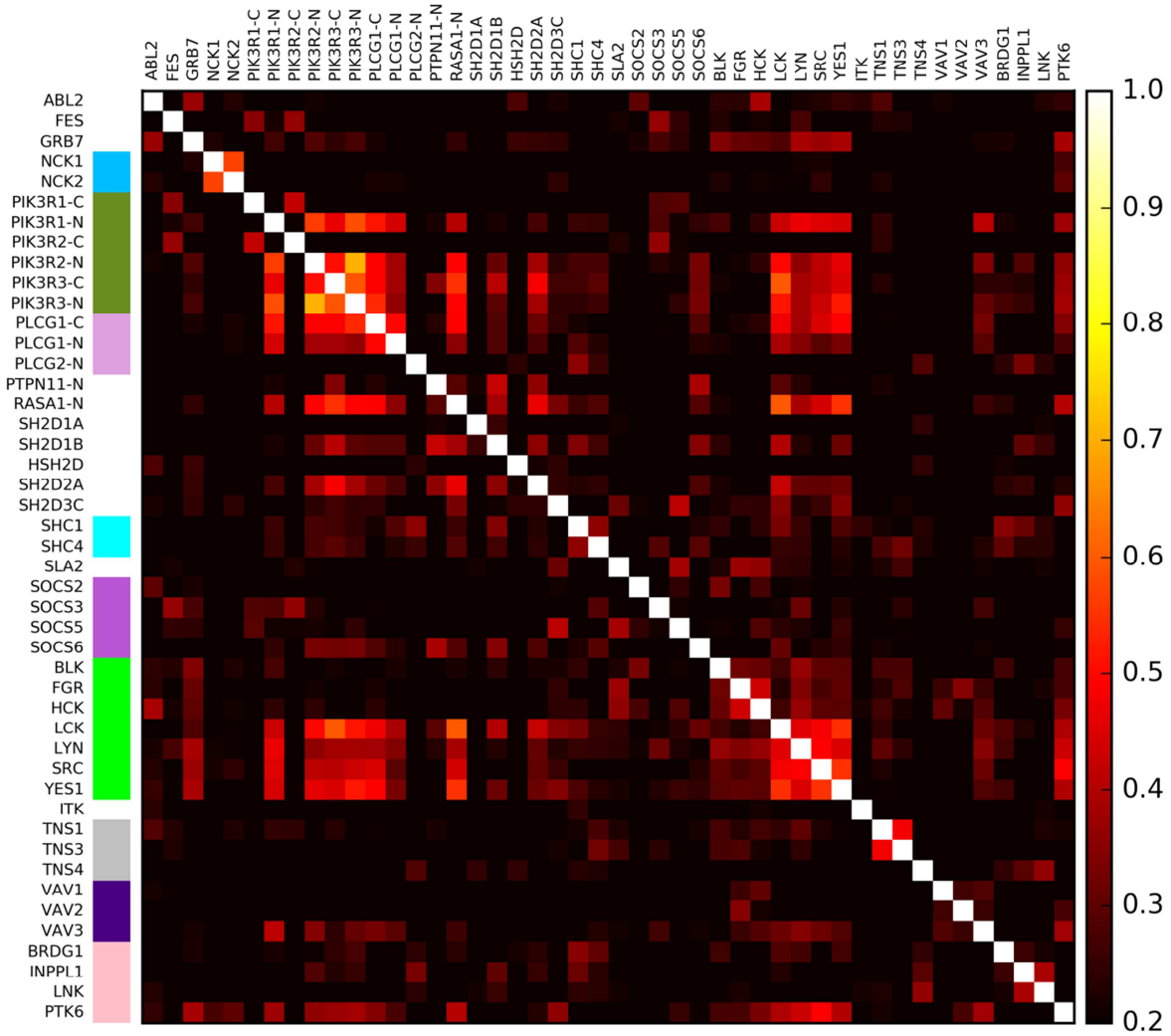

Figure 2.23: Domain Binding Similarity for a Subset of Protein Families.
Comparison of binding similarity for proteins with more than 10 positive interactions. Similarity is measured by the Jaccard similarity index and displayed as a heat map, with a highest similarity value of 1.0 displayed as white. Scale is a temperature –style scale with low similarity displaying as black, and increasing similarity displaying as higher temperature colors from red through orange, yellow and white. Protein families are marked with a color bar on the left. Blue – NCK. Olive – PIK3R. Light purple – PLCG1. Cyan – SHC. Purple – SOCS. Green – SRC. Gray – TNS. Dark purple – VAV. Pink – LNK.

The analysis compared 49 proteins from more than 10 different protein families. Overall, very few individual domains exhibited high similarity with other domains – only two domains tested were closer than 0.60 on the Jaccard similarly index (with 1.0 representing perfect match between binding members). This suggests that most proteins show significant divergence in binding. Some protein families showed highly similar binding patterns within the family. NCK,

PIK3R, SHC, and SRC had the highest within family similarity, but that similarly at best was around 0.70. Some families showed very interesting divergent behavior. For example, VAV family members had very low similarity to one another – about the same low similarity as to PIK3R and SRC. In the TNS family, TNS1 and TNS3 had similar binding, but neither were similar to TNS4. Interestingly, families like PLCG1, PIK3R, and SRC which are expected to be far apart in evolutionary distance (32), have very similar response across families.

Although in some cases, domain families exhibit similar binding profiles, in many cases they do not. Methods which rely universally on this assumption should be reconsidered in the light of these findings.

2.5.4 GST Affects Binding Profiles in a Non-Linear Manner

One surprising result from our revised analysis came from examining binding profiles of the subset of GST-tagged proteins. Although only a small number of measurements were made with GST-tagged proteins, comparison of the GST-tagged to the untagged proteins shows that GST-tagged proteins behave very differently than non-tagged proteins (Figure 2.24).

Four proteins were measured against the panel of peptides as tagged and untagged versions. The GST tag was also measured by itself. GST alone showed no positive interactions. For CRK, CRK-L, and GRB2 the GST tag generally prevented binding when compared to the untagged version. For example, all peptides that were positive interactions for GRB2 were also tested on the GST-labeled molecule but did not bind the GST-labeled molecule. Similarly for CRK, there were 8 peptides where CRK bound, but CRK-GST did not bind (and many more where CRK bound but were not tested on the GST-labeled molecule). Interestingly, one peptide bound CRK-GST that did not bind CRK. In contrast, GST-labeled SRC did not prevent binding. Instead, it seemed to radically change the affinity for almost all peptides.

There are several potential explanations for this phenomenon. First, GST is known to cause proteins to dimerize (33) which could result in interference with access to the binding pocket of the SH2 domain, or could affect the affinity for all or a subset of peptides by steric hindrance of certain peptides. However, in a handful of cases, affinity surprisingly increased for the GST-labeled molecule.

One high-throughput data set testing SH2 domain interactions had all protein domains labeled with GST (18). Although our results are based on FP data, they strongly suggest that SH2 domain affinity with GST-tagged protein is not representative of non-tagged protein.

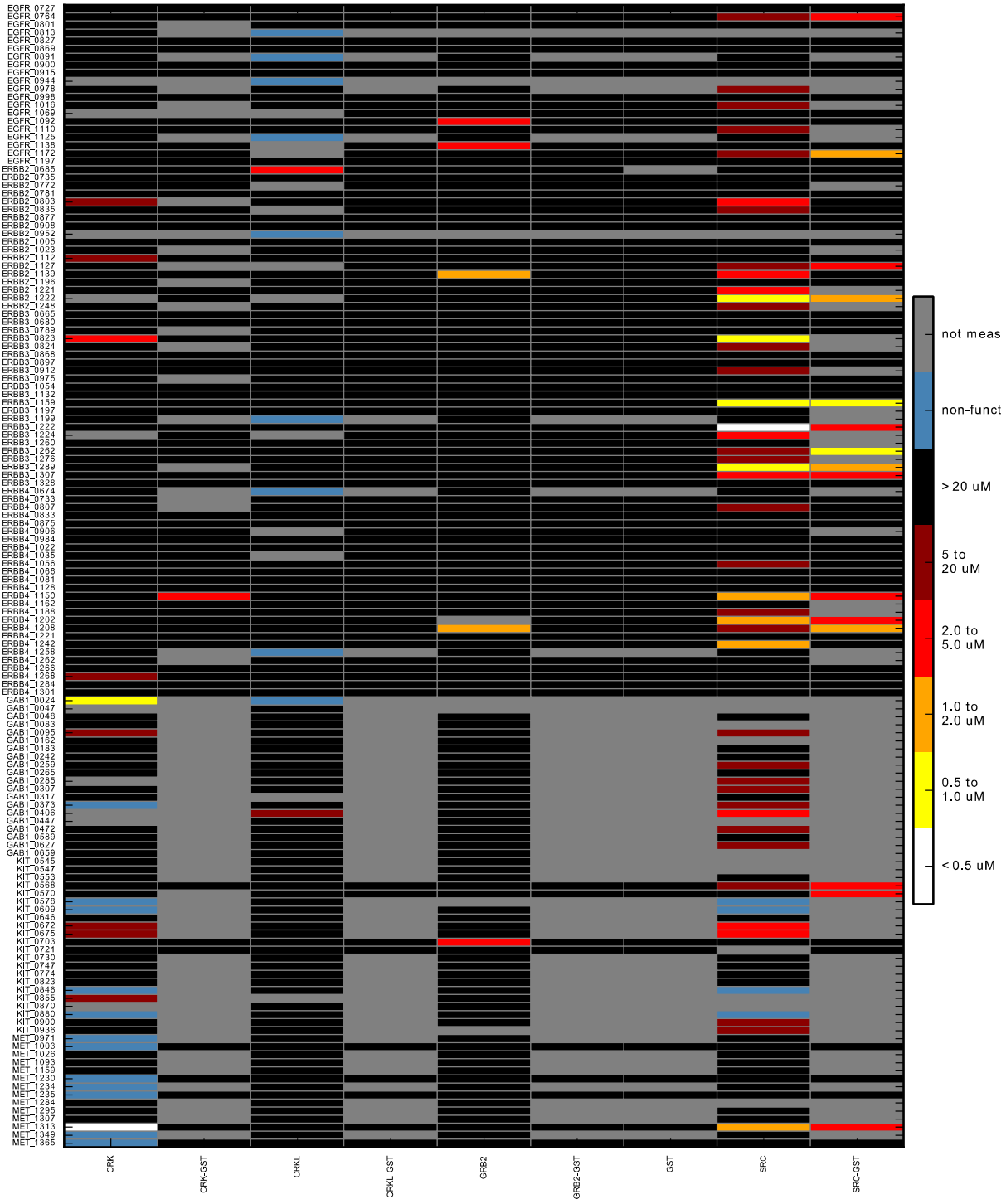

Figure 2.24: Effect of GST-Tagging on Interaction Affinity.
Comparison of binding profiles of GST-tagged protein to non-tagged protein, plotted as a heatmap of affinity (K_d). Gray values represent not measured interactions. Blue values represent non-functional protein (and thus inconclusive measurements.) Higher affinity interactions have higher temperature colors.

2.6 Discussion

In this work we have reviewed the raw data from a set of fluorescence polarization experiments designed to measure SH2 domain interaction with phosphorylated peptides. We found that the experimental design lacked important controls to establish protein functionality and peptide activity, and some ways to compensate for those problems. We also used different analysis techniques including fitting multiple models and better suited noise and model selection criteria which allowed us to improve the calls of both positive and negative interactions. Our qualitative calls show significant differences, and our quantitative results differ greatly from the original publication. We would like to comment on the factors which give credence that this revised analysis is better than the original published analysis, and discuss the implications for the use of other SH2 data in future research.

The fluorescence polarization method employed by Jones FP is likely to give accurate and sensitive results when conducting high-throughput measurements of affinity. Unlike protein microarray experiments where proteins are not maintained in solution and reactions are carried out on a surface, fluorescence polarization experiments are done completely in solution. In addition, the process of handling samples was done robotically, likely eliminating many manual errors in handling. This can be seen in the individual interaction measurements displayed low noise and the vast majority of results represented high-quality measurements when examined with the correct analytical tools. Although this was the only raw data available, it is an experimental method which is likely to yield an accurate reading of affinity.

Nevertheless, different runs in this experiment seemed to have a dramatic effect on calculated SH2 affinity. We identified patterns in the data acquisition which suggest that these

variations were likely due to protein degradation or preparation, and could very well be traced back to the difference between assumed concentration and actual active concentration of the protein. With the chosen affinity model, accurate measurements of affinity completely depend on accurate protein concentrations. Protein degradation over time, and varying presence of non-monomeric protein would almost certainly contribute significantly to the kind of variation observed. We implemented changes in analysis to account for these patterns, which should increase confidence that our revised data contains more useful information than the original analysis. Nevertheless, this underlying experiment displayed serious flaws limiting the accuracy of the affinity measurements. Our revised analysis should be seen as the most rigorous interpretation of the originally measured data, with caveats as to absolute accuracy limited by the experimental technique.

2.6.1 Implications for Other Work and Analysis

Although our conclusions were drawn only on this set of raw data, our findings are likely to apply more broadly. Our work has significant implications for analysis of other previously published data sets because both the general experimental design and specific analysis methods in this original publication are common to other published high-throughput SH2 data. Furthermore, published binding models derived from this and other published SH2 data will have inaccuracies if they depend on qualitative or quantitative versions of published SH2 data. Finally, we propose some recommendations to improve quality of measurements and analysis in future work.

Without explicit controls to avoid concentration inaccuracies, it would be reasonable to assume that other quantitative SH2 protein interaction data sets also suffer from these types of variations. Furthermore, without raw data to examine for these patterns, the use of previously

published data sets should be carefully considered. Concentration inaccuracies, combined with the tendency to report a mean value easily influenced by outliers, could very well explain the widely varying results with SH2 affinities reported between different research groups.

The analysis methods used in the original publication of this data were also used in several other previously published sets of high-throughput SH2 interaction data (10, 13, 14, 19). While the experiments in these other publications used protein microarrays, the improvements in model selection, model fitting, and noise evaluation from our analysis would equally apply to other data. We have demonstrated that improved quality metrics and model selection methods can improve the final quantitative results, and these methods are likely applicable to previously published data.

Although researchers can be careful to draw *future* conclusions from previously published SH2 data, many publications have already used this data to draw conclusions and to build models of SH2 interactions. We have demonstrated that models built from degenerate library data do not match previously published quantitative data, or our revised analysis of the Jones group FP data. Scansite (28), SMALI (6), NetPhorest (34), KinSpect (35), and DomPep (31) all rely on degenerate libraries to make their predictions of SH2 domain interactions. Despite the problems with availability and data scrambling from the Cesareni Group data (18), multiple predictors have used their results as inputs, including MSM/D (36), and NetSH (18). We demonstrated that the early (2006-2009) MacBeath group data (10–13) does not correlate well with later work (2013) from the MacBeath group (14). Several predictive models of binding were built using the earlier data (21, 37). One predictor from the Jones Group, PEBL (17), was built from the published Jones FP data (16, 17) alone. Each of these models may be affected by the quality of the data upon which it was built.

2.6.2 Suggestions for Future Measurements of Affinity

Accurately measuring interaction affinity between SH2 domains and phosphopeptides is a difficult undertaking, especially when attempting to do so in a high-throughput manner. All of the issues encountered in low-throughput measurements are present and further complicated by making a large quantity of measurements in a reasonable time at a reasonable cost. Thus, controls for protein function and peptide activity that are sufficient in a qualitative assay are unlikely to be sufficient in a quantitative assay. Measurements of interaction based on equilibrium affinity, as done in this analysis, depend heavily on accurate protein concentration. Since affinity is a function of the active protein concentration, the difference between assumed protein concentration and actual functional protein concentration must be minimized. Determination of functional protein is compounded in an exploratory experiment such as this, because interactors for many of the proteins tested are not known ahead of time.

In order to minimize concentration inaccuracies, one must be assured that the protein being measured is maximally active, and the concentration is accurately known. First, protein must be highly purified. Non-monomeric SH2 domain function may not be exactly the same as monomeric domain or may be completely non-functional. It should be eliminated with high-performance liquid chromatography (HPLC) and size-exclusion purification methods. Second, protein activity should be benchmarked to determine factors that affect degradation because only measurements of fully active protein will provide accurate affinity. Assuming one or more peptides have been identified as potential binders, protein expression, purification, and storage methods need to be varied to determine their effect on activity. Protein degradation can happen over hours and even minutes at room temperature, so understanding time-dependent modulations of activity would be critical. If no known interactors exist, then – before high-throughput

experiments can begin – exploratory experiments must be conducted to identify potential binding partners and their maximal activity. Finally, guaranteeing linear and stable peptide activity is also important. Although we did not identify any problems with peptide activity in this data, it is possible that it was due to insufficient data. Benchmarking peptide fluorescence magnitude and stability over time, and any tendency to form multimers would identify windows of accurate response from the reagents.

Alternatively, methods of determining affinity which do not depend so heavily on accurate protein concentration could be explored. A concentration-independent method of measuring interaction affinity might provide an attractive alternative to the cumbersome procedures required to ensure precise protein concentrations.

One such method was recently developed by the Stormo lab (38). In that method, a 2-color competitive fluorescence anisotropy assay measures the relative affinity of two interactions in solution. Although the experiment as published measured protein interaction with nucleotide oligomers, it could also be used to measure protein interaction with two peptides labeled with different color fluorophores. 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. Considering the problems with determining protein concertation accurately, and the significant impact they have on accuracy, a method such as this employing competition

and relative affinity combined with a careful measurement of absolute affinity seems to be a promising direction.

Chapter 3: Different Epidermal Growth Factor Receptor (EGFR) Agonists Produce Unique Signatures for the Recruitment of Downstream Signaling Proteins.

This research was originally published in the Journal of Biological Chemistry. Ronan,T., Macdonald-Obermann,J.L., Huelsmann,L., Bessman,N.J., Naegle,K.M. and Pike,L.J. (2016) Different Epidermal Growth Factor Receptor (EGFR) agonists produce unique signatures for the recruitment of downstream signaling proteins. *J. Biol. Chem.*, **291**, 5528–5540. © The American Society for Biochemistry and Molecular Biology.

3.1 Abstract

The EGF receptor can bind seven different agonist ligands. Although each agonist appears to stimulate the same suite of downstream signaling proteins, different agonists are capable of inducing distinct responses in the same cell. To determine the basis for these differences, we used luciferase fragment complementation imaging to monitor the recruitment of Cbl, CrkL, Gab1, Grb2, PI3K, p52 Shc, p66 Shc, and Shp2 to the EGF receptor when stimulated by the seven EGF receptor ligands. Recruitment of all eight proteins was rapid, dose-dependent, and inhibited by erlotinib and lapatinib, although to differing extents. Comparison of the time course of recruitment of the eight proteins in response to a fixed concentration of each growth factor revealed differences among the growth factors that could contribute to their differing biological effects. Principal component analysis of the resulting data set confirmed that the recruitment of these proteins differed between agonists and also between different doses of the same agonist. Ensemble clustering of the overall response to the different growth factors suggests that these EGF receptor ligands fall into two major groups as follows: (i) EGF, amphiregulin, and EPR; and (ii) betacellulin, TGFβ, and epigen. Heparin-binding EGF is

distantly related to both clusters. Our data identify differences in network utilization by different EGF receptor agonists and highlight the need to characterize network interactions under conditions other than high dose EGF. The EGF receptor is an intrinsic membrane protein composed of an extracellular ligand-binding domain connected to an intracellular tyrosine kinase domain by a single transmembrane α-helix. In the absence of ligand, the EGF receptor is thought to exist as a monomer, although inactive "pre-dimers" are known to form (39–43). Upon binding an agonist ligand, the EGF receptor dimerizes leading to the activation of its tyrosine kinase and the phosphorylation of tyrosine residues in the C-terminal tail of the receptor (44–46). The phosphorylated tyrosines on the EGF receptor serve as binding sites for a large number of signaling proteins that contain SH2 and/or phosphotyrosine-binding domains (2, 47). Some of these proteins, such as Cbl, possess an enzymatic activity (48). Others, such as Grb2 or Shc, serve as adapter proteins that bring other proteins into the EGF receptor-containing complex. For example, Grb2 recruits the scaffolding protein, Gab1, to the EGF receptor (49). Phosphorylation of Gab1 by the EGF receptor allows Gab1 to recruit additional proteins, such as Shp2 or PI3K-R1, to the signaling complex (50–53). The recruitment of these signaling proteins to the receptor ultimately triggers the activation of a variety of downstream signaling pathways, thereby mediating the intracellular effects of growth factor binding.

3.2 Introduction

The EGF receptor binds seven different agonist ligands, including some of high affinity (EGF, TGFβ, BTC,[6] and HBEGF) and some of low affinity (AREG, EPG, and EPR) (53). It has been reported that different EGF receptor ligands induce different responses when binding to the same cell line (54–57). Given that these agonists bind to the same receptor and stimulate similar downstream signaling molecules, it is difficult to explain how these divergent responses are achieved. We have previously used a luciferase fragment complementation system to assess the ability of EGF to induce dimerization of the EGF receptor (58–60). In this study, we use our luciferase fragment complementation assay to visualize the recruitment of a variety of signaling proteins to the EGF receptor. The fine temporal resolution and quantitative nature of the split luciferase complementation system allowed us to continuously monitor the association of Cbl, CrkL, Gab1, Grb2, PI3K-R1, p52 Shc, p66 Shc, and Shp2 with the EGF receptor in response to increasing concentrations of all seven different EGF receptor ligands. Principal component analysis was applied to this large dataset to determine how the response to these growth factors differed. The data demonstrate that each growth factor produces a unique signature for the recruitment of signaling proteins, and this signature differs at different doses of the same growth factor. This suggests that each growth factor utilizes the signaling network differently, preferentially promoting flux through some pathways over others, which could readily lead to a different net biological outcome.

3.3 Experimental Procedures

3.3.1 Materials

EGF was purchased from Biomedical Technologies. TGFβ and amphiregulin were from Leinco. Betacellulin was from ProSpec. Heparin-binding EGF was from Sigma. Epigen and epiregulin were synthesized and purified in the laboratory of Dr. Mark Lemmon (University of Pennsylvania). Fetal- Plex was from Gemini Bioproducts. The anti-EGF receptor antibody was from Cell Signaling. The PY20 anti-phosphotyrosine antibody was from BD Transduction Laboratories.

3.3.2 DNA Constructs

Full-length cDNA constructs for the signaling proteins were obtained from Addgene (CrkL PI3K-R1 and Shp2), Source Bioscience (Gab1), Thermo Fisher (p52 Shc, p66 Shc, and Grb2), or Sino Biologicals (c-Cbl). The stop codon in each was removed, and an in-frame BsiWI site was inserted through site-directed mutagenesis. The cDNAs were cut with BsiWI and fused to the C-terminal fragment of luciferase (CLuc). The construct was moved into the pcDNA3.1 Zeo expression vector where expression of the fusion protein is driven off the constitutive CMV promoter.

3.3.3 Cell Lines

CHO cells stably expressing the tetracycline-inducible EGF receptor C-terminally fused to the N-terminal fragment of firefly luciferase (EGFR-NLuc) (60) were used as the starting parental line. These cells were transfected with the pcDNA3.1 Zeo plasmids encoding the CLuc fusion of each of the eight signaling proteins. Eight (double) stable cell lines were selected by

growth in 5 mg/ml Zeocin. Quantitation of EGF receptor expression in each line by ^{125}I-EGF saturation binding indicated that the number of cell surface EGF receptors expressed in each line is within ±20% of the average level of receptor expression (data not shown). Cells were grown in Dulbecco's modified Eagle's medium supplemented with 10% FetalPlex, 100 µg/ml G418, 100 µg/ml hygromycin, and 100 µg/ml Zeocin and maintained in an incubator at 37 °C in 5% CO_2.

3.3.4 Luciferase Assays

Double stable CHO cells were plated into 96-well black-walled dishes 2 days prior to use in medium containing 1.5 µg/ml doxycycline to induce expression of the EGFR-NLuc fusion protein. For assay, cells were transferred into Dulbecco's phosphate-buffered saline supplemented with 5 mg/ml BSA and 20 mM MOPS, pH 7.2. Cells were incubated with 0.9 mg/ml D-luciferin for 30 min at 37 °C prior to the addition of growth factor and the start of imaging. Cell radiance (photons/s/cm^2/steradian) was measured every 30 s for 25 min using a cooled charge-coupled device camera in the IVIS50 or IVIS Lumina imaging system. Assays were performed in hextuplicate. The lines through the data were drawn using Equation 1, which represents the sum of a logistics association equation and an exponential dissociation equation.

$$Y = \frac{Y_0}{1+he^{-k_1 t}} + (plateau - bottom)e^{-k_2 t} + bottom \quad \text{(Eq. 1)}$$

where $Y = \frac{photons}{s}$ at time t, k_1 represents the association rate constant, and k_2 is the dissociation rate constant. This curve drawing was not part of the principal component analysis and was used only for visual presentation of the dose-response curves.

3.3.5 Western Blotting

CHO cells expressing the wild type EGF receptor were treated without or with 5 μM erlotinib or 10 μM lapatinib for 1 h and then stimulated with the indicated concentrations of EGF for 5 min. Lysates were prepared, and Western blotting with anti-EGF receptor and anti-phosphotyrosine antibodies were performed as described previously (59).

3.3.6 PCA and Enrichment Analysis

Computational analysis was performed using the Python programming language. PCA utilized the scikit-learn package (61). PCA was performed on a 280 x 44 matrix, with 280 unique combinations of protein, growth factor, and dose, each with 44 time points, normalized to the maximal response elicited for that agonist/protein pair. For PCA, a subset of five (out of seven) doses was chosen for each growth factor to bracket the EC_{50} value for the recruited signaling proteins as follows: for BTC and EGF, the doses ranged from 0.03 to 3 nM; for TGF, the doses ranged from 0.1 to 10 nM; for HB-EGF, the doses ranged from 0.3 to 30 nM; and for AREG, EPG, and EPR, the doses ranged from 3 to 300 nM. References to “low” doses of growth factor (as used in Figure 3.10 and Figure 3.14) represent the second dose in the five-dose series, and references to “high” doses (as used for Figure 3.14) represent the fourth dose in the five-dose series. Distances between protein pairs were calculated using Euclidean distance between the five-dimensional vector across doses in PC space. Top- and bottom-quartile enrichment was calculated using the hypergeometric test and Bonferroni-corrected.

For clustering of the growth factors based on protein recruitment across all doses, pairwise protein distances for each ligand were converted to a one-dimensional vector. The vectors for each ligand were then clustered using hierarchical clustering. An ensemble of 35

clustering results was generated by varying linkages (single, complete, average, and weighted) and distance metrics (Euclidean, Pearson correlation, city block, cosine, Bray-Curtis, Canberra, Chebyshev, and square Euclidean). The Euclidean metric was also used with median, centroid, and Ward linkage. The results for each ligand were assembled into a matrix and hierarchically clustered using single linkage and Euclidean distance. p66 Shc was not included in this analysis so as not to over-weight the results toward the contribution of Shc isoforms.

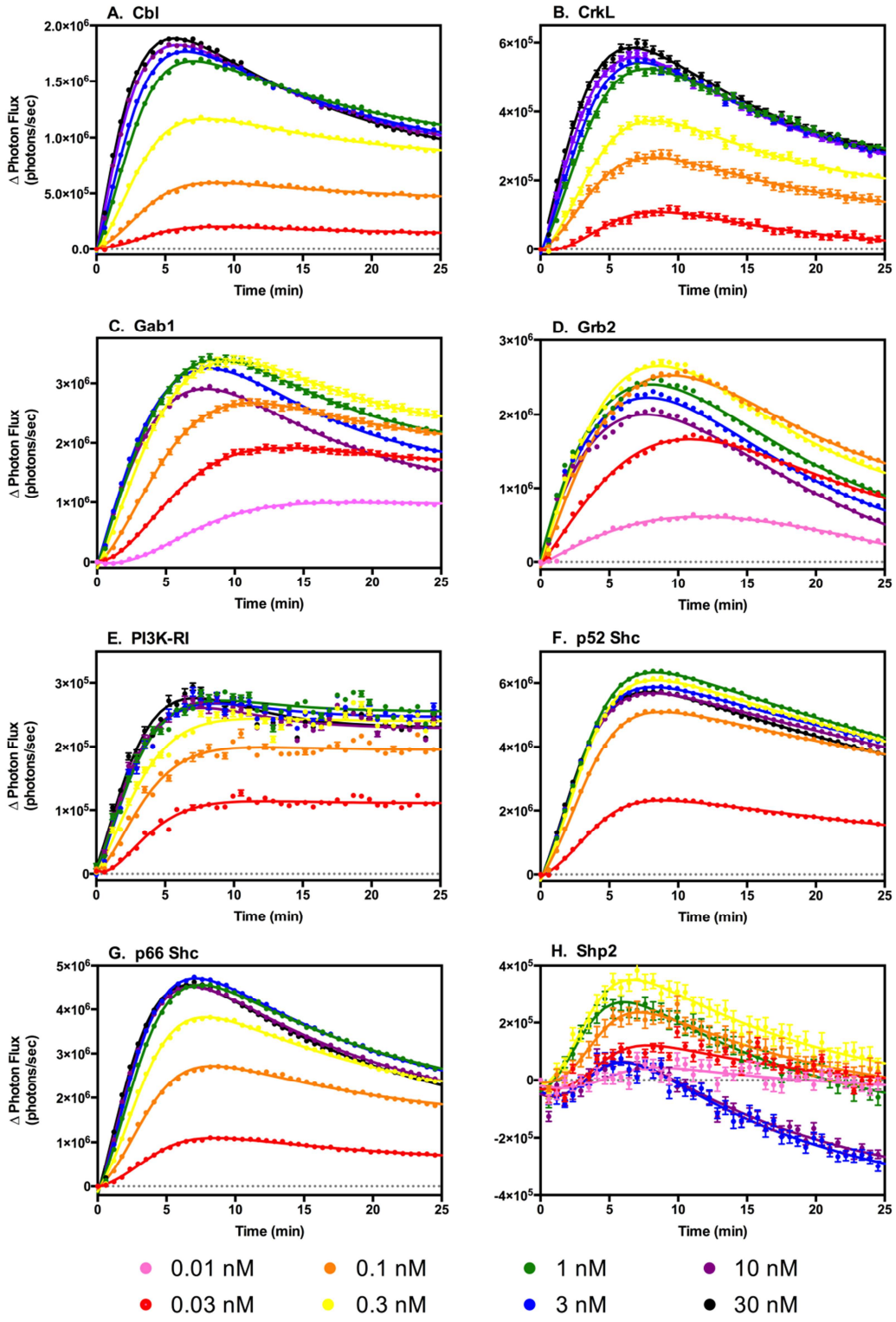

Figure 3.1: EGF-stimulated association of eight signaling proteins with the EGF receptor measured using luciferase fragment complementation imaging.
CHO cells stably co-expressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were assayed for EGF-stimulated light production in the presence of luciferin. Cells were stimulated with the indicated concentration of EGF at time t=0 and light production monitored for 25 min.

3.4 Results

3.4.1 Generation and Characterization of Stable Cell Lines

The split luciferase complementation assay utilizes an N-terminal (NLuc) and C-terminal (CLuc) fragment of firefly luciferase (62). Individually, the fragments have no enzymatic activity. However, when they are brought into proximity, they complement each other forming a catalytically active luciferase that produces light upon oxidation of luciferin. For our luciferase complementation assays, each of eight signaling proteins (Cbl, CrkL, Gab1, Grb2, PI3K-R1, p52 Shc, p66 Shc, and Shp2) was C-terminally fused to the CLuc fragment via a 16-amino acid flexible linker. The cDNA for the fusion protein was then transfected into aCHOcell line that stably expressed the EGF receptor C-terminally fused to the NLuc fragment (EGFR-NLuc) off a tetracycline-inducible promoter. Double stable cell lines were selected for use in these experiments. For assay, the CHO cells were cultured for 24 h in 1.5 μg/ml doxycycline to induce expression of the EGFR-NLuc fusion protein. The signaling proteins were constitutively expressed from a CMV promoter.

3.4.2 Luciferase Complementation between the EGF Receptor and Signaling Proteins

All eight signaling proteins yielded an EGFstimulated increase in luciferase activity when co-expressed in cells with EGFR-NLuc (Figure 3.1). EGF-stimulated complementation between the EGF receptor and these signaling proteins was seen as early as 30 s after the addition of EGF. At low concentrations of EGF, essentially all of the pairings exhibited a rapid rise in luciferase activity, which peaked by ~5–8 min. For some pairings, such as the EGF receptor and PI3K-R1 (Figure 3.1*E*), this level of complementation was maintained over the entire time course at all doses. In other pairings, such as Cbl (Figure 3.1*A*) or CrkL (Figure 3.1*B*), complementation

plateaued at low concentrations of growth factor but declined after an early peak at high concentrations of EGF. Still other proteins demonstrated a bimodal response across doses. For example, for Grb2 (Figure 3.1*D*) and Shp2 (Figure 3.1Figure 3.1*H*), the maximum complementation occurred at a relatively low dose of EGF, with higher doses of growth factor resulting in lower peak responses and a marked decrease at longer times.

In the EGF receptor/Shp2 pairing (Figure 3.1*H*), the luciferase activity observed at the highest concentrations of EGF actually fell below the basal level after about 15 min. These data imply that Shp2 associates with the EGF receptor under non-stimulated conditions. This association is apparently disrupted upon stimulation with high doses of growth factor.

If these signaling proteins are being recruited to the EGF receptor via phosphotyrosine-dependent interactions, then the associations visualized through luciferase complementation should be sensitive to inhibition of the EGF receptor tyrosine kinase. As shown in Figure 3.2, *A–H*, treatment of cells with 5 μM erlotinib (*green lines*) effectively inhibited EGF-stimulated complementation between the EGF receptor and each of these eight signaling proteins. Inhibition was essentially complete for all pairings with the exception of p52 Shc and p66 Shc, for which the inhibition was ~70%. The complementation between the EGF receptor and Shp2 actually showed an EGFstimulated decline in luciferase activity, again consistent with there being a basal level of association between the two proteins, which is disrupted after ligand binding.

Despite the fact that lapatinib appeared to inhibit EGF receptor autophosphorylation to the same extent as erlotinib (Figure 3.2*I*), pretreatment of the cells with 10 μM lapatinib (*red lines*) was far less effective than pretreatment with erlotinib at blocking the association of these signaling proteins with the EGF receptor. Although lapatinib was able to completely block

complementation between the EGF receptor and Cbl, CrkL, and Shp2, the other five signaling proteins all showed at least 20% residual EGF-stimulated luciferase activity in the presence of lapatinib. The association of Gab1 was particularly insensitive to lapatinib treatment. Thus, there is a significant difference between erlotinib and lapatinib in terms of their efficacy for inhibiting EGF-stimulated signaling complex assembly.

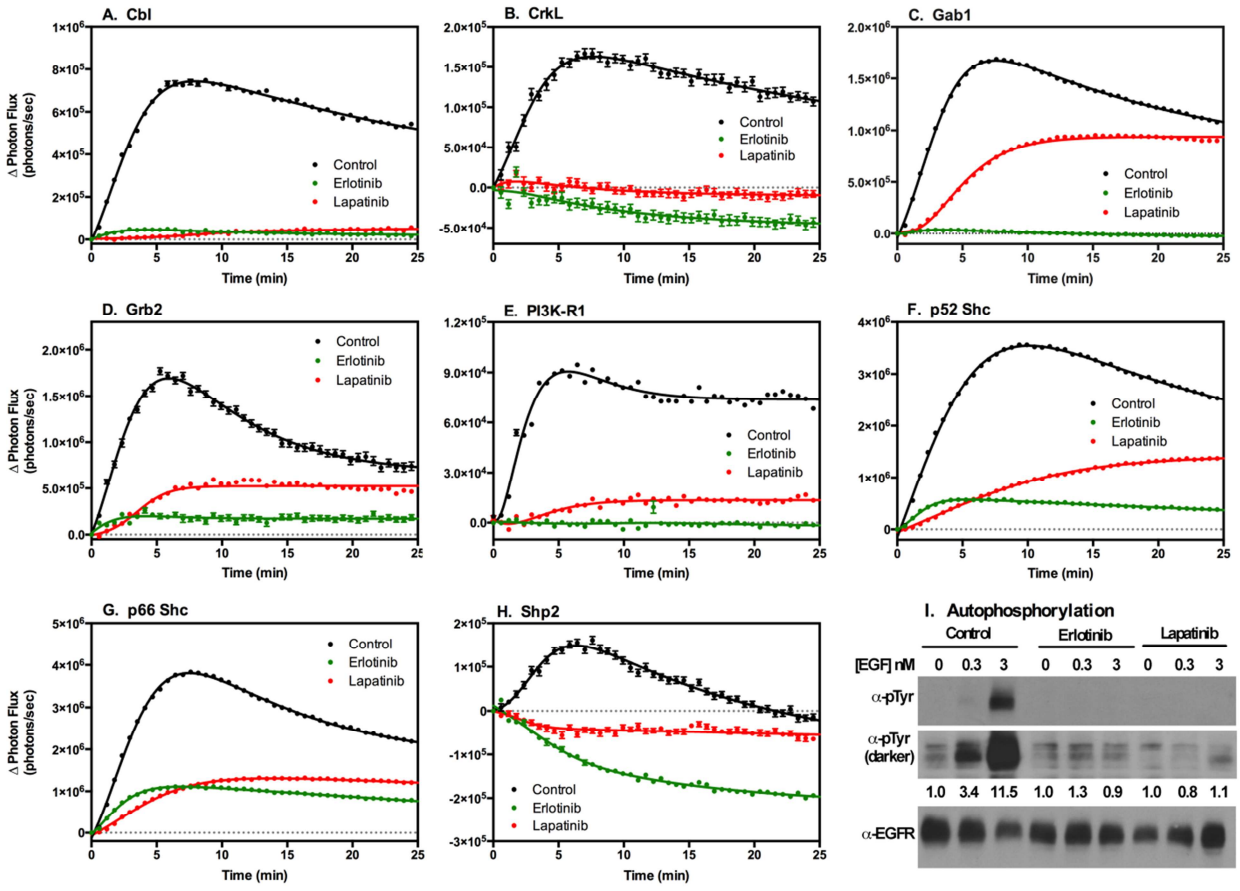

Figure 3.2: Effect of erlotinib and lapatinib on EGF-stimulated association of eight signaling proteins with the EGF receptor.
Panels A to H) CHO cells stably co-expressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were treated with 5 µM erlotinib (green lines) or 10 µM lapatinib (red lines) for 60 min prior to stimulation without or with 0.3 nM. EGF-stimulated light production was monitored for 25 min after addition of EGF. Panel I) CHO cells expressing wild type EGF receptor were treated with 5 µM erlotinib or 10 µM lapatinib for 60 min prior to stimulation with 0.3 or 3.0 nM EGF. Lysates were prepared and equal amounts of protein analyzed by SDS gel electrophoresis and Western blotting with an anti-phosphotyrosine antibody or an anti-EGF receptor antibody. Quantitation of anti-phosphotyrosine blot is shown.

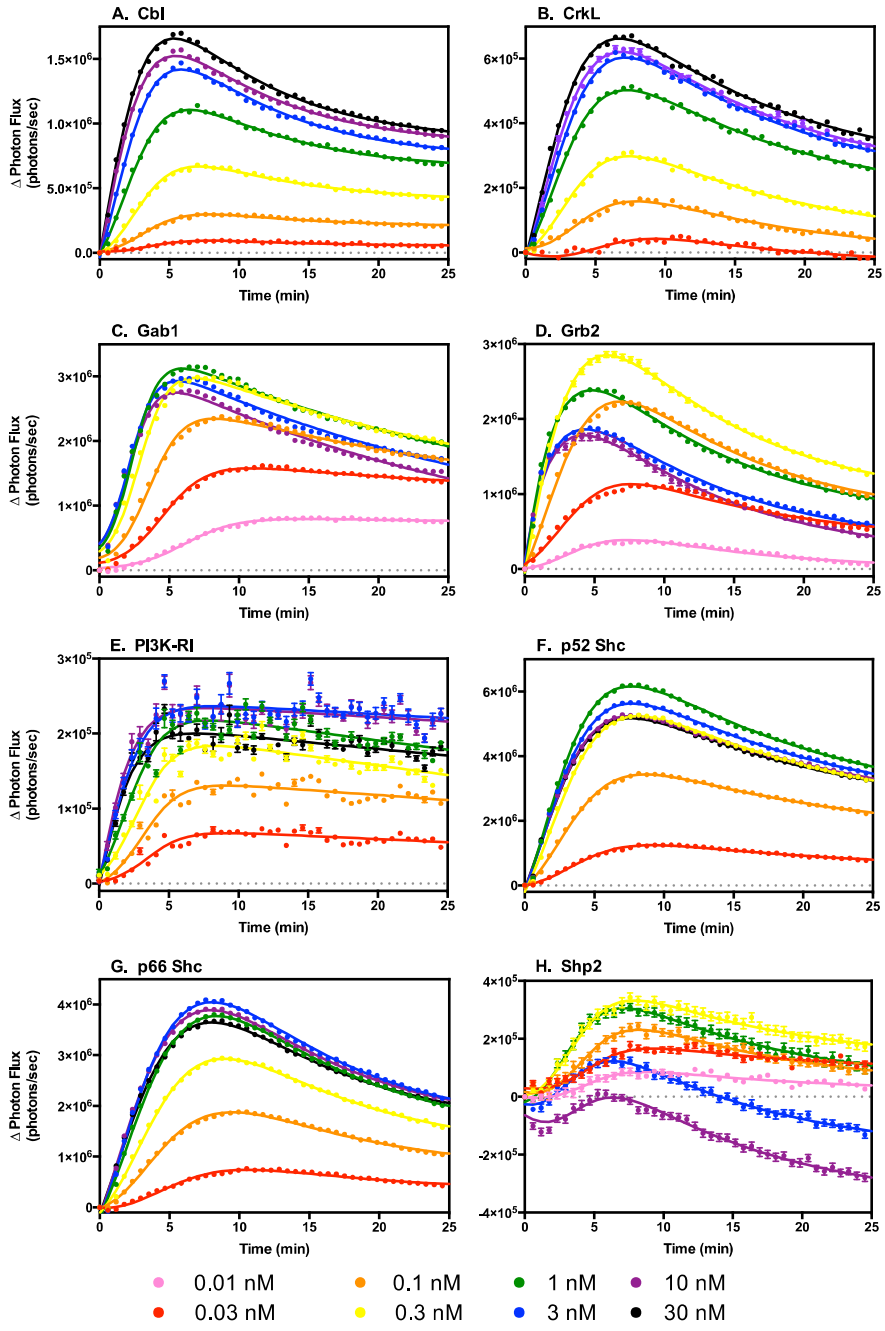

Figure 3.3: TGF-stimulated association of eight signaling proteins with the EGF receptor measured using luciferase fragment complementation imaging.
CHO cells stably coexpressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were assayed for TGF-stimulated light production in the presence of luciferin. Cells were stimulated with the indicated concentration of TGF at time t=0 and light production monitored for 25 min.

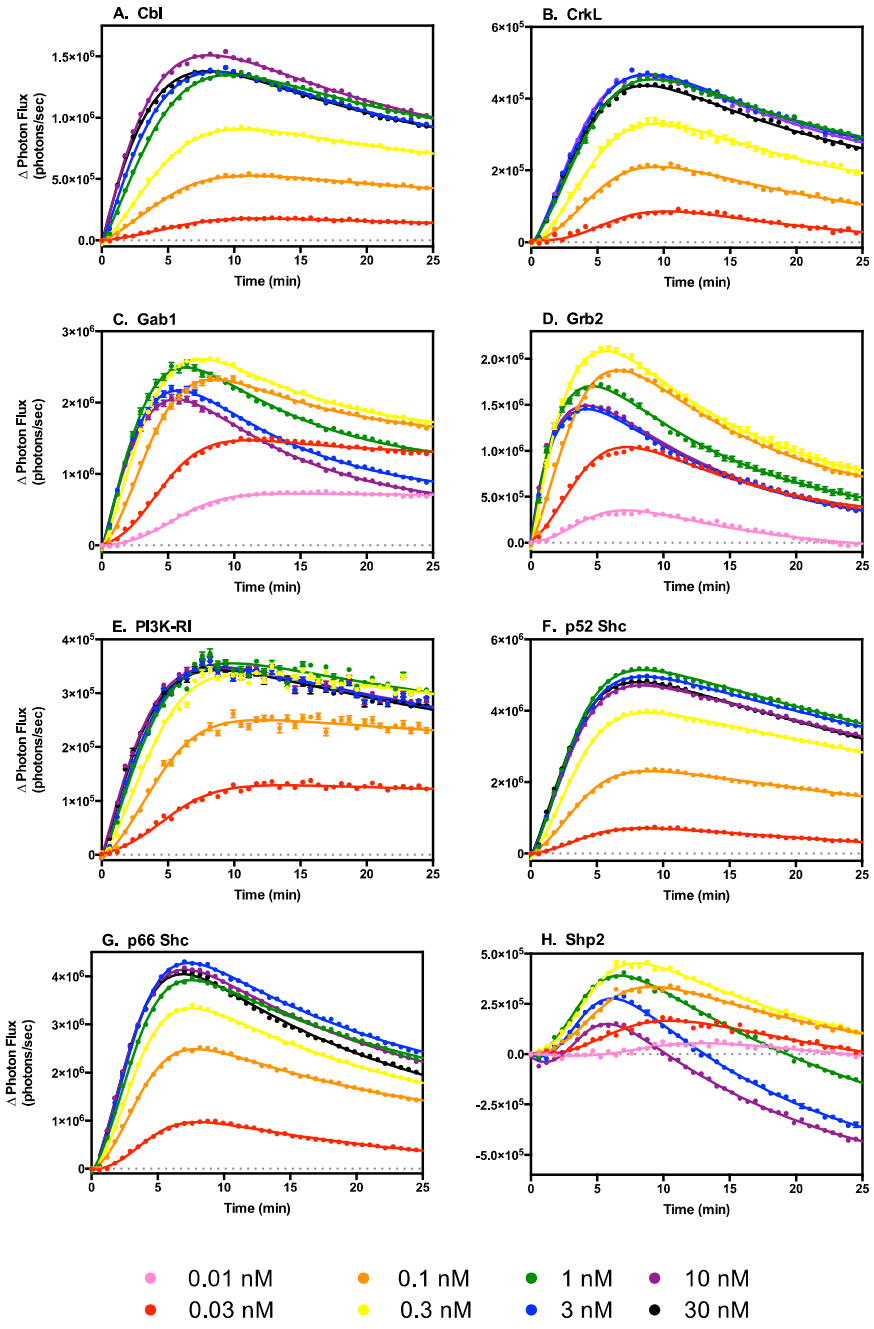

Figure 3.4: BTC-stimulated association of eight signaling proteins with the EGF receptor measured using luciferase fragment complementation imaging.
CHO cells stably coexpressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were assayed for BTC-stimulated light production in the presence of luciferin. Cells were stimulated with the indicated concentration of BTC at time t=0 and light production monitored for 25 min.

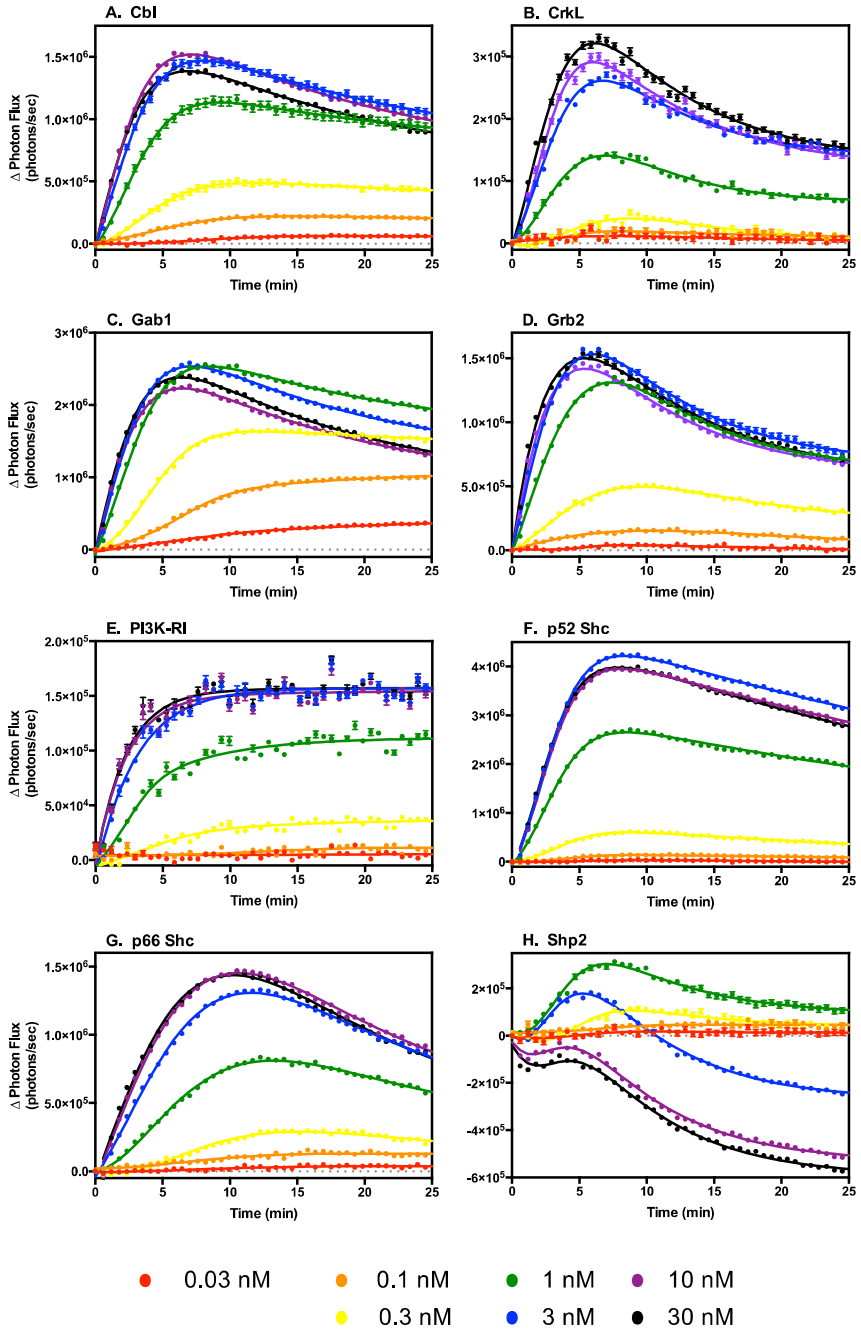

Figure 3.5: HB-EGF-stimulated association of eight signaling proteins with the EGF receptor measured using luciferase fragment complementation imaging.
CHO cells stably co-expressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were assayed for HB-EGF-stimulated light production in the presence of luciferin. Cells were stimulated with the indicated concentration of HB-EGF at time t=0 and light production monitored for 25 min.

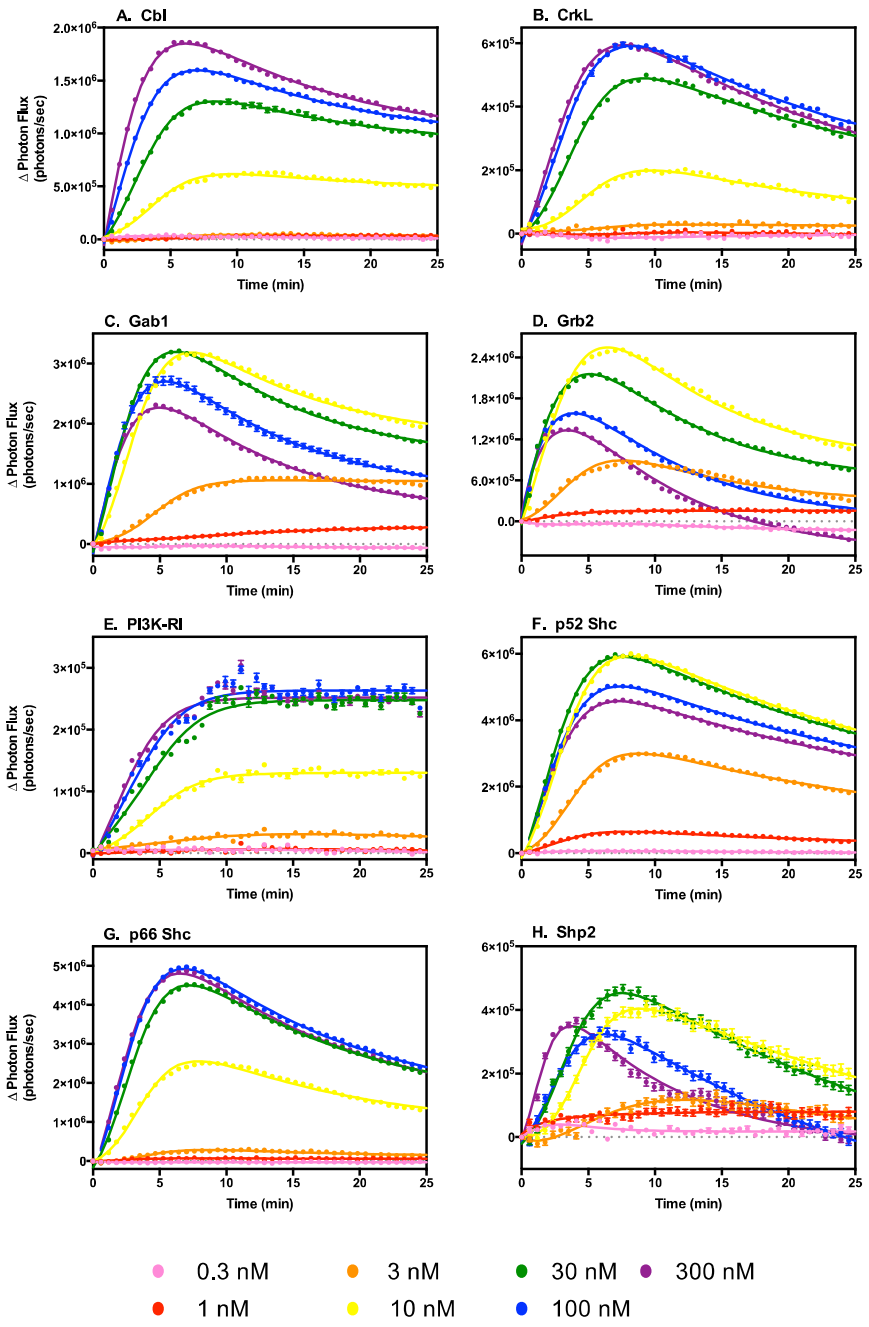

Figure 3.6: AREG-stimulated association of eight signaling proteins with the EGF receptor measured using luciferase fragment complementation imaging.
CHO cells stably coexpressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were assayed for AREG-stimulated light production in the presence of luciferin. Cells were stimulated with the indicated concentration of AREG at time t=0 and light production monitored for 25 min.

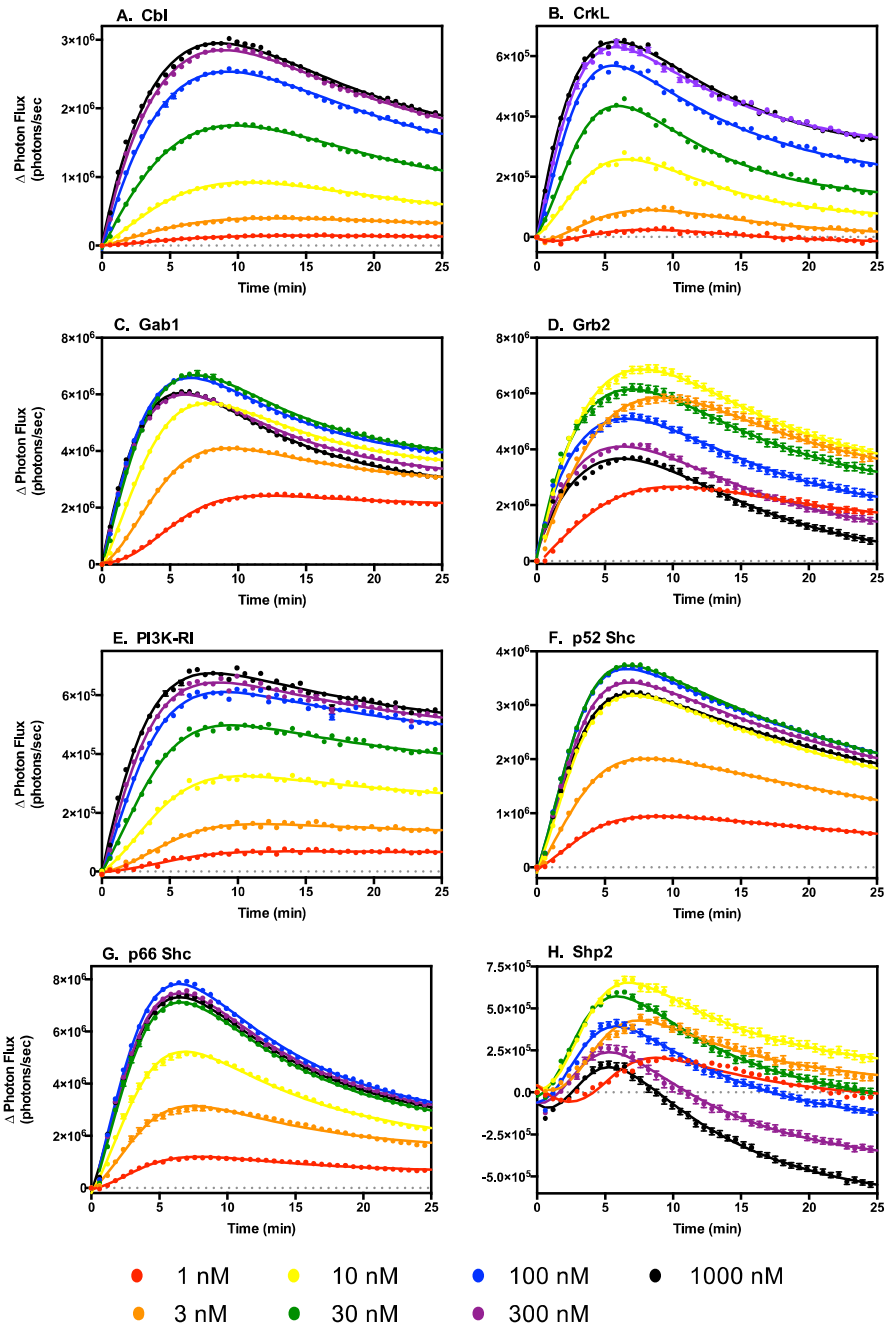

Figure 3.7: EPG-stimulated association of eight signaling proteins with the EGF receptor measured using luciferase fragment complementation imaging.
CHO cells stably coexpressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were assayed for EPG-stimulated light production in the presence of luciferin. Cells were stimulated with the indicated concentration of EPG at time t=0 and light production monitored for 25 min.

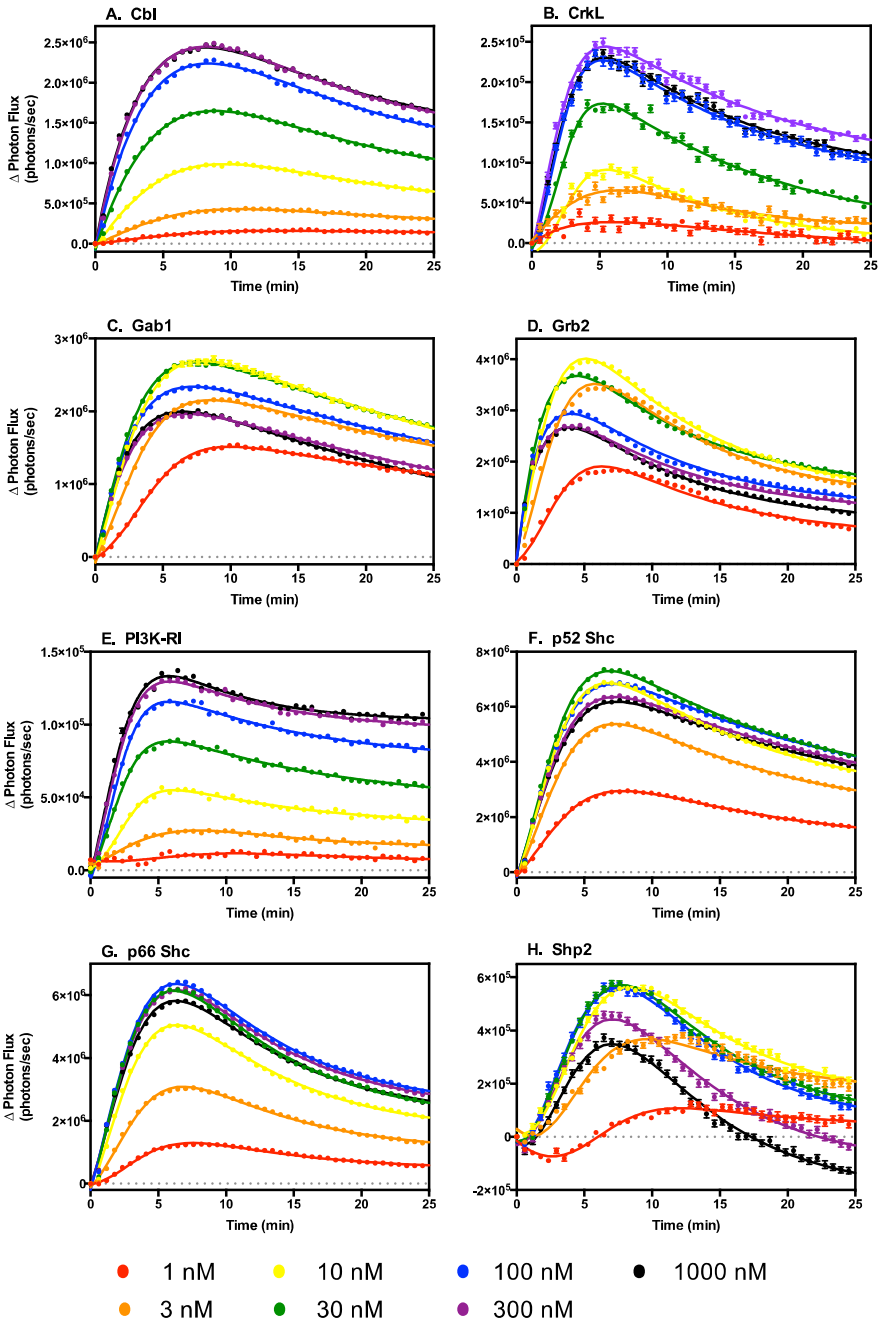

Figure 3.8: EPR-stimulated association of eight signaling proteins with the EGF receptor measured using luciferase fragment complementation imaging.
CHO cells stably coexpressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were assayed for EPR-stimulated light production in the presence of luciferin. Cells were stimulated with the indicated concentration of EPR at time t=0 and light production monitored for 25 min.

3.4.3 Recruitment Stimulated by Other EGF Receptor Ligands

The EGF receptor is activated by a family of homologous growth factors, including EGF, TGFβ, BTC, HB-EGF, AREG, EPG, and EPR (63). To quantify the similarities and differences in the response of cells to stimulation by each of these ligands, the luciferase complementation assay was used to monitor the recruitment of the eight different signaling proteins to the EGF receptor in response to each of these agonist ligands.

Figures 3.3 through 3.8 show the time courses of the recruitment of these eight signaling proteins to the EGF receptor in response to increasing doses of each of these additional six growth factors. Like EGF, all of these growth factors stimulated the recruitment of all eight signaling proteins in a dose-dependent manner. However, the patterns of the dose response curves for all seven growth factors for each individual pairing were similar. For example, for all growth factors, PI3K-R1 recruitment plateaued early, and the level of signal was maintained over the entire time course. Similarly, the bimodal response for the recruitment of Grb2 and Shp2 was observed for all growth factors.

Table 3.1 reports the estimated EC_{50} values for each ligand stimulating the recruitment of each protein. As expected from their low binding affinities, AREG, EPG, and EPR required ~30–100-fold greater concentrations of ligand to stimulate a maximal response than did EGF, TGFβ, BTC, or HB-EGF. Surprisingly, the EC_{50} values for a given growth factor for stimulating the recruitment of the different signaling proteins differed up to 18-fold.

EC_{50} (nM)	EGF	TGF	BTC	HB-EGF	AREG	EPG	EPR
Cbl	0.31	0.73	0.40	0.85	36.0	33.0	21.0
CrkL	0.14	0.40	0.25	1.5	26.0	19.0	21.0
Gab1	0.08	0.11	0.06	0.47	5.9	4.8	2.7
Grb2	0.03	0.06	0.04	0.64	4.0	1.8	2.7
PI3K-RI	0.03	0.24	0.11	1.2	21.0	22.0	18.0
p52 Shc	0.06	0.12	0.15	0.85	3.7	3.7	2.6
p66 Shc	0.10	0.21	0.05	1.7	13.0	6.1	4.0
Shp2	0.09	0.11	0.09	0.50	13.0	3.3	4.0

Table 3.1: EC_{50}'s for Agonist-Stimulated EGF Receptor/Protein Association
Table 1 compares the EC_{50}'s for each growth factor for stimulating the recruitment of the eight signaling proteins. These values were estimated based on the response to each dose of growth factor at t = 2.5 min. This largely eliminates the effects of the declines in signal at longer times and means that these values reflect mainly the initial association of the two proteins. The EC_{50}'s differed for the recruitment of different proteins by the same growth factor. So for example, EGF exhibited an EC_{50} of ~0.03 nM for recruiting Grb2 and PI3K-R1 but an EC_{50} about 10-fold higher for recruiting Cbl. EPG exhibited the widest range of EC_{50}'s (~18-fold difference) while HB-EGF showed the smallest range of EC_{50}'s (~3-fold).

Figure 3.9 compares the extent of recruitment of the eight signaling proteins in response to an optimal concentration of each of the seven growth factors. The concentrations compared were those that gave the maximal peak response for that particular pairing (Figure 3.1 and Figures 3.3 through 3.8). For most of the pairs, all seven ligands stimulated a similar maximal response. However, HB-EGF routinely elicited a slightly lower response than the other growth factors. The greatest difference in response was observed for the recruitment of Grb2 for which the response to EPG and EPR was ~30% higher than that to EGF, while the response to HB-EGF was ~30% lower than that to EGF. Consequently, there was nearly a 2-fold difference in the relative extent of Grb2 recruitment between the high of EPG/EPR and the low of HB-EGF.

Figure 3.10 compares the ability of a fixed (comparable) low dose of each growth factor to stimulate the recruitment of all eight signaling proteins. The responses have been normalized to the maximal response observed for that EGFR/protein pair at the optimal dose of that growth factor. For all growth factors, Grb2 appears to have the fastest relative response time. Cbl and

CrkL most frequently have the slowest relative response time. The recruitment of PI3K-R1 shows the most variability being similar to Cbl and CrkL for the low affinity ligands but closer to Grb2 and Gab1 for the high affinity ligands. Interestingly, the recruitment of p52 Shc and p66 Shc differs noticeably from each other. In many cases, p52 Shc shows a shorter relative response time than p66 Shc, often significantly shorter, as for AREG and HB-EGF. However, this order is reversed for BTC where p66 Shc is recruited more rapidly than p52 Shc. Thus, at the earliest times of signal transduction, differences in response to the different the growth factors can be identified and would contribute to a different biological outcome.

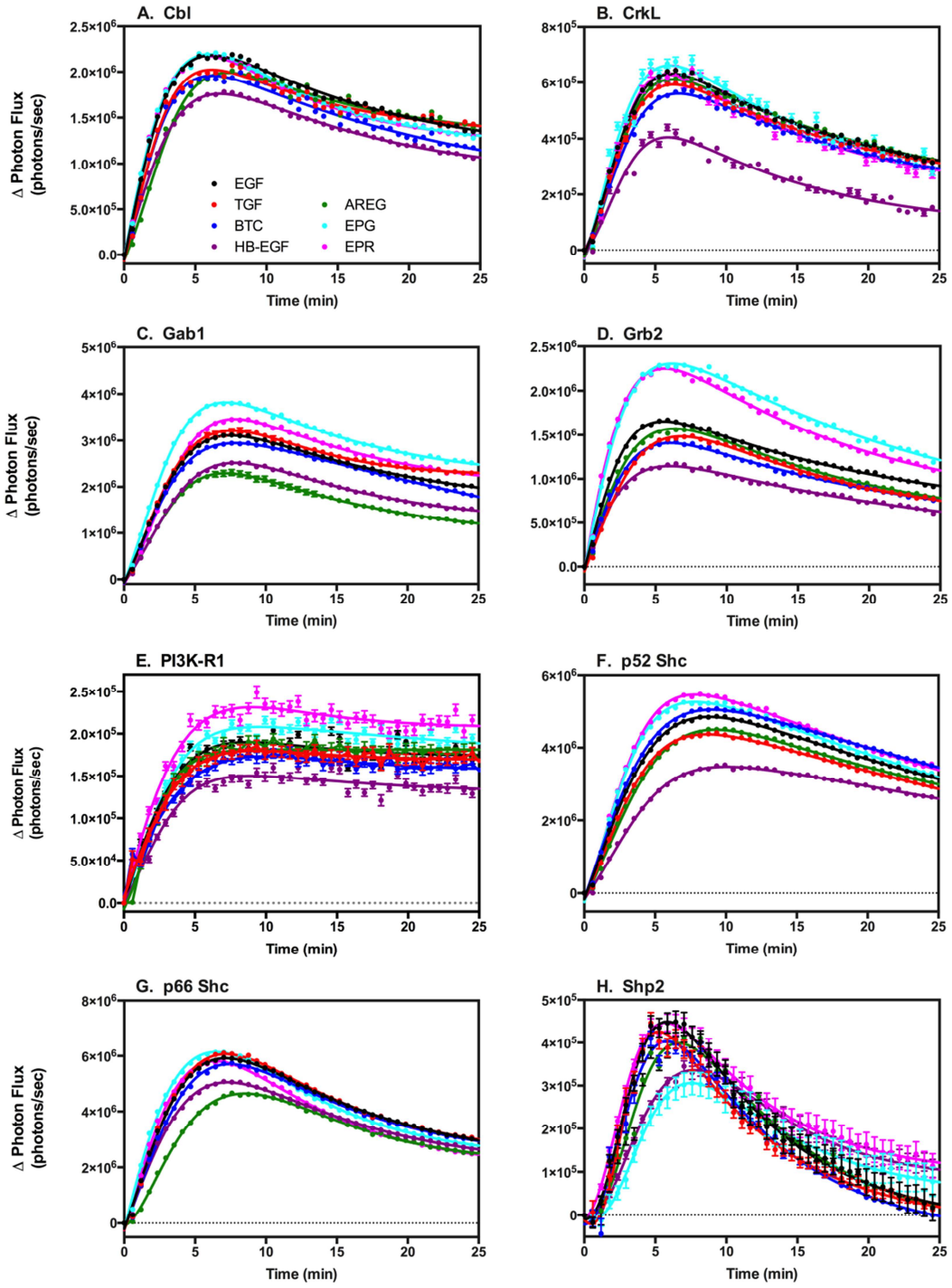

Figure 3.9: Comparison of the association of eight signaling proteins with the EGF receptor stimulated by optimal concentrations of the seven EGF receptor agonists.
CHO cells stably co-expressing EGFR-NLuc and the CLuc-fused version of one of eight signaling proteins were stimulated at time t=0 with the concentration of each growth factor that yielded maximal peak complementation for a given receptor/protein pair and light production monitored for 25 min.

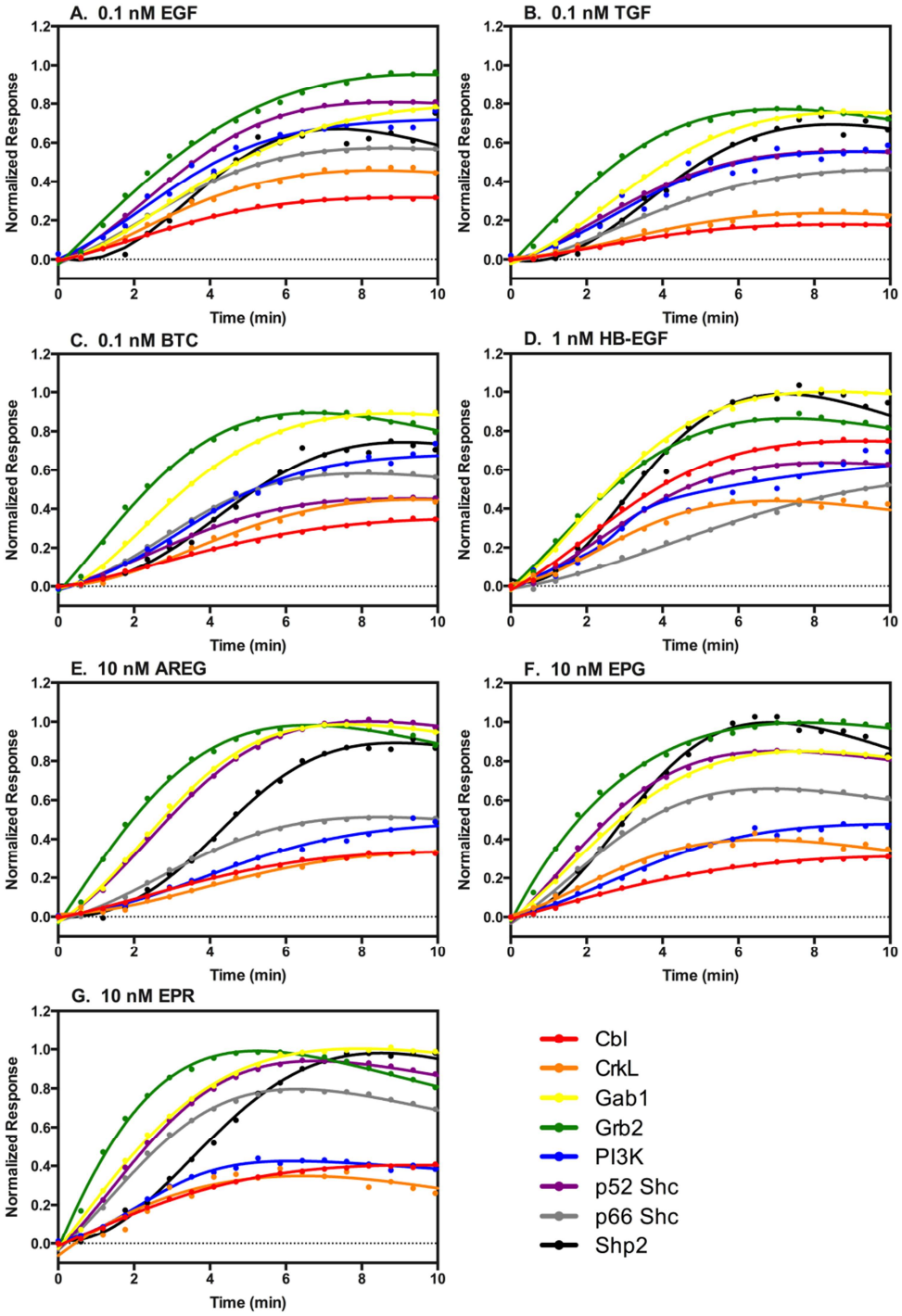

Figure 3.10: Relative response times for the recruitment of the eight signaling proteins by comparable low doses of each of the EGF receptor ligands.
The response to the indicated low concentration of each of the seven agonists was normalized to the maximal response elicited by that agonist for that EGF receptor/protein pair. The normalized responses for all signaling proteins stimulated by a single agonist were then plotted on the same graph.

3.4.4 Global Behaviors Observed via Reduced Dimensionality

The foregoing data represent an extremely rich set of measurements of the recruitment of eight different signaling proteins by the EGF receptor. Within this dataset, discovering relationships among the proteins and growth factors is difficult due to the high dimensionality of the problem. To reduce the dimensionality of the dataset, while keeping the relationships within the data intact, we used PCA.

For this analysis, a fixed subset of five (out of the seven) doses of each growth factor was used. The subset of doses was chosen so that we captured a comparable range of response above and below the EC50 values for each of the different growth factors. As a result, the doses that were not included in the analysis were either the very lowest concentrations that elicited a weak or no response or the very highest concentrations that were super-saturating. This approach allows us to compare the behavior of the same dose of a single growth factor across all eight signaling proteins and to compare the behavior of a single signaling protein at comparable concentrations across the different growth factors.

We could account for 97.0% of the co-variation within the entire dataset by projecting the original data into the first two dimensions of the principal component space. Principal component 1 (PC1) captures 87.6% of the variance, whereas PC2 captures 9.4% of the variance. As a result, each time series for one growth factor dose and signaling protein response can be plotted as a single point in two-dimensional PC space while retaining almost all of the variation that exists in the original 44 dimensions (*i.e.* the 44 time points per curve). The loading plots (Figure 3.11*A*) indicate that PC1 represents a positive integration of information across most time points. By contrast, PC2 negatively weights the earliest time points while positively weighting the latter half of the time course.

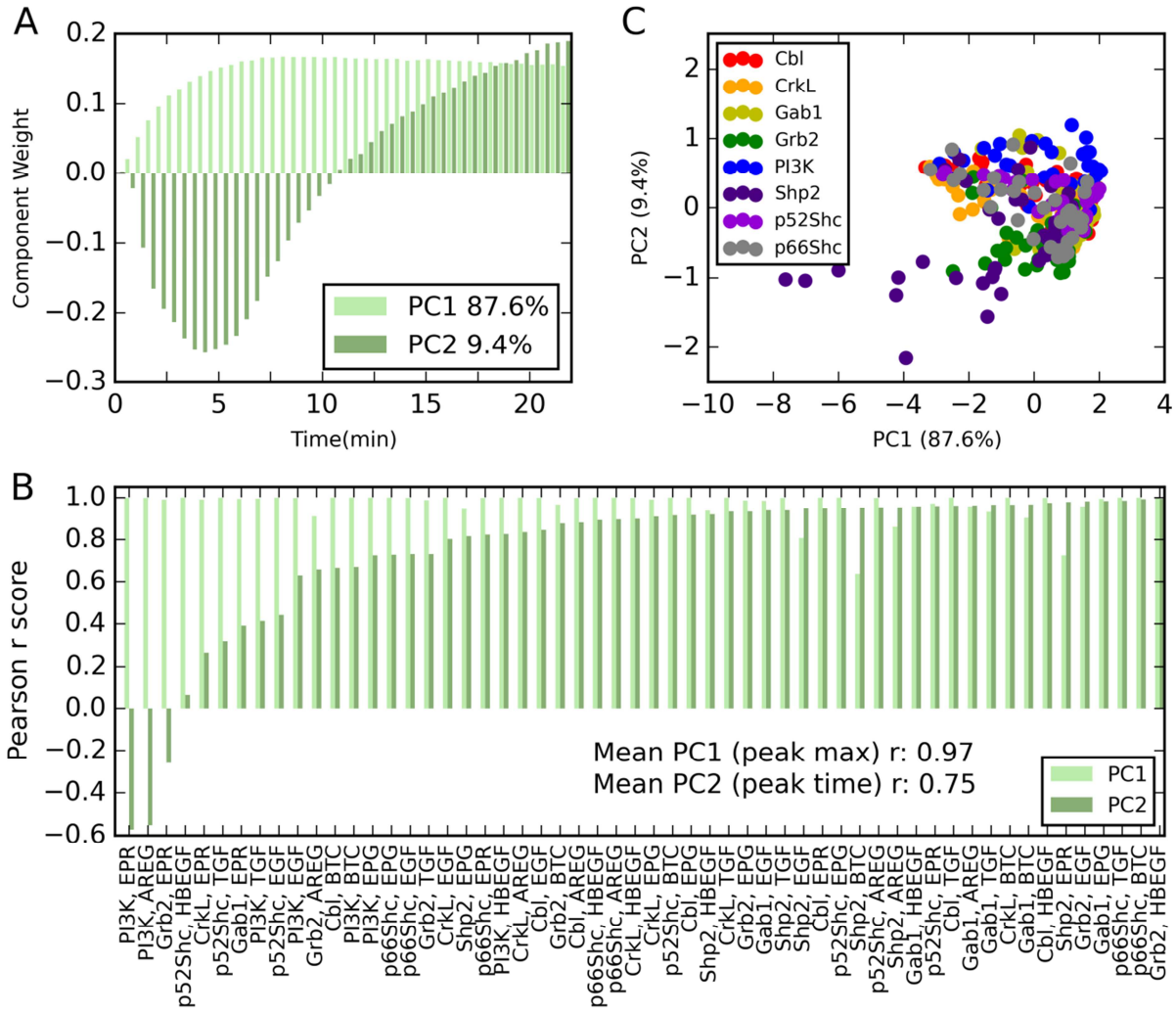

Figure 3.11: Dimensionality can be reduced using principal component analysis.
(A) The plot of the loadings of the first two principal components, which accounts for 97.0% of the covariance in the dataset (PC1 accounts for 87.6%, and PC2 accounts for 9.4%). (B) Correlation of PC1 with maximum peak magnitude and PC2 with peak time was calculated across 5 doses for each protein-growth factor pair using Pearson correlation. Mean correlation for PC1 with maximum peak magnitude was 0.97, while mean correlation for PC2 with peak time was 0.75. (C) The dataset is plotted based on projections onto the first two principal components, capturing 97.0% of the variance. Individual points are colored according to the signaling protein measured.

Based on our observations of the data, we thought the latent dimensions of the principal component analysis might describe physical features of the data, specifically information about the relative maximum signal achieved and the rate at which this signal was achieved. To test this hypothesis, for each protein ligand pair we determined the correlation between the dose response vector in PC1 and the magnitude of the peak for each dose in the original normalized data. We also determined the correlation between the PC2 dose-response vector and the time of peak

signal for each dose. As shown in Figure 3.11*B*, there is an extremely high correlation between the PC1 value and the relative magnitude of the peak response (mean $r = 0.97$). Similarly, with the exception of a few outliers in PC2 space, there is a high correlation between the value in PC2 space and the time to peak response in the original data (mean $r = 0.76$). The high correlation indicates that we can ascribe physical meaning to our principal component axes. Specifically, higher values along PC1 indicate that the signal achieves a higher relative maximum value. Higher values in PC2 space indicate that the signal achieves its maximum value at a later time. Lower values in PC2 space indicate that the signal achieves its maximum value at an earlier time.

Figure 3.11*C* shows the entire dataset reduced to the first two dimensions of PCA space. Each point represents a time course for a particular signaling protein at a single dose of a single growth factor. Points are colored to indicate the signaling protein being recruited to the EGF receptor. Points close together in PCA space represent responses that are similar to each other across the entire time course. The responses of Shp2 and PI3K-R1 are the most separated in both PCs indicating that they are the most different. The remaining points are densely packed in the intermediate region between the extremes of the PI3K-R1 and Shp2 signals.

To identify trends in the data, the measurements were organized into a dose series for each ligand/protein pair. This was visualized by connecting the PCA point representing the curve at the lowest dose of one growth factor to the point representing the curve at the next higher dose of that same growth factor with a directed arrow, continuing on for the five doses of each growth factor (see Figure 3.12 for an example). This approach reveals significant trends in the evolution of signals across the dose range, despite the density initially observed in PC space (Figure 3.13). Overall, the major mode of behavior for a given signaling protein is dominated by the identity

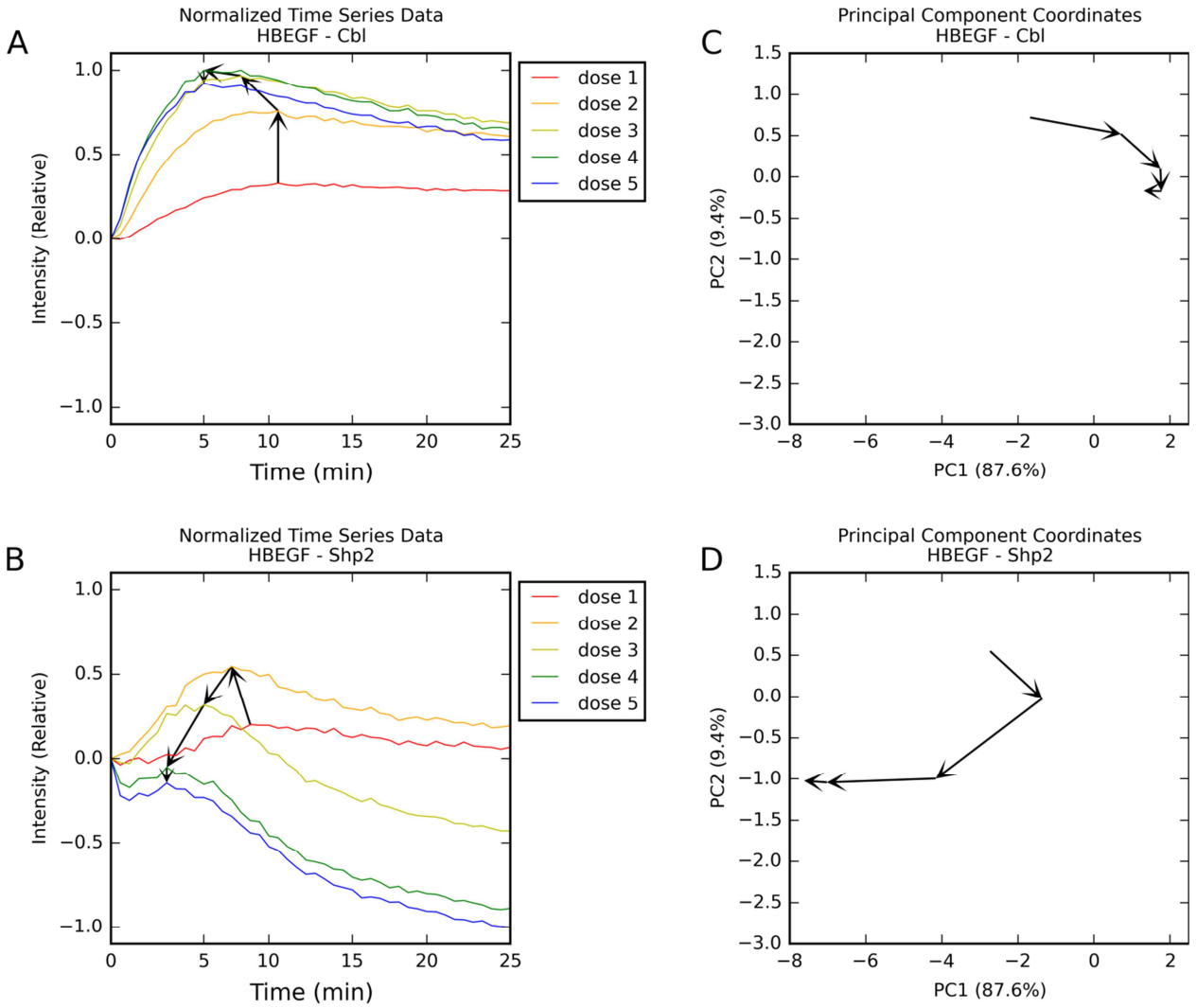

Figure 3.12: Data in Principal Component Space Correlate with Physical Trends in the Time Series.
The left panel shows normalized time series complementation data for HB-EGF with Cbl **(A)** and HB-EGF with Shp2 **(B)**. Superimposed on the plots for the five doses are lines with directed arrows connecting the peak value for each dose, from lowest dose to highest dose. The right panel shows the same time series data but now represented in principal component space, where each point represents an entire curve from the left panel. For both HB-EGF with Cbl **(C)** and HB-EGF with Shp2 **(D)**, PC1 correlates highly (r=0.97) with the magnitude of the peak at each dose (the peak y-axis value for each dose in the left panel). PC2 correlates highly (r=0.90) with the timing of the peak at each dose (the time of each peak from the x-axis of the left panel). Note that the axes are scaled differently and rotated in principal component plots, but the overall shape of the dose response has not changed from the time series to the principal component representation and preserves physical meaning from the time series data.

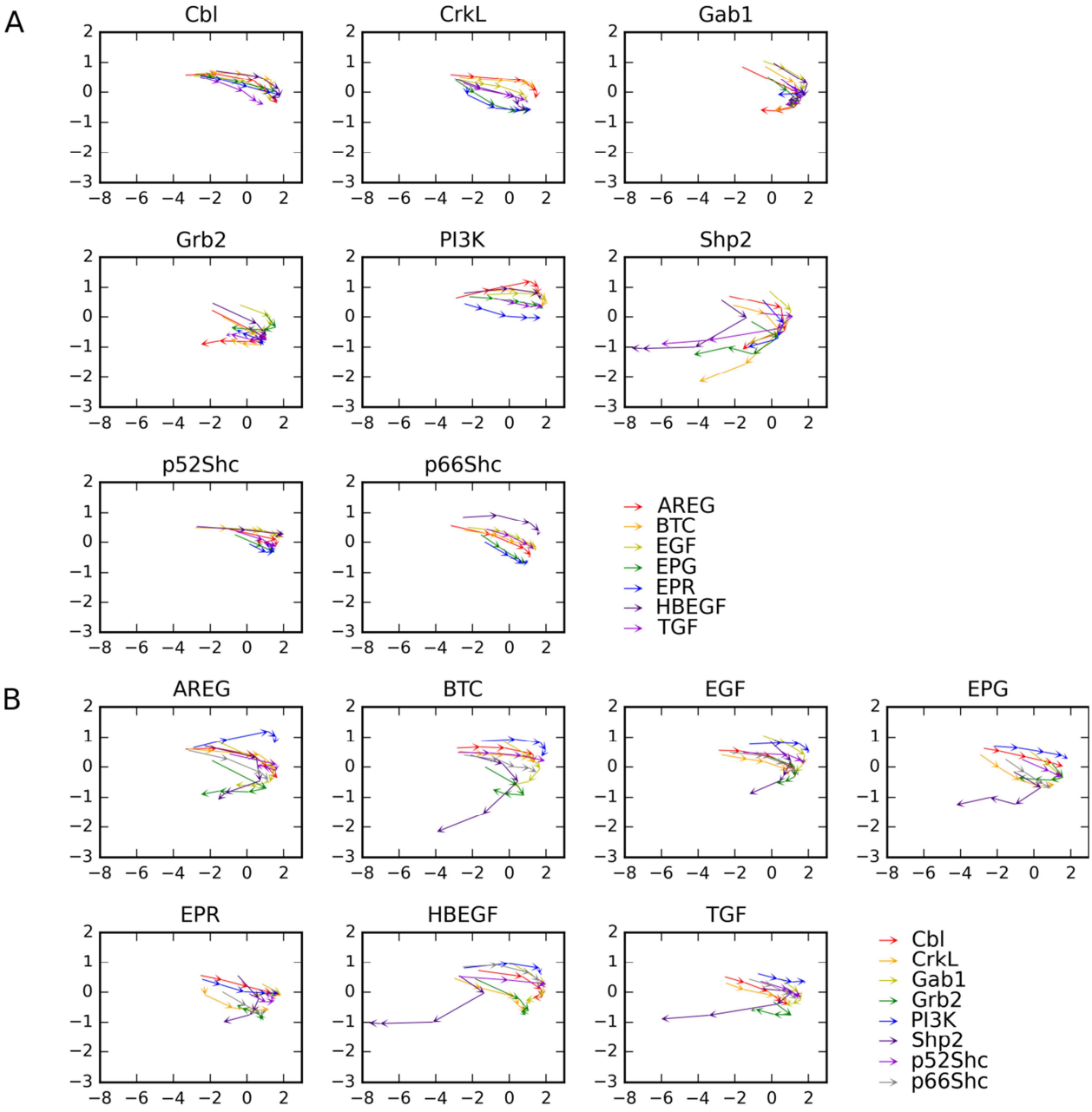

Figure 3.13: Global trends based on dose response.
The individual points in two-dimensional PC space representing a protein-ligand pair at a particular dose were organized into a dose response series for the five chosen doses, by connecting the response at lower doses to the next higher dose using a directed arrow. Panel A) The resulting vectors are grouped by signaling protein and colored according to the growth factor. Panel B) The resulting vectors are grouped by growth factor and colored according to signaling protein.

of the protein rather than the identity of the growth factor. Therefore, the curves describing the recruitment of the same signaling protein stimulated by any of the seven ligands (Figure 3.13*A*) are more similar to each other than they are to the curves that describe the recruitment of a different signaling protein stimulated by the same growth factor (Figure 3.13*B*). As is apparent from Fig. 6*B*, there are differences in how the individual protein responses evolve based on the stimulatory ligand.

Aside from these general observations, each PC shows two contrasting trends in a subset of proteins. First, for Cbl, CrkL, PI3K-RI, p52 Shc, and p66 Shc, there is a monotonic increase in PC1 (relative maximal signal) as the dose of growth factor increases. This is what is expected in a traditional dose-response curve, *i.e.* the signal increases with increasing dose. By contrast, Gab1, Grb2, and Shp2 show a bimodal response in PC1 space, reflecting an initial increase in response followed by a marked decrease in peak signal at the highest doses of most of the growth factors.

A second trend is that for most of the signals there is a monotonic progression down the PC2 axis. This indicates that the peak response is achieved more rapidly at higher concentrations of growth factor. An exception to this rule is the recruitment of p52 Shc in response to EGF, BTC, and HB-EGF, where there is little change in the time to peak response over the entire dose range tested.

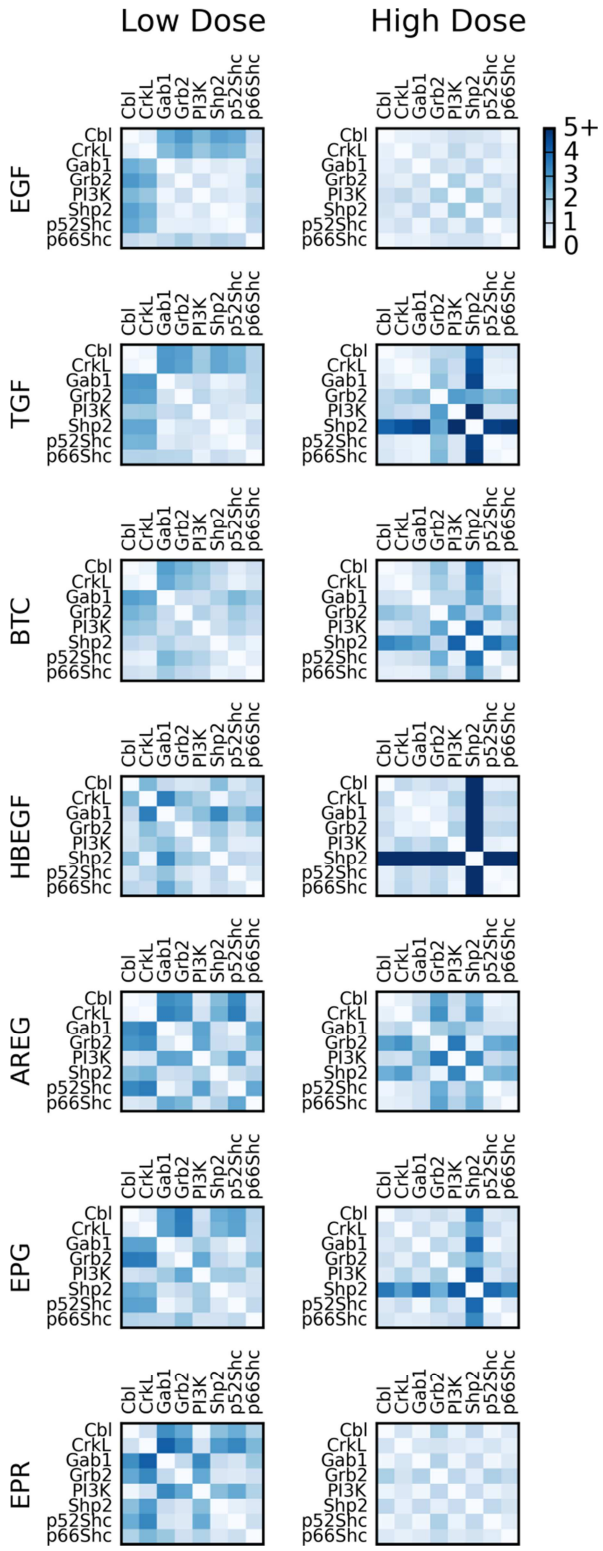

Figure 3.14: Signaling protein response varies by growth factor dose.
In order to quantify differences in protein response, Euclidean distances were calculated between proteins for each growth factor at both a low and high dose and visualized as a heat map. The response pattern for each growth factor at low dose is in the left column, and the response pattern for a high dose is in the right column.

3.4.5 Pairwise Interactions

The proteins chosen for this study were selected because they are involved in well documented interactions with the EGF receptor and with each other. Therefore, we would expect that the behaviors of some of these proteins should correlate in PC space. To quantify these relationships, we calculated the distances between interacting protein pairs in PC space for each dose of a single growth factor. Figure 3.14 shows heat maps of protein-pair distances for a low dose and a high dose for each growth factor. Two important features are immediately apparent from these heat maps. First, the patterns seen for the low and high doses of the same growth factor are distinctly different. This suggests that the same growth factor utilizes the network differently when applied at different concentrations. Second, the heat maps for each growth factor are very different, suggesting that the different growth factors activate the network in a manner that is specific to that growth factor.

To evaluate the similarity of protein recruitment dynamics across the different growth factors, distances were calculated between protein pairs in PC space across the five doses of each individual growth factor. The complete set of these cumulative distances was then rank-ordered, and both the top and bottom quartiles were probed for statistical enrichment for individual proteins or specific protein pairs. Several specific protein pairs were strongly represented in the top quartile. The pairwise distances of Cbl with CrkL and p52 Shc with p66 Shc were the most significantly enriched protein pairs in the top quartile ($p < 0.005$, Bonferroni-corrected), whereas the interaction of Gab1 with p52 Shc was also significantly enriched ($p < 0.05$, Bonferroni-corrected). Enrichment in the bottom quartile was also calculated to identify proteins and protein pairs that rarely exhibited similar dynamics. Shp2-based interactions as a group were identified as being enriched in this quartile ($p < 0.0005$, Bonferroni-corrected).

To compare the global response of the network to each of the seven different growth factors, we performed ensemble clustering on the pairwise distances between proteins, as described under “Experimental Procedures.” The percentage of time each growth factor clustered with another growth factor in the ensemble of clustering solutions is visualized in Figure 3.15as a heat map. The growth factors were then hierarchically clustered. As can be seen from this figure, BTC, EPG, and TGFβ form a strong cluster (the BTC cluster) because they cluster together in every clustering solution in the ensemble. AREG and EPR form a second strong cluster (the AREG cluster). EGF clusters most frequently with the AREG cluster (77%) but shares some membership in the BTC cluster (23%). HB-EGF is rather unique and is far from both the BTC and AREG clusters.

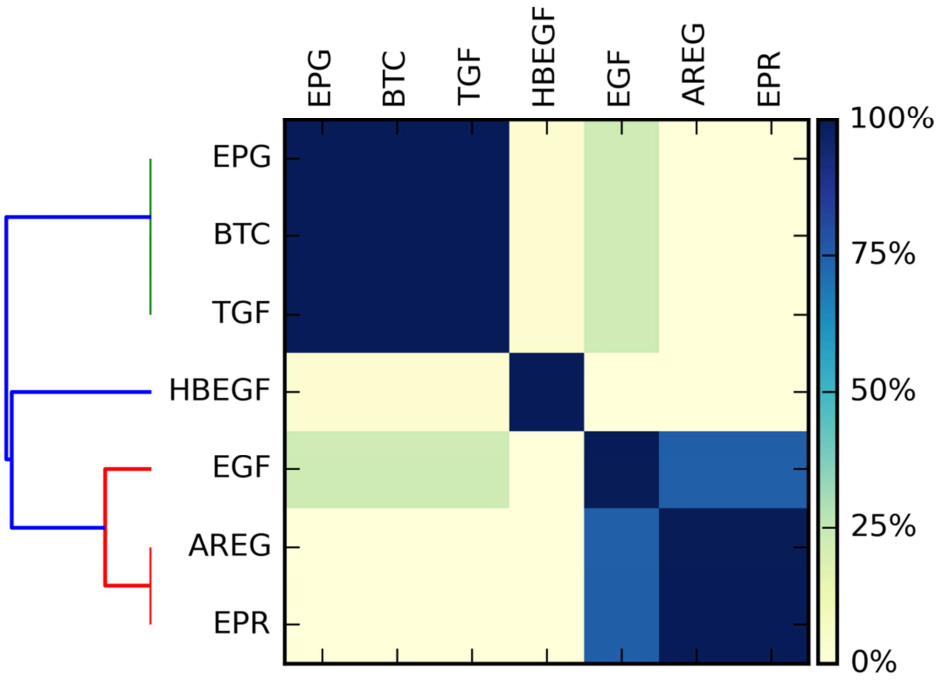

Figure 3.15: Heat map and dendrogram showing the results of clustering of the responses to the seven growth factors.
The pairwise protein distances for each ligand were converted to a vector and the vectors for each ligand were hierarchically clustered via the ensemble approach described in Experimental Procedures. The results are visualized via a heat map displaying the fraction of time each ligand pair clustered together across the ensemble.

3.5 Discussion

We report here on the use of luciferase fragment complementation to study the association of downstream signaling proteins with the EGF receptor. The advantages of this system include the ease of assay and the fact that it can be done in live cells with continuous monitoring. In addition, the signals generated from the eight signaling proteins examined here were robust, allowing detection of differences associated with changes in the concentration of growth factor. Finally, the approach is scalable and useful for screening applications. Using this system, we found that all eight of the selected signaling proteins are rapidly recruited to an EGF receptor containing complex, with association being apparent by 30 s. The peak extent of association occurred between 5 and 7 min, depending on the pairing. This is consistent with the time course of assembly of Shc-containing complexes after stimulation of cells with EGF, as documented through quantitative mass spectrometry (64).

In most of the pairings, the luciferase signal decreased slowly over time particularly at the higher doses of growth factor. As internalization of the EGF receptor begins almost immediately after addition of growth factor (65), it seems likely that at least part of the decrease in signal at longer times is due to internalization and degradation of the receptor and its associated signaling proteins. Nevertheless, at least some fraction of the agonist- induced increase in luciferase activity is maintained for as long as 25 min after the addition of EGF. These data imply that these signaling proteins remain associated with the EGF receptor even after it has been internalized. Thus, some aspects of signaling probably continue to occur well after the receptor has been removed from the cell surface. Receptor internalization is unlikely to account for the decrease in peak signal observed for the recruitment of Gab1, Grb2, and Shp2 at

high concentrations of all the growth factors. This decrease could reflect increased competition between signaling proteins for binding to sites on the EGF receptor when the signal is strong. It could also arise from depletion of a common pool of adapter or scaffold proteins when the stimulatory signal is maximal. Finally, it is possible that there is steric interference with luciferase complementation when the signal is strong, and Gab1, Grb2, and Shp2 bind to the EGF receptor in a multiprotein complex.

All EGF receptor/signaling protein pairs showed a dose dependence on the concentration of growth factor. However, the EC50 for any given growth factor varied as much as 18-fold for the recruitment of different proteins. Knudsen *et al.* (66) reported similar differences in the EC50 values of four EGF receptor ligands for inducing the phosphorylation of the EGF receptor and several signaling proteins. The molecular basis for this observation is not known, but it may reflect differences in the order or extent of phosphorylation of sites in response to these seven agonists (57, 67–70). Surprisingly, there were significant differences in the ability of saturating concentrations of erlotinib and lapatinib to inhibit the recruitment of these signaling proteins. This is likely due to differences in residual phosphorylation of the EGF receptor. These findings clearly identify erlotinib as a more effective inhibitor of signaling in this system than lapatinib and suggest that these complementation assays may be useful for identifying residual signaling pathways that could be targeted for therapeutic benefit.

The overarching message from the principal component analysis is that there are significant differences in signaling protein recruitment depending on both ligand and dose. These variable responses likely reflect different signaling protein recruitment strategies employed by the individual ligands over their entire dose range. Although the observed differences are subtle at the level of individual proteins recruited, collectively they could readily give rise to a

distinctly different biological outcome for each of the agonist ligands. *In vivo* levels of EGF and other ErbB family growth factors vary widely from tissue to tissue, being low in plasma but 10–100-fold higher in secretions such as saliva and tears (70–72). Given the differences in network behavior identified here, our data imply that therapeutic agents that target one particular node in the signaling pathway could be efficacious in one tissue but not in another, simply because of differences in network utilization based on the identity of the stimulating growth factor and/or the dose involved. This underscores the need to understand the signaling network at all doses of growth factor, as different tissues will likely be responding to vastly different doses of EGF or other EGF receptor agonists. Because many of the experiments that have defined our understanding of this network have been carried out using high dose EGF (73–75), our current appreciation of the network may not reflect the actual flux through the pathways under all physiological conditions.

Ensemble clustering of the responses to the growth factors demonstrated that the seven different EGF receptor ligands basically cluster into three groups as follows: (i) BTC, TGFβ, and EPG; and (ii) EGF, AREG, and EPR. HB-EGF is distantly related to both clusters. Thus, based on their ability to recruit these signaling proteins to the EGF receptor, these ligands do not fall neatly into groups defined by high *versus* low affinity nor do they fall into groups based on whether they bind only to the EGF receptor or to both the EGF receptor and ErbB4 (63). Whether there is some specific functional difference that distinguishes the two main groups of EGF receptor agonists, such as temporal or spatial differences in expression, remains to be determined.

With respect to similarities in the utilization of the network by the different growth factors, our analysis identified a strong correlation between the recruitment of Cbl and the

recruitment of CrkL to the EGF receptor. As CrkL is known to bind directly to Cbl (76, 77), the detection of a correlation between Cbl and CrkL binding to the EGF receptor suggests that the primary mechanism through which CrkL associates with the EGF receptor may be through binding to tyrosine-phosphorylated Cbl. The fact that this relationship is clearly observed in our dataset suggests that this analysis is capable of identifying interactions between proteins that associate within this signaling network. Viewed in this light, the significant correlation between p52 Shc and Gab1 suggests that this also represents a preferred interaction in this network. The direct binding of p52 Shc to Gab1 has been reported (78, 79). The finding that other canonical network interactions, such as Grb2/Shc or Grb2-Gab1, were not detected in this analysis likely reflects the complex and dynamic behavior of the network. Grb2 is an adapter protein that recruits a number of proteins, including Cbl, Gab1, and Shp2, to the EGF receptor. It can bind directly to the EGF receptor or indirectly through Shc. As a result, the interaction of Grb2 with the EGF receptor represents the summation of a multiplicity of different binding events. Variation in the dynamics of the different binding events, such as Grb2-Cbl *versus* Grb2-Gab1, could easily obscure any correlations between the binding of the individual partners in the protein pairs. Thus, it will be necessary to assess these interactions more directly to determine whether their association is differentially affected by the seven EGF receptor agonists. Ultimately, we would like to be able to determine which path through the network is used to recruit a particular protein to the EGF receptor signaling complex by a particular growth factor at a particular dose. Prediction on this level is likely to require careful modeling of network behavior. To this end, these data can be used, in conjunction with other information, to build mechanistic models of the network interactions to determine the dose-dependent network paths of a given signaling protein.

Chapter 4: Avoiding Common Pitfalls when Clustering Biological Data.

From Ronan,T., Qi,Z. and Naegle,K.M. (2016) Avoiding common pitfalls when clustering biological data. *Sci. Signal.*, **9(432)**: re6. doi: 10.1126/scisignal.aad1932. Reprinted with permission from AAAS.

4.1 Abstract

Clustering is an unsupervised learning method, which groups data points based on similarity, and is used to reveal the underlying structure of data. This computational approach is essential to understanding and visualizing the complex data that are acquired in high-throughput multidimensional biological experiments. Clustering enables researchers to make biological inferences for further experiments. Although a powerful technique, inappropriate application can lead biological researchers to waste resources and time in experimental follow-up. We review common pitfalls identified from the published molecular biology literature and present methods to avoid them. Commonly encountered pitfalls relate to the high-dimensional nature of biological data from high-throughput experiments, the failure to consider more than one clustering method for a given problem, and the difficulty in determining whether clustering has produced meaningful results. We present concrete examples of problems and solutions (clustering results) in the form of toy problems and real biological data for these issues. We also discuss ensemble clustering as an easy-to-implement method that enables the exploration of multiple clustering solutions and improves robustness of clustering solutions. Increased awareness of common clustering pitfalls will help researchers avoid overinterpreting or misinterpreting the results and missing valuable insights when clustering biological data.

4.2 Introduction

Technological advances in recent decades have resulted in the ability to measure large numbers of molecules, typically across a smaller number of conditions. Systems-level measurements are mined for meaningful relationships between molecules and conditions. Clustering represents a common technique for mining large data sets. Clustering is the unsupervised partitioning of data into groups, such that items in each group are more similar to each other than they are to items in another group. The purpose of clustering analysis of biological data is to gain insight into the underlying structure in the complex data – to find important patterns within the data, to uncover relationships between molecules and conditions, and to use these discoveries to generate hypotheses and decide on further biological experimentation. The basics of clustering have been extensively reviewed (80–82). Clustering has led to various discoveries, including molecular subtypes of cancer (83–86), previously unknown protein interactions (87), similar temporal modules in receptor tyrosine kinase cascades (88), metabolic alterations in cancer (89), and protease substrate specificity (90).

Although clustering is useful, it harbors potential pitfalls when applied to biological data from high-throughput experiments. Many of these pitfalls have been analyzed and addressed in publications in the fields of computation, bioinformatics, and machine learning, yet the solutions to these problems are not commonly implemented in biomedical literature. The pitfalls encountered when clustering biological data derive primarily from (i) the high-dimensional nature of biological data from high-throughput experiments, (ii) the failure to consider the results from more than one clustering method, and (iii) the difficulty in determining whether clustering has produced meaningful results. Biological systems are complex, so there are likely to be many

relevant interactions between different aspects of the system, as well as meaningless relationships due to random chance. Differentiating between a meaningful and a random clustering result can be accomplished by applying cluster validation methods, determining statistical and biological significance, accounting for noise, and evaluating multiple clustering solutions for each data set. The high-dimensional nature of biological data means the underlying structure is difficult to visualize, that valid but conflicting clustering results may be found in different subsets of the dimensions, and that some common clustering algorithms and distance metrics fail in unexpected and hidden ways. To address these issues, clustering parameters and methods that are compatible with high-dimensional data must be identified and implemented, the results must be validated and tested for statistical significance, and researchers should become used to applying multiple different clustering methods as part of routine analysis.

Some solutions to address these pitfalls require awareness of the issue and the use of appropriate methods, whereas other solutions require substantial computational skill and resources to implement successfully. However, one method – *ensemble clustering* (that is, clustering data many ways while making some perturbation to the data or clustering parameters) – solves multiple pitfalls and can be implemented without extensive programming or computational resources. We mention the uses of ensemble clustering, as appropriate, and provide an overview of ensemble clustering at the end.

4.3 High-Dimensionality Affects Clustering Results

Systems-level measurements and high-throughput experiments are diverse in the size of the data sets, number of dimensions, and types of experimental noise. Examples include measurements of the transcript abundance from thousands of genes across several conditions (such as in multiple cell lines or in response to drug treatments) (85, 91), measurements of changes in the abundance of hundreds of peptides over time after a stimulation (92), or measurements of hundreds of microRNAs across tissue samples (85). Because the dimensionality of the data in a clustering experiment depends on the objects and features selected during clustering, understanding how to determine dimensionality and its effects on clustering are prerequisites for approaching a clustering problem. As a rule of thumb, data with more than 10 dimensions should be considered high dimensional and should be given special consideration.

4.3.1 Determining Dimensionality

The dimensionality of a clustering problem is defined by the number features that an object has, rather than by the number of objects clustered. However, the definition of object and feature in a given clustering problem depends on the hypothesis being tested and which part of the data is being clustered. For example, in the measurement of 14,546 genes across 89 cell lines, as found by Lu, et al. (2005) (3), we can ask two with these data. First, what is the relationship between the genes based on their changes across the 89 cell lines? This case represents a gene-centric clustering problem with 14,456 observations, with each observation having 89 dimensions (Figure 4.1A). Second, what is the relationship between cells based on the changes in gene expression across all of the genes? This case represents a cell line clustering problem,

obtained by transposing the original data matrix such that the matrix now has 89 observations, with each observation having 14,456-dimensions (Figure 4.1B).

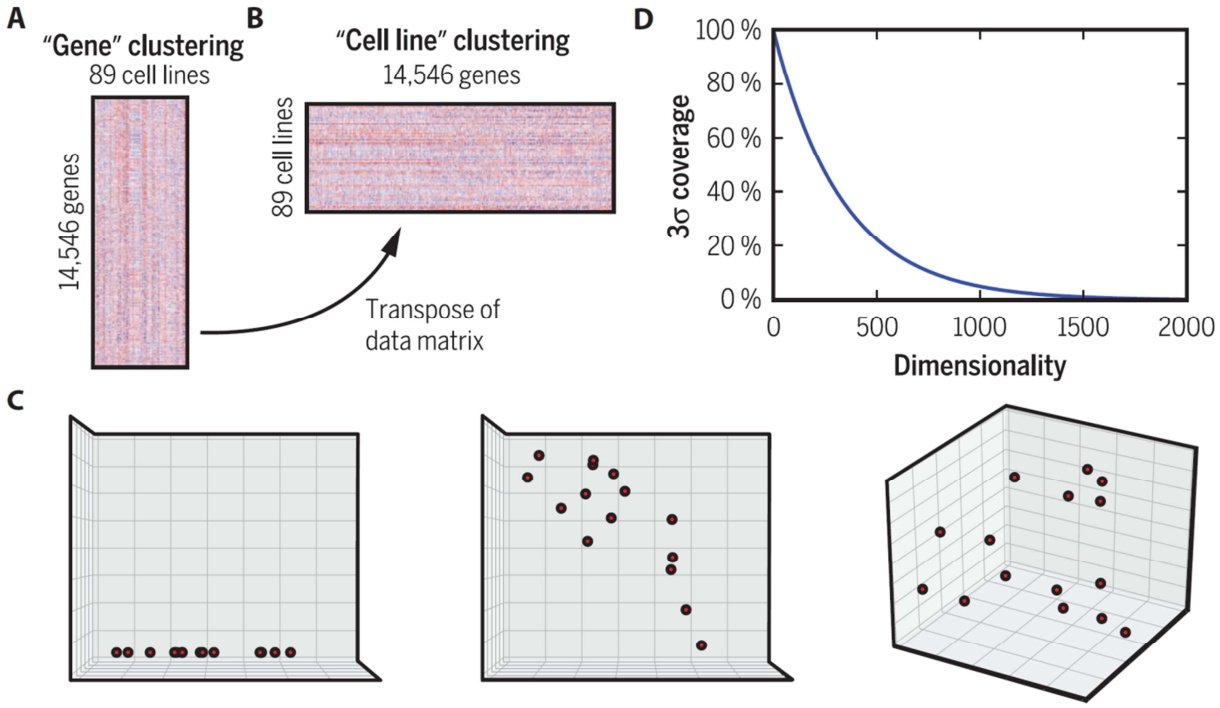

Figure 4.1: Determining the Dimensionality of a Clustering Problem.
(A) A representation of the Lu mRNA clustering problem, consisting of over 14,000 mRNA measured across 89 cell lines, from Lu, et al. (2005). When the mRNA are being clustered, the mRNA are the objects and each cell line represents a feature resulting in an 89-dimensional problem. (B) When attempting to classify normal and tumor cell lines using gene expression, the objects to be clustered are the cell lines and each mRNA is a features resulting in a clustering problem of thousands of dimensions. (C) For a fixed number of points, sparsity increases as dimensionality increases. (D) In a one-dimensional Gaussian distribution (represented by a typical bell curve) three standard deviations cover 99.7% of the data. In two- or three-dimensions, when independently distributed, this coverage is reduced slightly (to 99.5% and 99.2% respectively). In 10 dimensions, three standard deviations cover only 97.3%. By 100 dimensions, coverage is reduced to 76.3%, and by 1000 dimensions it is reduced to 6.7%.

The same data can be clustered in different ways to discover different biological insights, and these different ways of clustering the data can have large differences in dimensionalities. The gene-centric clustering problem represents a good basis for clustering because observations

greatly exceed the dimensions. However, even that problem represents a high-dimensionality situation. The cell line clustering problem is even more challenging because the relatively small number of observations (89) compared with the large dimensionality (>14,000) could be dominated by noise in the expression data. Without careful handling of sparsity and feature masking, clustering will almost certainly deliver poor or misleading results. Reliable and meaningful clustering results are likely achievable only with careful dimensionality reduction or subspace clustering.

4.3.2 Geometry and Distance in High-Dimensional Data

As dimensionality increases beyond two- and three-dimensional spaces, the effects of high-dimensionality, referred to as the “curse of dimensionality” come into play. These effects manifest in three key ways: (i) Geometry behaves nonintuitively in higher dimensions; (ii) sparsity is common in high-dimensional data sets; and (iii) relevant features tend to become masked by increasing numbers of irrelevant features.

As dimensionality increases , familiar relationships of distance, volume, and probability distributions behave nonintuitively (93). The method of determining distance between points in a clustering problem (the distance metric) influences the clustering result. With high-dimensional biological data, distance metrics fail to behave as expected based on our low-dimensional intuition. For example, as dimensionality increases beyond 16 dimensions, the nearest neighbor to a point can be the same distance away as the farthest neighbor to a point (for many common distance metrics) (94–96). Put another way, because it is likely that any two points are far apart in at least a few dimensions, all points approach being equally far apart (97). This means that for many types of distance functions in high dimensional space, all points are effectively equidistant from one another (96, 98). As a result, some common distance metrics and the concept of

"nearest neighbor" can be potentially meaningless in high-dimensional space (96, 99). Although data with more than 16 dimensions should be considered high-dimensional (96), data with as few as 10 dimensions can also exhibit non-intuitive high-dimensional behavior (93).

Some algorithms that were developed for lower-dimensional spaces do not generalize well to high-dimensional spaces. Many centroid and density-based algorithms, such as *k*-means (centroid) and DBSCAN (density-based) rely on defining a *nearest neighbor,* which only works well in lower-dimensional spaces (100). Thus *k*-means and DBSCAN (101) often fail to return useful results when used on high-dimensional data (94, 96). Furthermore, these algorithms will give no indication that they are not working as expected. Despite these problems, reliable and interpretable results with high-dimensional data can be achieved if ensembles of these clustering algorithms are used (102, 103).

Some algorithms are specifically designed to function with high-dimensional data. Hypergraph-based clustering methods draw on the field of graph theory, a method of representing pairwise relations between objects. Hypergraph-based clustering can be used to transform sparse data in high-dimensions to a problem in graph partitioning theory, drawing on unique methods from that field, to produce accurate and informative clustering (94).Grid-based clustering is a density-based approach and can fail to give meaningful results on some high-dimensional data sets. However, Optigrid is a grid-based clustering approach that specifically addresses the problems of distance and noise that confound other similar algorithms when applied to high-dimensional data sets (100). Some algorithms, such as NDFS, simultaneously solve problems with noise and find subgroups (104), which can improve the accuracy of clustering of high-dimensional data. Others, like OptiGrid, alternate rounds of grouping and dimensionality reduction to cluster high-dimensional data (105). When working with high-

dimensional data, clustering algorithms that are suited for the dimensionality of the clustering problem should be used.

4.3.3 Sparsity in High-Dimensional Data

Sparsity is a problem with high-dimensional data because clusters are most clearly defined when there are many objects in each cluster, resulting in robust representation of each cluster. Given a fixed amount of data, increasing dimensionality causes the data to become spread out in space. For example, given a set of data points in one dimension, data may be densely packed (Figure 4.1C). As the dimensionality increases to two or three dimensions, each unit volume of the space becomes less and less populated. Extrapolating to an even more sparsely populated high-dimensional data set, it becomes increasingly difficult to ascertain whether a distant data point represents a noisy point far from one cluster or a new cluster that has only a few members and thus is difficult to identify. Effectively, random noise dominates clustering results based on sparse data.

Sparsity also affects our low-dimensional intuition for statistical rules-of-thumb. We typically use three standard deviations (SDs, $\pm 3\sigma$) to determine a reasonable threshold for statistical significance. In a one-dimensional Gaussian distribution (represented by a typical bell curve) 3 SDs cover 99.7% of the data. In two- or three-dimensions, when independently distributed, this coverage is reduced slightly (to 99.5% and 99.2% respectively). In 10 dimensions, 3 SDs cover only 97.3%. By 100 dimensions, coverage is reduced to 76.3%; by 1000 dimensions it is reduced to 6.7% (Figure 4.1D) (95). This means that, in high-dimensional space, our rules-of-thumb for interpreting variance and what threshold should be considered statistically significant need to be reconsidered.

4.3.4 Masking Relationships in High-Dimensional Data

With high-dimensional data, the signal can easily get lost in the noise. Biological noise and variation contribute to irrelevant features, which can mask important features, and ultimately influence cluster membership in the full-dimensional space (106). In the example of gene expression analysis across 89 cell lines, measuring tens of thousands of transcripts guarantees that transcripts with no relevance to the biology under study will also be measured. However, the noise that is present in those transcripts from biological or technical variation can dominate clustering results because clustering algorithms treat the noise as if they are true features in the data. The background noise resulting from cell-cell variation might swamp the most relevant features necessary to identify differences between cell lines, or background noise can even masquerade as significant differences between cell lines when those differences are due to random processes in the cell.

Strong relationships among only a subset of features can mask other relationships. For example, metastatic cells and normal cells of a particular type may have similar gene expression profiles, so cell lines might tend to cluster by cell type (rather than normal versus metastatic) when the entire gene expression data are used. However, when only expression data from a subset of genes are used, the metastatic cells may exhibit similar expression patterns and thus would be grouped apart from the normal cells.

To exemplify how a signal can be detected in a lower-dimensional data set and lost in a higher-dimensional data set, we turn to a study by Lu and colleagues (85) in which they measured the relative abundances of microRNA and mRNA in 89 cell lines. Eighteen of 20 gastrointestinal cell lines clustered together with the microRNA data, which had 217 dimensions, but this relationship was lost when the mRNA data with ~14,000 dimensions were clustered.

Understanding that higher-dimensionality data can result in loss of the signal in the noise, we agree with the authors when they suggested that the loss of signal might result from the "the large amount of noise and unrelated signals that are embedded in the high-dimensional mRNA data". Fortunately strategies are available for unmasking the hidden signals in biological data.

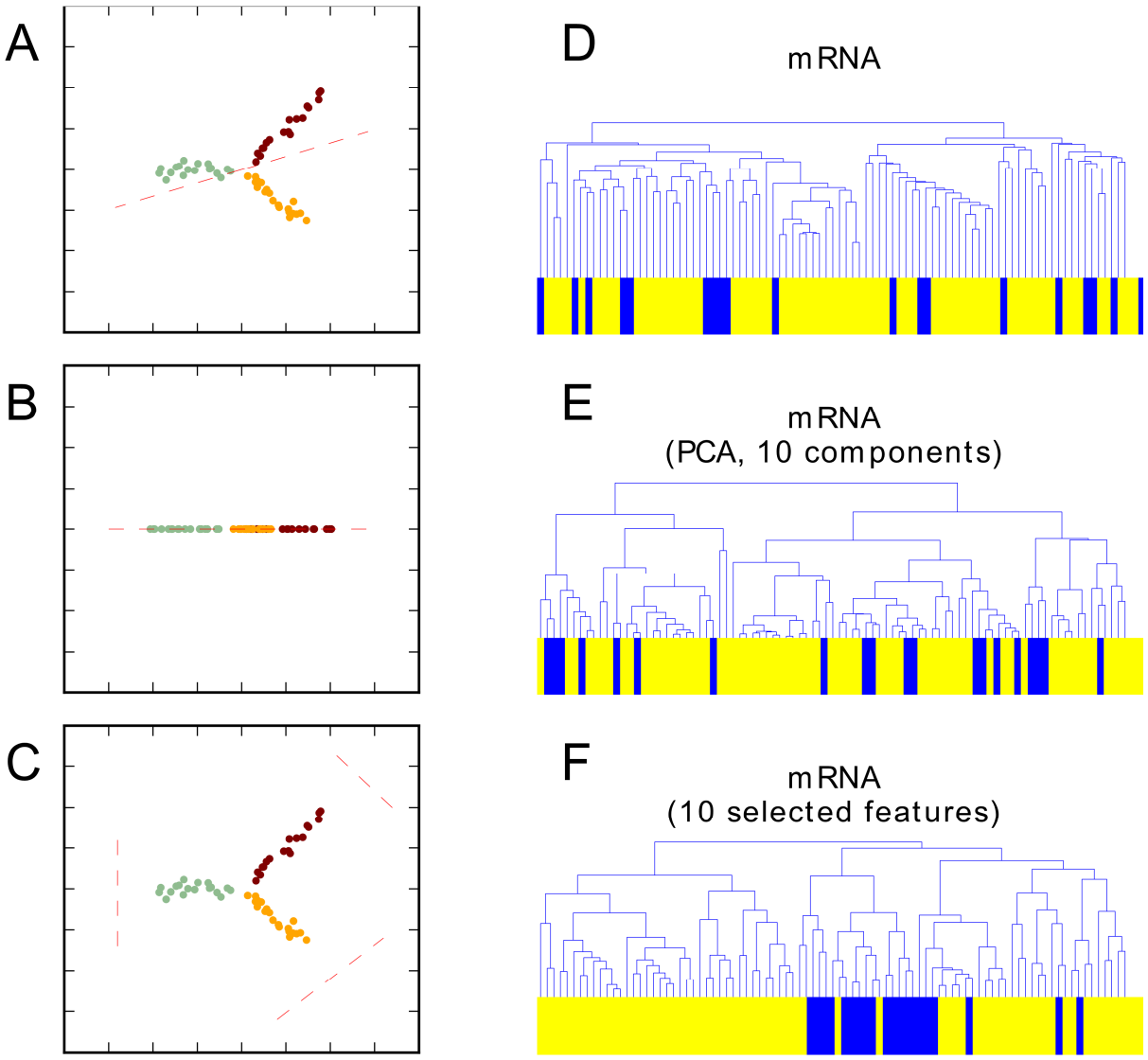

Figure 4.2. Dimensionality Reduction Methods and Effects.
Comparison of PCA and subspace clustering. (A) Three clusters are plotted in two dimensions. PCA determines the direction of greatest variance (A, dashed red line). (B) Clusters after dimensionality reduction by PCA. (C) Three subspaces (red dashed line) are identified upon which to project the data. (D through F) A comparison of the original clustering results of 89 cell lines in ~14,000 dimensional mRNA data (D), to clustering results after PCA (E), and after subspace clustering (F). Blue bars represent GI cell lines, yellow bars represent non-GI cell lines.

There are competing schools of thought for addressing the masking of relevant features by irrelevant features. If only a few features (dimensions) of the data are likely to contain the most relevant information, some advocate applying techniques that reduce the dimensionality. Dimensionality reduction involves the selection of only a subset of the features; the selection of which is based on a criterion such as predictive power or variance. When only a few features are expected to contain signal and the rest are expected to be noise, Principal Component Analysis (PCA) is often used. PCA reduces high-dimensional data to fewer dimensions that capture the largest amount of covariance in the data. Another method of reducing dimensionality is to remove features with low values, as is often done in microarray analysis where transcripts below a threshold, or transcripts changing only a very small amount between conditions, are removed from the data set. Alternatively, if multiple independent signals exist in the data, selection of different subset of features to cluster (107) may reveal different relationships among the data

These three commonly applied methods for reducing dimensionality will produce clustering results, but each has limitations and, by definition, eliminates some features of the data. Dimensionality reduction techniques can degrade clustering results for high-dimensional datasets (108) when that is present only in some subset of the data is eliminated as a result of the dimensionality reduction. To illustrate this problem, we use a toy data example in which three clusters are readily apparent when the data are presented in two dimensions (Figure 4.2A). A single projection onto any one dimension (determined by PCA, defined by the dotted line) (Figure 4.2B) reduces the separability compared with the two-dimensional representation such that the identity of at least one cluster is lost. This “local feature relevance” suggests that dimensionality reduction techniques, such as PCA, may be inappropriate if different clusters lie

in different subspaces, and that one should avoid the application of a single instance of feature selection (106) to avoid missing important structures within the data.

In contrast, subspace clustering is a method that searches different subsets of original features to find all valid clustering results, no matter which subset they are found in. Subspace clustering does not use graph theory or partitioning methods. Because it involves the selection of a subset of features, it also does not rely on the density of the data. In some data, objects will cluster only in subspaces of the full data space. For example, in the toy problem, when three different subspaces are chosen, at least two of the three clusters would be well separated when projected on to any of the highlighted subspaces (when projected onto the dashed lines) (Figure 4.2C). Subspace clustering must be carefully tailored to high-dimensional data to produce valid results (109). As a computationally intensive method, it can hit limitations due to a combinatorial explosion of potential subspaces in high dimensions (94). However, subspace clustering can reveal the multiple, complex relationships in biological data. Although information is lost in each subspace, multiple subspaces are considered; therefore, subspace clustering results are informed by information from all relevant dimensions.

The problematic effects of dimensionality reduction and the efficacy of subspace clustering can be seen on the expression data from Lu *et al*. (85). The clustering results from the original study (clustering 89 cell lines using the MRNA data, ~14,000 dimensions) (Figure 4.2D) were compared to the clustering results after using PCA to reduce the dimensionality to the 10 most relevant features (Figure 4.2E) and after using subspace clustering to reduce the dimensionality to the 10 most informative features (Figure 4.2F). As Lu *et al*. found, we do not see significant grouping of cell lines of gastrointestinal origin when we cluster the data in the full feature space (Figure 4.2D), or if we reduce dimensionality using the first 10 principal

components as features from PCA (Figure 4.2E). However, a selection of one 10-dimensional subspace shows strong clustering for GI cell lines (Figure 4.2F) – almost as strong as the results that Lu and colleagues presented for much lower-dimensional microRNA data (as discussed below, and as shown in (Figure 4.2D). This analysis suggests that, although the reduced dimensions of principal component space may have uncovered a structure we do not understand, PCA was not informative when attempting to group cells based on their tissue of origin. However, there is a subspace (a subset features from the original dimensions) for the mRNA data in which we can successfully group cells by their origins. This example illustrates how irrelevant features in the high-dimensional space masked the grouping of cells by origin, despite the data including expression measurements of genes that reflect tissue origin. The key was finding the right approach to cluster the data.

4.4 Effects of Clustering Parameters on Clustering Results

In addition to issues related to the high-dimensionality of biological data, clustering parameters also affect the clustering result. Given the same data, varying a single parameter of clustering such as the transformation, the distance metric, or the algorithm, can drastically alter the clustering solution. Unfortunately, there is often no clear choice of best metric or best transformation to use on a particular type of biological data. Each choice can mask or reveal a different facet of the organization within the data. Therefore, in addition to applying different methods of clustering and different approaches to addressing dimensionality, it is essential to consider results from multiple parameters when evaluating clustering solutions in biological data.

4.4.1 Transformations and Distance Metrics

Data is often transformed as part of analysis and processing. For example, transcriptional microarray data are commonly $\log_2$-transformed. This transformation expands the information for genes with low expression variation across samples and simplifies the identification of genes with differential expression. Similarly, in proteomics datasets, data may be centered and scaled by autoscaling or z-scoring (*25*) to make relative comparisons between signals for which magnitude cannot be directly compared. Although transformation can improve the ability to draw useful biological insight (110, 111), transformation also generates a new dataset with altered relationships that may reveal or mask some underlying biological relationships in the data (Figure 4.3, A and B).

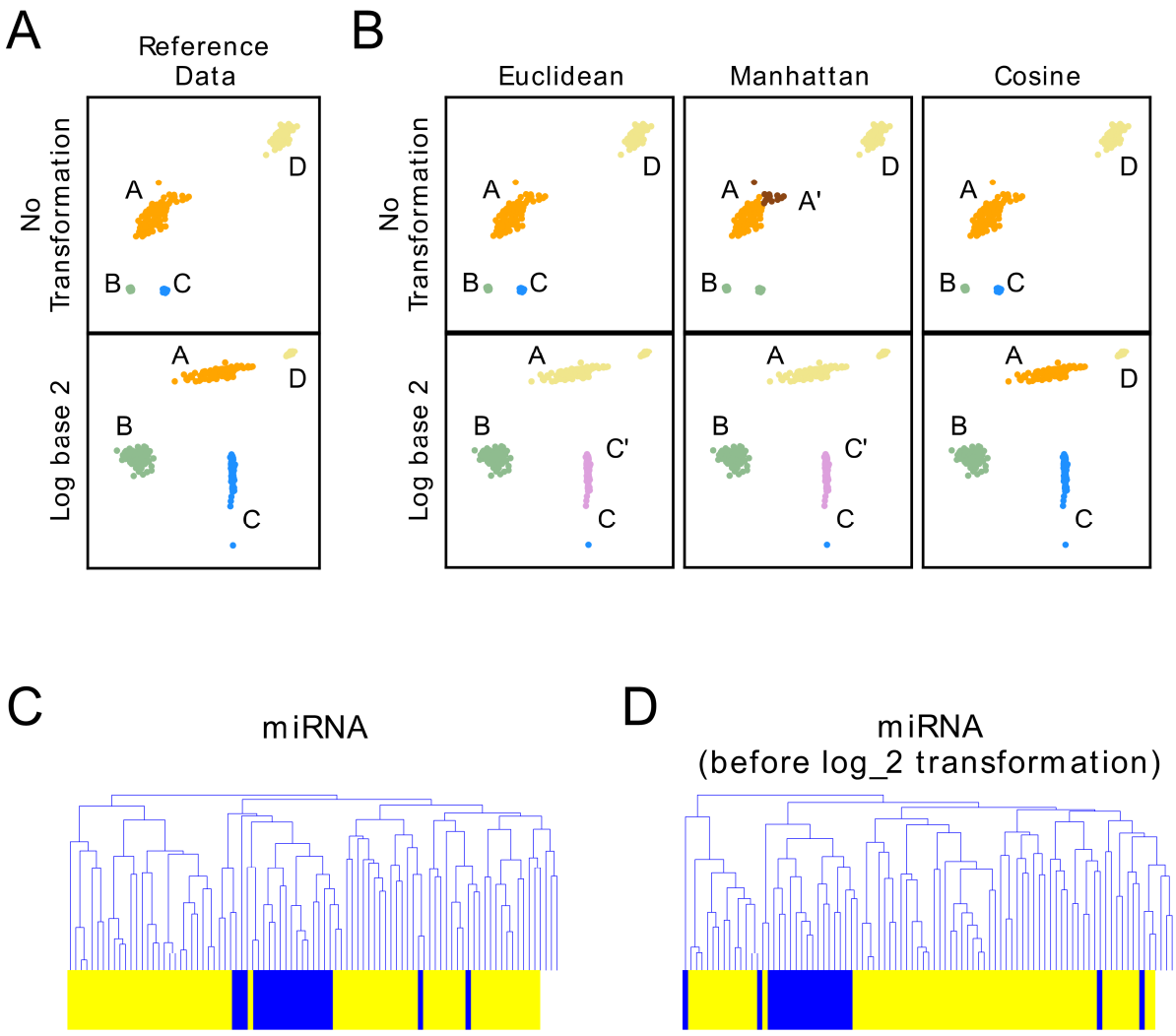

Figure 4.3. Transformations and Choice of Distance Metric Can Affect Clustering Results.
(A) A demonstration of how transformations affect the relationship of data points in space. A toy dataset (reference set) was clustered into four clusters with agglomerative clustering, average linkage, and Euclidean distance. The four reference clusters without transformation (upper panel) and after $\log_2$ transformation (lower panel). (B) Transformations and distance metrics change clustering results when compared to the reference clustering. With no transformation (upper panels), Euclidean and cosine distance do not change cluster identity, but with Manhattan distance, a new cluster A' is added, and cluster C is merged into cluster B. With the $\log_2$ transformation (lower panels), the Euclidean and Manhattan metrics cause cluster C' to emerge and cluster D to be lost. (C) The dendrogram from the microRNA clustering experiment result from Lu covering 89 cell lines and 217 microRNA. Gastrointestinal-derived cell lines (blue bars) predominantly cluster together in the full dimensional space. Note: The data was log2 transformed as part of the pre-clustering analysis. (D) The same microRNA data as in *(C)* but without log2 transformation. Although transformations can have a large effect on clustering results, here the effect is relatively minor. The GI cell lines still predominantly cluster together, although one GI cell line (leftmost blue bar) leaves the main cluster, and the two most distant cell lines become even more separated from the main cluster (rightmost two blue bars).

Although transformation of data is routine as a part of preclustering analysis, because of the effects on density and distance in the data set, any transformations done on the data at any point should be explicitly considered as a clustering parameter during the clustering process. We found that, compared with using different distance metrics or clustering algorithms (111), transformations often had the greatest impact on a clustering result (Figure 4.3B). On other cases, transformation has little impact on clustering results (Figure 4.3, C and D).

The choice of a distance metric also greatly affects clustering results, because different distance metrics accentuate different characteristics of the data (Figure 4.3B). Thus, to avoid missing information in the data, different distance metrics and transformations should be applied as a routine part of clustering analysis.

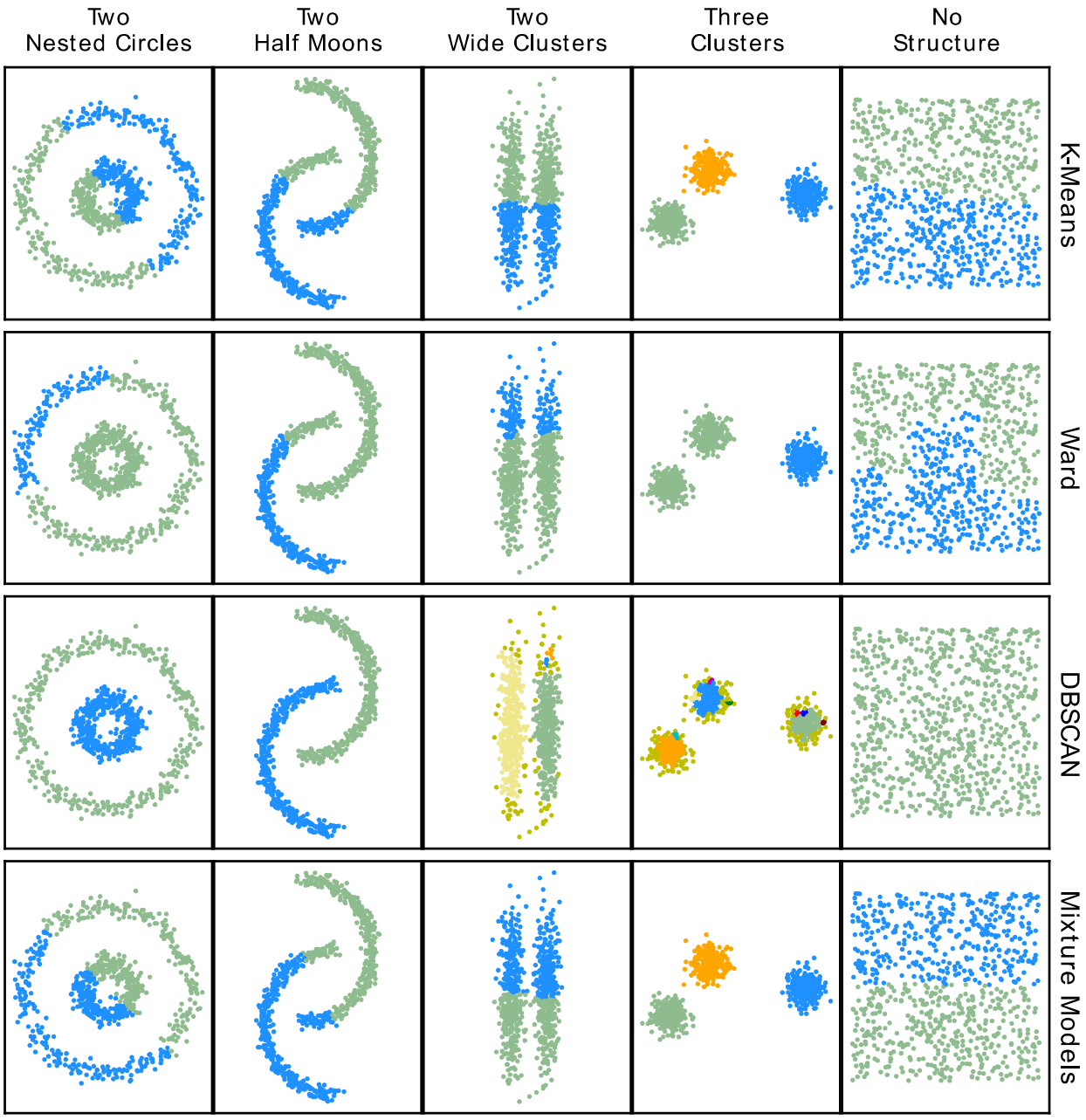

Figure 4.4: The Choice of Algorithm Can Affect Clustering Results.
Four toy datasets demonstrate effects of different types of clustering algorithms on various structures in two-dimensional data. The k-means algorithm depends heavily on selection of the correct value for *k* and tends to find spherical clusters (col 1). It performs poorly on irregularly shaped data (rows 3-5). The Ward algorithm (col 2) can produce different results depending on the threshold for similarity, highlighting hierarchical relationships. For example, at different thresholds, the green cluster (row 2, two lower groups) might take on separate cluster identities, indicating that the group is composed of two subgroups. The DBSCAN algorithm is a density-based clustering algorithm. Since it does not rely on a particular cluster shape, it can capture more complex structures in low-dimensional data (rows 3-5), and does not tend to find clusters in uniform data (row 1). However, variations in density can cause it to find additional clusters not found by other algorithms (rows 2 and 3). Statistical methods, like Gaussian mixture models (col4) fit statistical distributions to the data, but have limited success on non-normally distributed data (rows 3-5).

4.4.1 Clustering Algorithms

The choice of a clustering algorithm is based on several factors: (i) the underlying structure of the data, (ii) the dimensionality of the data, (iii) the number of relevant features for the biological questions being asked, and (iv) the noise and variance in the data. Each algorithm incorporates different assumptions about the data and can reveal different relationships among the data (Figure 4.4). There four primary classes of clustering algorithms are hierarchical, centroid, density, and statistical. The choice of clustering algorithm depends on the predicted structure of the data, and each algorithm class produces clusters with different properties. *Hierarchical clustering* is useful when data are expected to contain clusters within clusters, or the objects are expected to have a nested relationship. The Ward algorithm (Figure 4.4, column 2) can produce different results depending on the threshold for similarity, highlighting hierarchical relationships. For example, at different thresholds, the green cluster (Figure 4.4, row 2, column 2, two lower groups) might take on separate cluster identities, indicating that the group is composed of two subgroups. *Centroid clustering*, such as *k*-means clustering, assigns membership in a cluster based on the distance from multiple centers, resulting in roughly spherical clusters even when the underlying data is not spherically distributed. The *k*-means algorithm depends heavily on the selection of the correct value for *k* and tends to find spherical clusters (Figure 4.4, column 1). It performs poorly on irregularly shaped data (Figure 4.4, rows 3 to 5, column 3). *Density-based* algorithms, such as DBSCAN (101), connect groups of densely-connected points with regions of lower density separating clusters. Because it does not rely on a particular cluster shape, it can capture more complex structures in low-dimensional data (Figure 4.4, rows 3 to 5, column 3) and does not tend to find clusters in uniform data (Figure 4.4, row 1, column 3). However, variations in density can cause it to find additional clusters not found by

other algorithms (Figure 4.4, rows 2 and 3, column 3) *Statistical methods*, such as self-organizing maps (112) and Gaussian mixture models (Figure 4.4, column 4), fit statistical distributions to data in order to identify multiple groups of observations, each belonging to their respective distribution, but have limited success on non-normally distributed data (Figure 4.4, rows 3 to 5).

Whereas the data in Figure 4.4 are toy examples, real-world examples are often high-dimensional and difficult to plot. Often, the underlying structure is not known for most biological data, and it is likely that a complex biological data set will have multiple structures – non-spherical distributions, widely varying density scales, and nested relationships – that will only be revealed by applying multiple clustering algorithms.

4.5 Evaluating Clustering Results

How can you tell when the clustering result of biological data is meaningful? Because of the complexity of biological systems, there are likely to be many valid clustering solutions each revealing some aspect of underlying biological behavior. Unfortunately, there are likely to be many meaningless relationships simply due to random chance because the data are complex. Most clustering algorithms will find clusters, even if there is no true underlying structure in the data (as exemplified by Figure 4.4, top row). Therefore, clusters must be evaluated for biological relevance, stability, and cluster fitness. Understanding and accounting for noise and uncertainty in the data should be also considered when determining whether a clustering result is meaningful.

4.5.1 Cluster Validation

Validation metrics are a measure of clustering fitness. They can be used to determine if the result represents a well-defined structure within the dataset, using concepts such as cluster compactness, connectedness, separation, or combinations of these attributes. Much like distance metrics, each validation metric accentuates different aspects of the data to account for the final score (Table 4.1). The compactness of each cluster can be measured by the root-mean square standard deviation (RMSSTD) method (113), the *r*-squared (RS) method (114), and Hubert's Γ-statistic (115). Connectedness within a cluster can be measured by *k*-nearest neighbor consistency (116) or Handl's connectivity metric (117). Separateness is measured by the SD validity index (114), which compares average scattering and total separation between clusters. Some metrics combine measures of compactness and separation, such as the pseudo F-statistic (also known as the Calinski-Harabasz index) (118), the Dunn index (119), Davies-Bouldin index (120), silhouette width (121), or the gap statistic (122). When there are different numbers of clusters between the clustering results, or predominantly different membership in each cluster, some metrics might fail. The Rand index (123) is appropriate for validating these complex clustering differences.

To illustrate the differences in validation metrics that measure compactness, connectedness, and separation, we applied different metrics to the clustering results Figure 4.4. The results can be found in Table 4.2. For the toy data, DBSCAN generally performs better on the more complex structures, with some exceptions. DBSCAN correctly identifies no structure (Figure 4.4, row 1), the half-moons (Figure 4.4, row 4) and nested circles (Figure 4.4, row 5). This is also reflected in the validation metric results for connectedness, which is a better measure than compactness for non-spherical data shapes. The lack of structure in row 1 is also suggested

by the "zero" result to the *RS* measure of compactness and the 'not-available' result for separation from the SD validity index. For the spherical clusters, *k*-means and mixture models perform the best (Figure 4.4, row 2), which is reflected in the best scores for validation metrics measuring both compactness and separation. No algorithm performs well on the parallel lines data (Table 4.2). Although DBSCAN appears to do better with the center of the clusters (Figure 4.4, row 3), the errant results at the far left and right result in a poor score for validation metrics – on par with those algorithms which chop the horizontal parallel lines into vertical clusters.

Metric	References	Measures
Root-mean-square standard deviation (RMSSTD)	(113)	Compactness
R-squared (RS)	(114)	
Modified Hubert Gamma (Γ) statistic	(115)	
k-nearest neighbor consistency	(116)	Connectedness
Connectivity	(117)	
Determinant Ratio Index	(124)	
SD Validity Index	(114)	Separation
Pseudo-F statistic (Calinski-Harabasz)	(118)	Combination of compactness and separation
Dunn index	(119)	
Silhouette width	(120)	
Davies-Bouldin index	(120)	
Gap Statistic	(122)	
Rand index	(123)	Similarity between solutions

Table 4.1: Validation Metrics.
A number of validation metrics can be used for testing the quality of a clustering solution. While some focus on compactness, or connectedness, others use a combination of compactness and separation. The Rand index is particularly useful when there is a wide range of *k* between solutions, or large differences in cluster membership between clustering solutions.

		Compactness		Connectedness	Separation
	Validation Metric	RMSSTD	r-squared	Determinant Ratio Index	SD Validity Index
Algorithm	**Structure**	min	max	min	min
k-means	No structure	0.38	0.23	4.04	6.67
Ward	No structure	0.40	0.18	2.58	7.56
DBSCAN	No structure	0.44	0.00	1.00	N/A
Mixture models	No structure	0.38	0.23	4.09	6.67
k-means	Spherical clusters	0.78	0.81	249.08	0.32
Ward	Spherical clusters	1.18	0.56	26.43	0.18
DBSCAN	Spherical clusters	0.58	0.86	1683.53	2.52
Mixture models	Spherical clusters	0.78	0.81	249.08	0.32
k-means	Long parallel clusters	0.89	0.27	2.77	0.79
Ward	Long parallel clusters	0.93	0.21	2.10	0.75
DBSCAN	Long parallel clusters	0.84	0.21	22.36	4.32
Mixture models	Long parallel clusters	0.89	0.27	2.79	0.79
k-means	Half Moons	0.55	0.35	4.23	1.77
Ward	Half Moons	0.57	0.30	3.77	1.98
DBSCAN	Half Moons	0.60	0.21	3.03	2.61
Mixture models	Half Moons	0.56	0.33	5.21	1.93
k-means	Nested Circles	0.52	0.17	2.71	4.04
Ward	Nested Circles	0.53	0.13	1.97	3.30
DBSCAN	Nested Circles	0.57	0.00	1.00	4059.71
Mixture models	Nested Circles	0.52	0.17	2.71	3.99

Table 4.2: Validation Metrics.
Results of validation metrics measuring compactness (RMSSTD (113), r-squared (114)), connectedness (Determinant Ratio Index (124)), and separation (SD Validity Index (114)), applied to the data from Error! Reference source not found.. The validation metrics indicate the type of best score (max or min). The green cells represent the best score for each structure and validation metric.

The validation metric chosen should match the characteristics being tested. One approach is to use a metric that matches the selected algorithm. For example, when an algorithm optimizes connectedness, a metric that evaluates connectedness instead of one that evaluates compactness should be used (Table 4.1). An alternate approach is use to a panel of validation metrics, each measuring a specific aspect of the data. For example, when creating an ensemble of distance metrics and algorithms, a panel of metrics measuring compactness, connectedness, and

separation will eliminate inappropriate combinations of parameters. The worst pitfall is to not use any metric to assess the clustering outcome.

4.5.2 Clustering Stability

Because complex biological data can demonstrate multiple and independent valid clustering results, any single clustering result can be inconclusive. Validation metrics provide an assessment of the quality of a clustering result with regard to specific aspects of the data. In contrast, stability of the clustering result defines how robust the clustering is to perturbation, that is, how many ways of assessing the data produce a similar clustering result. Stability analysis works under the premise that clustering results represent the underlying structure of the data and should be relatively invariant to perturbations in the analysis. Stability analysis can identify robust clustering solutions using a variety of perturbations, such as accounting for (i) noise (125–127), (ii) projections of high dimensional data onto fewer dimensions (109), (iii) differences in algorithms, distance metrics, and data transformations (111), and (iv) the effect of selecting different random starting positions in nondeterministic algorithms (128).

Stability analysis assumes that there is a single “true” structure within the data; however, high-dimensional biological data can have multiple true structures that reveal distinct biological insights. Therefore, it may be equally valid to assume that there are multiple structures within the data. Multiple Clustering Analysis Methodology (MCAM) (111) is an approach that tests multiple clustering results. The approach in MCAM is based on the assumption that some perturbations to clustering parameters can uncover clustering results that reveal different biological insights. For example, some transformations uncovered shared binding partners of phosphorylation sites, while others highlighted shared biological function within a cell signaling pathway (111, 125).

Application of multiple clustering methods can reveal a single stable and robust result or uncover additional underlying structures in the data. Applying only a single clustering approach runs the risk of drawing conclusions based on noise or missing interesting and useful structures within the data.

4.5.3 Accounting for Noise and Measurement Uncertainty

Uncertainty in data means uncertainty in clustering results. However, many clustering algorithms do not account for noise or error when determining relationships between observations or when calculating distance. Although the molecular and cellular biology research community has adopted a set of rules that enable meaningful interpretation of differences in molecular measurements between pairs of conditions, this kind of standard has not been adopted for clustering complex, high-dimensional data. For example, when comparing means of measured data, we tend to require a minimum of triplicate measurements and the use of Student's *t* test to determine statistically significant differences within biological data (129, 130). Surprisingly, similar standards and rules do not exist for identifying differential patterns or groupings from clustering results, despite the data having the same measurements and biology as low-dimensional data between pairs of conditions.

Several methods account for the uncertainty in the data and properly propagate that uncertainty into clustering results. One ensemble approach is exemplified by Kerr and Churchill's work (131) in which the researchers performed repeated clustering on multiple datasets created by sampling gene expression from a statistical model that incorporates the variance of the measurements. From this analysis, the variation in the clustering result due to the variability in the original data is determined; thus producing a range of clustering results representing the variability. An alternative approach is to use model-based clustering algorithms

that account for the variance in the replicate measurements during the formations of clusters (127, 132, 133).

Because model-based approaches are not commonly accessible to molecular biologists in standard software packages, we explored ways to account for noise with ensembles (125) without necessarily relying on a statistical model of the data. This exploration led us to two major conclusions about the effects of noise. First, when data are well-separated, robust clusters can still be found, even in data with noisy distributions. Thus, clustering robustness cannot be predicted on the basis of the noise in the data. Second, the noise and variance in data are useful and contain information that can be revealed from the ensemble analysis. Specifically, we found that as signals propagate in time from a receptor tyrosine kinase to the mitogen-activated protein kinase (MAPK) pathway, there was high variability in an intermediate signaling state in the MAPK pathway. The effect of this noise in the clustering results reflected the relationship of this intermediate signal (a singly phosphorylated kinase) with both the upstream signal (the receptor) and downstream signal (the doubly phosphorylated form of the kinase) and represented a meaningful biological relationship. Consequently, we recommend against prefiltering data to remove those with high variance; instead, noise should be addressed in the clustering analysis, even in the absence of replicates (as detailed in (125)).

4.5.4 Determining Biological and Statistical Significance of Clustering Results

A primary challenge of clustering analysis is deriving biological insight. The most successful analyses often result from combining clustering with prior biological understanding of the system. However, unless the process of attaching biological meaning to the clusters is done with statistical care, we can often over-interpret relationships that “make sense” to us. The two

most common pitfalls related to this are (i) using anecdotal observations instead of a statistical test, and (ii) failing to account for the increase in false positives when multiple statistical tests are performed.

Ultimately, to avoid the first pitfall, a researcher must determine the likelihood of having made a particular observation by random chance. This "null model" or "background model" can then be used to assess how likely the relationship uncovered by clustering is truly related to the biological information under consideration, as opposed to the likelihood of it occurring by random chance. We demonstrate this process with the microRNA clustering results from the analysis of the 89 cell lines (Figure 4.3C) (85). To test the statistical hypothesis, we asked how likely it is that the 15 gastrointestinal cell lines would have occurred in the cluster of 20 cell lines by random chance. We used the theoretical hypergeometric distribution to calculate the probability using a right-tailed test. The resulting *P* value is approaches zero, indicating that this clustering result is unlikely to be due to random chance alone. In contrast, the random chance of any two gastrointestinal cell lines appearing in a cluster of 20 cell lines is likely due to random chance alone ($P \leq 0.9$).

In many biological data sets, we are not simply testing one label in a single cluster, but rather multiple labels across multiple clusters. A common example in cell biology and signaling research is to test for enrichment in any cluster of Gene Ontology terms (134) or biological pathway assignments from other sources.

Testing for the enrichment of such biological properties or classifications across the clustering results represents multiple hypothesis testing and thus requires "multiple hypothesis correction." For example, a *P* value cutoff of 0.01 for a single tested hypothesis yields the

prediction that false-positive results due to random chance occur rarely – only $1/100^{th}$ of the time. However, as is common is biological dataset analysis, we often test thousands of hypotheses (for example, asking if any one gene out of 10,000 is differentially expressed). In this case, even with a *P* value cutoff of 0.01, we would expect to find 100 false positive results due to random chance. If we identify 105 differentially expressed genes, but we know 100 are false positives, we cannot separate which five are likely to be the true positives without multiple hypothesis correction.

When choosing a procedure for multiple hypothesis correction, reducing false positives may simultaneously increase false-negatives (the elimination of real positives), and can result in missing biological insight. To reduce the frequency of false-negatives, we recommend the false discovery rate (FDR) correction procedure introduced by Benjamini and Hochberg (135), which is often applied in microarray analysis (136, 137), rather than the Bonferroni correction (138). Regardless of the type of correction chosen, when asking multiple questions, one must implement some form of multiple hypothesis correction to prevent overinterpreting the results.

4.6 Ensemble Clustering: A Solution to Many Pitfalls

Ensemble clustering refers to the act of clustering the data many times while making some perturbation – either to the data matrix or to clustering parameters – and then accounting for all of the clustering results across the ensemble. The goal of ensemble clustering is to improve the quality and robustness of clustering results when compared to any single clustering result.

Why do ensembles improve quality and robustness? In short, it is because uncorrelated "noise" cancels across clustering results. In ensemble clustering, noise in the clustering results occurs when there are strong biases due to the selected algorithms or because data contains poorly clustering points. Fortunately, if the errors made in each clustering result are not correlated, and the errors pertain to only a subset of the data, then the shared decisions made across the ensemble will dominate, resulting in convergence to the robust clustering result. A combination of diverse clustering results strengthens the underlying signal while filtering out the individual noise from each clustering result. Ensembles enable a more robust determination of the data structure than that acquired from a single clustering result obtained through analysis of the data without perturbation.

Perturbation	**Reason Behind Perturbation**	**References**
k	Stability can identify the optimum number of clusters	(126, 139, 140)
Noise	Biological and experimental noise should not change strong relationships within the data	(125, 127, 141)
Starting point (non-deterministic algorithms)	Identify those partitions that are independent of starting position or identify set of minima	(102, 128)
Projections into lower dimensions	Increase robustness to clustering noise as a result of the curse of high dimensionality	(109, 139, 142–144)
Subsampling	Subsets of the data should cluster consistently if the relationships are real	(145–147)
Parameters of clustering	Unique biological information can be uncovered by perturbing solution space	(111)

Table 4.3: Ensemble Perturbations.
Major perturbations applied to the data or to the clustering parameters in ensemble clustering and the motivating ideas for their use.

4.6.1 Ensemble Generation, Finishing, and Visualization

The process of performing ensemble clustering involves selection an appropriate perturbation (Table 4.3), collecting the clustering results based on the perturbation, and then either combining the individual clustering results into a single clustering result or exploring the ensemble of clustering results for information about the underlying structure of the data. To illustrate an ensemble of clustering solutions, we used random toy data and created an ensemble of clustering results using *k*-means clustering with an increasing number of clusters (*k*) (Figure 4.5).

There have been many techniques proposed to combine results of individual clustering results in the ensemble into one final clustering result. We refer to these methods as finishing techniques. Most finishing techniques use agreement across the ensemble to build a final clustering result. One method is to calculate the co-occurrence (or consensus) matrix. A co-occurrence matrix is an *m x m* matrix, where each entry, $C_{i,j}$, represents the number of times object *i* clusters with object *j* across all of the clustering results in the ensemble.

We clustered the co-occurrence matrix using hierarchical clustering and Ward linkage and plotted the result as a heatmap (Figure 4.5B). The clusters are formed on the basis of creating maximal in-group co-occurrence frequency and minimum co-occurrence with members outside the group. This representation reveals a wealth of detail about the relationships between data points and highlights data points that cocluster robustly (that is, frequenctly) with each other across the ensemble.

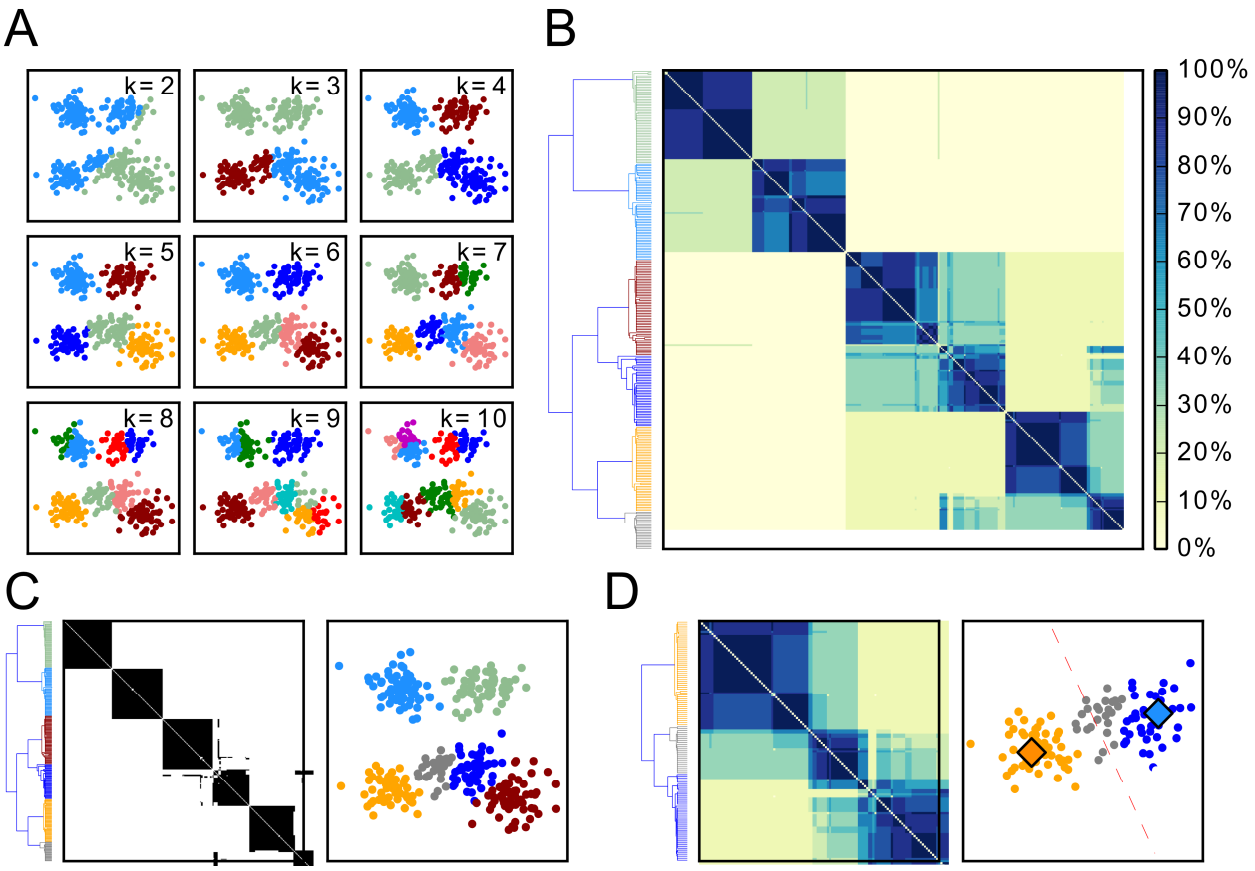

Figure 4.5: Ensemble Clustering Overview.
(A) A set of clustering results obtained using the k-means algorithm with various values of *k* (a *k*-sweep). (B) The hierarchically clustered (Ward linkage) co-occurrence matrix for the ensemble of results in (A). The heat map represents the percentage of times any pair of objects co-clusters across the ensemble. (C) A majority vote was taken using a threshold of 50% threshold on the co-occurrence matrix (left panel). The six strong clusters (see dendrogram color groupings) resulting from the majority vote are shown in the right panel. (D) The grey cluster provides an example of partially fuzzy clustering because it shares membership with the orange and dark blue clusters. These three clusters are isolated (D, left panel) and the co-occurrence matrix is reexamined. Rather than considering the gray cluster as distinct, one can consider it to have a partial membership in either the orange or blue clusters with a probability based on the co-occurrence matrix value.

Others have used majority voting, also based on the ideas of robustness (140, 148), as demonstrated in Figure 4.5C. We subjected the co-occurrence matrix above to majority voting (Figure 4.5C, left) and then plotted the ensemble majority-voting cluster assignments (Figure 4.5C, right). Identifying the most robust clustering result is useful for generating hypotheses for further experimental testing because hypotheses can be ranked on the basis of strength of the co-occurrence (90% of the time compared to 10% of the time).

Another finishing technique to identify robust portions of the ensemble is to apply graph theory. For example, if we assumed that the co-occurrence matrix represented edge weights (a numerical value indicating the strength of a connection) connecting the data points, we could traverse these weights to find maximally connected subgraphs and provide a different representation of robust clusters (149). With this concept of graph theory representation of ensemble clustering, we discovered that robustly clustered dynamic tyrosine phosphorylation data uncovered molecular-level interactions (87).

The finishing techniques mentioned above uniquely assign each observation to one cluster, thereby creating hard partitions within the data. However, a benefit of ensemble clustering is the ability to identify the probability of relationships (fuzzy partitioning), which can be applied to the entire data (probability is calculated for membership of any observation to any cluster) or a mixture of hard partitioning and probability based assignment. An examination of the portion of the heatmap representation of the co-occurrence matrix containing the blue, gray, and orange clusters demonstrates fuzzy partitioning (Figure 4.5D, left). The heatmap indicates that the gray cluster members share partial membership with the blue and orange clusters (Figure 4.5D, right). Rather than considering the gray cluster as distinct, one can consider it to have a

partial membership in either the orange or blue clusters with a probability based on the co-occurrence matrix value.

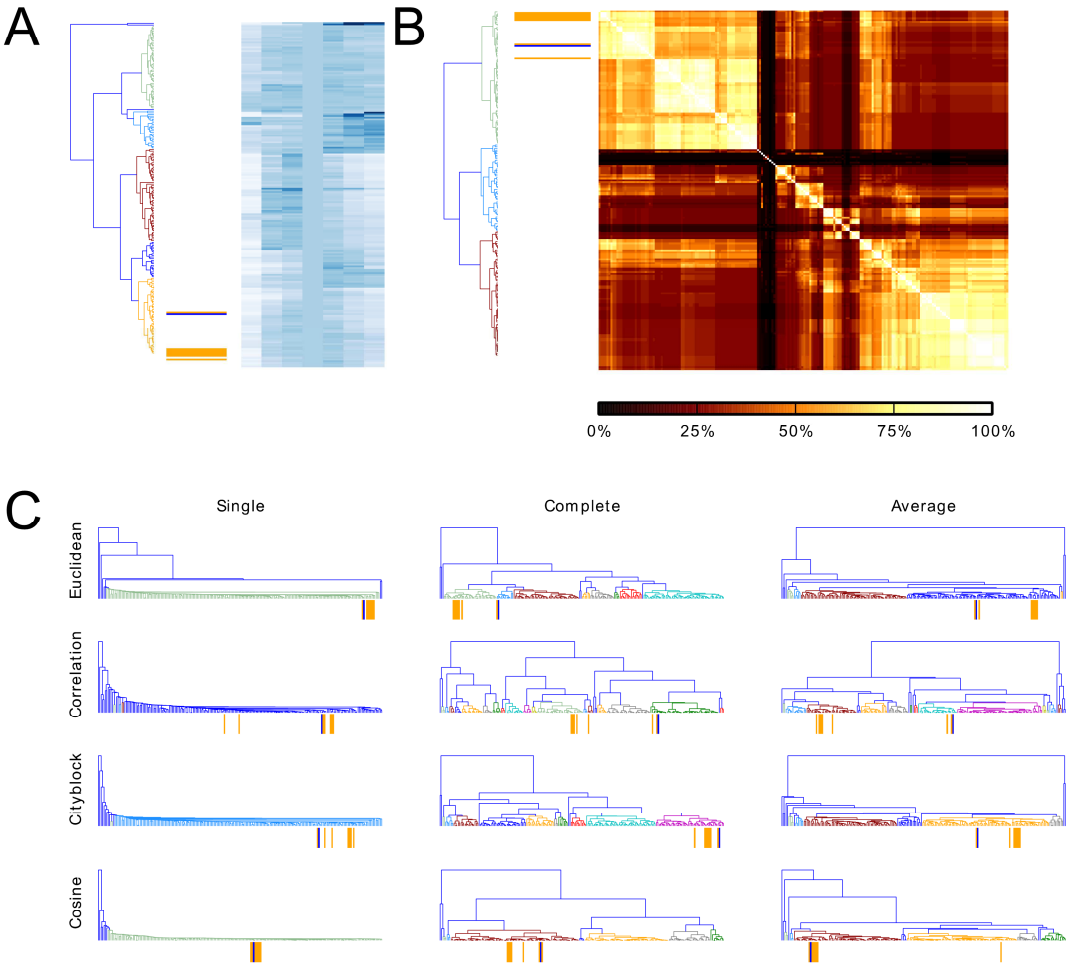

Figure 4.6: Ensemble Clustering on Phosphoproteomic Data.
(A) A single clustering solution showing known interactors with EGFR (orange bars) and PDLIM1 (blue bar) co-clustering in the phosphoproteomic data (blue heatmap). (B) The co-occurrence matrix heatmap demonstrating robust clustering of interactors with EGFR. The known interactors with EGFR (orange bars) and PDLIM1 (blue bar) are found in a single, robust cluster (upper left). (C) A subset of clustering results across multiple distance metrics and clustering algorithms. Under the dendrogram, known interactors with EGFR are marked with orange bars and PDLIM1 is marked with a blue bar.

4.6.2 Ensembles for Robust Clustering Results

Ensemble clustering can reveal unexpected results. When an algorithm with limited capabilities is combined into an ensemble, the clustering result of the ensemble can have new capabilities. For example, although a k-means algorithm can only identify spherical clusters, when used as part of an ensemble with majority voting, the ensemble can identify half-moons and spirals (140). This is possible because signals from relatively weak relationships in each clustering result are combined to improve the strength of the pairwise relationship between points in the ensemble clustering results. Ensemble clustering can assess the impact of perturbations to clustering parameters on the clustering results, revealing when transformations to the data can have a larger impact on a clustering solution than the algorithm or distance metric does (111).

As an example of ensemble clustering, we describe a subset of the results of an ensemble approach that we used to cluster the dynamics of tyrosine phosphorylation in the epidermal growth factor receptor (EGFR) network measured by Wolf-Yadlin and colleagues (92). From this analysis, we identified previously unknown protein interactions (87). We show a subset of the full analysis to illustrate this process (Figure 4.6). PDLIM1, a protein not previously reported as part of the EGFR network, had similar phosphotyrosine dynamics with many other proteins known to directly interact with phosphorylated tyrosines residues of EGFR (Y1197 and Y1192) (Figure 4.6; blue bar, PDLIM1; orange bar, EGFR interactors).To identify a more robust representation of the clustering behavior of the system, we generated an ensemble of clusters by varying distance metrics and clustering algorithms. Across the ensemble, known interactions had a much higher tendency to cluster together than with non-interactors. A visualization of the ensemble results – a co-occurrence matrix – places the interactors of these EGFR

phosphotyrosine sites in one of the clusters (Figure 4.6B, upper left and orange bars), along with the potential interactor PDLIM1 (blue bar). On the basis of these results, we experimentally tested and validated PDLIM1 as a protein that interacted with EGFR (87). It is important to note that, in many of the ensemble clustering results, PDLIM1 did not cluster with all known EGFR phosphorylated at Y1197 and Y1192. Rather, PDLIM1 tended to cluster with a subset of known interactors. Furthermore, in many ensemble clustering results, the known interactors of EGFR did not all cluster together (Figure 4.6C). This demonstrates the value of robust clustering since a single clustering solution might have missed this important relationship (87).

In-Depth Review Area	References
Specific clustering algorithms, their tradeoffs, and how they function	(81, 82, 150–152)
Analysis of the effects of different distance metrics on clustering gene expression data	(153, 154)
Practical and mathematical implications of high-dimensionality on clustering.	(94, 95)
A thorough review of validation metrics	(114, 117, 124, 155)
The most common multiple hypothesis correction (MHC) procedures including Bonferroni correction and False Discovery Rate (FDR) correction.	(135, 138)
The effects of specific distances on data clustering for lower-dimensional spaces	(156)
The effects of specific distances on data clustering for high-dimensional spaces	(94, 157)
Ensembles of some algorithms incompatible with high-dimensional data can be useful on higher-dimensional data, even when a single clustering solution is uninformative.	(102, 103)
A more in-depth analysis of ensembles, including evaluating the results of multiple clustering runs and determining consensus	(139, 148)

Table 4.4: Summary of In-Depth Review Articles.
A collection of reviews for more in depth coverage of each topic.

4.7 Conclusion

Clustering biological data involves a number of choices, many of which are critical to meaningful results. Evaluation of a dataset should be performed to make sure it has a sufficient number of observations (data points) that and the dimensionality of those observations informs subsequent clustering choices. For each new dataset, dimensionality and any transformations applied should influence the choice of appropriate distance metrics and algorithms for clustering. Data sets of more than 10 dimensions often behave unexpectedly, and clustering can produce meaningless results. Using only a single clustering result from any dataset can lead to wasted time and resources resulting from erroneous hypothesis testing. When possible, noise and variance should be accounted for in the clustering method directly rather than simply taking averages at each data point. Once clustering results are obtained, their validity should be evaluated using the appropriate metrics. The statistical significance of clustering results should also be evaluated, and multiple hypothesis correction should be applied when necessary. For robust results, ensemble clustering over a range of distance metrics, transformations, and other clustering parameters is effective. Following these steps will result in obtaining robust and reliable results from clustering, and will provide a basis for solid generation of testable hypotheses.

References

1. Liu,B.A., Shah,E., Jablonowski,K., Stergachis,A., Engelmann,B. and Nash,P.D. (2011) The SH2 Domain-Containing Proteins in 21 Species Establish the Provenance and Scope of Phosphotyrosine Signaling in Eukaryotes. *Sci. Signal.*, **4**, ra83-ra83.

2. Pawson,T. (2004) Specificity in Signal Transduction: From Phosphotyrosine-SH2 Domain Interactions to Complex Cellular Systems. *Cell*, **116**, 191–203.

3. Pincus,D., Letunic,I., Bork,P. and Lim,W. a (2008) Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 9680–9684.

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

5. Sen,B. and Johnson,F.M. (2011) Regulation of SRC family kinases in human cancers. *J. Signal Transduct.*, **2011**, 865819.

6. Li,L., Wu,C., Huang,H., Zhang,K., Gan,J. and Li,S.S.C. (2008) Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. *Nucleic Acids Res.*, **36**, 3263–3273.

7. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: Proteome-wide prediction of cell signalling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.

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

9. Yoon,S.O., Soltoff,S.P. and Chao,M. V (1997) A Dominant Role of the Juxtamembrane Region of the TrkA Nerve Growth Factor Receptor during Neuronal Cell Differentiation. *J. Biol. Chem.*, **272**, 23231–23238.

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

11. Kaushansky,A., Gordus,A., Budnik,B.A., Lane,W.S., Rush,J. and MacBeath,G. (2008) System-wide Investigation of ErbB4 Reveals 19 Sites of Tyr Phosphorylation that Are Unusually Selective in Their Recruitment Properties. *Chem. Biol.*, **15**, 808–817.

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

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

14. Koytiger,G., Kaushansky,A., Gordus,A., Rush,J., Sorger,P.K. and MacBeath,G. (2013) Phosphotyrosine Signaling Proteins that Drive Oncogenesis Tend to be Highly Interconnected. *Mol. Cell. Proteomics*, **12**, 1204–1213.

15. Engelmann,B.W., Kim,Y., Wang,M., Peters,B., Rock,R.S. and Nash,P.D. (2014) The Development and Application of a Quantitative Peptide Microarray Based Approach to Protein Interaction Domain Specificity Space. *Mol. Cell. Proteomics*, **13**, 3647–3662.

16. Hause,R.J., Leung,K.K., Barkinge,J.L., Ciaccio,M.F., Chuu,C. pin and Jones,R.B. (2012) Comprehensive Binary Interaction Mapping of SH2 Domains via Fluorescence Polarization Reveals Novel Functional Diversification of ErbB Receptors. *PLoS One*, **7**.

17. Leung,K.K., Hause,R.J., Barkinge,J.L., Ciaccio,M.F., Chuu,C.-P. and Jones,R.B. (2014) Enhanced Prediction of Src Homology 2 (SH2) Domain Binding Potentials Using a Fluorescence Polarization-derived c-Met, c-Kit, ErbB, and Androgen Receptor Interactome. *Mol. Cell. Proteomics*, **13**, 1705–1723.

18. Tinti,M., Kiemer,L., Costa,S., Miller,M.L., Sacco,F., Olsen,J. V., Carducci,M., Paoluzi,S., Langone,F., Workman,C.T., *et al.* (2013) The SH2 Domain Interaction Landscape. *Cell Rep.*, **3**, 1293–1305.

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

20. Gordus,A. and MacBeath,G. (2006) Circumventing the problems caused by protein diversity in microarrays: Implications for protein interaction networks. *J. Am. Chem. Soc.*, **128**, 13668–13669.

21. Kundu,K., Costa,F., Huber,M., Reth,M. and Backofen,R. (2013) Semi-Supervised Prediction of SH2-Peptide Interactions from Imbalanced High-Throughput Data. *PLoS One*, **8**.

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

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

24. Mazerolle,M. (2004) 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*.

25. Blinov,M.L., Faeder,J.R., Goldstein,B. and Hlavacek,W.S. (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Biosystems.*, **83**, 136–51.

26. Chen,W.W., Schoeberl,B., Jasper,P.J., Niepel,M., Nielsen,U.B., Lauffenburger,D.A. and

Sorger,P.K. (2009) Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.*, **5**.

27. Ronan,T., Macdonald-Obermann,J.L., Huelsmann,L., Bessman,N.J., Naegle,K.M. and Pike,L.J. (2016) Different Epidermal Growth Factor Receptor (EGFR) agonists produce unique signatures for the recruitment of downstream signaling proteins. *J. Biol. Chem.*, **291**, 5528–5540.

28. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: Proteome-wide prediction of cell signalling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.

29. Kaneko,T., Sidhu,S.S. and Li,S.S.C. (2011) Evolving specificity from variability for protein interaction domains. *Trends Biochem. Sci.*, **36**, 183–190.

30. Li,S.S.-C. (2005) Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. *Biochem. J.*, **390**, 641–653.

31. Li,L., Zhao,B., Du,J., Zhang,K., Ling,C.X. and LiShawn,S.S.C. (2011) Dompep-a general method for predicting modular domain-mediated protein-protein interactions. *PLoS One*, **6**.

32. Liu,B.A., Engelmann,B.W. and Nash,P.D. (2012) The language of SH2 domain interactions defines phosphotyrosine-mediated signal transduction. *FEBS Lett.*, **586**, 2597–2605.

33. Ladbury,J.E., Lemmon,M.A., Zhou,M., Green,J., Botfield,M.C. and Schlessinger,J. (1995) Measurement of the binding of tyrosyl phosphopeptides to SH2 domains: a reappraisal. *Proc Natl Acad Sci U S A*, **92**, 3199–3203.

34. Miller,M.L., Jensen,L.J., Diella,F., Jørgensen,C., Tinti,M., Li,L., Hsiung,M., Parker,S. a, Bordeaux,J., Sicheritz-Ponten,T., *et al.* (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.*, **1**, ra2.

35. Creixell,P., Palmeri,A., Miller,C.J., Lou,H.J., Santini,C.C., Nielsen,M., Turk,B.E. and Linding,R. (2015) Unmasking Determinants of Specificity in the Human Kinome. *Cell*, **163**, 187–201.

36. Alquraishi,M., Koytiger,G., Jenney,A., Macbeath,G. and Sorger,P.K. (2014) A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. *Nat. Genet.*, **46**, 1363–1371.

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

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

39. Clayton,A.H.A., Walker,F., Orchard,S.G., Henderson,C., Fuchs,D., Rothacker,J., Nice,E.C. and Burgess,A.W. (2005) Ligand-induced dimer-tetramer transition during the activation of the cell surface epidermal growth factor receptor-A multidimensional microscopy analysis. *J. Biol. Chem.*, **280**, 30392–30399.

40. Hofman,E.G., Bader,A.N., Voortman,J., Van den Heuvel,D.J., Sigismund,S., Verkleij,A.J., Gerritsen,H.C. and Van Bergen En Henegouwen,P.M.P. (2010) Ligand-induced epidermal growth factor receptor (EGFR) oligomerization is kinase-dependent and enhances internalization. *J. Biol. Chem.*, **285**, 39481–39489.

41. Moriki,T., Maruyama,H. and Maruyama,I.N. (2001) Activation of preformed EGF receptor dimers by ligand-induced rotation of the transmembrane domain11Edited by B. Holland. *J. Mol. Biol.*, **311**, 1011–1026.

42. Saffarian,S., Li,Y., Elson,E.L. and Pike,L.J. (2007) Oligomerization of the EGF receptor investigated by live cell fluorescence intensity distribution analysis. *Biophys J*, **93**, 1021–1031.

43. Poirier,M.G., Eroglu,S. and Marko,J.F. (2002) Ligand-independent Dimer Formation of EGFR is a step Separable from Ligand-induced EGFR signaling. *Mol. Biol. Cell*, **13**, 2170–2179.

44. Garrett,T.P.J., McKern,N.M., Lou,M., Elleman,T.C., Adams,T.E., Lovrecz,G.O., Zhu,H.-J., Walker,F., Frenkel,M.J., Hoyne,P.A., *et al.* (2002) Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor alpha. *Cell*, **110**, 763–73.

45. Ogiso,H., Ishitani,R., Nureki,O., Fukai,S., Yamanaka,M., Kim,J.H., Saito,K., Sakamoto,A., Inoue,M., Shirouzu,M., *et al.* (2002) Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell*, **110**, 775–787.

46. Zhang,X., Gureasko,J., Shen,K., Cole,P. a and Kuriyan,J. (2006) An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell (Cambridge, MA, U.S.)*, **125**, 1137–1149.

47. Citri,A. and Yarden,Y. (2006) EGF–ERBB signalling: towards the systems level. *Nat. Rev. Mol. Cell Biol.*, **7**, 505–516.

48. Levkowitz,G., Waterman,H., Ettenberg,S.A., Katz,M., Tsygankov,A.Y., Alroy,I., Lavi,S., Iwai,K., Reiss,Y., Ciechanover,A., *et al.* (1999) Ubiquitin ligase activity and tyrosine phosphorylation underlie suppression of growth factor signaling by c-Cbl/Sli-1. *Mol. Cell*, **4**, 1029–1040.

49. Holgado-Madruga,M., Emlet,D.R., Moscatello,D.K., Godwin,A.K. and Wong,A.J. (1996) A Grb2-associated docking protein in EGF- and insulin-receptor signalling. *Nature*, **379**, 560–564.

50. Gu,H. and Neel,B.G. (2003) The 'Gab' in signal transduction. *Trends Cell Biol.*, **13**, 122–130.

51. Rodrigues,G.A., Falasca,M., Zhang,Z., Ong,S.H. and Schlessinger,J. (2000) A novel positive feedback loop mediated by the docking protein Gab1 and phosphatidylinositol 3-kinase in epidermal growth factor receptor signaling. *Mol. Cell. Biol.*, **20**, 1448–1459.

52. Simister,P.C. and Feller,S.M. (2012) Order and disorder in large multi-site docking proteins of the Gab family--implications for signalling complex formation and inhibitor design strategies. *Mol Biosyst*, **8**, 33–46.

53. Wang,W., Xu,S., Yin,M. and Jin,Z.G. (2015) Essential roles of Gab1 tyrosine phosphorylation in growth factor-mediated signaling and angiogenesis. *Int. J. Cardiol.*, **181**, 180–184.

54. Saito,T., Okada,S., Ohshima,K., Yamada,E., Sato,M., Uehara,Y., Shimizu,H., Pessin,J.E. and Mori,M. (2004) Differential activation of Epidermal Growth Factor (EGF) receptor downstream signaling pathways by betacellulin and EGF. *Endocrinology*, **145**, 4232–4243.

55. Shin,H.S., Lee,H.J., Nishida,M., Lee,M.S., Tamura,R., Yamashita,S., Matsuzawa,Y., Lee,I.K. and Koh,G.Y. (2003) Betacellulin and amphiregulin induce upregulation of cyclin D1 and DNA synthesis activity through differential signaling pathways in vascular smooth muscle cells. *Circ. Res.*, **93**, 302–310.

56. Streicher,K.L., Willmarth,N.E., Garcia,J., Boerner,J.L., Dewey,T.G. and Ethier,S.P. (2007) Activation of a Nuclear Factor KB/Interleukin-1 Positive Feedback Loop by Amphiregulin in Human Breast Cancer Cells. *Mol Cancer Res*, **5**, 847–62.

57. Wilson,K.J., Mill,C., Lambert,S., Buchman,J., Wilson,T.R., Hernandez-Gordillo,V., Gallo,R.M., Ades,L.M.C., Settleman,J. and Riese,D.J. (2012) EGFR ligands exhibit functional differences in models of paracrine and autocrine signaling. *Growth Factors*, **30**, 107–16.

58. Macdonald-Obermann,J.L., Adak,S., Landgraf,R., Piwnica-Worms,D. and Pike,L.J. (2013) Dynamic analysis of the epidermal growth factor (EGF) receptor-ErbB2-ErbB3 protein network by luciferase fragment complementation imaging. *J. Biol. Chem.*, **288**, 30773–30784.

59. Macdonald-Obermann,J.L. and Pike,L.J. (2014) Different epidermal growth factor (EGF) receptor ligands show distinct kinetics and biased or partial agonism for homodimer and heterodimer formation. *J. Biol. Chem.*, **289**, 26178–26188.

60. Yang,K.S., Ilagan,M.X.G., Piwnica-Worms,D. and Pike,L.J. (2009) Luciferase fragment complementation imaging of conformational changes in the epidermal growth factor receptor. *J. Biol. Chem.*, **284**, 7474–7482.

61. Pedregosa,F. and Varoquaux,G. (2011) Scikit-learn: Machine learning in Python.

62. Luker,K.E., Smith,M.C.P., Luker,G.D., Gammon,S.T., Piwnica-Worms,H. and Piwnica-Worms,D. (2004) Kinetics of regulated protein-protein interactions revealed with firefly luciferase complementation imaging in cells and living animals. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 12288–12293.

63. Wilson,K.J., Gilmore,J.L., Foley,J., Lemmon,M.A. and Riese,D.J. (2009) Functional selectivity of EGF family peptide growth factors: Implications for cancer. *Pharmacol. Ther.*, **122**, 1–8.

64. Zheng,Y., Zhang,C., Croucher,D.R., Soliman,M.A., St-Denis,N., Pasculescu,A., Taylor,L., Tate,S.A., Hardy,W.R., Colwill,K., *et al.* (2013) Temporal regulation of EGF signalling networks by the scaffold protein Shc1. *Nature*, **499**, 166–171.

65. Beguinot,L., Lyall,R.M., Willingham,M.C. and Pastan,I. (1984) Down-regulation of the epidermal growth factor receptor in KB cells is due to receptor internalization and subsequent degradation in lysosomes. *Proc. Natl. Acad. Sci. U. S. A.*, **81**, 2384–2388.

66. Jeppe Knudsen,S.L., Wai Mac,A.S., Henriksen,L., Deurs,B. van and Grøvdal,L.M. (2014) EGFR signaling patterns are regulated by its different ligands. *Growth Factors*, **32**, 155–163.

67. Gilmore,J.L., Scott,J.A., Bouizar,Z., Robling,A., Pitfield,S.E., Riese,D.J. and Foley,J. (2008) Amphiregulin-EGFR signaling regulates PTHrP gene expression in breast cancer cells. *Breast Cancer Res. Treat.*, **110**, 493–505.

68. Curran,T.G., Zhang,Y., Ma,D.J., Sarkaria,J.N. and White,F.M. (2015) MARQUIS: A multiplex method for absolute quantification of peptides and posttranslational modifications. *Nat. Commun.*, **6**, 5924.

69. Guo,L., Kozlosky,C.J., Ericsson,L.H., Daniel,T.O., Cerretti,D.P. and Johnson,R.S. (2003) Studies of ligand-induced site-specific phosphorylation of epidermal growth factor receptor. *J. Am. Soc. Mass Spectrom.*, **14**, 1022–1031.

70. Grayson,L.S., Hansbrough,J.F., Zapata-Sirvent,R.L., Dore,C.A., Morgan,J.L. and Nicolson,M.A. (1993) Quantitation of cytokine levels in skin graft donor site wound fluid. *Burns*, **19**, 401–405.

71. Lemos-González,Y., Rodríguez-Berrocal,F.J., Cordero,O.J., Gómez,C. and Páez de la Cadena,M. (2007) Alteration of the serum levels of the epidermal growth factor receptor and its ligands in patients with non-small cell lung cancer and head and neck carcinoma. *Br. J. Cancer*, **96**, 1569–1578.

72. Marti,U., Burwen,S.J. and Jones,A.L. (1989) Biological effects of epidermal growth factor, with emphasis on the gastrointestinal tract and liver: An update. *Hepatology*, **9**, 126–138.

73. Dengjel,J., Akimov,V., Olsen,J. V, Bunkenborg,J., Mann,M., Blagoev,B. and Andersen,J.S. (2007) Quantitative proteomic assessment of very early cellular signaling events. *Nat. Biotechnol.*, **25**, 566–568.

74. Olsen,J. V., Blagoev,B., Gnad,F., Macek,B., Kumar,C., Mortensen,P. and Mann,M. (2006) Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell*, **127**, 635–648.

75. Zhang,Y., Wolf-yadlin,A., Ross,P.L., Pappin,D.J., Rush,J., Lauffenburger,D.A. and White,F.M. (2005) Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol. Cell. Proteomics*, **4**, 1240–50.

76. Kyono,W.T., Jong,R. de, Park,R.K., Liu,Y., Heisterkamp,N., Groffen,J. and Durden,D.L. (1998) Differential Interaction of Crkl with Cbl or C3G, Hef-1, and γ Subunit Immunoreceptor Tyrosine-Based Activation Motif in Signaling of Myeloid High Affinity Fc Receptor for IgG (FcγRI). *J. Immunol.*, **161**, 5555–5563.

77. Sattler,M., Salgia,R., Shrikhande,G., Verma,S., Uemura,N., Law,S.F., Golemis,E.A. and Griffin,J.D. (1997) Differential signaling after ??1 integrin ligation is mediated through binding of CRKL to p120(CBL) and p110(HEF1). *J. Biol. Chem.*, **272**, 14320–14326.

78. Ingham,R.J., Holgado-Madruga,M., Siu,C., Wong,A.J. and Gold,M.R. (1998) The Gab1 protein is a docking site for multiple proteins involved in signaling by the B cell antigen receptor. *J Biol Chem*, **273**, 30630–30637.

79. Lecoq-Lafon,C., Verdier,F., Fichelson,S., Chrétien,S., Gisselbrecht,S., Lacombe,C. and Mayeux,P. (1999) Erythropoietin induces the tyrosine phosphorylation of GAB1 and its association with SHC, SHP2, SHIP, and phosphatidylinositol 3-kinase. *Blood*, **93**, 2578–2585.

80. Jain,A., Murty,M. and Flynn,P. (1999) Data clustering: a review. *ACM Comput. Surv.*

81. Xu,R. and Wunsch,D.C. (2010) Clustering algorithms in biomedical research: A review. *IEEE Rev. Biomed. Eng.*, **3**, 120–154.

82. Andreopoulos,B., An,A., Wang,X. and Schroeder,M. (2009) A roadmap of clustering algorithms: Finding a match for a biomedical application. *Brief. Bioinform.*, **10**, 297–314.

83. Sørlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 10869–10874.

84. Lapointe,J., Li,C., Higgins,J.P., van de Rijn,M., Bair,E., Montgomery,K., Ferrari,M., Egevad,L., Rayford,W., Bergerheim,U., *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 811–816.

85. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A., *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.

86. Verhaak,R.G.W., Hoadley,K.A., Purdom,E., Wang,V., Qi,Y., Wilkerson,M.D., Miller,C.R., Ding,L., Golub,T., Mesirov,J.P., *et al.* (2010) Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.

87. Naegle,K.M., White,F.M., Lauffenburger,D.A. and Yaffe,M.B. (2012) Robust co-regulation of tyrosine phosphorylation sites on proteins reveals novel protein interactions. *Mol. Biosyst.*, **8**, 2771.

88. Wolf-Yadlin,A., Kumar,N., Zhang,Y., Hautaniemi,S., Zaman,M., Kim,H.-D., Grantcharova,V., Lauffenburger,D.A. and White,F.M. (2006) Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol. Syst. Biol.*, **2**, 54.

89. Jain,M., Nilsson,R., Sharma,S., Madhusudhan,N., Kitami,T., Souza,A.L., Kafri,R., Kirschner,M.W., Clish,C.B. and Mootha,V.K. (2012) Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science*, **336**, 1040–4.

90. Miller,M.A., Barkal,L., Jeng,K., Herrlich,A., Moss,M., Griffith,L.G. and Lauffenburger,D.A. (2011) Proteolytic Activity Matrix Analysis (PrAMA) for simultaneous determination of multiple protease activities. *Integr. Biol. (Camb).*, **3**, 422–438.

91. Mootha,V.K., Lindgren,C.M., Eriksson,K.-F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstråle,M., Laurila,E., *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

92. Wolf-Yadlin,A., Hautaniemi,S., Lauffenburger,D.A. and White,F.M. (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc Natl Acad Sci U S A*, **104**, 5860–5865.

93. Köppen,M. (2000) The curse of dimensionality. *5th Online World Conf. Soft Comput. Ind. Appl.*, **1**, 4–8.

94. Steinbach,M., Ertoz,L. and Kumar,V. (2004) The Challenges of Clustering High Dimensional Data. *New Dir. Stat. Phys.*, 10.1007/978-3-662-08968-2_16.

95. Zimek,A., Schubert,E. and Kriegel,H.P. (2012) A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min.*, **5**, 363–387.

96. Beyer,K., Goldstein,J., Ramakrishnan,R. and Shaft,U. (1999) When is ‘nearest neighbor’ meaningful? In *International Conference on Database Theory*.pp. 217–235.

97. Aggarwal,C.C., Wolf,J.L., Yu,P.S., Procopiuc,C. and Park,J.S. (1999) Fast algorithms for projected clustering. *ACM SIGMOD Rec.*, **28**, 61–72.

98. Pestov,V. (2000) On the geometry of similarity search: dimensionality curse and concentration of measure. *Inf. Process. Lett.*, **73**, 47–51.

99. Hinneburg,A., Aggarwal,C.C. and Keim,D. a (2000) What is the Nearest Neighbor in High Dimensional Spaces? *Proc. 26th VLDB Conf.*

100. Hinneburg,A. and Keim,D. a (1999) Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. *Int. Conf. Very Large Databases*, 1-55860-615-7.

101. Ester,M., Kriegel,H.P., Sander,J. and Xu,X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Second Int. Conf. Knowl. Discov. Data Min.*

102. Kim,E.-Y., Kim,S.-Y., Ashlock,D. and Nam,D. (2009) MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics*, **10**, 260.

103. Amaratunga,D., Cabrera,J. and Lee,Y.-S. (2015) Resampling-Based Similarity Measures for High-Dimensional Data. *J. Comput. Biol.*, **22**, 54–62.

104. Byeon,B. and Rasheed,K. (2008) Simultaneously removing noise and selecting relevant features for high dimensional noisy data. *Proc. - 7th Int. Conf. Mach. Learn. Appl. ICMLA 2008*, 10.1109/ICMLA.2008.87.

105. Hou,C., Nie,F., Yi,D. and Tao,D. (2014) Discriminative Embedded Clustering: A Framework for Grouping High-Dimensional Data. *IEEE Trans. Neural Networks Learn. Syst.*, 10.1109/TNNLS.2014.2337335.

106. Kriegel,H.-P., Kröger,P. and Zimek,A. (2009) Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. *ACM Trans. Knowl. Discov. Data*, **3**, 1–58.

107. Tang,J., Alelyani,S. and Liu,H. (2014) Feature Selection for Classification: A Review. In Aggarwal,C.C.. (ed), *Data Classification: Algorithms and Applications*. Chapman and Hall/CRC, pp. 37–64.

108. Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.

109. Parsons,L., Haque,E. and Liu,H. (2004) Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explor. Newsl.*, **6**, 90–105.

110. van den Berg,R.A., Hoefsloot,H.C.J., Westerhuis,J.A., Smilde,A.K. and van der Werf,M.J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **7**, 142.

111. Naegle,K.M., Welsch,R.E., Yaffe,M.B., White,F.M. and Lauffenburger,D.A. (2011) MCAM: Multiple clustering analysis methodology for deriving hypotheses and insights from high-throughput proteomic datasets. *PLoS Comput. Biol.*, **7**.

112. Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 2907–2912.

113. Sharma,S. (1996) Applied Multivariate Techniques.

114. Halkidi,M., Batistakis,Y. and Vazirgiannis,M. (2001) On clustering validation techniques. *J. Intell. Inf. Syst.*, **17**, 107–145.

115. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

116. Ding,C. and He,X. (2004) K-nearest-neighbor consistency in data clustering: Incorporating local information into global optimization. *Proc. 2004 ACM Symp. Appl. Comput.*

117. Handl,J., Knowles,J. and Kell,D.B. (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

118. Calinski,T. and Harabasz,J. (1974) A Dendrite Method for Cluster Analysis. *Commun. Stat. - Simul. Comput.*, **3**, 1–27.

119. Dunn,J.C. (1974) Well-Separated Clusters and Optimal Fuzzy Partitions. *J. Cybern.*, **4**, 95–104.

120. Davies,D.L. and Bouldin,D.W. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1**, 224–227.

121. Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

122. Tibshirani,R., Walther,G. and Hastie,T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, **63**, 411–423.

123. Rand,W.M. (1971) Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

124. Desgraupes,B. (2013) Clustering Indices. *CRAN Packag.*

125. Sloutsky,R., Jimenez,N., Swamidass,S.J. and Naegle,K.M. (2013) Accounting for noise when clustering biological data. *Brief. Bioinform.*, **14**, 423–436.

126. Ben-Hur,A., Elisseeff,A. and Guyon,I. (2002) A stability based method for discovering structure in clustered data. *Pac. Symp. Biocomput.*, **17**, 6–17.

127. Yeung,K.Y., Medvedovic,M. and Bumgarner,R.E. (2003) Clustering gene-expression data with repeated measurements. *Genome Biol.*, **4**, R34.

128. Kuncheva,L.I. and Vetrov,D.P. (2006) Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**, 1798–1808.

129. Naegle,K., Gough,N.R. and Yaffe,M.B. (2015) Criteria for biological reproducibility : What does ‘n’ mean ? *Sci. Signal.*, **8**, 2–5.

130. Ryder,E.F. and Robakiewicz,P. (2001) Statistics for the molecular biologist: group comparisons. Ausubel,F.M. (ed).

131. Kerr,M.K. and Churchill,G. a (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 8961–8965.

132. Medvedovic,M., Yeung,K.Y. and Bumgarner,R.E. (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**, 1222–1232.

133. Cooke,E.J., Savage,R.S., Kirk,P., Darkins,R. and Wild,D.L. (2011) Bayesian hierarchical

clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinformatics*, **12**, 399.

134. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

135. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

136. Tsai,C.-A., Hsueh,H. and Chen,J.J. (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, **59**, 1071–1081.

137. Pounds,S.B. (2006) Estimation and control of multiple testing error rates for microarray studies. *Brief. Bioinform.*, **7**, 25–36.

138. Aickin,M. and Gensler,H. (1996) Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *Am. J. Public Health*, **86**, 726–728.

139. Monti,S., Tamayo,P., Mesirov,J. and Golub,T. (2003) Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.

140. Fred,A.L.N. and Jain,A.K. (2002) Data clustering using evidence accumulation. *Proc. 16th Int. Conf. Pattern Recognit.*, **4**, 276–280.

141. Dougherty,E.R., Barrera,J., Brun,M., Kim,S., Cesar,R.M., Chen,Y., Bittner,M. and Trent,J.M. (2002) Inference from clustering with application to gene-expression microarrays. *J. Comput. Biol.*, **9**, 105–126.

142. Agrawal,R., Gehrke,J., Gunopulos,D. and Raghavan,P. (1998) Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD Rec.*, **27**, 94–105.

143. Goil,S., Nagesh,H. and Choudhary,A. (1999) MAFIA: Efficient and scalable subspace clustering for very large data sets. *Tech. Rep. Number CPDC-TR-9906-019, Cent. Parallel Distrib. Comput. Northwest. Univ.*

144. Zimek,A. and Vreeken,J. (2013) The blind men and the elephant: on meeting the problem of multiple truths in data from clustering and pattern mining perspectives. *Mach. Learn.*, 10.1007/s10994-013-5334-y.

145. Bellec,P., Rosa-Neto,P., Lyttelton,O.C., Benali,H. and Evans,A.C. (2010) Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *Neuroimage*, **51**, 1126–1139.

146. Levine,E. and Domany,E. (2001) Resampling method for unsupervised estimation of cluster validity. *Neural Comput.*, **13**, 2573–2593.

147. Lange,T., Roth,V., Braun,M.L. and Buhmann,J.M. (2004) Stability-based validation of clustering solutions. *Neural Comput.*, **16**, 1299–1323.

148. Strehl,A. and Ghosh,J. (2002) Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.*, **3**, 583–617.

149. Mimaroglu,S. and Yagci,M. (2012) CLICOM: Cliques for combining multiple clusterings. In *Expert Systems with Applications*.Vol. 39, pp. 1889–1901.

150. Jain,A.K., Murty,M.N. and Flynn,P.J. (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.

151. Jain,A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.*, **31**, 651–666.

152. Berkhin,P. (2002) Survey of Clustering Data Mining Techniques. *Techincal Report, Accrue Softw*.

153. Jaskowiak,P.A., Campello,R.J.G.B. and Costa,I.G. (2014) On the selection of appropriate distances for gene expression data clustering. *BMC Bioinformatics*, **15 Suppl 2**, S2.

154. Giancarlo,R., Lo Bosco,G. and Pinello,L. (2010) Distance Functions, Clustering Algorithms and Microarray Data Analysis. In *Learning and Intelligent Optimization*.Vol. 6073, pp. 125–138.

155. Liu,Y., Li,Z., Xiong,H., Gao,X. and Wu,J. (2013) Understanding of Internal Clustering Validation Measures. 10.1109/ICDM.2010.35.

156. Mohamad,I. Bin and Usman,D. (2013) Standardization and its effects on K-means clustering algorithm. *Res. J. Appl. Sci. Eng. Technol.*, **6**, 3299–3303.

157. Aggarwal,C.C., Hinneburg,A. and Keim,D.A. (2001) On the surprising behavior of distance metrics in high dimensional space. *Database Theory – ICDT 2001*, 10.1007/3-540-44503-X_27.