INTEGRATION OF PROTEIN BINDING INTERFACES AND ABUNDANCE DATA REVEALS EVOLUTIONARY PRESSURES IN PROTEIN NETWORKS

by
David Orestis Holland

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland
January, 2018

Abstract

Networks of protein-protein interactions have received considerable interest in the past two decades for their insights about protein function and evolution. Traditionally, these networks only map the functional partners of proteins; they lack further levels of data such as binding affinity, allosteric regulation, competitive vs noncompetitive binding, and protein abundance. Recent experiments have made such data on a network-wide scale available, and in this thesis I integrate two extra layers of data in particular: the binding sites that proteins use to interact with their partners, and the abundance or "copy numbers" of the proteins. By analyzing the networks for the clathrin-mediated endocytosis (CME) system in yeast and the ErbB signaling pathway in humans, I find that this extra data reveals new insights about the evolution of protein networks. The structure of the binding site or *interface* interaction network (IIN) is optimized to allow higher binding specificity; that is, a high gap in strength between functional binding and nonfunctional mis-binding. This strongly implies that mis-binding is an evolutionary error-load constraint shaping protein network structure. Another method to limit mis-binding is to balance protein copy numbers so that there are no "leftover" proteins available for mis-binding. By developing a new method to quantify balance in IINs, I show that

the CME network is significantly balanced when compared to randomly sampled sets of copy numbers. Furthermore, IINs with a biologically realistic structure produce less mis-binding under balanced concentrations, when compared to random networks, but more mis-binding under unbalanced concentrations. This implies strong pressure for copy number balance and that any imbalance should occur for functional reasons. I thus explore some functional consequences of imbalance by constructing dynamic models of two poorly balanced subnetworks of the larger CME network. In general, I find that balanced copy numbers provide higher protein complex yield (number of complete complexes), but imbalance may allow cells to "bottleneck" a functional process, effectively turning complex formation on or off via spatial localization of subunits. Finally, I find that strongly binding proteins are more likely to be balanced, as these "sticky" proteins would be more likely to engage in mid-binding otherwise.

Dissertation Advisor: Dr. Margaret E. Johnson

Committee Members: Dr. Joel S. Bader, Dr. Elijah Roberts, Dr. Feilim Mac Gabhann

Acknowledgements

I start by thanking all the friends who kept me sane during my 6.5 years completing this thesis. Locally, these include Matthew Kerr, Brian Hu, Irma Zhang, Ernest Le, Lindsay Wendell, Melanie Zile, Joshua Banks, Michael Lysak, Sam Sklar, Roger Messick, Molly Casewit, Janaka Senarathna, and many others. From further away: Sarah Seid, David Rhodenbaugh, Peter Trauernitch, Justin McKee, Emily Guinivan, and my siblings Marilyn Holland and John Holland Jr. From my lab I'd like to thank Osman Yogurtcu, Dariush Mohammadyani, Sewwandi Rathnayake, Matthew Varga, Ben Shapiro, Patrick Xue, and Daisy Duan.

I thank my advisor, Dr. Margaret Johnson, as well as all the other professors who helped me along this path including Drs. Joel Bader, Rebecca Schulman, John Wierman, Pablo Iglesias, Andre Levchenko, Elijah Roberts, Rachel Karchin, Feilim Mac Gabhann, and Daniel Naiman. I thank Hong Lan and Nancy Foltz for handling administration. I also thank whomever it was that suggested, when my last advisor left Hopkins, to look for a lab in the Biophysics department to join.

Finally, I thank God for the ability and opportunity to complete this work.

Dedication

This dissertation is dedicated to my late father, Dr. John Francis Holland (Feb. 25, 1945 – Oct 31, 2012).

"Say not in grief 'he is no more' but live in thankfulness that he was."

~Hebrew proverb

Contents

List of Tables

List of Figures

Chapter 1. Introduction

This chapter will provide an overview of the topics to be discussed in this thesis: copy number balance, dosage sensitivity, protein misinteractions, and protein network science including interface-interaction networks.

1.1 Copy Number Balance And Dosage Sensitivity

Expression levels, along with binding affinity and binding partners, are an important determinant of protein function. Copy numbers of proteins in a cell range from a few to well over a million[1,2]. It is believed that for obligate protein complexes the subunits should be expressed at roughly the same level to avoid wasteful leftovers. This is referred to as the "dosage balance hypothesis" (DBH) [3-5]. The DBH was formed from observations that duplicating a single chromosome in flowering plants and fruit flies was detrimental or lethal, whereas duplicating an entire genome was viable[6]. It was hypothesized that an imbalance in the stoichiometric relationships of regulatory gene products could lead to cell death. Further experiments found that ~15% of genes in *S. cerevisiae* were sensitive to overexpression[8,9], and the negative effects of overexpression have also been observed in several eukaryotic species including maize[5], flies[11], and worms[12]. Gene copy number variation in humans has been linked to a number of diseases, most notably Down syndrome but also neurodegenerative diseases[13,14] and some cancers[15]. Besides cell waste, an overexpressed core or "bridge" subunit may sequester periphery subunits, paradoxically lowering the final number of complete

complexes[4,16,17]. Excess free proteins are prone to misinteractions (discussed below) with nonfunctional partners. Underexpression carries its own dangers: a single underexpressed subunit will become a bottleneck for the whole complex. In addition, copy numbers of weakly expressed proteins are noisier[18] and thus less reliable for the cell.

Veitia et al. performed theoretical work[3] showing that copy number imbalance in an A-B-A protein complex could lead to formations of half complexes (A-B or B-A) and fewer full complexes. An extension of this concept to heterotrimeric complexes like A-B-C would mean that too much B could result in A-B and B-C complexes, without enough extra A or C to complete these half complexes. This sequestering effect has been observed in several real complexes, such as MAPK scaffolds[17]. The cell contains several dosage compensation mechanisms to maintain balance. One such mechanism is the upregulation or downregulation of gene expression levels[12], known to happen in male animals compensating for the lack of a second X chromosome[12,19] and in females through X-inactivation. Another example is tagging free subunits or incomplete complexes for degradation[20]. A complete complex would mask such a degradation signal, whereas unbound binding interfaces may display them. Similarly, a complete complex may be more thermodynamically stable.

If true, the hypothesis that incomplete complexes are a primary cause of dosage sensitivity meant that a simple feature – complex membership – could be used to predict which gene duplications are risk factors for disease. In support of this hypothesis were observations that low haploinsufficiency fitness (i.e. when a

diploid organism has only a single functional copy of a gene, effectively halving gene expression) correlates with complex membership[21], that complex members tend to be co-expressed[21], and that complex members are less likely to survive gene duplication[22]. However, a later study by Sopko et al. [8] failed to find a correlation between dosage sensitivity and complex membership in yeast. Furthermore, there was no general relationship between the phenotypes for gene overexpression and haploinsufficiency, suggesting unique side effects of overexpression beyond imbalance. To explain this apparent lack of correlation with complex membership, other studies have proposed position in complex topology[16,23] determines sensitivity. However, neither core nor peripheral proteins were found to be particularly dosage sensitive[24], highlighting the need for additional hypotheses. Overexpression could, for example, drain cell resources[25-27], or leads to hyper-activation of signaling pathways, compromising cell regulation[8,25].

Vavouri et al. [28] tested 27 genomic and experimental features to predict dosage-sensitive genes in *D. melanogaster* (fly) and *C. elegans* (worm). Intrinsic disorder, using three different metrics, was found to have the strongest correlation, followed by number of binary partners. In contrast, number of protein complex interactions was a poor predictor, as were abundance and aggregation load. This suggests that misassembly of protein complexes does not cause dosage sensitivity. This led the group to coin the Interaction Promiscuity Hypothesis (IPH), which states that misinteractions are the primary culprit. Intrinsically disordered regions of proteins – i.e. regions lacking a stable domain structure – are inherently promiscuous. They contain short linear motifs, which can bind to structured regions

on other proteins. These motifs have many more potential off target interactions, since they are short and degenerate[29]. The ability to bind linear motifs was also found to be a predictor of dosage sensitivity, as well as a high number of low-affinity binding partners. Both of these are consistent with the IPH. The study also looked at yeast and found that disorder content, linear motif content, number of binary protein interactions, and the ability to bind linear motifs are all predictors of dosage sensitivity. Dosage sensitive genes are tightly regulated and rapidly degraded in yeast, likely to prevent this sort of harmful promiscuity. A later series of experiments, which increased gene expression to 1% of cell protein content, similarly found that proteins with disorder or membrane-protruding regions resulted in a high fitness cost [27]. The former cause misinteractions and toxicity, whereas the latter proteins are quite large and thus overload cell resources (ribosomes, chaperones, amino acids, etc.)

The datasets tested by Vavouri et al. were from "absolute overexpression" experiments, where a gene was overexpressed by swapping the native promoter region with a strong promoter – that of the gene *GAL1* in the case of yeast experiments [25]. This makes the fold-increase of overexpression variable. A natively low abundance protein might be increased 1000-fold in abundance, while a natively high abundance protein might be increased merely 2-fold. Hence, these experiments bias natively low abundance proteins as dosage sensitive. Makanae et al. ran a "relative overexpression" experiment using a method they term "genetic tug-of-war" [9]. By attaching a nutrient gene to the gene of interest, this study was able to roughly estimate the gene copy number increase at which the overexpression costs

outweighed the benefit of the nutrient. Genes with a low number of copies were deemed dosage sensitive. 80% of genes were robust against a 100-fold increase or higher, whereas ~14% (786 genes) were deleterious after a 10-fold increase or lower. Of these, only 161 overlapped with the gene set from Sopko et al., reflecting different biases. Natively *high* abundance proteins were enriched as dosage sensitive, as were protein complex members. However, this set also found intrinsically disordered proteins to be dosage sensitive, despite that they tend to be natively weakly expressed[30]. Thus both types of experiments support the IPH.

In summary, there is support across several experiments that promiscuity/misinteractions are a primary cause of dosage sensitivity. Both the IPH and DBH imply that proteins copy numbers should be in balance, either to avoid having leftover proteins that misinteract or to avoid incomplete complexes. However, these do not guarantee balance, as proteins may also be out of balance for functional reasons. In signaling networks underexpression of bottleneck proteins can modulate pathway activation[26]. Overexpression may compensate for low binding affinity[31]. Imbalance may aid kinetic assembly of protein complexes by minimized undesired structures[32]. Finally, while a recent study of 5,400 human proteins found that strongly bound complexes are indeed balanced[10], weakly bound complexes are not. Although copy number balance has been studied in obligate complexes, balance at a network-wide level remains untested, and is a primary component of this dissertation.

1.2 Protein Misinteractions

A protein binding to an incorrect partner is called a misinteraction; also referred to as nonspecific interactions or promiscuity in the literature. Broadly speaking, misinteractions are low strength interactions not selected for by evolution that confer no functional benefit to the cell. They create functionless aggregates that occupy cell space and waste resources, often sequestering other proteins from their functional partners[33]. These aggregates are responsible for neurodegenerative disease such as Parkinson's and Alzheimers[34,35]. Zhang et al. estimates that in yeast cells roughly 22% of proteins not in specific complexes are engaged in misinteractions[36]. (It should be noted that functional interactions occur at a wide range of affinities, from millimolar to femtomolar[37], so one cannot define misinteractions based on low affinity alone. However, protein concentrations tend to be related to their affinities, and functional partners may be localized together in a cell. Hence, context can help identify misinteractions.) The strength of a protein's functional interactions relative to its misinteractions is called the protein's *specificity*.

Misinteractions occur for three reasons. Firstly, the physio-chemical nature of binding sites makes misinteractions possible, even if they occur with low strength. The strength of interactions are shaped by three major forces: electrostatic interactions between residues, hydrophobicity, and shape complementarity[38]. Hydrophobic pockets will attract other hydrophobic pockets, and amino acid residues can only be arranged in a finite number of ways[39]. The limited number of domain types and linear motifs commonly used by biology compounds this issue. A

common domain type will have many possible "off-target" interactions. For example, an SH3 domain may easily misbind to several proline-rich regions[40].

Secondly, the vast amount of proteins in a single cell leads to a crowded environment: macromolecules occupy 5-40% of cell volume[41,42]. This dense environment has some biological benefits: it promotes oligomerization by keeping partners close together[42,43]. The equilibrium constants can be increased by two to three orders of magnitude compared to environments where proteins can diffuse freely. But conversely, the lack of free diffusion and proximity to nonfunctional partners can lead to misinteractions and aggregation. Indeed, crowding increases the likelihood of amyloid fibril formation[44], an aggregate implicated in Parkinson's. This effect has been observed in organisms as simple as *E. coli*, where crowding isolates aggregates to certain regions of the cell[45]. Given that cells are crowded, proteins will contact their incorrect partners with great frequency.

Thirdly, many proteins are designed for multi-specificity, i.e. the ability to bind to multiple partners. Disordered regions of proteins provide a structural plasticity that allows multiple binding partners[37]. However, even interfaces on structured regions can be designed in such a way that there are many physical "solutions" for a partner to bind. This has been observed on immunoglobulins, which can bind very different targets with high affinity[46,47]. This multi-specificity strategy has an important benefit – allowing more functions with less proteins – but inadvertently leads to more possible off-target sites, with disordered regions being at particular risk.

Cells employ various methods to limit misinteractions. Eukaryotic cells divide proteins into various compartments (cytosol, nucleus, mitochondria). Zhang et al. calculated a limit on protein diversity within a compartment due to misinteractions and found that protein diversity in yeast was very close to that upper limit[36]. Cells face a conundrum where increased diversity and protein concentrations allow more function, but both can lead to more misinteractions. Negative design is another strategy: an expected off-target reaction will be blocked due to the protein's structure. One example of this is the proline-rich region on the protein PBS2 in yeast. Rather than having evolved to bind as strongly as possible to its intended partner (the SH3 domain on SHO1), it is optimized to bind moderately strong to SHO1 while binding very weakly to other SH3 domains[40]. A third strategy is allosteric regulation. A binding site may be hidden or weakened until an event elsewhere on the protein, such as phosphorylation or binding to a ligand, causes it to become accessible[48].

Finally, there is evidence that proteins more likely to participate in misinteractions are kept at low expression levels. Proteins with high intrinsic disorder or that are aggregation-prone tend to be weakly expressed[30,49]. Levy et al. created a "stickiness" scale for the 20 amino acids by measuring their frequency in protein binding sites[50]. In a study of proteins across *E. coli*, *S. cerevisiae*, and human, they found highly expressed proteins to be less likely to have these "sticky" residues on their surfaces. This gives them less binding strength overall but decreases the propensity for misinteractions. Similar studies have been performed substituting stickiness for hydrophobicity[51] as binding interfaces tend to be hydrophobic. It is

also known that highly expressed proteins evolve slowly[52-54], the so-called "E-R anticorrelation" (**E**xpression level vs evolutionary **R**ate). Misfolding avoidance is not sufficient to explain this correlation since misfolding would affect the evolution of internal residues rather than surface residues. Instead Yang et al. shows slower evolution prevents the addition of hydrophobic residues, which would increase propensity for misinteractions[54]. Ciryam et al. found that the proteins shared in three major neurodegenerative diseases (Huntington's, Parkinson's, and Alzheimer's) were "super-saturated", having a high abundance to solubility ratio[35]. This highlights the cell's need to limit aggregate-prone proteins.

Although it is tempting to view all misinteractions as deleterious, this imperfection in protein networks is a possible mechanism for the evolution of new protein interactions[55]. Thus some level of misinteractions may confer a long-term evolutionary benefit. Unlike complex membership, misinteractions must be studied as a network-wide phenomenon to truly understand the costs.

1.3 Protein Network Science

It is easy to see how proteins should be balanced at the level of single complexes. But proteins often participate in multiple complexes, utilizing multiple binding interfaces. To truly quantify balance, one must study protein abundance at the network level.

Once the interactions between different proteins in a cell are known, they can be analyzed on a global scale using a protein-protein interaction network (PPIN). In a PPIN, each protein serves as a *node* and each interaction as an *edge*.

These networks (or "graphs") can be used to deduce the global design of protein interactions. It is known, for example, that a small number of *hub* proteins have many functional interactions, whereas the majority of proteins have only a few functional partners. The *degree* of each protein, i.e. the number of partners each protein has, usually has a distribution akin to a power-law. Thus protein networks are often referred to as being power-law or *scale-free* (referring to a property of power-law distributions)[56]. This is in contrast to random Erdos-Renyi networks, which have a Poisson degree distribution. Many real networks – not just PPINs but social and information networks – share this power-law-like structure[57].

Another peculiarity of protein networks is their *motifs*. A network motif is a subgraph that appears more often than would be expected in a random network. Motifs in protein networks include feedforward loops, bi-fans, bi-parallels, and triangles[58-60]. Protein networks also have the "small-world" property: a high clustering coefficient (i.e. frequency of triangles) and a short typical path, meaning the average difference between any two chosen nodes is small[58]. Finally, protein networks have a modular design[61], meaning that proteins can be grouped into sub-networks with interactions enriched within sub-networks and depleted between them. Taken together, all these various features are referred to as the structure or *topology* of protein networks.

Protein network science has a variety of applications. Hub proteins tend to be essential[62] and thus may be controllers for network function. Modeling signaling networks is useful for drug discovery, especially for modeling combination therapy (the use of multiple drugs to treat a disease)[60,63]. Network modules (based on the

network structure alone) correlate both with "functional modules" and "disease modules"[64], meaning that they can be used to suggest novel functions of proteins. In transcription networks, motifs have been used to model transcriptional regulation[65]. Despite these advances, however, classic PPINs are lacking in many features, such as binding affinities, temporal or allosteric regulation, abundance, and identification of binding sites.

Binding sites, or *interfaces*, are the regions that proteins use to interact with their partners. Although domain-domain interactions and interface interactions are often used interchangeably in the literature, the two are not synonymous. A domain may contain multiple interfaces, and disordered regions may contain interfaces (via linear motifs) as well. Recently, studies have begun integrating structural information to add binding sites to protein networks, creating what we refer to as interface-interaction networks (IINs)[26,62,66,67]. By mapping which interfaces proteins use to bind, IINs can be used to track competitive vs noncompetitive binding[26], predict the effects of domain mutations on disease[67-69], identify linear motifs and promiscuous regions[28], and study the structure and dynamics of multi-protein complexes[70]. For a simple example, many proteins bind to themselves, which would be represented as a self-edge on a PPIN. But if the self-binding is mediated by two different interfaces – as in the case of the "barbed" end of Actin binding to the "pointed" end – the protein can polymerize into long fibers. In contrast, if one interface binds to itself, the protein may merely dimerize. IIN construction is still in its infancy, but IINs appear to have their own unique properties, including fragmentation, little to no clustering, and sparseness[71,72].

IINs are difficult to construct. High-throughput arrays of protein-protein interactions only reveal which proteins bind, not how they bind. Low-throughput experiments testing which domain or region of a protein binds are the most reliable source of data. Crystal structures may also be used to approximate the regions proteins use to bind. However, data is still lacking, and many studies have used homology modeling[73,74] to assign interfaces. Homology modeling uses a "template" interface and searches for a similar sequence of amino acids on the protein. There are several issues with this method however. Not only are these interfaces only putative, it is not necessary for proteins to use "similar" interfaces to bind to the same partners. As noted above, multispecificity of interfaces (in which a single interface can bind two very different partners) is an important feature of protein design[37], especially for human antibodies[46]. The Interactome3D approach uses several criteria to improve accuracy in predicting binding interfaces, but recovered acceptable models for only ~64% of interactions in their database[73]. Homology modeling also does not capture short linear motifs – described above – which are an important mediator of protein-protein interactions[75], especially for proteins with high intrinsic disorder.

IINs are necessary for the study of copy number balance network-wide. A protein that binds noncompetitively with two partners must have equal expression with each partner to be balanced. But a protein that binds competitively with two partners must have an abundance equal to the sum of the two partners. Hence, interface-resolved networks are essential for showing whether or not protein expression is balanced at a network-wide level.

1.4 Research Summary

In this project I extend the concept of copy number balance from obligate complexes to the network level, with particular focus on the role in limiting misinteractions. The results are divided into three chapters.

Chapter 2 focuses on the topology of IINs. By studying two manually curated IINs from the literature, one of the clathrin-mediated endocytosis (CME) system in yeast, and one of the ErbB signaling network in humans, I show that their topologies are consistent with a balance between a limit on interface diversity and a need for physio-chemical binding complementarity.

Chapter 3 shows evidence that the copy numbers of proteins in the CME network are balanced, though not perfectly, according to the underlying IIN. Proteins that are out of balance are analyzed for their functional benefit. I also show in this chapter that the topology of IINs, when compared to random networks, are robust against misinteractions when copy numbers are balanced but more prone to misinteractions when copy numbers are not balanced. This suggests a joint optimization of protein network topology and protein expression level.

Chapter 4 focuses on functional outcomes of misinteractions and imbalance, mainly by studying their effects on vesicle formation in a dynamic model of clathrin-mediated endocytosis.

In the **Conclusion** I provide a general summary, discuss medical applications, and suggest future directions.

Chapter 2. Protein Binding Specificity Constrains Interface-Interaction Network Topology

Chapter adapted from:

Holland DO, Shapiro BH, Xue P, Johnson ME. Protein-protein binding selectivity and network topology constrain global and local properties of interface binding networks. Sci Rep. 2017;7(1):5631.

Protein-protein interactions networks (PPINs) are known to share a highly conserved structure across all organisms. What is poorly understood, however, is the structure of the child interface interaction networks (IINs), which map the binding sites proteins use for each interaction. In this study we analyze four independently constructed IINs from yeast and humans and find a conserved structure of these networks with a unique topology distinct from the parent PPIN. Using an IIN sampling algorithm and a fitness function trained on the manually curated PPINs, we show that IIN topology can be mostly explained as a balance between limits on interface diversity and a need for selective binding complementarity. This complementarity must be optimized both for functional interactions and against mis-interactions, and this selectivity is encoded in the IIN motifs. To test whether the parent PPIN shapes IINs, we compared optimal IINs in biological PPINs versus random PPINs. We found that the hubs in biological networks allow for selective binding with minimal interfaces, suggesting that binding specificity is an additional pressure for a scale-free-like PPIN. We confirm through phylogenetic analysis that hub interfaces are strongly conserved and

rewiring of interactions between proteins involved in endocytosis preserves interface binding selectivity.

2.1 Introduction

Interface interaction networks (IINs), also referred to as structural interaction networks[62,67], domain-domain interaction networks[26,76], or structurally annotated pathways[73], are a map of the binding sites proteins use for various interactions. Such a map can be used to model how competition modulates signal transduction[26,77]; predict the effects of domain mutations on disease[67-69,78] and the immune response[79], predict dosage sensitivity by identifying linear motifs and promiscuous regions[28], and study the structure and dynamics of multi-protein complexes[70]. For example, actin can form long fibers because it has a "barbed" end that binds to a "pointed" end of another actin protein. On a typical protein-protein interaction network (PPIN) map, this interaction would appear as a self-edge, whereas more accurately, they are two distinct binding sites with their own share of possible partners.

We ask four major questions in this work. First, is the structure of IINs conserved across PPINs? Second, does this structure reflect any selective constraints on protein interactions? Third, do the presence of hubs in the PPIN network affect the types of IIN structures possible? And fourth, do hubs in the PPIN provide an advantage (relative to random networks) in producing selective interface

interactions with minimal interfaces, suggesting a new benefit for scale-free PPINs? The answer is yes in each case.

We analyze the structure of four PPINs with IINs defined: two smaller manually curated networks (621 total interactions) and two larger automatically constructed networks (6,893 interactions). Little work has been done on IIN structure, in large part due to the paucity of experimental and crystallography data identifying where proteins bind to one another. The Protein Data Bank[80] provides an excellent resource for automated computational assignment of interfaces. However, with limited crystal structures of proteins in complexes, homology modeling[73,81,82] is needed to help infer domains and interfaces used for interactions. Interfaces assigned through homology modeling are only putative, however, as this approach is limited in accuracy. The binding sites discovered will depend on the experimental templates used. Even if the sites have similar sequence there is no guarantee of an interaction[82]. Stein et al., using known PPIs from six organisms including humans, estimated that less than 30% have templates for comparative modeling.[83] The Interactome3D approach uses several criteria to improve accuracy in predicting binding interfaces, but recovered acceptable models for only ~64% of interactions in their database[73]. Homology modeling will also miss many short linear motif (SLiM)-mediated interactions[84], both due to their rapid evolution[85] and low affinity, which has hindered experimental detection[75]. As we describe below, limited accuracy in automatically predicted interfaces significantly alters the structure of the inferred IIN, although major features are still visible.

With manual curation, in contrast, putative interfaces can be refined, corrected, or rejected, and the many protein interactions that lack homology models can be assigned based on detailed biochemical approaches, functional studies, and analysis of disordered regions and SLiMs. Two such IINs constructed to this standard are considered: the Clathrin-mediated endocytosis network in yeast[66], and the ErbB signaling network in humans[26]. (Fig 2.1) Despite being independently constructed by different research groups, the two share similar features: fragmentation into multiple components, little clustering, and a high frequency of square and hub motifs. With the exception of the presence of hubs, these features differ from their parent PPINs, and thus display a unique topology that we show results from different selective forces.

Regarding selective constraints on protein interactions, we propose that one of the selective forces shaping IIN structure is the need to maintain high binding specificity. Due to the chemical nature of binding sites, occasionally nonspecific misinteractions will occur. Avoiding these misinteractions has been demonstrated to be a fundamental force limiting the number of distinct proteins in an organism[36,86], protein expression levels[28,87,88], binding strengths[50], and interface interaction motifs[71,86]. Regarding IIN motifs, an amino acid residue optimization model demonstrated that specific motifs (and not others) and a fragmented IIN structure were needed to optimally design protein interfaces for high specificity[71,86]. We first compare IIN structures to randomized versions to demonstrate the biological networks' clear departure from the statistically most probable IIN structure. We then construct a trainable fitness function to reproduce the observed

biological IIN. This fitness function favors network motifs that have been shown to improve the binding selectivity of interfaces[71] while penalizing high interface diversity. The former constraint makes the network easier to optimize, and also reflects motifs abundant in the biological IIN, whereas the latter lowers the number of possible misinteractions that must be optimized against (order of n^2)[36] where n represents the number of interface species. Because the search space for possible IINs of a given PPIN is enormous (quantified below), we used a Monte Carlo sampling algorithm combined with a fitness function (See Methods section 2.4.1) to sample optimized IINs at various parameterizations, similar to previous work optimizing spatial networks.[89]

Because the automatically constructed[62,67] IINs contained systematic errors, largely due to missing SLiMs as binding partners and false positives, we restricted our training and sampling procedure to the two manually curated networks. However, this outcome highlighted a powerful advantage of visualizing the IINs: the network motifs can be used to identify erroneous domain-domain interaction predictions. Disagreements over the evolution of proteins and their networks can often be attributed to variability and poor overlap in PPIN datasets[90]. Boosting domain assignment accuracy by identifying errors in automatically constructed networks using network motifs, as we demonstrate here, improves these crucial resources for understanding protein function and evolution.

To learn how the presence of protein hubs affects the IIN sampling space (our third question), we combined both analytical and computational sampling approaches to characterize the structure of IINs as a function of varying PPIN

structure. PPINs feature a degree distribution that is approximately power-law or "scale-free", meaning (loosely speaking) that a few proteins act as hubs, while the majority of proteins are restricted to only a few interaction partners[91]. This same basic structure describes airport networks, and is the optimal structure for maximizing transport with minimal costs[92]. By considering the possibility of a random PPIN, we can then compare whether this alternative structure is different and possibly worse than a scale-free PPIN in terms of IINs possible. For example, a well-known advantage of scale-free PPINs relative to random networks is their ability to maintain connectivity under attack[93]. Because IINs have not been studied in the context of their parent PPINs, we first establish how the ensemble of possible IINs varies with PPIN structure, showing that hubs do alter the space of IINs in specific ways.

Finally, we sought to test whether observed PPINs were any better for developing selective binding than the random PPINs. We applied our data-trained fitness function at its optimized parameters to sample IINs for scale-free versus random PPINs of the same size. Random PPINs proved more difficult to optimize, requiring the evolution of significantly more interfaces (penalized in our fitness function) in order to achieve the same level of binding complementarity encoded in the IIN motifs. This runs counter to the parsimonious use of domains across species, where new domain combinations rather than new domains drive functional divergence[94]. Ultimately our results suggest an additional pressure favoring a scale-free-like PPIN. It is a cheaper (fewer interfaces) design for maintaining a multitude of selective binding interactions.

Our model emphasizes that selectivity in interface binding is critically conserved across IINs, and that hubs in the PPIN provide an advantage in this regard, largely because these hub proteins contain hub interfaces. As a final analysis we use phylogenetic analysis to test whether interface binding selectivity is conserved as protein-protein interactions are rewired throughout evolution[95]. We use this analysis to test whether, despite this rewiring, hub interfaces are nonetheless conserved, providing a new physico-chemical argument supporting the conservation of hub proteins.

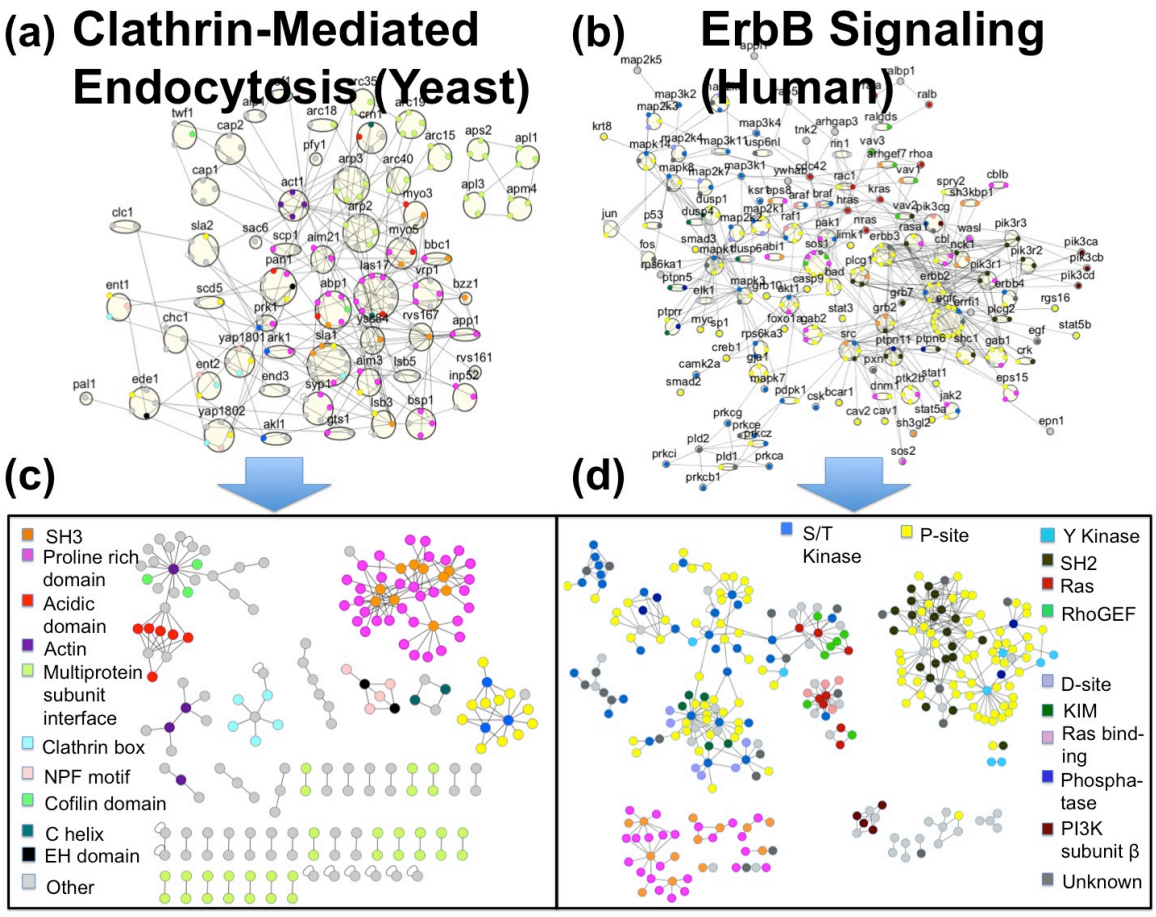


Figure 2.1. PPINs and their IINs have distinctive topologies. We analyze the PPINs of the manually curated yeast endocytosis **(a)** and human ErbB networks **(b)** with all domains and interfaces identified and shown here colored by domain type

(see Appendix B). The resulting interface interaction networks (IINs) in **(c)** and **(d)**, respectively, have highly distinct topologies that reflect the needs of interfaces to achieve strong functional binding and minimize non-functional interactions. Both IINs break into multiple components with a selection of hub interfaces, and they contain an abundance of hub and square motifs with a minimal (or zero) number of triangle motifs. Both PPINs contain hub proteins.

2.2 Results

2.2.1 IINs for the Biological PPINs Have Highly Specialized Features Sensitive to Rewiring. To determine if IIN structure is conserved across PPINs, we first characterize the manually curated PPINs from yeast and humans shown in Figure 2.1, which involve different protein sets but both exhibit scale-free-like topologies. Analysis of both their IINs (Fig 2.1c,d) demonstrates that they both share highly similar features to one another. They have fragmented structure, almost no triangle motifs (low C_{global}), a higher fraction of hub versus chain motifs, and a significant fraction of square motifs (Table 2.1). In contrast, expected values for these features, calculated by randomly rewiring the interface interactions while keeping the PPIN structure intact, have no similarities. Rewired IINs organized into a giant component with many chains, increased triangles (higher clustering coefficient C_{global}), and minimal squares. The lack of hub interfaces in these rewired IINs is reflected by the low preferential attachment exponent (P.A.E.), which varies from 0 for random networks to ~1 for scale-free networks (See section 2.4.3).

The structure of the two automatically constructed IINs[62,67] had similarities to the manually curated IINs, but were closer to randomly rewired networks. Similar to the manually curated networks, they had large PAEs, indicating hub interfaces in the network, and a similar fraction of square motifs (Table 2.1). They

also had correspondingly more hub motifs in the network than would be observed in a random network. A significant difference was the degree of fragmentation. The manually curated networks are nearly fully connected at the PPIN level, and yet the

Table 2.1. Comparison of properties of the IINs from two manually curated PPINs and two automatically constructed IINs.

	Yeast CME IIN	Human ErbB IIN	Human SIN	Yeast SIN[a]
Proteins	56	127	3626	167
PPIN Edges	186	268	6585	308
Interfaces	195 [200]	297 [411]	5494	308
IIN Edges	206 [207[b]]	415 [420 [b]]	11,466	539
Self Loops	10	2	3414	0
IIN PAE	0.8 [0.09±0.09]	0.7 [0.24±0.07]	1	1
LC[c] (PPIN)	92%	100%	43%	36%
LC[c] (IIN)	23% [82±4.0%]	35% [96±2%]	33%	35%
C Global	0 [0.016±0.01]	0.002 [0.01±0.005]	0.17	0.21
Tetramers	2,743 [819±92]	10,856 [4312±280]	$2.5x10^6$	16,530
Squares	0.061 [0.002±0.002]	0.066 [0.005±0.001]	0.0210	0.0557
Hubs	0.56 [0.26±0.020]	0.58 [0.27±0.01]	0.461	0.339
Chains	0.37 [0.73±0.02]	0.36 [0.72±0.01]	0.374	0.455

Bracketed values are expected values for IIN properties with standard deviations, see Supplemental Methods in Appendix A for further details on calculations.
[a]Only the cytoplasmic proteins used in (Deeds et al, 2012)[70]
[b]Edges numbers were capped when sampling to prevent continuous growth.
[c]Percent of nodes in largest component of network

IINs contain a largest connected component of only 23-35% of nodes. In contrast, the automatically constructed Human Structural Interaction Network (SIN)[67] is already fragmented at the PPIN level (43% of nodes in the largest component), and the IIN fragmentation is therefore more strongly driven by the PPIN fragmentation. The Yeast SIN[62] shares these features. The reason for this higher connectivity is the larger ratio of chain to hub motifs, as chain motifs prevent fragmentation into many distinct modules (see Appendix C, Fig C.1). The number of triangle motifs, which is quantified by the clustering coefficient C_{global}, is also significantly higher in these networks than in the manually curated networks. Does the increased randomness of these IIN connections occur due to mis-identification of interaction interfaces? By following up on this implication by investigating the many unexpected triangles in the automatically curated IINs, we found this was true (Fig C.1).

We found mis-assignments of interface interactions can be largely attributed to a lack of linear motifs included as potential binding partners, and a permissive decision-making algorithm. Applying of the INstruct website[96] to predicting CME protein interface interactions produces only 44 interactions (versus 206 for the manually curated network of Fig 2.1a[66]). Of these 44 predicted interactions, only 1 involves the correct domains (Fig C.1). This method predicts a disproportionate abundance of homo-dimers. Many interactions are predicted to be SH3-SH3 interactions (including in the Human SIN[67] (Fig C.1)), but even in the crystal structures, SH3 domains form homo-dimers only in special cases when mediated by a ligand (such as a Proline Rich Region) [97]. We also note that some structured domains (such as kinase domains) must be recognized as containing multiple

protein binding interfaces. Many kinase domains, for example, form dimers through distinct interfaces and can still perform catalysis[98].

2.2.2 Network Motifs in the IINs Indicate Suppression of Nonfunctional Interactions. For our second question, we connect the special conserved structure of the biological IINs (Fig 2.1 and Table 2.1) to constraints on binding selectivity. Previous work, using Monte Carlo based optimization of amino acid sequences in small networks, showed that mediating interface interactions by hub or pair motifs, and not chain motifs, increased the binding selectivity of the interfaces[86]. Thus the level of achievable binding selectivity is encoded by basic motifs; which include hubs, squares, and pairs. Subsequently, it was shown that IINs were also more selective if they were highly fragmented into modules[71]. In both cases this is because it is easier to optimize the interfaces for both strong specific interactions, and against non-functional mis-interactions. All of these trends are clearly present in the biological IINs, but not in random IINs (Table 2.1). In Fig 2.2 we further illustrate how, for the same reason, square motifs are beneficial to selectivity, and triangle motifs are detrimental. While it is perfectly possible to design interfaces that will bind strongly in any motif configurations, the real challenge is to simultaneously suppress the nonfunctional interactions possible for those motifs. For the chain motif, the challenge is preventing the interaction between the two ends of the chains. For the triangle motif, in order for all three distinct domains to attract one another, they must all be similar to one another. If an interface binds a very similar interface to itself, it will likely also bind to itself. Thus, triangle motifs are only

consistent with high-selectivity optimization if their interfaces also self-binding. We found that for the one triangle present in the ErbB IIN, this was indeed the case. Two kinase domains form not only a heterodimer with a shared target, but also homo-dimers[98]. Hence we added these previously absent self-interactions to the network.

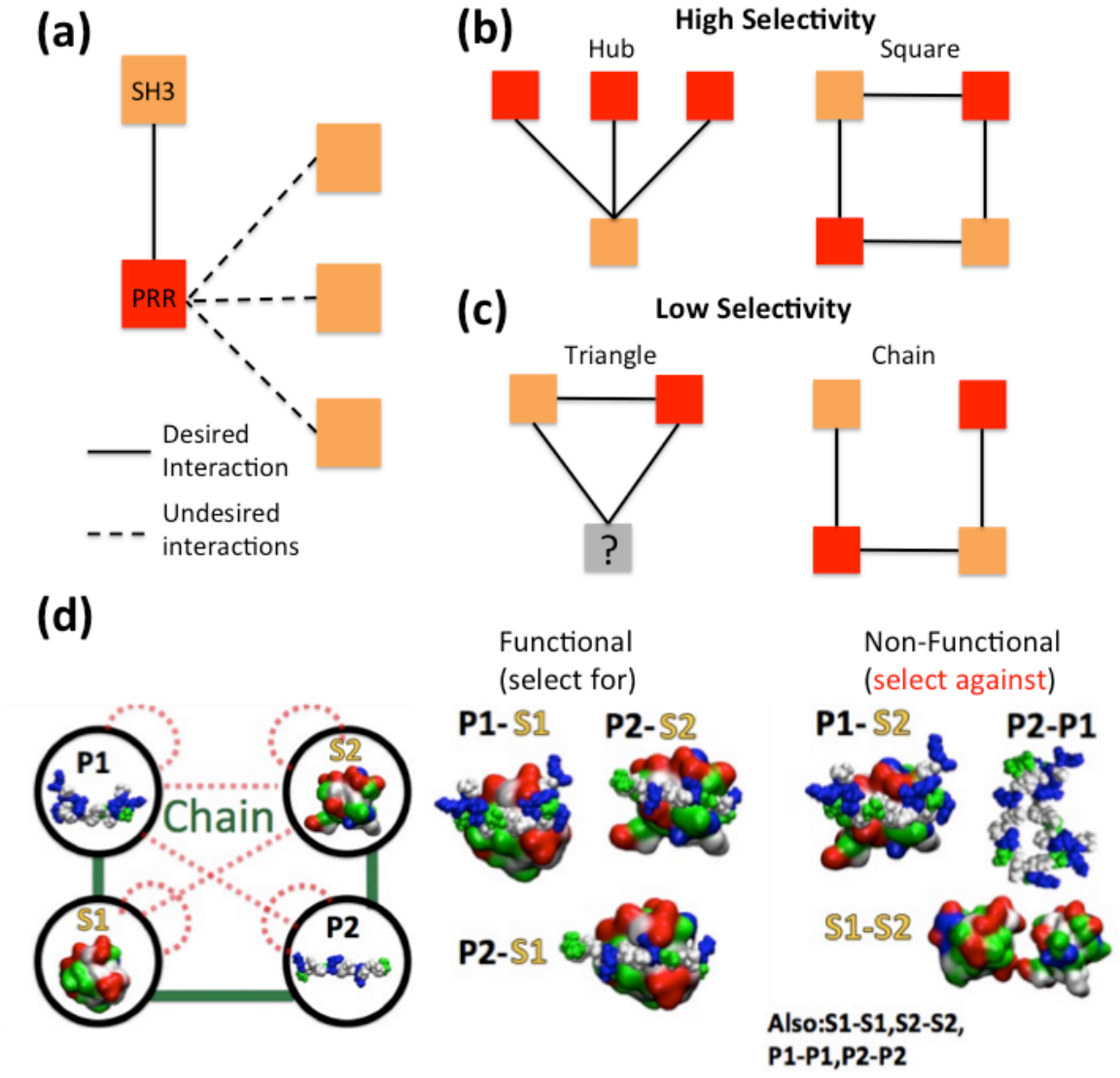


Figure 2.2. Subgraphs uncommon in the biological IINs due to poor interface binding selectivity. (a) Binding site optimization in proteins are subject to both positive design constraints (strengthening desired interactions) and negative design constraints (weakening undesired interactions). In the case of the common SH3 domain to proline-rich-region (PRR) binding pair, sites must be optimized such that off-target interactions with the wrong SH3 or PRR are minimized. **(b)** IIN subgraphs that confer high selectivity. Binding partners may achieve structural and chemical

complementarity with few constraints. **(c)** IIN subgraphs with poor selectivity due to the difficulty of constructing a three-way competitive binding set of interfaces in the case of the triangle and the negative design constraint in the case of the chain. **(d)** Example of desired (solid line) and undesired (dashed line) interactions in the case of the chain subgraph, illustrated with SH3 domains and PRRs. The interaction between P1 and S2 in particular is difficult to optimize against. Binding surfaces are colored by residue as non-polar (white), polar (green), acidic (red), and basic (blue). Example structures from 2RPN, 2LCS, 1UWH, 3OMV.pdb. Truly non-functional interactions (i.e. PRR-PRR) are just illustrations.

2.2.3 The Space of Possible Interface Networks for a PPIN Is Enormous and Varies with Protein Degrees. Our third question considers how the PPIN structure might constrain the IINs accessible. While a PPIN and its interface interaction network (IIN) must evolve together, it is not obvious how one constrains the other, given that a protein can use one or many interfaces for its various partners. To illustrate properties of IINs constrained to a PPIN, in Fig 2.3 we enumerate the 8 possible IINs for the simple PPIN of three proteins binding. The total number of possible interface networks is determined by the number of interactions (degree, k) per protein and quantified through the Bell number B_k (See Appendix A.4 for details). Bell numbers grow rapidly and hence high-degree hub nodes can dramatically increase the number of possible IINs, meaning a scale-free PPIN will have significantly more IINs possible than a random PPIN because of its hubs. We calculate 10^{166} IINs for the clathrin-mediated endocytosis (CME) PPIN in Fig 2.1a, and 10^{143} for a similarly sized random PPIN, more than the number of atoms in the known universe! Both types of PPINs produce IINs with an expected degree distribution that is random, not scale-free. This is because configurations that create hub interfaces, which are necessary to produce a scale-free IIN, are rare. However, hub proteins do cause several subtle shifts in the properties of the IINs

possible, including slightly fewer expected interfaces, more 4-node motifs (tetramers) and more hub interfaces. Since these are the features are important in the biological IINs, this is indicates that the hub proteins found in scale-free PPINs may promote more selective IINs.

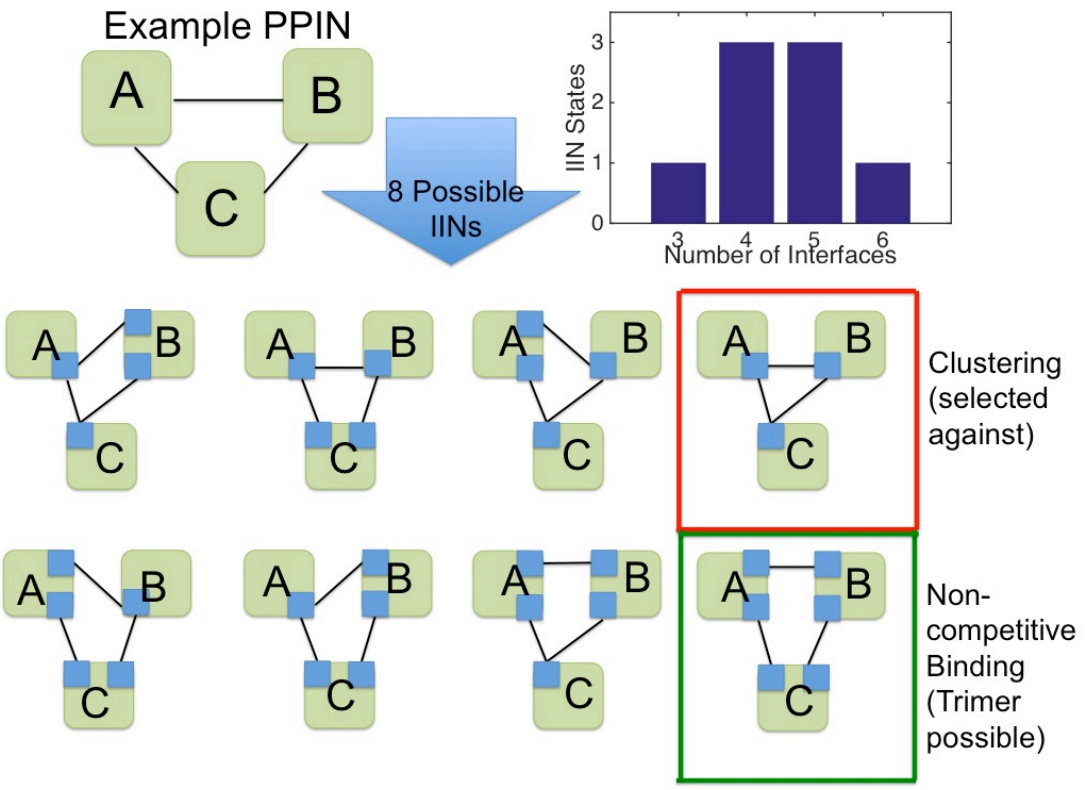


Figure 2.3. Each PPIN has many possible IINs, and only some are good for promoting selectivity. For the simple PPIN with three interacting proteins, there are 8 possible IINs with either 3, 4, 5 or 6 interfaces (blue squares). Because each IIN has different motifs present, only a subset will be favored in biological networks. The top row contains IINs with chain motifs or a triangle motif (red box), which are bad for promoting selectivity and less common in biological IINs. The bottom row contains favorable motifs, and in the green box is the only IIN that allows a true protein trimer to form. IINs with 4 or 5 interfaces are most common, as counted in the histogram. The same trend holds for much larger PPINs, with the sparse and dense IINs becoming increasingly rare, and hub interfaces less common.

2.2.4 Strong Motif Biases Are Needed to Reproduce Biological IINs. To address question four, whether the PPIN structure influences the ability to produce biologically optimal IIN structures, we first needed to be able to sample biologically realistic IINs given a PPIN. To do so we created a fitness function and trained it to reproduce the networks of Fig 2.1. Due to the inaccuracies of the automatically constructed IINs, we did not include them to avoid biasing the fitness functions towards erroneous network structures. The fitness function is biologically motivated to penalize features that promote mis-interactions, to not penalize features that promote strong interactions, and to capture physical size constraints of proteins. We therefore included a bias against triangle subgraphs without self-loops (parameterized by β) and chain subgraphs (parameterized by κ), which are difficult to optimize for structural and chemical complementarity as explained above (Fig 2.2). These two separate terms resolved a problem we found with our previous fitness function[71] that penalized chain subgraphs and also penalized biologically realistic square subgraphs. Our current fitness function does not penalize squares. We introduced a third parameter, μ, to penalize having large numbers of interfaces in the network, both because this increased diversity leads to more possible misinteractions[36] and because proteins have limited volume for extra interfaces. Finally, in the biological IINs, protein pairs can interact through multiple domains, resulting in a significant increase in edges from the PPIN to the IIN (Table 2.1). Our fourth and final term thus allowed new duplicate edges in the IIN but limited their growth by a parameter ω.

We optimized the four parameters of our fitness function to locate the biological IINs out of an enormous space of possible IINs (e.g. 10^{166}). We found that the key to generating realistic IIN features required a balance between creating new fragmented modules and avoding introducing too many interfaces. To do so required re-using interfaces that would generate either isolated star hubs (e.g. turquoise nodes in Fig 2.1c) or hubs connected in square clusters (e.g. orange and pink nodes in Fig 2.1c,d). Fig 2.4 shows how these networks were most sensitive to the parameters κ, which penalizes chains, and μ, which penalizes the creation of new interfaces (Methods section 2.4.1). Star hubs, like squares, result from pressure to avoid chains and hence are favored by increasing κ (Fig C.2). Our trained fitness function samples IINs with properties similar to to the observed CME network (Fig 2.4d, Fig C.2; Table 2.2) with parameters κ=2, μ=0.42, β=4 and ω=0.1. Comparable parameters applied to the ErbB PPIN (κ=2.3, μ=0.45, β=4) except we lowered ω to 0.02 to account for the much greater frequency of edge duplication. In the discussion we consider ways to further improve the agreement.

2.2.5 PPINs Need Hubs to Minimize New Domain Interfaces. To determine the effects of the parent PPIN on optimized IIN structure, we used our trained fitness function to sample IINs for a variety of PPIN topologies and sizes. We compared the CME and ErbB PPINs with PPINs of the same size but a random degree distribution, as well as for new PPINs both more and less densely connected than these (Fig C.3). Regardless of the size of the PPINs, we found that because random PPINs lack hub proteins, they cannot produce selective domain modules without significant

addition of new interfaces (Fig 2.5). Thus random PPINs have the disadvantage that evolving more interfaces is more costly for mediating protein-protein interactions than re-using domains already optimized for selectivity.

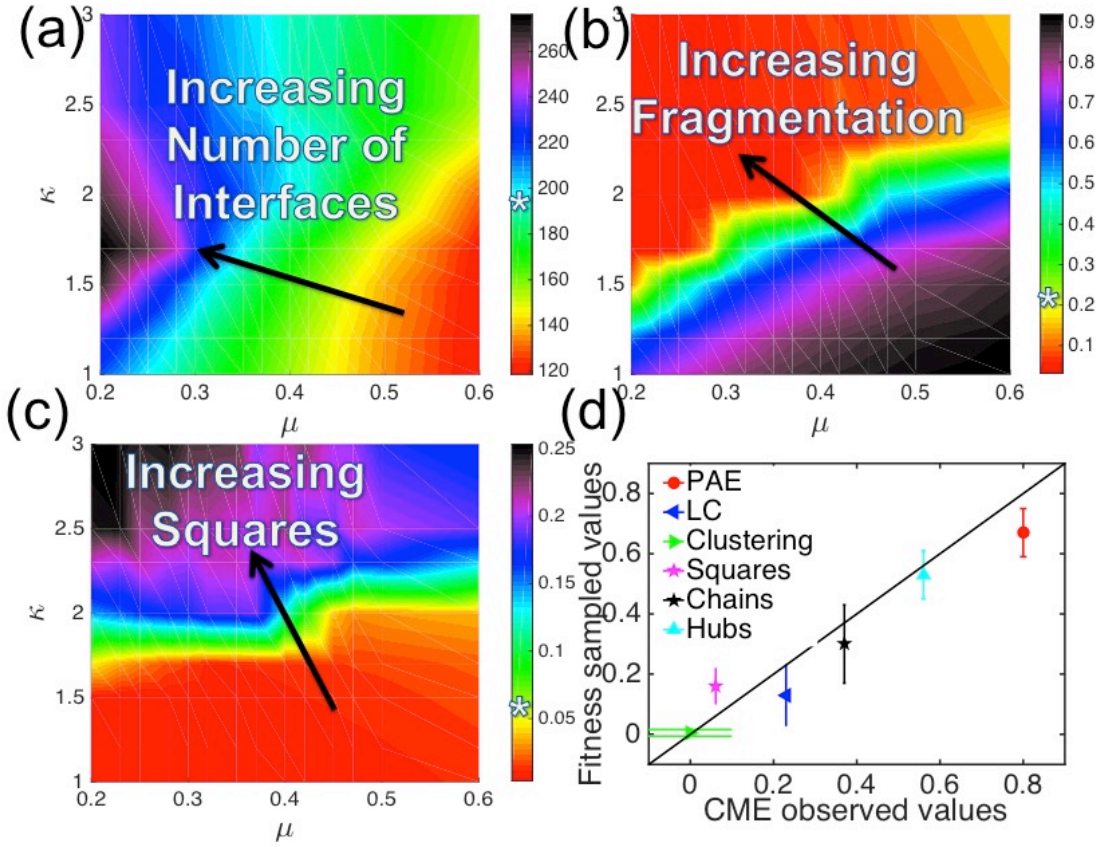


Figure 2.4. Learning how to select biologically realistic IINs for a PPIN using a parameterized fitness function. Because biological IINs are so distinct from a randomly generated IIN, we needed a four parameter fitness function to bias the sampling towards the correct: **(a)** number of interfaces; **(b)** size of the largest module/fragment; **(c)** frequency of square motifs in the IINs, as well as other properties. The results were most sensitive to variation in the parameters κ and μ (on the axes) that regulated the square-to-chain ratios and number of interfaces, respectively, in the fitness function. White stars on color bars indicate observed values of the CME PPIN (Fig 2.1a). **d.** By training the fitness function, we achieved very good agreement between the properties of the sampled IINs and observed CME IIN with optimal fitness parameters κ=2, μ=0.42, β=4 and ω=0.1. Comparable parameters applied to the ErbB PPIN (κ=2.3, μ=0.45, β=4) except we lowered ω to 0.02 to account for the much greater frequency of edge duplication.

Table 2.2. Sampling with a fitness function reproduces properties of biological IINs.

	Yeast CME Protein Network				Human ErbB Signaling Network				
	Unbiased Shuffle[a]	Unbiased Sampling[b]	CME IIN	Fitness Sampling	Unbiased Shuffle[a]	Unbiased Sampling[b]	ErbB (Reduced) IIN	Fitness Sampling	ErbB Full IIN
Interfaces	195	193.7 ± 2.46 (Max 196)	195	192.1 ± 5.4	195	303.7±0.55 (Max 304)	297	308.1 ± 7.13	377
Edges	206	207.0 ± 0.17 (Max 207)	206	209.9 ± 2.8	206	419.8±0.38 (Max 420)	415	424.8 ± 13.1	540
Pref. Attach. Exp.	0.50 ± 0.043	0.086 ± 0.090	0.8	0.67 ± 0.08	0.49 ± 0.004	0.24 ± 0.07	0.7	0.81 ± 0.05	-
Largest Component (%)	75% ± 3.4%	82% ± 4.0%	23%	13% ± 10%	92% ± 2.4%	96% ± 2.0%	35%	13% ± 3.3%	38%
C_{Global}	0.015 ± 0.0099	0.016 ± 0.011	0	0.0048 ± 0.011	0.0044 ± 0.003	0.011 ± 0.0054	0.0015	0.0020 ± 0.0036	0.0014
Tetramers	1,306 ± 126	819.2 ± 91.9	2,743	1,139.7 ± 620	6,177 ± 273	4,312 ± 280.4	10,856	14,750 ± 3,254	38,626
Square	0.0033 ± 0.002	0.0021 ± 0.0016	0.061	0.16 ± 0.06	0.017 ± 0.020	0.0054 ± 0.0012	0.066	0.18 ± 0.026	0.03
Chain	0.67 ± 0.021	0.73 ± 0.021	0.37	0.30 ± 0.13	0.65 ± 0.01	0.72 ± 0.011	0.36	0.36 ± 0.044	0.16
Hub	0.32 ± 0.021	0.26 ± 0.020	0.56	0.53 ± 0.08	0.33 ± 0.0098	0.27 ± 0.010	0.57	0.45 ± 0.023	0.8
Other Tetramer	0.013 ± 0.009	0.0095 ± 0.0081	0.0	0.005 ± 0.01	0.0032 ± 0.0024	0.0094 ± 0.0050	6e-4	0.0019 ± 0.0037	5e-4

[a]Shuffling of edges while keeping the number of interfaces on each protein constant. No bias from a fitness function.
[b]Sampling without bias from a fitness function, i.e. random sampling of IINs for a given protein network.

The main advantage of hub proteins in a PPIN is that they are capable of more highly connected hub interfaces in the IIN. Although hub interfaces are still possible for a random PPIN of sufficient density (Fig C.4a), the reduced size and frequency of these hubs limits how many square motifs can form (Fig C.4b). Square cluster components are a prominent feature of the biological IINs and they are critical for maintaining selectivity with a minimum number of interfaces. Without access to these motifs, random PPINs require more interface splitting to instead produce selective star hubs. These results were robust to changes in the fitness

function that allowed larger fluctuations in interfaces per protein (Fig 2.6e,f; Table C.1). Ultimately, our results suggest that a scale-free-like PPIN is beneficial to evolving specificity in interface binding interactions.

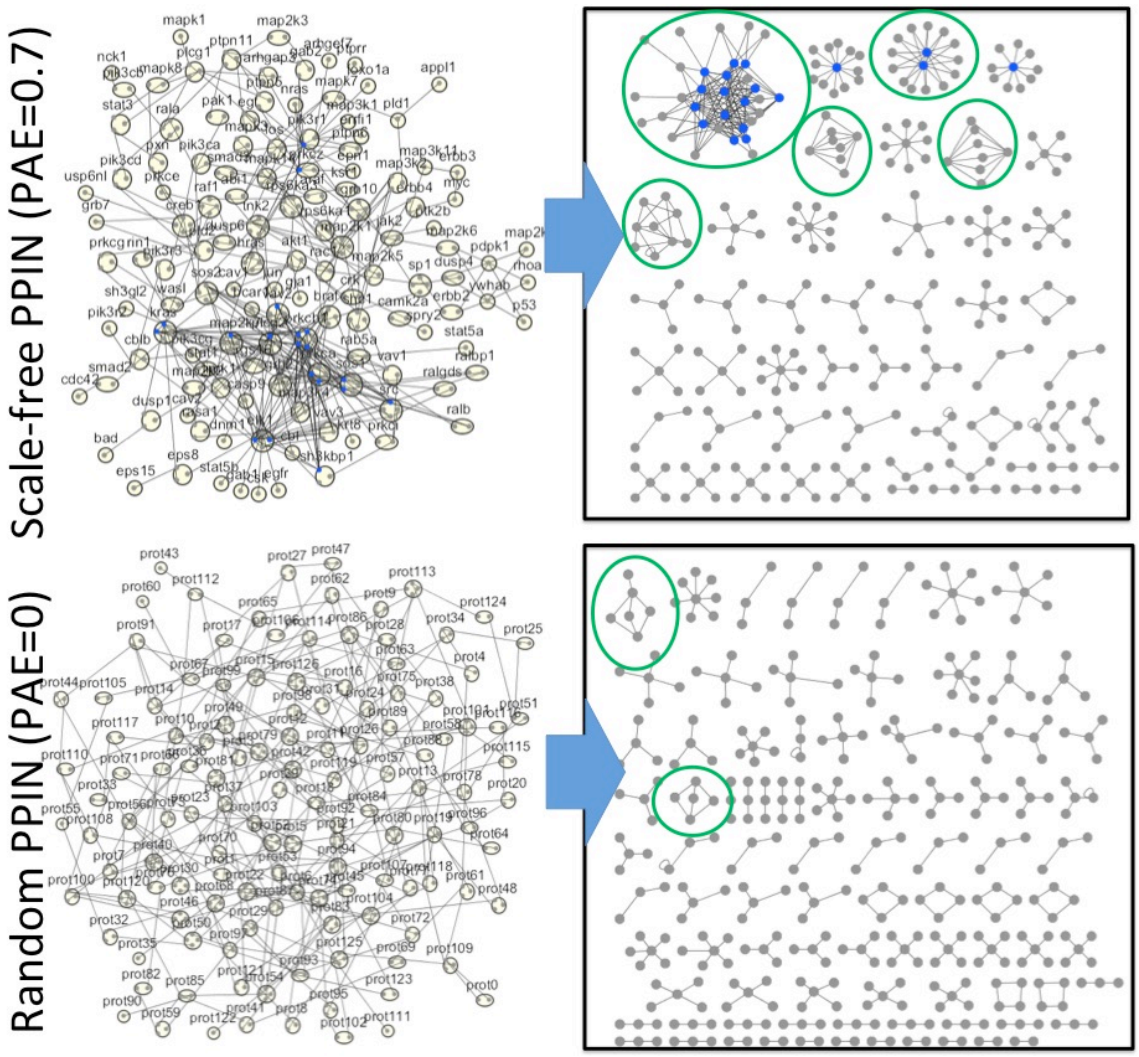


Figure 2.5. Scale-free PPINs produce fitter IINs than random PPINs. We performed fitness sampling for selective IINs on the ErbB scale-free like PPIN (top) and a random network with the same number of proteins and PPIs (bottom). For the scale-free like PPIN (top) fewer interfaces (n=290) were needed to produce selective motifs, including 2000 squares (in green circled modules). Without hub proteins, the random PPIN (bottom) produced only 12 squares, and introduced many additional interfaces (n=356) in order to maintain selective motifs. The same trends held with the CME PPIN. IINs discovered with random PPINs were also less fit than those found with scale-free PPINs (Table C.1). Nodes with >9 partners are shown in blue.

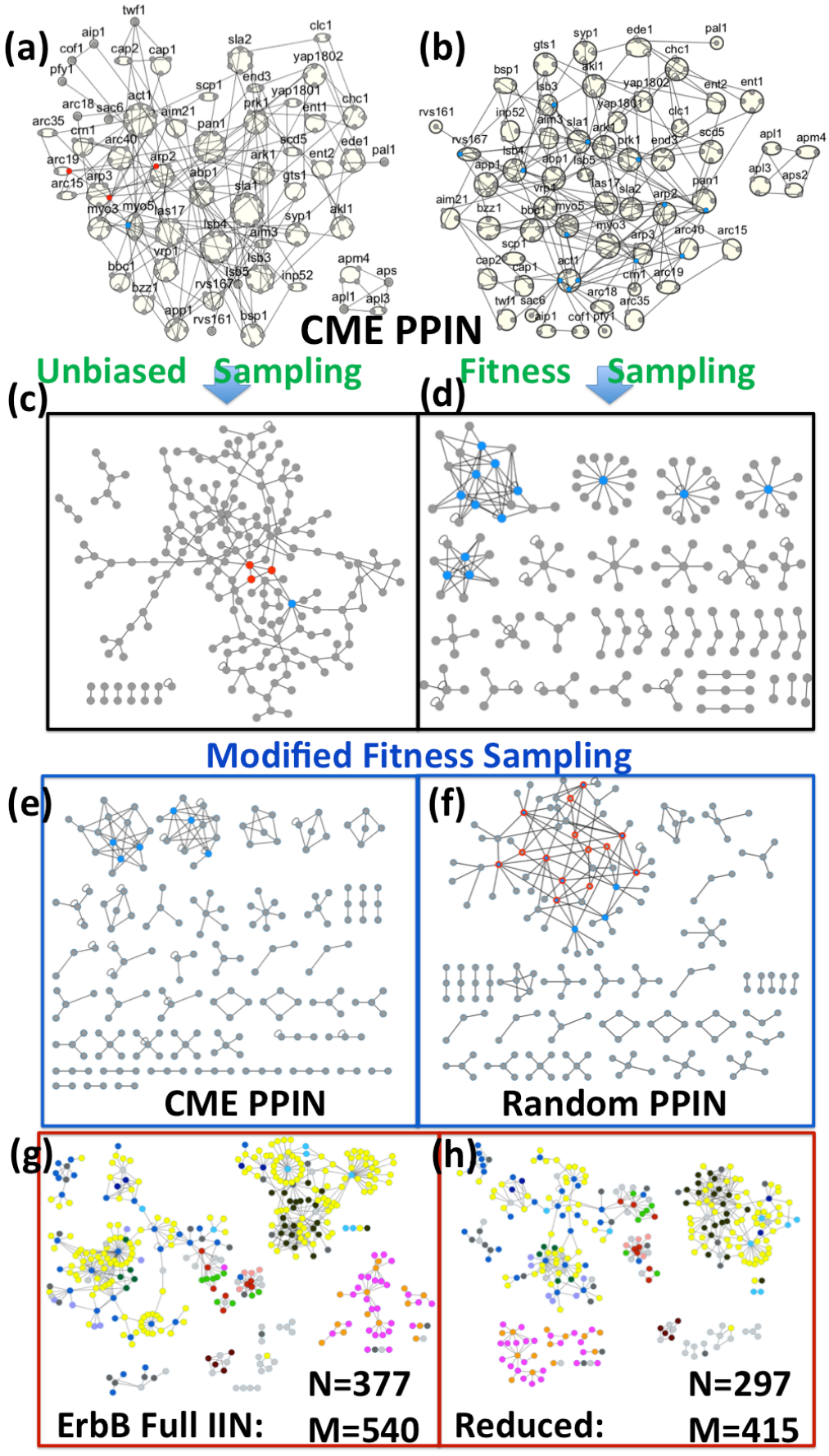
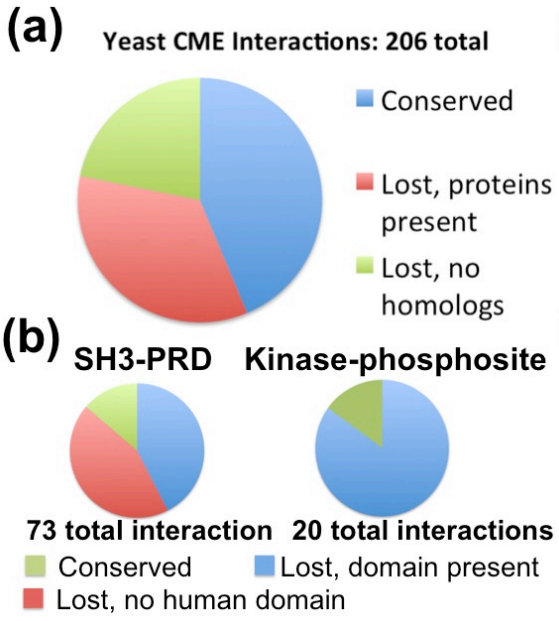

(a)
(b)
CME PPIN
Unbiased Sampling
Fitness Sampling
(c)
(d)
Modified Fitness Sampling
(e)
(f)
CME PPIN
Random PPIN
(g)
(h)
N=377
M=540
ErbB Full IIN:
N=297
M=415
Reduced:

Figure 2.6. IIN structures from distinct sampling approaches have distinct structures. The CME PPIN in both **a** and **b** is identical, but the number and distribution of interfaces on the proteins is different due to **(a)** unbiased sampling of interface networks and **(b)** fitness sampling of the interface networks. **c, d.** The interface interaction networks (IINs) of A and B are shown separated from the protein network. Unbiased (random) sampling of interface networks in **(c)** and fitness sampled result shown in **(d)**. Unbiased sampling shares no features in common with the biological IINs of Fig 2.1, but with fitness sampling we can reproduce nearly all the properties of the biological IINs. **(e)** Modified fitness sampling produced similar results for the CME PPIN. **(f)** However, on the random PPIN, modified fitness sampling limits on total interfaces resulted in a significant number of triangle motifs (red nodes). Hub nodes (k>7) are colored blue. **g,h** The ErbB IIN with domains colored as in main text has similar properties whether repeated interfaces are kept separate **(g)** or grouped **(h).** Network figures were all prepared with Cytoscape [99], and site graphs required the AutoAnnotate App. Interactive files are available from our website: https://hollandnetworkmotif.wordpress.com along with associated data files.

2.2.6 Network Rewiring Maintains Selectivity. Our results imply that selectivity in interface interactions is highly conserved across various protein networks. Therefore, if we compare IINs across evolution, we should find that rewiring of interactions between species is not random (as they are treated in growth models) but correlated and constrained to maintain this selectivity. Orthologous proteins with similar domain sets may change protein interactions but should preserve domain partners, as has been experimentally observed in SH3 domain interactions between worms and yeast[100]. By comparing the yeast CME PPIN with a human CME PPIN constructed (Methods section 2.4.7) from 64 proteins with recognized functional homology [101], we find that rewiring events are highly correlated and attributable to specific binding domains (Fig 2.7). From yeast to human, about half of the interactions are conserved. Of those that are lost, 39% are due to lack of a homologous protein, and 98% of the remainder involved at least one domain that retained no interaction partners (Fig 2.7c). A major source of divergence was

domains targeting the linear motif proline rich regions (PRRs) and phospho-sites (Fig 2.7b). SH3-PRR interactions accounted for over half the losses from yeast to humans. The divergence of these interactions can be attributed to the biological distinctions between yeast and metazoan CME: in yeast the actin cytoskeleton is required to deform the stiffer cell membrane and the SH3 containing proteins link the cytoskeleton to the clathrin-coated vesicle[101]. New interactions gained within the human PPIN were concentrated in a few proteins, most significantly in the AP-2 complex. The source of these new interactions is an added domain to the human AP-2 complex that interacts with a range of diverse binding partners[102]. Without this hub domain, the yeast AP-2 complex evolved with few binding partners, accounting for the minimal interaction conservation between the homologs.



(c)

Domain 1→ Domain 2 ↓	No Human homolog	Domain loses all partners	Domain loses some partners
No Human homolog	3	25	17
Domain loses all partners		24	45
Domain loses some partners	Yeast → Human 116 interactions lost		2

(d)

Yeast Protein	Human homolog	Inter-actions in yeast	Inter-actions in human	Conserved Inter-actions
ABP1	DBNL	16	2	1
MYO3	MYO1E/F	10	3 (1E)	1
MYO5	MYO1E/F	14	1 (1F)	1
YSC84	SH3YL1	14	0	0
LSB3	SH3YL1	14	0	0

Figure 2.7. Network rewiring between yeast and human CME networks is correlated and controlled by specific domains. (a). Comparison of the CME interactome of 56 yeast proteins with that of their 64 human homologs reveals the majority of interactions are either conserved or lost from yeast to humans due to a missing homolog in the human network. Analysis of changes from the human to the yeast interactome in Fig C.5. **(b).** The interactions present in yeast and absent in

humans were highly correlated, with most being absent due to a full protein homolog being absent, or a domain losing all binding partners. **(c).** Most absent interactions involved SH3 and proline rich region (PRR) interactions, or kinase-phosphosite interactions, highlighting the fluidity of linear motif driven interactions. (**d).** Some yeast proteins conserved almost no interactions with human counterparts, and these proteins contain SH3 domains.

2.2.7 Hub Interfaces in the CME and ErbB Networks Are Strongly Conserved.

Our results also emphasize the importance of hub interfaces to avoid the need for new domain innovation. We thus predict hub interfaces should be preferentially conserved throughout evolution. With all the domain information available for the two manually curated networks (Fig 2.1), we can isolate the contribution of hub interfaces to hub protein evolution. Hub proteins may evolve more slowly[103], and one rationale (among other[104]) is that it is harder to change with so many binding partners. However, a conflicting observation is that hub proteins also have more disordered regions[105], which evolve more rapidly[106]. Furthermore, a distinction between evolutionary rates of different hub types (date vs party hubs) may actually be attributable to expression levels[107,108], which, along with number of translational events[109] are the strongest predictors of evolutionary rates[110]. Our analysis (Methods 2.4.6) of residue conservation demonstrates that hub interfaces (defined in two independent ways) are significantly more likely to be conserved than other binding interfaces, with almost 90% being strongly conserved, compared with 70% of non-hub interfaces (Table C.2). Because we evaluate conservation on both hub and non-hub interfaces of the very same proteins, the effects of protein expression level variation on conservation are explicitly included. Whether a protein has high or low expression, its hub interfaces are more strongly conserved than its non-hub

interfaces. It is the interfaces that bind to the hub interfaces that are more likely to have weaker conservation (Table C.2), hence facilitating the rewiring of hub interfaces. This analysis thus suggests how hub proteins can participate in more rewiring events[95] while still evolving slowly: the partners evolve to achieve binding.

Because hub proteins have more disorder, and because most interactions result from the binding of a structured domain (e.g. SH3, kinase) to a short motif (e.g. PRR), disordered regions may be a source of novel interactions for hubs. In the CME network, the hub LAS17 interacts with PRRs or acidic domains for 78% of its interactions. ABP1 uses disordered interfaces for ~46% of its interactions. In contrast, the kinase PRK1 uses its structured kinase domain for ~83% of its interactions. In the ErbB signal transduction network, protein hubs either have several phosphosites (e.g. EGFR) or a kinase or SH2 domain that binds to several such sites (e.g. MAPK1, PIK3R1). Thus, hub proteins exist on a stratum between having several unstructured binding regions and having a few versatile structured binding domains. Figure C.6 shows the correlation between number of interfaces and percent of interactions mediated by disordered regions in hubs (k>9), with R=0.67 and R=0.61 for the CME and ErbB IIN respectively.

2.3 Discussion

PPINs feature a scale-free-like topology. Much like airport networks, a few proteins act as hubs, while the majority of proteins have only a few interaction partners. Stochastic growth models[111-113] provide a simple explanation for how protein networks evolve towards a scale-free topology. Hubs are generated via protein

genes duplicating and diverging[111,114], where at least one of the duplicated proteins retains an original interaction as they sub-functionalize[115,116]. While gene duplication and divergence are undoubtedly sources of evolutionary change of protein interactions, the network growth models of duplication and divergence have an unrealistic portrayal of rewiring, usually performing only one rewiring per duplication event, and without incorporating any physico-chemical or evolutionary basis for the rewiring. Rewiring happens on a much faster evolutionary timescale than gene duplication: the human interactome has been estimated to rewire 1000 times per million years[95,114], whereas gene duplication is estimated to occur at a rate of 2 to 30 events per million years[117,118] (assuming 20,000 genes), with the majority of these duplications being deleted by natural selection[119]. Orthologous proteins are often highly rewired, as a recent study comparing the yeast and worm SH3 interactome found[100]. Additionally, growth models ignore homo-dimers despite their prevalence[120] and influence on evolving new interactions[121].

Biological rewiring is capable of abolishing the majority of interactions from one species to another[122], and creating and destroying interactions involving transcription factor[123] and protein hubs such as AP-2[124] between species[101]. If the rewiring were random, it would destroy any scale-free structure created by gene duplication. Yet scale-free topology is conserved, and this suggests rewiring is not random and hubs are preferentially conserved[114]. A scale-free topology is known to provide benefits relative to a random network in that it fortifies communication across networks by centralizing connections into hubs[93]. We propose that our results provide another advantage of hubs in PPINs: they improve binding

selectivity and thus avoid misinteractions. This selection pressure is of molecular origin and reflects the primary physico-chemical requirements of proteins to fold into stable structures and bind to other molecules. Hub proteins allow the creation of hub interfaces, which facilitate chemical and structural complementarity and selectivity with the fewest number of interfaces needed.

We note that the actual IINs were not the most optimal solutions in any fitness landscape. Raising the temperature allowed us to sample more randomized versions of the optimal solutions, but the real IINs departed from the optimum in specific, rather than random ways, suggesting additional selective pressure acting on the network structure. In particular, the observed IINs had a smaller number of isolated modules. Each large module corresponds to a particular binding mode; e.g. SH3 to PRR or Ras to GEF interactions. Cells have a limited number of domain types to work with, but our model was only concerned with network topology and so created more modules. Limiting the number of modules in our sampling process could solve this. The same motif structure in fewer modules would better match the observed biological IIN structure and also mimic the limited number of domain types used by proteins. We also note that some sub-optimal interactions are not constitutive as our model treats them: they can be turned on or off by phosphorylation or allostery. This is especially true of "bridge" interfaces that connect otherwise separate modules. The ARC40 subunit of the ARP2/3 complex acts as a bridge node in the CME IIN that can be inhibited from binding actin.[125] This, as well as other functional consequences of protein interactions, may shape IIN topology. However, it is difficult to select for functional constraints without knowing

the true function of every protein in the network, and even then function is not a generic constraint; it would have to be selected for in a targeted way. It is noteworthy however that we are able to reproduce key features of the IINs without the need for incorporating protein function.

Finally, it is estimated that at least 40% of proteins bind to themselves, and the majority of these interactions involve a homo-dimer using the same interface[120]. In networks, however, these interactions produce self-loops that are often ignored when calculating network properties and simulating network growth, despite providing a justification for frequent paralog interactions in growth models[121]. They are ignored because having another unique edge type increases the combinatorial complexity of network structures, but we found here that they are critical in correctly capturing motif selectivities. This is best illustrated by the triangle motif in Fig 2.2 that switches from low to high specificity with the introduction of multiple self-interactions. The optimal selectivity for a self-binding interface is as an isolated node, or as part of a pair of hetero-dimer forming homo-dimer interfaces, as is clearly evident in the CME IIN (Fig 2.1a). Self-binding nodes are least selective as hub interfaces because suppressing non-functional interactions grows more difficult with more partners that are not self-binding. These distinctive motif preferences for self-binding interfaces present another important consideration for curating domain assignments in PPINs, in this case suggesting both potential mis-assignments and missing assignments.

2.4 Methods

2.4.1 Fitness function to sample IINs on a PPIN: Given a fixed PPIN, we used Monte Carlo sampling in the space of IIN structures with networks structures accepted or rejected via the Boltzmann weight $e^{-(f_{new}-f_{old})/k_BT}$. The four parameter ($\omega,\beta,\kappa,\mu$) fitness function given by

$$f = e^{\omega(M_{IIN}-M_{PPI})} + \sum_{i,(k_i>1)}^{N_{\text{int}}} (e^{\beta C_{i,3}} + e^{\kappa(1-C_{i,4s})} - 2) + \sum_{p}^{N_{pro}} e^{\mu N_{\text{int},p}} \qquad \text{Eq. 2.1}$$

controlled the numbers of interfaces N_{int} and edges M_{IIN} in the IINs, as well as the triangle motifs and square-to-chain motif ratio via the local clustering and grid coefficients[126] , $C_{\text{i,3}}$ and $C_{\text{i,4s}}$.

$$C_{i,3} = \frac{2N_{triangle,i}}{k_i(k_i-1)} \qquad \text{Eq. 2.2}$$

$$C_{i,4s} = \frac{1+N_{square,i}}{1+k_i^{2nd}k_i(k_i-1)/2} \qquad \text{Eq. 2.3}$$

where k_i is the degree of node "i", $k_{\text{i}}^{\text{2nd}}$ is the number of nodes two steps away from "i", and $N_{triangle,i}$ and $N_{square,i}$ are respectively the number of triangles and squares which pass through "i". A dummy square (numerator +1 term) in the grid coefficient is used to penalize having a high number of chains even when $N_{square,i}$ equaled zero. Triangles on which at least two of the nodes had self-edges were ignored, since this is not a constraint against high specificity. The fitness function penalizes having a

high clustering coefficient (many triangles), a low grid coefficient (many chains), a high number of interfaces, and it penalizes duplicating too many edges (Fig 2.8).

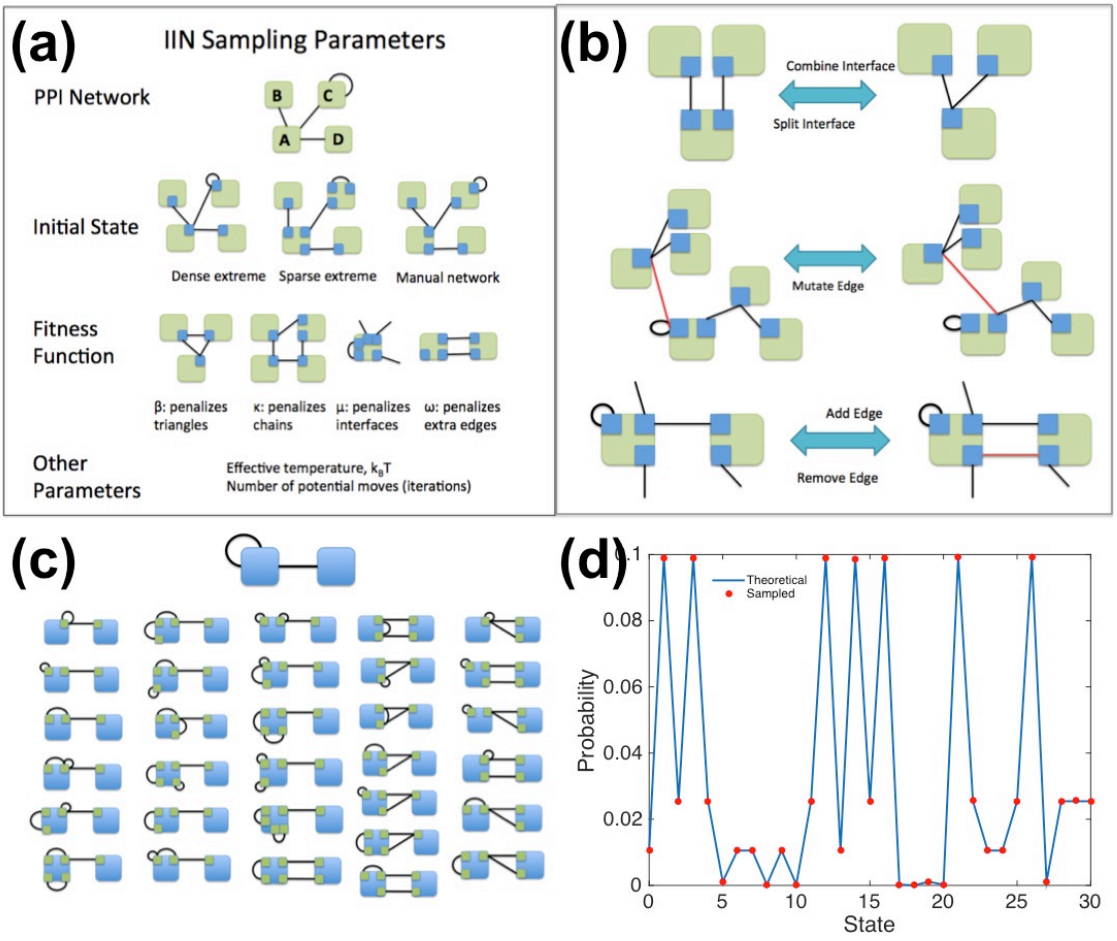


Figure 2.8. Interface networks for a given protein network can be sampled via Monte Carlo methods with or without bias. (a) Inputs and parameters for our stochastic IIN sampling model for a given PPIN that is not altered. (**b)** Monte Carlo reversible move sets (5 moves possible) to transition between IIN structures. (**c)** A two protein network with 2 PPIs can be enumerated as 31 distinct IINs when one extra edge is allowed. Moves between states were enumerated as a Markov chain to determine the factors necessary for detailed balance. (**d)** Proof of detailed balance in the toy model (C). The probability of being in a given state is proportional to its propensity $e^{-f/kBT}$, where "f" is the assigned fitness penalty (low "f" = more fit) and k_BT is set to 2. The blue line is the theoretical stationary distribution based on propensities, and the red circles are the MC sampled results.

2.4.2 Monte Carlo sampling of networks We first initialized the IIN structure to either the dense extreme (one interface per protein), the sparse extreme (new interface per each edge), or the known IIN structure. Moves (illustrated in Fig 2.8) were accepted or rejected based on the Boltzmann criteria, where we were careful to ensure detailed balance given the different probabilities of generating forwards and reverse moves (p^{gen}) via the acceptance probability:

$$p^{accept}_{IIN_{old}\rightarrow IIN_{new}} = \min(1, \frac{p^{gen}_{reverse}}{p^{gen}_{forward}} e^{-\frac{(f_{new}-f_{old})}{k_B T}}) \quad \text{Eq. 2.4}$$

where f is the fitness of the IIN defined in Eq. 2.1, and $k_B T$ is the effective temperature. We verified our implementation for a small test network in Fig 2.8. The entire space of possible IINs could be sampled by setting $k_B T=\infty$. For the fitness sampled IINs, we found a range of $k_B T$=0.1-1 to be optimal. Modified versions of sampling to test the robustness of our network properties are described in Appendix A.

Simulations were allowed to equilibrate for the first 1/5 of the total number of iterations, (usually ~1 million iterations) after which the statistics of each network sampled was recorded so as to record average statistics favored by the fitness function. The best-fit (lowest fitness penalty) network discovered was also recorded.

2.4.3 Quantifying network degree distributions We generated the spectrum of networks ranging from homogenous to scale-free using a single parameter (α) via

the method of Goh et al [127]. We term this parameter α the "preferential attachment exponent" (PAE) of the network. A PAE=0 corresponds to a Poisson (random) network with λ=<*k*>, and PAE=1 roughly corresponds to a power-law (scale-free) network with γ=2. We reverse fit the degree distribution of our sampled IINs by generating networks with specific P.A.E.s for comparison. Degree distributions for 11 values of the P.A.E. (0, 0.1, 0.2 … 1) were generated by building 30 networks (per P.A.E.) with the same number of nodes and edges as the IIN. Least χ^2 distance was used to choose the best-fit P.A.E. for the degree distribution of the given IIN. We had to modify the algorithm of Goh et al [127] to generate networks that did not contain orphan nodes, and this procedure is detailed in Appendix A.

2.4.4 Statistic for identifying 'date' vs 'party' hubs The distribution of interfaces for a protein is calculated by normalizing the Stirling numbers of the second kind (see Appendix A for definitions). We use this probability distribution to generate a statistic for identifying proteins with an unusually high (party hubs) or an unusually low (date hubs) number of interfaces. For a protein with degree *k* and *U* interfaces, we can calculate a *p*-value using a two-tailed test, given by

$$p\text{-value} = \Pr\left(t \le \frac{(k+1)}{2} \;\middle|\; U \le \frac{(k+1)}{2}\right) + \Pr\left(t \ge \frac{(k+1)}{2} + \middle|\; U \ge \frac{(k+1)}{2}\right) \qquad \text{Eq. 2.5}$$

where *t* can take only integer values [1:k]. If $U=(k+1)/2$, *p*-value$\equiv 1$.

In Table C.3 we report these *p*-values per protein, indicating which proteins have an unusually small or large number of interfaces.

2.4.5 Generation of alternate PPIN structures Five variations of the CME network [66] were used to test PPIN constraints on IIN sampling: a "dense" network with the same P.A.E. where 186 edges were added to the existing CME network, a "sparse" network also with a comparable P.A.E. where 93 edges were deleted, and a random version of each of the preceding three networks with the same number of proteins and PPIs using the Erdos-Renyi algorithm[128]. Finally, a random version of the ErbB PPIN [26] was also used.

2.4.6 Phylogenetic analysis of yeast CME proteins and human ErbB proteins To determine the evolutionary conservation of domains in the 56 yeast CME proteins and 127 human ErbB proteins, we collected orthologs of each protein, ran multiple sequence alignments with MAFFT [129], and analyzed residue conservation with the ConSurf [130] rate4site program (or web-server). To assign a conservation score to each domain, the average over all residues in the domain were taken. Orthologs were constructed from BLAST [131] searches against the UniRef90 clustered sequence database with an E-value cutoff of 0.0001. This approach to use BLAST searches against UniRef90 to identify orthologs across all species is the same as used in other conservation calculation approaches [130,132]. Consistent with these approaches[132], we kept only sequences that were similar in length to the query sequence (25% longer or shorter) and shared sequence identity of 35%-95% before performing the multiple sequence alignment (MSA).

Hub interfaces were defined in two independent ways: firstly, as any interface with 5 or more interactions. Secondly, we used the statistic defined in Eq 5

to identify proteins with an unusually low number of interfaces given their connectivity, implying the presence of hub interfaces. The statistics were almost identical, with 89% and 71% of hub and non-hubs, respectively, being more conserved than average.

2.4.7 Network rewiring between yeast CME proteins and human CME proteins

We constructed the CME interaction network for human homologs of the yeast proteins using the review of Weinberg et al [101] as a guide to functional homologs in metazoans. Most human homologs were identified directly from this review[101], and in a few cases we supplemented this with human orthologs identified from the EggNOG database [133], which were supported by BLAST searches of the yeast proteins against exclusively human proteins. Nine yeast proteins lacked human homologs (as was previously documented[101]) and the remaining 45 yeast proteins were matched with 64 human homologs. Interactions between these 64 proteins were then extracted from BioGRID. We also added 9 interactions involving actin or the Arp2/3 complex and removed 11 involving the Arp2/3 complex to be consistent with the publications used to make the interface assignments in yeast [66] that involved crystal structures of metazoan homologs.

The yeast CME network contained 18 PPIs that were mediated through multiple duplicate binding modes (Fig C.5a). These interactions were found to be slightly more conserved than single binding mode interactions, with 9 conserved interactions, 4 lost due to a lost homolog, and 5 lost despite both proteins retaining homologs and domains.

Further Methods may be found in Appendix A. Manually curated network interactions may be found in Appendix B. Further figures and data may be found in Appendix C. Code for network sampling and analysis is available from github https://github.com/mjohn218/network_sampling_MC

Chapter 3. Stoichiometric Balance of Protein Copy Numbers Is Measurable in a Protein-Protein Interaction Network for Yeast Endocytosis

Chapter adapted from:

Holland, DO, & ME Johnson (2018) "Stoichiometric balance of protein copy numbers is measurable and functionally significant in a protein-protein interaction network for yeast endocytosis." In revision at PLoS Comp. Biol. Available at bioRxiv https://doi.org/10.1101/205674

Stoichiometric balance, or dosage balance, implies that proteins that are subunits of obligate complexes (e.g. the ribosome) should have copy numbers expressed to match their stoichiometry in that complex. Establishing balance (or imbalance) is an important tool for inferring subunit function and assembly bottlenecks. We show here that these correlations in protein copy numbers can extend beyond complex subunits to larger protein-protein interactions networks (PPIN) involving a range of reversible binding interactions. We develop a simple method for quantifying balance in any interface-resolved PPIN based on network structure and experimentally observed protein copy numbers. By analyzing such a network for the clathrin-mediated endocytosis (CME) system in yeast, we found that the observed protein copy numbers were significantly more balanced in relation to their binding partners compared to randomly sampled sets of yeast copy numbers. The observed balance is not perfect, highlighting both under and overexpressed proteins. We evaluate a potential cost to imbalance in the form of misinteractions between

'leftover' proteins without remaining functional partners. We find that networks with biological features have lower misinteraction frequency under balanced concentrations but higher misinteraction frequency under imbalanced concentrations. This suggests that evolution favors balanced protein abundance and that any conserved imbalance should occur for functional reasons. Strong-binding proteins are also susceptible to misinteractions regardless of balance, suggesting an upper limit on protein binding strengths as well.

3.1 Introduction

Protein copy numbers in yeast vary from a few to well over a million per cell[1,2]. Expression level, binding partners, and corresponding affinities, reflect a protein's function within the cell. In the context of multiprotein complexes – especially obligate complexes such as the ribosome – protein concentrations are thought to be balanced according to the stoichiometry of the complex. This is referred to as the dosage balance hypothesis (DBH)[3-5].

For obligate complexes, dosage balance means that there are no leftover subunits, as these would be a waste of cell resources. However, even for proteins in non-obligate complexes a number of deleterious effects could be caused by imbalance. An overexpressed core or "bridge" subunit may sequester peripheral subunits, paradoxically lowering the final number of complete complexes[4,16]. Excess proteins may be prone to misinteractions, also called interaction promiscuity, with nonfunctional partners. Numerous studies have identified proteins with high intrinsic disorder as sensitive to overexpression[9,27,28], and these proteins have low,

tightly regulated native expression levels[30,49] indicating that misinteraction propensity and abundance are related. Underexpression carries its own dangers: a single underexpressed subunit is a bottleneck for the whole complex. In addition, copy numbers of weakly expressed proteins are noisier[18] and thus less reliable for the cell. Male (XY) animal cells are known to employ "dosage compensation" mechanisms to increase the expression of X-chromosomal genes to be on par with female cells[12,19], though for other genes it is the female cell that inactivates one of the X chromosomes[134], indicating that the cell preserves an optimized set of expression levels.

Optimized does not necessarily mean balanced, however. Imbalance may be necessary for functional reasons: signaling networks utilize underexpressed hubs to regulate which pathways are active at a given time[26]. Recent models show imbalance can be beneficial to complex assembly when affinity and kinetics are taken into account[31,32]. A study of over 5,400 human proteins by Hein et al. found that strong interactions forming stable complexes are correlated with balance, but weak interactions are not, which may mean that the network as a whole is not balanced [10]. Finally, the DBH relies on the assumption that proteins reach an equilibrium state of complex yield, but few things in the cell are at equilibrium and deviations from balance could have benefits in non-equilibrium models.

Here, we test the hypothesis that protein expression levels are significantly biased towards balance, even for complex PPINs that include weak and transient interactions. This first required us to develop a method to quantify stoichiometric balance in an arbitrary PPIN, given known binding interfaces and some observed

copy numbers. Copy number correlations thus are evaluated beyond direct binding partners to the more global network of interactors. We then can quantify the consequences of imbalance relative to perfect balance according to two criteria: 1) the deleterious consequences and cost of forming misinteractions, and 2) the potentially beneficial control of specific functional outcomes by modulating which complexes, given known binding affinities, actually assemble. Applied to the 56-protein, manually curated, interface-resolved CME PPIN [66], two of its sub-networks, as well as the ErbB PPIN[26], we find that stoichiometric balance in observed copy numbers is often significant, and observed imbalances, particularly of underexpressed proteins, could permit fine-tuning of functional outcomes.

One consequence of imbalance we evaluate, misinteractions cost, has an indirect effect on function by allowing unbound proteins to bind to non-functional partners, sequestering components and thus affecting formation of specific complexes. Misinteractions are believed to cause dosage sensitivity[25,27,28], and avoiding them has been shown to be an evolutionary force limiting protein diversity[36,39], expression levels[51,54], binding strengths[50], and protein network structure[39,71]. Misinteractions, not being selected for by evolution, are weak and generally unstable, but there are far more ways for proteins to misinteract than bind to their few functional partners [36,39]. Cells have evolved a variety of mechanisms to increase specificity, such as allostery[48,135], negative design[37,40], compartmentalization[36], and temporal regulation of expression[65]. Copy number balance would be another such mechanism, as protein binding sites would saturate their stronger-binding functional partners

Quantifying balance in protein networks can lead to new insights, as unbalanced proteins may serve as assembly bottlenecks, or maintain alternate cellular functions outside of the network module being analyzed[31]. Dosage balance is also important for understanding dosage sensitivity[5,25], a phenomenon where overexpression of a gene is detrimental or even lethal to cell growth. Studies estimate ~15% of genes in *S. cerevisiae* are dosage sensitive[8,9], and negative effects of gene overexpression have been observed in several eukaryotic species including maize[5], flies[11], and humans[13-15]. Studying balance at a network-wide level is challenging because it requires resolved information about the interfaces proteins use to bind. A protein that binds noncompetitively with two partners requires equal abundance to its partners. But if the binding is competitive – i.e. the same interface is used to bind two different partners – the protein's abundance must equal the sum of that of its partners to have no leftovers. Classic protein-protein interactions networks (PPINs) lack this resolution, but recent studies have begun to add this information, creating what we refer to as interface-interaction networks (IINs)[26,66,72]. An IIN tracks not just protein partners but also the binding sites that proteins use to bind.

Our study of stoichiometric balance in larger, interface resolved PPINs is organized in three parts. In section 3.2 we define a metric for quantifying stoichiometric balance and how noise in protein expression levels can incorporated. We apply this metric to the CME PPIN [66,72] and the ErbB PPIN [26], highlighting which proteins are over- and underexpressed relative to perfect balance. Although this analysis excludes temporal expression and binding affinity, it provides a starting

point for the analysis of these features in the subsequent sections. In section 3.3, we switch to generalized interface-interaction network (IIN) topologies and network motifs to focus exclusively on how our first evaluation criteria, the cost of misinteractions under imbalance, is worse for strong binding proteins and for network topologies that resemble biological networks. Finally, in chapter 4, we will return to the interface-resolved CME PPIN to evaluate the observed degree of stoichiometric balance in two smaller sub-networks of the CME network.

3.2 Balancing Interface-Resolved Protein Networks

3.2.1. Stoichiometric balance is measureable in large PPINs when interfaces are resolved

For a multi-subunit complex such as the ribosome or ARP2/3 complex, all subunits bind together non-competitively to assemble a functional complex. Stoichiometric balance is defined as sufficient abundance of each subunit to form complete complexes, with no subunit in excess (see Fig C.7). Quantifying balance in a general protein-protein interaction network is more challenging because some proteins will bind competitively, using the same interface for multiple interactions. Such proteins require higher concentrations to saturate their binding partners. Thus, to establish stoichiometric balance in a PPIN the binding interfaces must be known. In previous work we analyzed several interface-resolved PPINs, including the 56-protein clathrin-mediated endocytosis (CME) network in yeast [66,72], and the 127-protein ErbB signaling network in human cells[26].

To balance a network, a number of desired complexes may be assigned to each edge and then the number of required interface copies directly solved for. This is constrained with a starting set of copy numbers, C_0, otherwise the solution would be arbitrary. However, the inclusion of multiple interfaces per protein introduces a new constraint: interfaces of the same protein should have identical copy number. This constraint often makes nontrivial solutions (i.e. when none of the proteins are set to zero) impossible (see Methods section 3.5.1). Therefore, we treat it as a soft constraint, using a parameter "α" to balance its influence. A high α allows more variation of interface copy numbers on the same protein. We constructed and minimized an objective function using quadratic programming (Methods 3.5.1), which produces a new, optimally balanced set of copy numbers, C_{balanced}. For any given interface-resolved PPIN, there may be multiple locally optimized solutions of balanced copy numbers.

The benefit of this method is that the distance between C_0 and C_{balanced} gives you a relative estimate of how "balanced" C_0 already is, and thus a metric from which to evaluate the significance of balance in the observed copy numbers. Using real copy numbers taken from Kulak et al.[2], C_{real}, as C_0, we calculated both chi-square distance (CSD) and Jensen-Shannon distance (JSD) between C_{real} and C_{balanced} (Methods 3.5.3). The CSD measure looks at differences between absolute values and penalizes high deviations more strongly than low deviations, whereas JSD converts both vectors to distributions and measures the similarity between them. We do not expect any networks to have C_{real} that is already perfectly optimized, such that $C_{\text{real}}=C_{\text{balanced}}$. To establish the significance of both distance metrics, we generated

5,000 sets of random C_0 vectors, sampled from a yeast concentration distribution. We then measured the CSD and JSD from C_0 to C_{balanced} for each of these random copy number vectors. If C_{real} is balanced, its distance metrics should have significant p-values relative to yeast copy numbers selected randomly from the yeast distribution.

3.2.2 Accounting for noise in observed copy number measurements

Even constitutively expressed genes do not have a constant abundance; they vary due to both extrinsic and intrinsic noise[136]. Taniguchi et al. found that the abundance of a single protein in *E. coli* fits a gamma distribution[18]. Therefore, one reason copy number balance should not be expected to be perfectly matched is due to inherent fluctuations in protein copy numbers. Our algorithm, however, ultimately assigns a single copy number to each interface in the network to optimize perfect balance, when realistically a distribution of values would be more appropriate.

Our method does provide one mechanism to allow a range of copy number values for a single protein, and that is through allowing interfaces on a single protein to have non-identical values. This range can be tuned through our parameter α, which biases solutions towards equivalent interface copies per protein when set to zero. As the α parameter increases, more variation is permitted. For example, one interface may be assigned 200 copies and another on the same protein 300 copies. If the protein is usually expressed within the 150-350 copy range, this solution is more realistic than enforcing both copy numbers to be exactly 250.

We therefore systematically characterized how variations in α changed the "noise", or variability in interface copy numbers on each protein. Taniguchi et al. found that yeast proteins with high abundance (~1,000 or more copies) had a noise (σ^2/μ^2) upper limit of about 0.5 with ungated data (i.e. without background noise removed) and 0.1 with gated data[18]. For α≤0.03, we found that proteins with mean interface copy numbers above 1,000 had noise less than 0.1, indicating that such a solution is possible. (Fig 3.1). Low abundance proteins exhibit higher noise in terms of expression level[18,137], and this feature is also observed in our model. We therefore used values of α in the 0.01 to 2 range based on this analysis.

3.2.3 Protein copy numbers in yeast clathrin-mediated endocytosis are balanced

As Fig 3.2a,b shows, at α=1 the p-value for JSD was found to be statistically significant (p=0.0054) but the p-value for chi-square distance was not (p=0.157). We analyzed the real copy numbers before and after balancing and found that the protein cofilin was highly overexpressed (Fig 3.2c) meaning that it had to be greatly lowered to achieve balance. This resulted in a skewed CSD for C_{real}, which the change in cofilin dominated. We therefore re-tested balance when cofilin was removed from the network. At α=1, both JSD (p=0.0012) and CSD (p=0.022) were statistically significant (Fig 3.2d), indicating that these 55 proteins are balanced compared to random copy numbers. These results were robust to changes in α, but the p-values tended to be lowest when α was in the 0.01 to 2 range. The absolute distance from C_{real} to C_{new} decreased as α was raised, plateauing when α≥10.

Because protein complexes that strongly bind are thought to be more balanced than complexes involving weak interactions, we repeated the analysis on

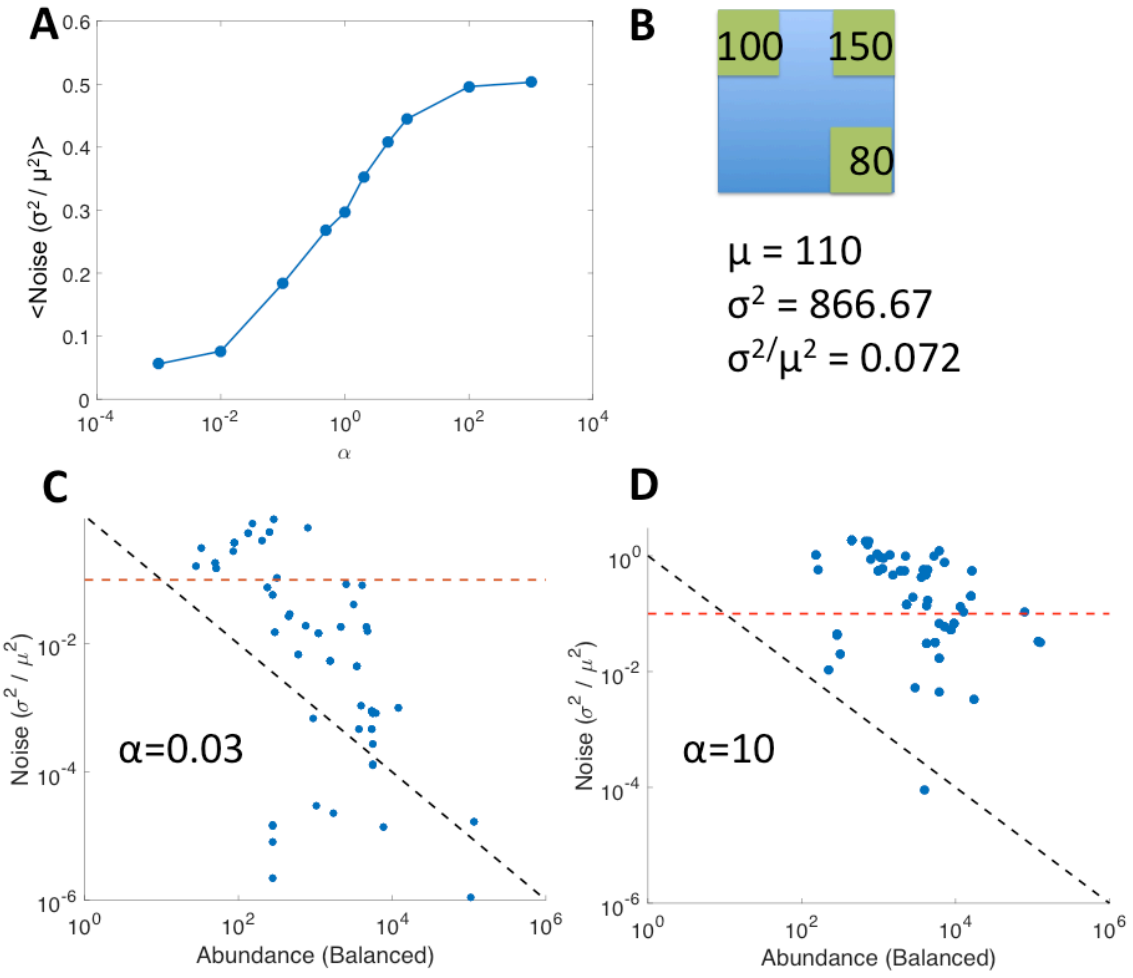


Figure 3.1. Effects of the α parameter on interface copy number noise. (A) Noise is calculated as the variance of the copy numbers assigned to interfaces on the same protein divided by the square of their average copy number. It does not refer to expression level noise. A high "α" parameter allowed greater variance, but even a low α could not remove noise entirely because there are no balanced solutions where all proteins can have interfaces of equal copy number. Noise had a sigmoidal relationship with log(α). **(B)** Example protein interface noise. **(C,D)** Scatter plot of protein interface copy number noise vs a protein's balanced "abundance", the average of their interface copy numbers. The black line is where noise is inverse of abundance. The red line is noise = 0.1, which is expected to be the upper limit of noise when abundance exceeds ~1000 copy numbers[18]. For a low α, proteins varied widely in the amount of noise they have, though high-abundance proteins tended to have less noise, and were below the 0.1 threshold. As α was raised, proteins approached the same level of noise.

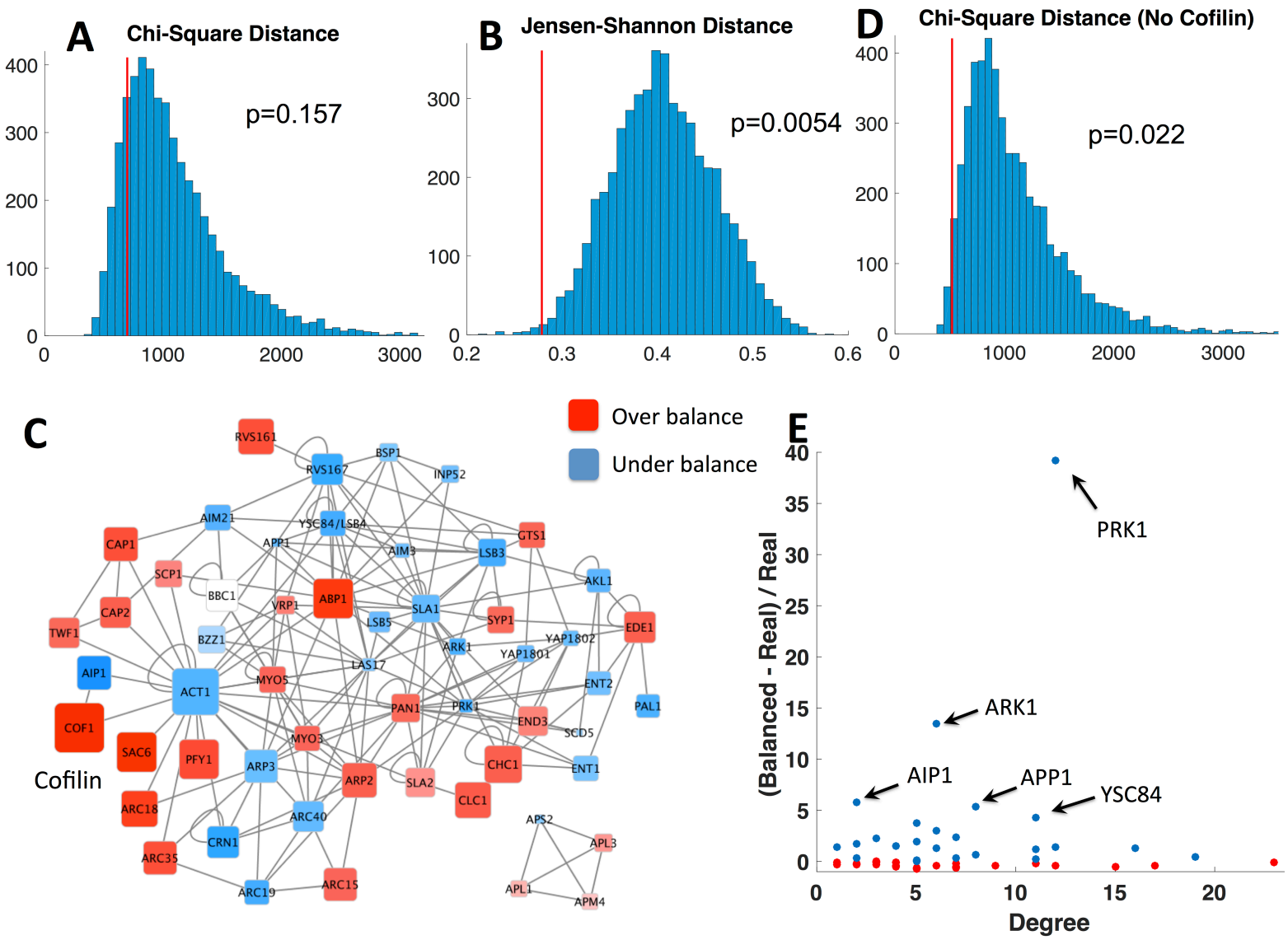

A
Chi-Square Distance
p=0.157
B
Jensen-Shannon Distance
p=0.0054
D
Chi-Square Distance (No Cofilin)
p=0.022
C
Over balance
Under balance
Cofilin
RVS161
RVS167
BSP1
INP52
AIM21
YSC84/LSB4
APP1
AIM3
GTS1
LSB3
CAP1
SCP1
BBC1
ABP1
AKL1
CAP2
VRP1
SLA1
LSB5
SYP1
EDE1
TWF1
BZZ1
YAP1802
ARK1
YAP1801
AIP1
LAS17
ACT1
MYO5
ENT2
PAN1
PRK1
END3
PAL1
COF1
SCD5
MYO3
SAC6
PFY1
ARP3
CHC1
ENT1
ARP2
SLA2
ARC18
CLC1
ARC40
APS2
APL3
CRN1
ARC35
APL1
APM4
ARC19
ARC15
E
(Balanced - Real) / Real
Degree
PRK1
ARK1
AIP1
APP1
YSC84

Figure 3.2. Clathrin-mediated endocytosis proteins are balanced. **(A,B)** Histograms for chi-square distance and Jensen-Shannon distance between the observed protein copy numbers and their copy numbers after balancing. Compared to 5,000 sets of random sampled copy numbers, the real copy numbers had a statistically significant Jensen-Shannon distance, but not chi-square distance. **(C)** Graph of CME network, showing which proteins were overexpressed (red) or underexpressed (blue) compared to the balanced copy numbers. Cofilin was highly overexpressed, which led to a high chi-square distance. **(D)** Histogram for chi-square distance when cofilin was removed from the network. It is now statistically significant, indicating that the other 55 proteins are balanced compared to random copy numbers. **(E)** The five most underexpressed proteins were two kinases (PRK1 and ARK1), one phosphatase (APP1), and two partners of Actin (AIP1 and YSC84). The former three bind transiently to their partners with no functional need for balance. The latter two are discussed in the text.

the full 56-protein network after removing one of two modules from the network: the four protein subunits of the AP complex, and the seven proteins in the ARP2/3 complex. Without the former, the p-value increased to 0.0088 for JSD and 0.197 for CSD, indicating less overall balance. Removing only the ARP2/3 complex similarly increased the p-values to 0.023 and 0.24. This trend held when cofilin was also removed.

The four AP subunits that form the obligate AP-2 complex are fairly close in abundance, as are the clathrin heavy chain and clathrin light chain proteins, which is consistent with the pressure for strong binding proteins to be more tightly balanced.

3.2.4 Stoichiometric Balance Is Not Measured Without Proper Interface Binding Interactions

To test whether balance depended mostly on protein network structure rather than the child interface interaction network (IIN) structure, we repeated our analysis using random IINs for the same parent protein network, again excluding cofilin. We

randomized whether proteins bind competitively or noncompetitively, using a rewiring method from Holland et al.[72]. For 20 random IINs, we found that the real copy numbers were significantly less balanced. For α=1, the same analysis obtained p-values of 0.44 ± 0.12 for CSD and 0.24 ± 0.13 for JSD. Thus the protein copy numbers are balanced according to the underlying interface network.

3.2.5 Observed protein imbalances can highlight functional relationships

Finally, by looking at the relative change between C_{real} and C_{balanced}, we could examine which proteins are underexpressed in the network relative to the predictions of balance. As Fig 3.2e shows, the five most underexpressed proteins are PRK1 (by a factor of nearly 40), ARK1, AIP1, APP1, and YSC84. PRK1 and ARK1 are both kinases; they form transient interactions with their partners for the purpose of phosphorylation. Since a single kinase can phosphorylate many proteins relatively quickly, rather than form stable complexes with each target, there is a sensible functional explanation for why these proteins can be underexpressed relative to their partners by such a large margin. Similarly, APP1 is a phosphatase. The protein AIP1 is an actin binding protein that targets a binding surface of actin without any competition from other actin binders, and also binds the highly expressed cofilin. Its low abundance relative to actin and cofilin could indicate it acts as a bottleneck in regulating cofilin-actin interactions, or perhaps more simply, that functionally it is not needed at a 1:1 stoichiometry with the highly abundant actin protein. YSC84 has 13 binding partners, and 10 of these partners all bind the YSC84 SH3 domain, including the relatively highly expressed ABP1. Although many of these binding

partners (all proline rich regions-PRRs) also have additional partners of their own, ABP1's PRR is specific to YSC84's SH3 domain[72]. As we return to in the discussion, underexpression could indicate a functional regulatory role for this protein, or indicate transient interactions with partners. Identifying underexpressed proteins, as well as their relatively overexpressed partners, may reveal the temporal dynamics of such proteins within the cell.

Actin is overexpressed compared to its partners, excluding cofilin, due to its primary role as a member of the cell cytoskeleton. Clathrin, another protein that polymerizes, is also overexpressed, the reasons for which are analyzed in section 3. The overexpression of cofilin is curious and Kulak et al. also found it highly expressed in HeLa cells and *S. pombe*[2]. The protein acts to sever actin filaments, without which the cytoskeleton cannot reorganize[138] and cells cannot migrate[139]. Perhaps it is cofilin's high expression that makes rapid reorganization of the cytoskeleton possible.

3.2.6 Ras and MAP3K proteins in the ErbB network are underexpressed

We applied our algorithm to another IIN from the literature: that of the 127 protein human ErbB signaling network, characterized by Kiel et al.[26]. Our algorithm optimizes copy numbers to the full network structure even if not all individual target copy numbers are available. Thus we measured the distance between the real (C_{real}) and optimized ($C_{balanced}$) copy numbers for the 115 of the 127 proteins for which we could assign expression levels from HeLa cells (Methods 3.5.2). We

compared results to copy numbers randomly sampled from a HeLa protein concentration distribution.

Because this is a signaling network where the majority of interactions are phosphorylation, we expected these transient interactions to bias the copy numbers against significant balance. However, while the JSD was not found to be significant ($p=0.274$), the CSD was ($p=0.022$). This result held when copy numbers were shuffled rather than randomly sampled (JSD: $p=0.120$; CSD: $p=0.019$). As stated above, CSD is dominated by large deviations. Thus, while the network as a whole is not balanced, there appears to be no dramatic overexpression.

The three Ras proteins (HRAS, NRAS, and KRAS) were found to be underexpressed (Fig 3.3), confirming the findings of Kiel et al. using simpler comparisons of Ras copy numbers to all binding partners [26]. Also found to be underexpressed were all five MAP3K proteins (RAF1, MAP3K1, MAP3K11, MAP3K2, and MAP3K4) in the network. MAP3K proteins are the top layer in MAPK cascades, a signaling motif consisting of three proteins (a MAP3K, MAP2K, and MAPK) occasionally bound together via a scaffold protein[17]. The membrane-bound receptors ErbB2 and ErbB3 were similarly underexpressed. These results suggest strategic underexpression of certain upstream proteins, potentially to control specific outputs from diverse inputs[26], and to amplify a signal as it travels "downstream" in a signaling network. Underexpression of upstream proteins is not a universal rule, however, and may depend on the type of interaction and the dynamics of the signaling network.

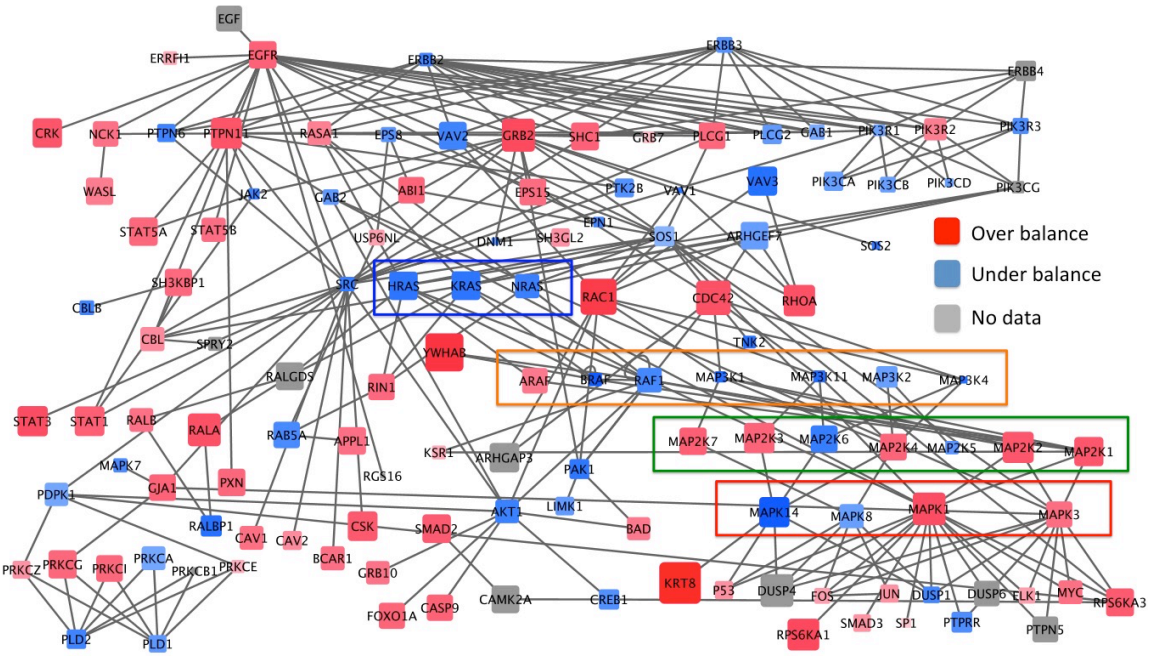


Figure 3.3. Ras and MAP3K proteins in the ErbB network are underexpressed. The ErbB network, which consists mainly of phosphorylation interactions, was not found to be statistically balanced based on the Jensen-Shannon divergence. However, certain proteins of note were found to be underexpressed, such as the three Ras proteins (HRAS, KRAS, and NRAS), and the MAP3K layer (RAF1, BRAF, ARAF, MAP3K1, MAP3K2, MAP3K4, and MAP3K11). Also underexpressed were the ErbB receptors and the hub SRC. These suggest a strategic imbalance of upstream proteins (in the case of MAPK cascades) or network bottlenecks (Ras proteins or SRC). Highlighted are the Ras proteins (blue), MAP3Ks (orange), MAP2Ks (green), and MAPKs (red). Proteins are arranged in approximate signaling cascade order – receptors at the top, targets at the bottom.

3.3 Co-optimization of Network Topology and Protein Concentrations

In this second part, we investigate how the cost of imbalance, measured solely in terms of misinteractions, depends on general properties of proteins, including binding affinity and number of binary partners. In a stoichiometrically balanced network, proteins will be driven to saturate their stronger-binding functional partners. Any “leftover” proteins, however, may misinteract, or form non-functional complexes that, while weak, are combinatorially numerous.

3.3.1 Misinteractions are minimized under balanced copy numbers and are largely independent of network motif structure

Complex formation and misinteractions must be evaluated at the level of individual protein binding interfaces, and we thus study small network motifs that have been previously characterized in real biological interface interaction networks (IINs) to control binding specificity [72]. Of these five motifs (Fig 3.4a), the hub and square motif are the most common in biological IINs relative to random networks [72]. The chain, triangle, and flag motif are selected against due to the challenges in optimizing such binding interfaces for strong selective binding and against misinteractions.[39,71,72] The motif defines the functional or "specific" interactions, which we allow at equal binding strengths. However, all other possible protein-protein interactions were allowed as misinteractions, which occur at weaker strength than the specific interactions. Because each node represents an interface (each on its own protein in this case), all binding was competitive.

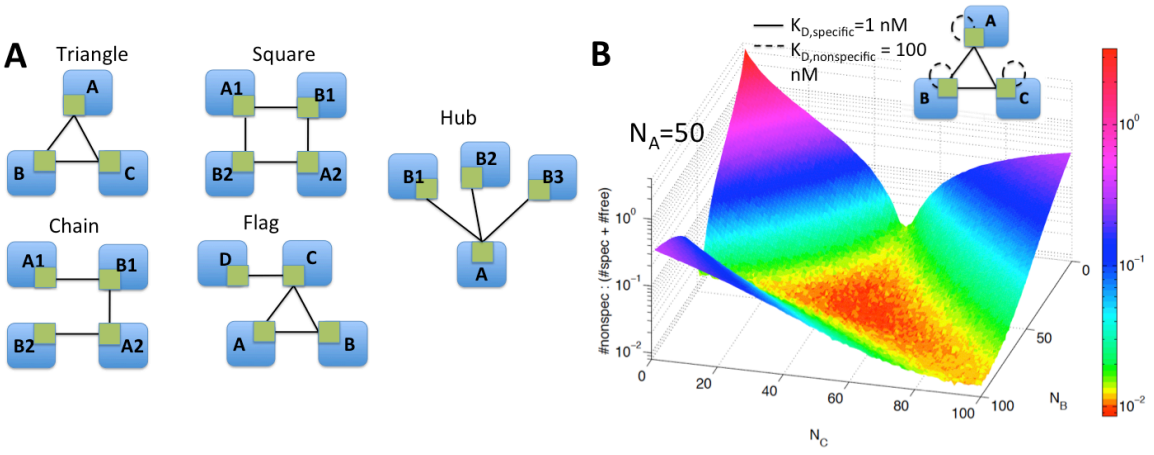


Figure 3.4. Misinteractions in network motifs from biological IINs (A) Five network motifs that have been shown to impact specificity of binding in biological IINs were tested for the effects of imbalance on misinteractions. **(B)** Surface plot obtained for the triangle network. The z-axis is the frequency of misinteractions at steady-state (Cost: Eq. 3.1) averaged across 1000 runs. The x and y axes are the number of B and C proteins; the number of A proteins is fixed at 50. As one protein becomes overexpressed, misinteractions increase exponentially.

Balanced copy numbers are relatively easy to design for these simple network motifs, and the optimization of the previous section is not necessary. We studied imbalanced copy numbers by simply varying the copy numbers of two proteins in each network over a wide range while keeping the remaining proteins constant. For each set of copy numbers, we equilibrated the system using the Gillespie algorithm[140]. We could then measure the total number of specific and non-specific complexes formed ($N_{specific}$, $N_{nonspecific}$), as well as unbound proteins (N_{free}), and use this to evaluate the cost of being out-of-balance in terms of misinteraction frequency:

$$Cost(C_0) = \frac{N_{nonspecific}(C_0)}{N_{specific}(C_0) + N_{free}(C_0)} \qquad \text{Eq. 3.1}$$

averaged across 1,000 runs, where C_0 is the vector of initial copy numbers.

The frequency of misinteractions is lowest when the protein copy numbers are balanced. Fig 3.4b shows the results for the triangle network. For example, when all three proteins have equal abundance of 50 copies, about 25 of each specific complex are formed, and minimal proteins are leftover. Cost also remains low when two proteins are equally overexpressed, as these excess proteins can bind to each other. The instances where misinteractions are the most frequent are when one protein is overexpressed, as this protein has no specific partners left and thus will

self-bind, a misinteraction for this motif. Similar surface plots were obtained for all five network motifs (See Appendix C, Fig C.8).

Notably, with balanced copy numbers, the frequency of misinteractions is almost entirely dependent on the relative strength, or energy gap, between specific and nonspecific binding (Fig 3.5a) and there was little difference among the five networks. The slope of this relationship varies slightly from one motif to another, and we confirmed that it can be calculated relatively accurately based on the ratio of specific versus non-specific interactions possible for that motif. Furthermore, the results were similar when we varied the absolute strength of specific binding from 10μM to 1 pM; under balanced conditions it affects the number of free proteins (N_{free}) relative to total complexes formed. Thus, under balanced copy numbers, the cost of misinteractions is not strongly dependent on specific binding affinities.

3.3.2 Misinteractions for imbalanced copy-numbers are worse for biologically common motifs and strong binding proteins

Unlike the similar cost of misinteractions under balanced copy numbers, the five networks noticeably differ in sensitivity to imbalanced copy numbers. In general, as copy numbers become more imbalanced, the misinteraction cost grows. To quantify this rate for each network motif, we measured the percent change in cost as one travels along the principal components away from the balanced copy numbers (Fig 3.5c; Fig C.8). The hub and square motifs were found to be the most sensitive, showing a rapid increase in cost of misinteractions as imbalance grows, whereas the flag and triangle motifs were found to be the least. (Fig 3.5d). The triangle motif has

the least sensitivity and it also has the fewest misinteractions possible; it can form 3 specific complexes and only 3 misinteracting complexes. The robustness of this module also then extends to the flag motif, which contains a triangle.

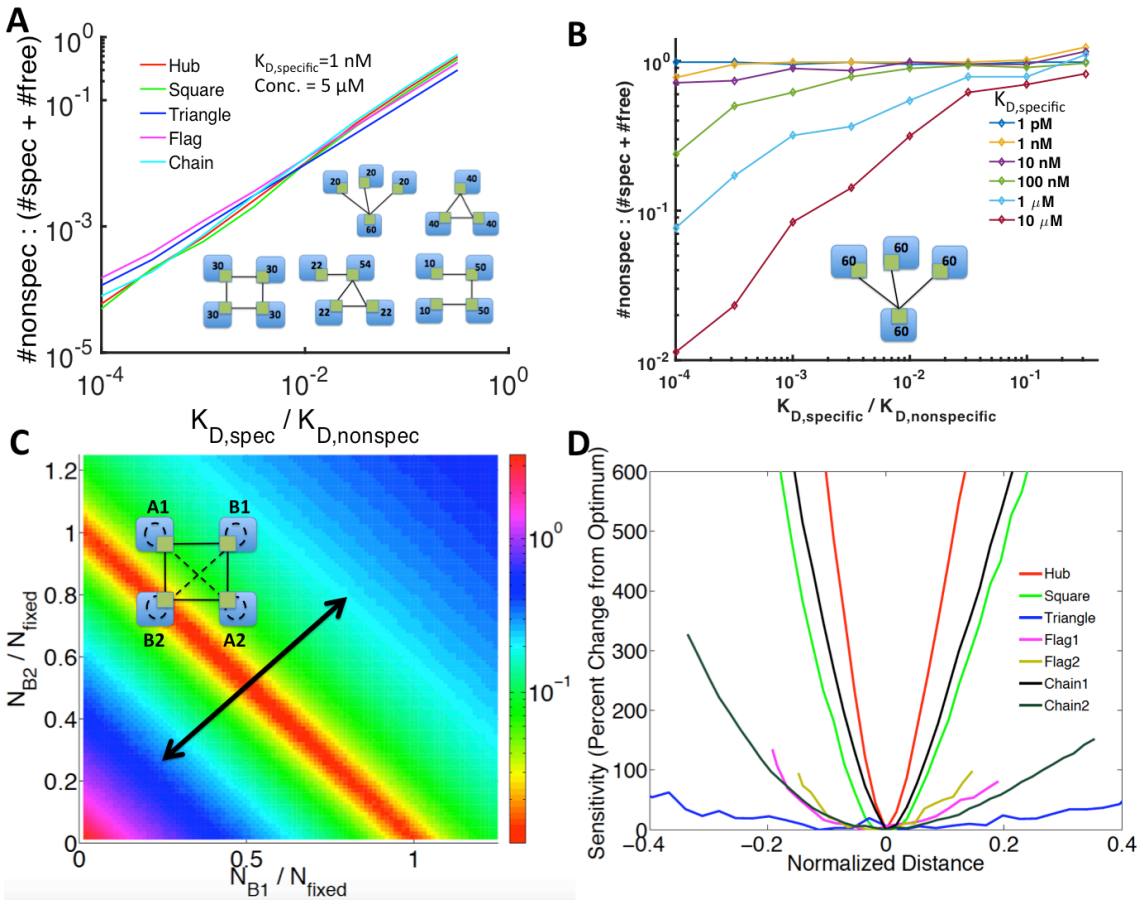


Figure 3.5. Misinteractions are motif dependent only when concentrations are imbalanced (A) At balanced concentrations, misinteraction frequency increased linearly with the ratio of $K_{D,specific}$ to $K_{D,\ nonspecific}$. It was also roughly equal for all five network motifs. **(B)** At unbalanced concentrations, misinteractions can occur even at a low energy gap, unless the overall binding is weak. **(C)** Surface plot for the square network, measuring the ratio of (#nonspecific complexes : #specific complexes + free proteins) when A1 and A2 are fixed while B1 and B2 are varied. The principal component (black line) is shown across the region of lowest misinteraction frequency. **(D)** Cost sensitivity to concentration imbalance varies significantly between motifs. The "distance" is measured along the principal component of the surface plots moving away from the optimal region. Two different pairs of fixed proteins were analyzed for the chain and flag networks. The hub and square networks were the most sensitive to imbalance, while the flag and triangle were the least.

The motifs most sensitive to imbalance, the hub and square motif, are also the motifs most common in biological networks [71,72]. In previous work, we demonstrated that these motifs are evolutionarily selected for in biological networks because binding interfaces that interact through these specific motifs are much easier to simultaneously design for high specificity (strong $K_{D,specific}$) and for weak nonfunctional interactions (weak $K_{D,nonspecific}$) [71,72]. Although these motifs thus produce more selective binding interfaces, our results show that there is more pressure to maintain copy number balance in these biologically common motifs to prevent misinteractions.

Importantly, unlike the results for balanced copy numbers, strong binding proteins are highly prone to misinteractions under imbalanced conditions (Fig 3.5b). Weak-binding proteins form minimal complexes overall, and thus imbalances in copy numbers do not strongly influence their binding patterns. Strong binding proteins, on the other hand, are driven to bind to any unbound interface, even when the gap separating specific and non-specific binding is large, because their misinteractions are stable and there is a larger pool of nonspecific partners. Thus leftover copies of these proteins frequently misinteract. This supports the observations that strong binding proteins should be tightly regulated to maintain stoichiometric balance[10], and therefore avoid misinteractions. For weak binding proteins, on the other hand, misinteraction cost is not a significant pressure favoring copy number balance.

3.3.3 Larger networks with biological topologies produce more misinteractions under copy number imbalance

Our analysis of network motifs above demonstrated that topologies common in biological IINs are actually more prone to misinteractions when copy numbers are imbalanced. We find here that the same trend applies to much larger networks that again exhibit biological topologies (Fig 3.6). To show this, we analyzed 500 IINs that differed in three properties: motif frequencies; degree distribution; and density, which was determined by the size of the network (90-200 proteins for 150 edges). The biological-like IINs have motif frequencies biased to hub and square motifs; they have a degree distribution that is power-law like or "scale-free", meaning, broadly speaking, that a few "hub" proteins have many connections while the majority are specialized for a few interactions; and they tend to be sparse, with interfaces in the CME IIN having an average degree of only 2.06[72]. For simplicity, here we will assume each interface is on its own protein, such that the PPIN is the same as the IIN. Balanced copy numbers are assigned to each network using our optimization method described above based on network structure (Methods 3.5.5), and imbalanced copy numbers are defined by randomly sampling copy numbers from the yeast distribution. Specific and non-specific K_D values for each possible binding interaction were initially taken from a previous study[71], where the gap between specific and non-specific binding was optimized based on selecting amino-acid sequences for each interface [71].

As expected, when copy numbers are balanced rather than imbalanced via random assignments, all networks produced fewer misinteractions. The networks

that, under balanced copy numbers, produced the fewest misinteractions were the networks most like biological IINs: they were sparse networks and they had optimized topologies favoring square and hub motifs (Fig 3.6). Because these IINs also had larger energy gaps separating $K_{D,\text{Specific}}$ and $K_{D,\text{Nonspecific}}$ [71], we verified that

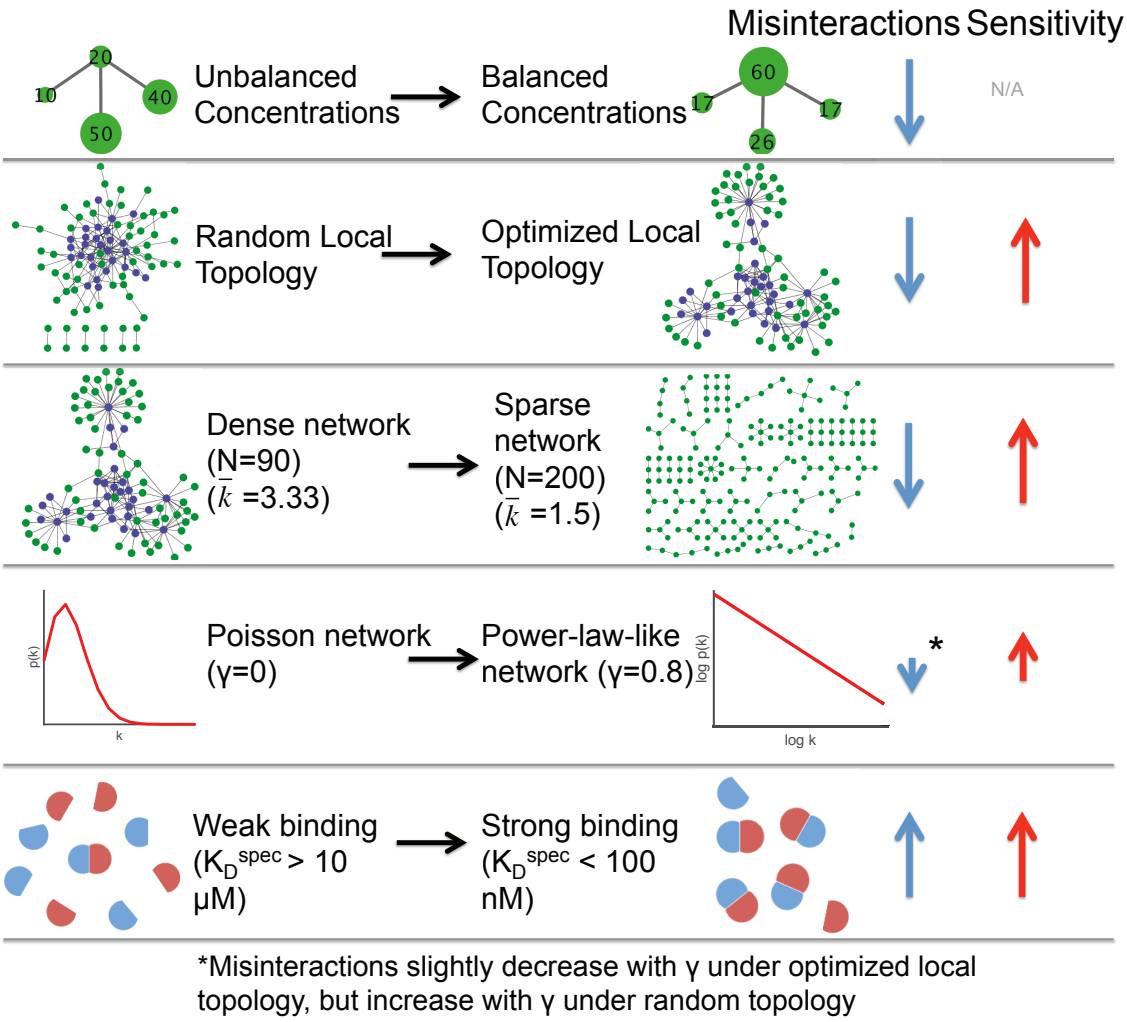


Figure 3.6. Biological IIN topologies have more misinteractions under imbalance. Shown are trends in misinteraction frequency under balanced concentrations (blue arrows) and sensitivity to imbalance (red arrows). Several features that make networks perform better under balanced concentrations make them perform worse under unbalanced concentrations: sparseness, a topology that matches with real interface networks, and a power-law degree distribution. Strong average binding caused both increased misinteractions and increased sensitivity.

when all networks were assigned the same $K_{D,\text{Specific}}$ and $K_{D,\text{Nonspecific}}$ (1000-fold different), the biological IINs indeed produced fewer misinteractions under balanced copy numbers (Fig 3.7a), although the difference was relatively small. Hence, overall, the results are similar to the findings with motifs, that for balanced copy numbers, misinteractions are not strongly influenced by network structure.

Once copy numbers were imbalanced, however, the biological-like IINs produced a sharper increase in misinteractions (higher sensitivity-Fig 3.6). This is consistent with the trends from the previous section, where the biological motifs of hub and square motifs were also more sensitive to imbalance. Sparse networks are more sensitive to imbalance because they have more interfaces (N) that can possibly misinteract (order N^2). The only network feature that did not have a significant trend in controlling misinteractions either for balanced or unbalanced copy numbers was the degree-distribution. For power-law network topologies compared to Poisson networks, misinteractions could be higher or lower depending on the local motifs or the network sparseness (Fig 3.6; Fig 3.7). Thus local topology and density were more important than the overall degree distribution.

Finally, because highly abundant proteins are thought to have low average affinity to avoid misinteractions, we increased the absolute strength of $K_{D,\text{Specific}}$, while keeping the gap between $K_{D,\text{Specific}}$ and $K_{D,\text{Nonspecific}}$ constant. Stronger affinity did indeed lead to both more nonspecific complexes and higher sensitivity to copy number imbalance. This result is consistent with the previous section and confirms that strong binding affinities can be paradoxically deleterious to specific complex formation.

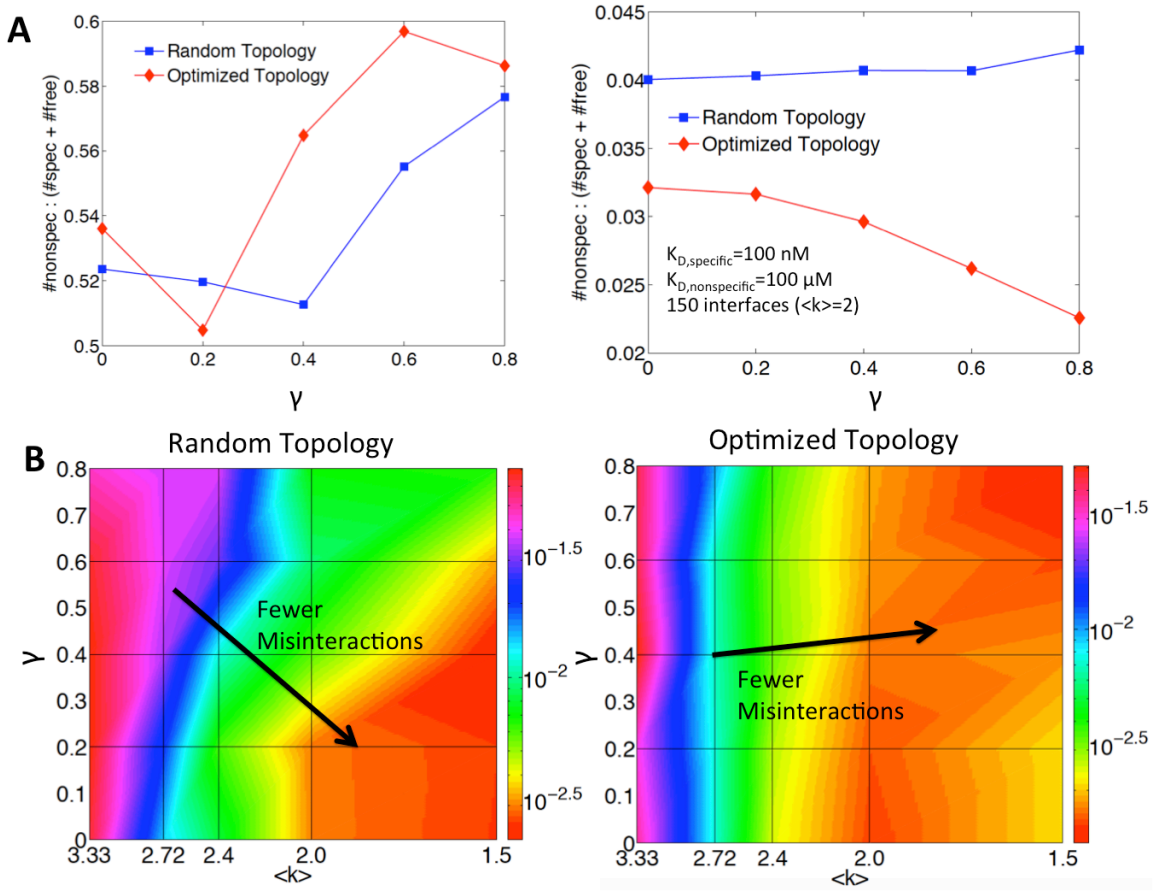


Figure 3.7. Effects of optimized local topology on misinteractions. (A) Misinteraction frequency of networks under randomly sampled (left) and balanced copy numbers (right) when fixed energy gaps were used ($K_{D,\text{Specific}}$=100nM, $K_{D,\text{nonspecific}}$=100μM). Networks with optimized topology and a power-law-like distribution (γ=0.8) performed best under balanced copy numbers but worse under imbalance. **(B)** Heat map of misinteraction frequency under balanced copy numbers vs degree distribution and network density. Denser networks always had more misinteractions, but the effects of degree distribution depended on whether the local topology was optimized or not.

3.4 Discussion

3.4.1 Measuring stoichiometric balance in protein-protein networks determines unexpected correlations in protein expression levels

The metric we have developed objectively determines whether a protein is under or overexpressed relative to not only its direct binding partners, but also to a larger

network including partners of partners. This global evaluation is thus sensitive to the size of the network, and captures how the multiple binding interfaces of a protein can control its competition for binding partners. In the interface-resolved CME network, we have shown evidence of imperfect, but statistically significant, stoichiometric balance. However, the original 56-protein network was overall unbalanced due to the high overexpression of the actin binding protein cofilin. It appears that removing or adding certain proteins can improve or deteriorate the measured balance of the network. Imbalance may also indicate possible missing interactions in our network. Despite the simplicity of our metric, our method was still able to highlight both correlated concentrations and proteins that violate balance for functional reasons, such as the kinase PRK1. Furthermore, the observed balances can suggest possible mechanisms of assembly, for example, that can then be studied using kinetic modeling, as we did here. What our results emphasize is that correlations are highly important: functionality can be obliterated with significant imbalance, and misinteractions can also be overwhelming due to significant imbalance.

Although we only applied our stoichiometric balance analysis to the 56 protein CME network, two smaller modules of this network, and the 127-protein ErbB network, these networks are significantly larger than the obligate complexes previous studied for copy number balance[4,16]. Our networks also contain a much larger variety of binding interaction strengths and competitive and non-competitive interactions. As we showed above, balance depended on the protein network's underlying IIN. While it would be beneficial to repeat this analysis on a larger

network, there is a paucity of manually curated IINs in the literature. There are various larger automatically constructed IINs generated by homology modeling[67,73], but our previous work found these automatic IINs suffer from various inaccuracies and differ significantly from manually curated IINs in topology[72].

3.4.2 Limitations of measuring stoichiometric balance for larger PPINs

Our metric for evaluating stoichiometric balance only accounts for the binding interface network structure and observed copy numbers. A missing feature of our stoichiometric balance metric is that proteins within a network can be expressed with both spatial and temporal variation. For a small binding network this is not a major concern, since proteins in the same complex tend to be co-expressed[21] and co-localized so they may bind. But as network size is scaled up, the probability of all proteins being equally present is reduced. Such temporal and spatial variations could be taken into account in the construction of the network, leaving out proteins that are not functional at the same time.

A natural extension to our measure of stoichiometric balance would be to also account for binding affinities of interactions in addition to the binding interface network structure and observed copy numbers. Our results here and previous studies[10] indicate that balance should be more tightly constrained for strong binding proteins. However, affinity data is in even more limited availability than binding interface data; hence we could not include affinities in our networks. Our existing metric can thus be much more easily applied to a variety of networks. Furthermore, by picking out highly correlated expression levels, our method can then indicate

which interactions might be quite strong, or vice-versa, which may be transient or weak.

3.4.3 Noise and variability in experimental copy number measurements can limit observed balance

In this study we used yeast copy numbers from Kulak et al.[2] because it was the most comprehensive. The other three studies we used for comparison did not cover all 56 proteins in our network. However, for the proteins we could compare, we found significant discrepancies between relative abundances. Light chains are weakly expressed in other studies, for example[1,141,142] A few possible reasons for this exist. One is that fluorescence data is inherently noisy. Experimentalists must deal with background noise, interference with protein localization due to the large fluorescent tags, and cross interactions with other proteins[143]. Cell lines can also accrue mutations over time that decrease or increase gene expression, a phenomenon observed with HeLa cells[144]. Finally, cells may alter gene expression for regulatory reasons, so the environment in which cells are grown may alter gene expression.

3.4.4 Perfect balance is not observed, even if it would prevent misinteractions

We found that copy number imbalance can lead to misinteractions and the features of biological IINs (power-law-like degree distribution, square and hub motifs, sparseness) typically have fewer misinteractions under balanced copy numbers but more misinteractions under imbalance. These networks thus should require more tightly controlled balance to avoid misinteractions. But

misinteractions are of course not the only pressure on copy numbers. For multi-protein assemblies and complex processes there may be functional benefits to imbalance. I explore two possible benefits in Chapter 4: increasing complex assembly yield and bottlenecking the endocytosis process.

3.5 Methods

3.5.1 Defining stoichiometric balance in a PPIN with interfaces resolved

A stoichiometrically balanced network has the copy numbers of each interface matched to the copy numbers of all pairwise complexes it participates in. Balanced copy numbers are obtained by assigning a number of desired complexes to each edge in the interface binding network. The balanced copy numbers of each interface can then be calculated from the equation

$$Ax = C \qquad \text{Eq. 3.2}$$

where "*A*" is a binary matrix with N_{int} rows (one for each interface) and M_{edge} columns (one for each pairwise interaction). $A_{i,j}$ =1 if the interface *i* is used in the interaction *j*, or 2 if a self-interaction, and 0 otherwise. "*x*" is the vector of desired pairwise complexes (M_{edge} x 1), and "*C*" is the number of interface copy numbers (N_{int} x 1). In Fig 3.8 we illustrate this procedure for a small toy network.

If desired pairwise complexes, *x*, is specified, interface copy numbers, *C*, can directly be solved for using Eq. 3.2, but if interface copy numbers, *C*, are specified, *x* will not, in general, have an exact or nontrivial solution unless *C* is balanced. This is because all entries of *x* must be >0 or some other minimum value, as negative copies cannot exist. This produces a hard constraint on *x*. Given a vector *C*, an optimal

solution to x must be solved for using quadratic programming rather than linear least-squares.

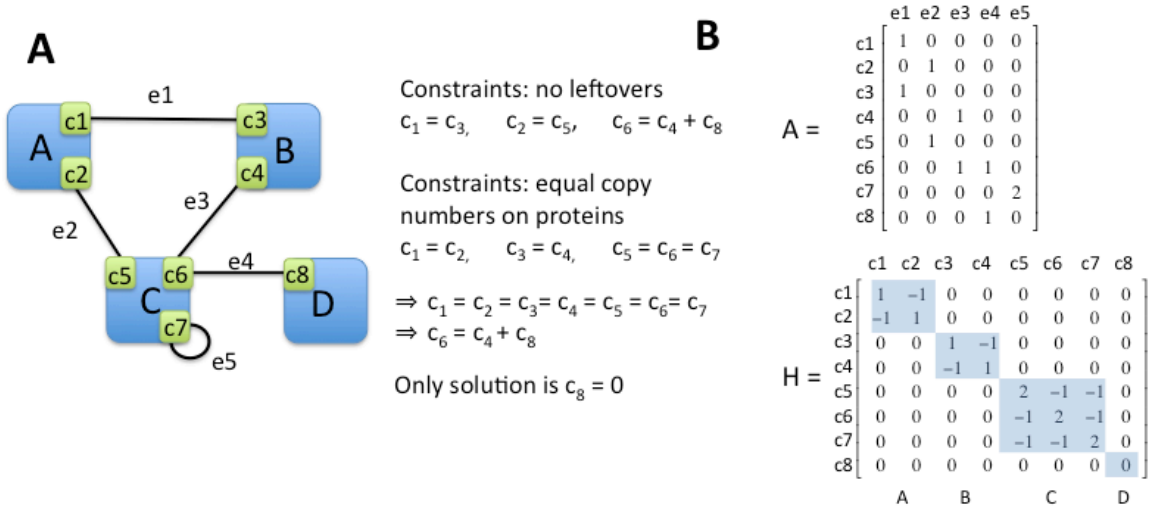


Figure 3.8. Example network. **(A)** An example network that has no nontrivial balanced solution when all constraints are applied. **(B)** The "A" and "H" matrices for the left network.

Our goal is to select for an optimal x given an input set of copy numbers "C_0". This is a soft constraint on the optimal x, because the input C_0 may not be balanced. Once an optimal x is found, forward solving Eq. 3.1 will in general not perfectly recover C_0. C_0 can constrain all interfaces or a subset of them. To constrain a protein is to constrain all interfaces on it, s we introduce a third constraint on the optimal x that the copy numbers of interfaces on the same protein should be equal. This often makes nontrivial solutions impossible (Fig 3.8), so it is also a soft constraint. Combining all of these constraints, the optimal desired number of complexes "x" can be found by minimizing the equation:

$$\min_x[\alpha(Ax-C_0)^T Z(Ax-C_0) + (Ax)^T H(Ax)], x \geq 0 \qquad \text{Eq. 3.3}$$

Where each variable is defined as follows:

A: N_{int} x M_{edge} matrix defining which interfaces are used in which interaction, i.e. pairwise complex.

x: M_{edge} x 1 vector of desired pairwise complex copy numbers

$\boldsymbol{C_0}$: N_{int} x 1 vector of constrained copy numbers.

Z: N_{int} x N_{int} diagonal matrix that selects which interfaces are constrained. Entries = 1 if the interface is constrained and =0 otherwise. If all interfaces are constrained, Z equals the identity matrix.

H: N_{int} x N_{int} permutated block diagonal matrix with positive and negative entries such that HC=0 if interfaces on the same protein have equal copy numbers. Each block corresponds to a protein (Fig 3.8).

$\boldsymbol{\alpha}$: 1x1 scaling parameter which determines the relative weight of the C_0 soft constraint vs the equal interfaces soft constraint.

For any vector x, Eq. 3.3 produces a positive scalar value. The equation was minimized using the OOQP (object-oriented quadratic programming) 0.99.26 package for C++[145]. Quadratic programming is necessary due to the constraint of x≥0. Eq. 3.3 can be converted into a quadratic equation of the form

$$\frac{1}{2}x^T Qx + d^T x + r \qquad \text{Eq. 3.4}$$

Using:

$Q = 2\alpha A^T Z A + 2A^T H A$

$d^T = -2\alpha C_0^T Z^T A$

$r = \alpha C_0^T Z C_0$

"r" can be ignored by the solver when minimizing the equation since it is a constant term.

Once $x_{\min}$ is found via Eq. 3.4, the optimized interface copy numbers can obtained by forward solving $Ax_{\min} = C_{\text{balanced}}$. Interfaces on the same protein will not necessarily have equal copy numbers due to the competing constraints of Eq. 3.3. We can assign a single copy number to each protein by averaging over all interface copy numbers on that protein to give C_{balanced}, a vector of protein copy numbers. These values were used when calculating which proteins were over or underexpressed in the networks. The distance from C_0 to C_{balanced} was used as a metric to determine relative balance (see below).

3.5.2 Biological protein copy numbers

For the yeast CME network, C_0 was used to constrain all 56 proteins because copy numbers from Kulak et al. were available[2]. For the ErbB signaling network, only 115 out of 127 proteins with available expression level data were constrained. 100 of these proteins were constrained with HeLa copy number estimations from Kulak et al. [2], while estimated copy numbers for 15 additional proteins were added from four additional studies[10,146-148], leaving 12 proteins with unknown expression data.

3.5.3 Measuring the degree of stoichiometric balance in observed concentrations

Using the optimized copy numbers, $C_{balanced}$, we can then ask, how close are the original, biologically observed copy numbers to these optimally balanced values? If the original copy numbers are already perfectly balanced, then they will match the optimal copy numbers. If they are imperfect, than the two distributions will differ. We use two metrics to quantify the distance between the observed and optimized concentrations: chi-square distance (CSD)

$$\sqrt{\sum_i \frac{(X_i - Y_i)^2}{(X_i + Y_i)}} \qquad \text{Eq 3.5}$$

and Jensen-Shannon Distance (JSD)

$$\sqrt{\frac{1}{2}(D_{KL}(x \parallel z) + D_{KL}(y \parallel z))} \qquad \text{Eq. 3.6}$$

Where $x = X/\Sigma(X)$, $y=Y/\Sigma(Y)$, $z=(x+y)/2$ and D_{KL} is the Kullback-Leibler divergence

$$D_{KL}(x \parallel y) = \sum_i x_i \log \frac{x_i}{y_i} \qquad \text{Eq. 3.7}$$

Both CSD and JSD are metrics.[149]

For cases where Z≠I (i.e. not all interfaces were constrained) only distance between constrained interfaces was measured.

3.5.4 Small network motifs

Binding for the five 3- or 4-node network motifs; triangle, chain, square, 4-node hub, and flag; was simulated using the Gillespie algorithm[140]. Besides the specific binary interactions, nonspecific interactions were allowed at a strength determined by an "energy gap" between binding energies, though in practice we defined the ratio

nonspecific K_D to specific K_D by factors of 10. This corresponded to a linear difference in free energies via the equations:

$$K_{D,specific} = e^{-\Delta E_1 / K_B T}$$

$$K_{D,nonspecific} = e^{-\Delta E_2 / K_B T}$$

$$\frac{K_{D,specific}}{K_{D,nonspecific}} = e^{-(\Delta E_1 - \Delta E_2) / K_B T}$$

The networks were simulated under various initial concentrations. The steady-state ratio of Eq. 3.1 was recorded, where $N_{\text{nonspecific}}$ is the number of nonspecific binary complexes, N_{specific} is the number of specific binary complexes, and N_{free} is the number of free proteins. Ratios were averaged across 5,000 runs.

To generate surface plots, two proteins were chosen to be variable while the remaining proteins were given fixed copy numbers. Because the flag motif produced asymmetric plots, two different choices of variable proteins were used. (Fig C.8) Surface plots were generated using Matlab.

We calculated sensitivity by determining the principal component of the surface plot data (i.e. the vector of greatest variance) and measuring the percent change in ratio from the optimum along this vector. For better comparison, we normalized distance along the surface plots via dividing the abundance of the variable proteins by the abundance of the fixed proteins.

Motifs with purely noncompetitive interactions were not considered, because the interface network would consist entirely of pairs and not provide meaningful insights about the effects of IIN topology on misinteractions

3.5.5 Analysis of complex IIN topologies

For the large network analysis we used the 500 networks from Johnson, *J Phys Chem B* 2013[71]. 25 sets of 10 networks each were randomly generated using two parameters: number of nodes (90, 110, 125, 150, 200), keeping the number of edges fixed at 150; and the preferential attachment exponent "γ" from Goh, 2001[127]. γ=0 corresponds to a binomial, Erdos-Renyi network, whereas γ=1 corresponds to a power-law or "scale-free" network. Values of 0, 0.2, 0.4, 0.6, and 0.8 were used. Finally, a local topology optimization algorithm that decreased the frequency of chain and triangle motifs and increased hub motifs was applied to each network, for 500 networks in total. All networks assume competitive (binary) binding.

Rather than assign an arbitrary specific and nonspecific K_D for the networks, we used the relative binding energies determined for each network in the source paper. This was determined by a physics-based Monte Carlo optimization scheme of amino acid residues, as described in Johnson, 2011[39]. The minimum energy gap between specific and nonspecific interactions could be measured as a relative metric of the network's propensity for misinteractions. Because the binding strengths were relative, we could alter the average binding strength to determine the effects on misinteractions. This was varied between 7 values of 1 nM to 1 mM, using factors of 10. Finally, to obtain results more comparable to the simple

networks, we also ran simulations where each specific interaction had K_D=100 nM and each nonspecific interaction had K_D=100 μM.

Networks were simulated to steady state using the Gillespie algorithm[140] under five differing sets of copy numbers (CNs) for free proteins: equal CNs for each protein, random CNs sampled from a yeast protein concentration distribution (performed 20 times) and three forms of balanced CNs using the network architecture. Any set of CNs without leftovers – i.e. having exactly enough proteins to create a certain number of specific complexes – is considered "balanced", and thus there are infinite solutions. The first balanced set assumed an equal number of each type of specific complex, which results in protein CNs proportional to the protein's number of partners. The remaining balanced CNs were determined by finding "x" to minimize a simplified form of Eq. 3.3:

$$\min_X (Ax\text{-}C_0)^T (Ax\text{-}C_0) \qquad \text{Eq. 3.8}$$

Here there is only one interface on each protein, and all the proteins are constrained, so there is no need for a Z matrix, the α scaling parameter, or the second term. C_0 is either equal copy numbers or randomly sampled copy numbers. After x_{min} is found via quadratic programming (see above), the balanced CNs are obtained by forward solving $C_{balanced} = Ax_{min}$.

To measure nonspecific complex formation, a modified ratio was used:

$$Cost(C_0) = \frac{2N_{nonspecific}(C_0)}{2N_{specific}(C_0) + N_{free}(C_0)} \qquad \text{Eq. 3.9}$$

to compare total individual proteins in each bound or unbound state, rather than number of unbound or bound states. To measure sensitivity, the ratio under

unbalanced CNs (C_0) divided by the ratio under balanced CNs (C_{balanced}) was calculated. A higher ratio indicates higher sensitivity to CN balancing.

Network maps were generated using Cytoscape[150]. Plots were generated in MATLAB. Future figures may be found in Appendix C.

C++ code for the network balancing algorithm is available at https://github.com/mjohn218/StoichiometricBalance, and may be applied to any interface-resolved network.

Chapter 4. Functional Significance of Copy Number Balance for Yeast Endocytosis

Chapter adapted from:

Holland, DO, & ME Johnson (2018) "Stoichiometric balance of protein copy numbers is measurable and functionally significant in a protein-protein interaction network for yeast endocytosis." In revision at PLoS Comp Biol. Available at bioRxiv https://doi.org/10.1101/205674

The formation of specific functional complexes depends on relative copy numbers. We constructed simple kinetic models of two sub-networks in the yeast clathrin-mediated endocytosis network to assess multi-protein assembly of the ARP2/3 complex and a minimal, nine-protein clathrin-coated vesicle forming module. We find that the observed, imperfectly balanced copy numbers are less effective than balanced copy numbers in producing fast and complete multi-protein assemblies. Further, we speculate that strategic imbalance in the vesicle forming module allows cells to tune spatially where endocytosis occurs, providing sensitive control over cargo uptake via clathrin-coated vesicles. Our results provide insight into how network design, expression level regulation, and cell fitness are intertwined.

4.1 Introduction

Protein copy numbers are often found to be stoichiometrically balanced for subunits of multi-protein complexes[10]. Imbalance is believed to be deleterious because it lowers complex yield (the dosage balance hypothesis) and increases the risk of

misinteractions, but imbalance may also provide unexplored functional benefits. In the previous chapter, I showed that biological networks are more robust to misinteractions than random networks when balanced, but are more sensitive to misinteractions under imbalance. This suggests evolutionary pressure for proteins to be balanced and that any conserved imbalance should occur for functional reasons.

Our previous simulations only studied binary, competitive interactions. But proteins often bind noncompetitively into higher complexes, and they may interact weakly and thus form few complexes, in which case imbalance may have functional benefits[31,32]. Furthermore, the previous models looked at equilibrium results, whereas many biological systems exhibit non-equilibrium dynamics. Underexpressed proteins could provide tuning knobs for functional outcomes, for example. We created kinetic models of two modules from the CME network[66] with observed imbalances: the ARP2/3 complex and a simplified vesicle forming protein subset. Simulating higher complex formation is challenging because of the exponentially large number of possible species, so we used NFSim[151], a stochastic solver of chemical kinetics that is rule-based, enabling an efficient tracking of higher-order complexes as they appear in time.

The direct consequence of imbalance we evaluate is that changes to copy numbers control which specific and functionally necessary complexes can form. When the central clathrin protein is knocked out in cells, for example, clathrin-mediated endocytosis (CME) is terminated, as clathrin is functionally irreplaceable[152]. The plasma membrane lipid $PI(4,5)P_2$ is also essential for CME, as it

is required for recruiting the diverse cytosolic clathrin-coat proteins to the membrane to assemble vesicles[153]. Many clathrin-coat proteins, however, can be knocked out without fully terminating CME[154]. As the CME network illustrates (Fig 2.1), most of these proteins have multiple domains mediating interactions involving both competitive and non-competitive interactions. Adaptor proteins (proteins that bind to the membrane, to transmembrane cargo, and often to clathrin as well) exhibit redundancy in their binding partners that can partially explain how knock-outs to one protein can be rescued by the activity of related proteins. With simulation of simple kinetic models, we can then test these hypotheses. Although these models are far too simple to recapitulate the complexities of CME *in vivo*, they are nonetheless useful in highlighting potential bottlenecks in assembly due to copy numbers or binding affinities.

Our simulations of (non-spatial) kinetic models demonstrate that stoichiometric balance does, in fact, improve multi-protein assembly relative to observed copy numbers. We speculate that the observed imbalances in clathrin adaptor proteins could offer a mechanism for making the vesicle formation process more tunable, since adaptor proteins are responsible for selecting cargo for endocytic uptake, which is the ultimate purpose of CME.

4.2 Results

4.2.1 The ARP2/3 Complex has higher yield under stoichiometric balance

One unexpected imbalance we found, from the work in Chapter 3, was that of the isolated, 7-component ARP2/3 complex. The complex has one highly

underexpressed subunit, ARC19. ARC19 is a core subunit, binding to five other subunits (Fig 4.1a). Because of this, it is more likely to form misinteractions (due to its five interfaces) and be a part of incorrect complexes (e.g. complexes of the form ARC19 – ARC40 – ARP2 – ARC19 are incorrect because they contain two ARP19 proteins). Thus we tested whether the observed copy numbers might improve formation of complete ARP2/3 complexes.

Ultimately, we found that balanced copy numbers always improved formation of complete ARP2/3 complexes relative to the observed copy numbers, whether or not misinteractions were modeled (Fig 4.1b). We simulated simplified complex assembly using arbitrary rate constants and two sets of copy numbers: those observed from Kulak et al. and stoichiometrically balanced (in this case equal) copy numbers for each subunit. We measured "yield" as the number of proteins in full complexes divided by the number of proteins in all complexes, including misassembled or incomplete complexes. Some cooperativity was allowed in that if three proteins in a trimer were held together by two binding events, the third binding event could occur at a faster rate (due to all three subunits being localized together). Binding to the core subunit ARC19 was also set to be 10-fold stronger than peripheral bindings, as this increased yield. But no matter what parameter ranges we used, we could not increase the yield of the Kulak copy numbers (max ~13%) versus the balanced copy numbers (max ~50%). Because ARC19 has ~5-fold underexpression compared to the other 6 subunits, incomplete complexes dominate. The results held when we also allowed ARC19 to form misinteractions.

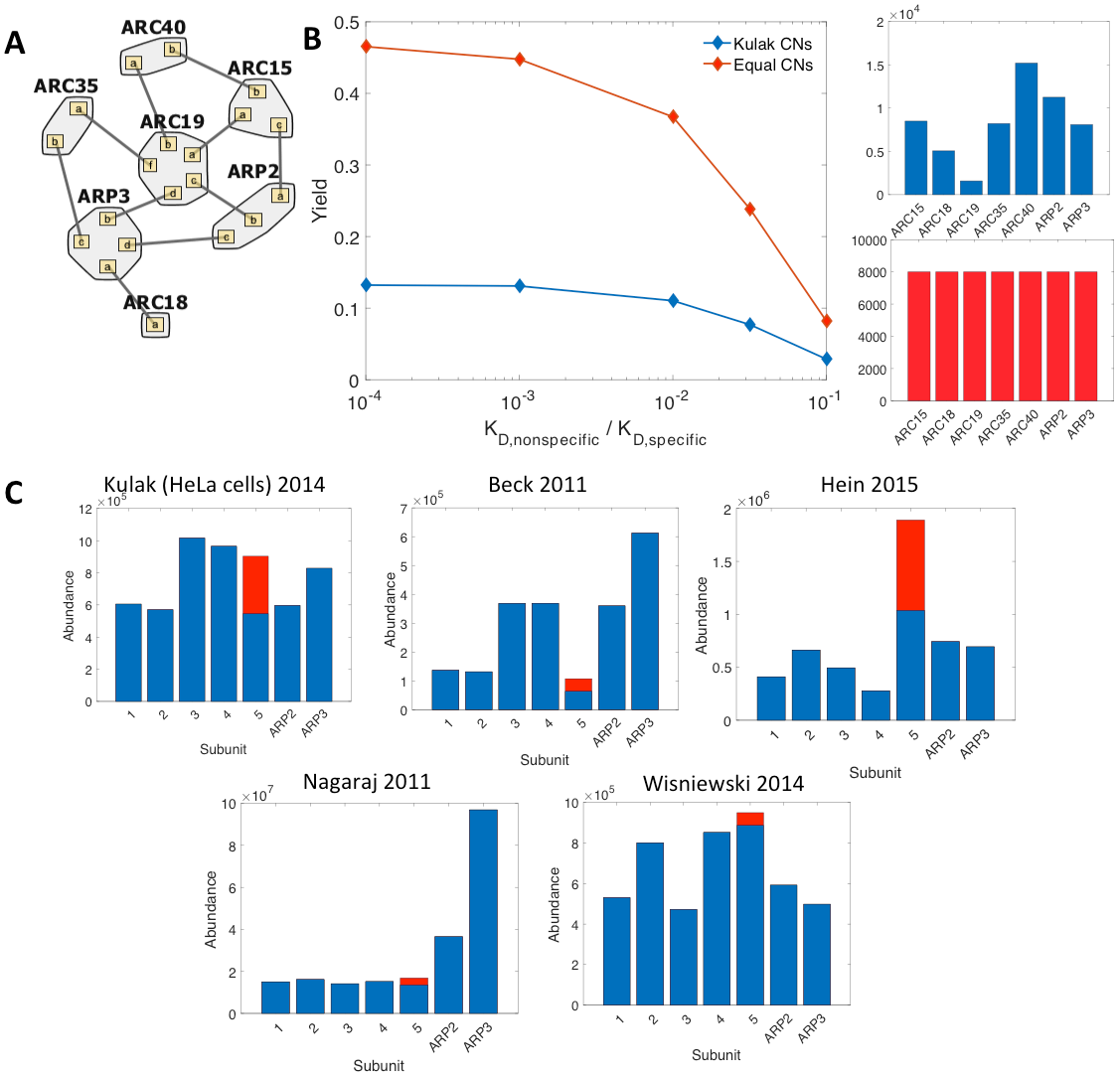


Figure 4.1. ARP2/3 complex has higher yield under balanced copy numbers. **(A)** Contact map of the seven subunits of the complex, generated with RuleBender[7] **(B)** Under varying misinteraction strengths, the yield for the balanced copy numbers was always higher than for the observed copy numbers from Kulak et al.[2] Yield was measured as N_{desired} / (N_{desired} + $N_{\text{undesired}}$), which refer to the number of proteins in either desired (complete) complexes or undesired (incomplete or misassembled) complexes. **(C)** The observed copy number distribution was not found to be conserved between studies in either yeast or humans. Bar plots are from five studies of the ARP2/3 subunits in human cells. The red bar is for the addition of the "subunit 5-like" protein. Only one study (Hein et al.) found ARC19's equivalent, subunit 4, to be underexpressed[10].

Imbalances in copy numbers have been shown to actually improve the yield for self-assembly, but the optimal copy numbers must take on specific ratios of components to optimize yield[32]. Here, we see that the ARP2/3 subunits do not exhibit optimal expression for yield in our model. One possible explanation is that the ARC19 subunit has distinct thermodynamics or kinetics that are critical for controlling the ARC19 subunit has distinct thermodynamics or kinetics that are critical for controlling assembly. This would suggest that this subunit has conserved expression across all organisms. However, this is not the case. We compared the expression levels of the seven subunits with data from three other studies – two also found ARC19 to be underexpressed[141,142], whereas one[1] found it to be overexpressed. However, Chong et al. also found ARP2 to be underexpressed, whereas Kulak et al. found it to be overexpressed. We also compared the abundance of human homologs from five studies[2,10,146-148] and found similar issues with noise, though only one found ARC19's homolog to be underexpressed. (Fig 4.1c) Thus, no conservation of subunit expression levels is observed. Without a more structurally and biochemically accurate model for the ARP2/3 components, it is difficult to assess whether the low expression of ARC19 does provide some benefit in assembly yield. As we return to in the discussion, several other factors may explain the imbalance, such as noise in expression levels or in measurements of expression levels, or additional roles in the cell for some ARP2/3 subunits.

4.2.2 A simplified clathrin-coated vesicle forming model enables a kinetic study of imbalance effects on non-equilibrium assembly

For our final analysis, we tested the effects of copy number balance on a more complex, non-equilibrium model of clathrin-coat assembly for vesicle formation. Our minimal model for vesicle formation includes nine cytoplasmic proteins plus the plasma membrane lipid recruiter $PI(4,5)P_2$, with the biochemical parameters taken from the literature for all known binding interface interactions (Fig 4.2; Table 4.1). In clathrin-mediated endocytosis, clathrin triskelia consisting of three heavy chains (CHC1) and three lights chains (CLC1) are recruited to the membrane via adaptor proteins that bind lipids (ENT1 & 2, SYP1, SLA2, YAP1801) and in some cases also transmembrane cargo (ENT1 & 2, YAP1801). Clathrin polymerize to form a hexagonal clathrin cage of ~100 triskelia [155] that helps deform the plasma membrane into spherical membrane vesicles of ~100 nm in diameter. Additional non-membrane-binding scaffold proteins help stabilize the assembly (EDE1, YAP1802). Importantly, the assemblies do not have to exhibit a perfect stoichiometry of components, unlike the ARP2/3 complex, in order to function, with variable compositions shown to produce clathrin-coated structures *in vitro*[154,156,157] . To measure vesicle formation in our model, we therefore make the assumption that completed vesicles contain 100 triskelia [155] in a complex on the membrane. Once a completed model vesicle is formed, all components that are a part of this complex are recycled, unbound, back to the cytoplasm, keeping total protein concentrations fixed.

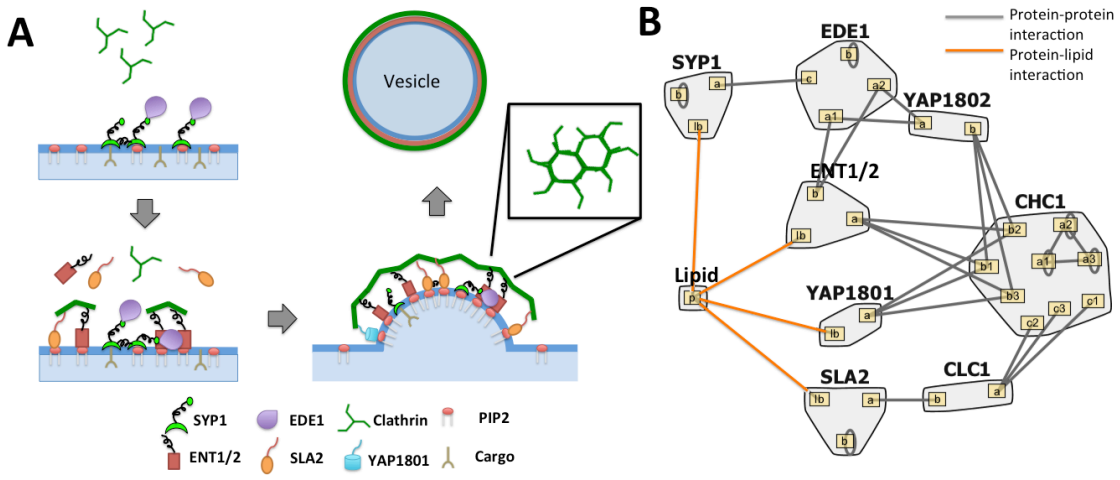


Figure 4.2. Clathrin membrane recruitment model. **(A)** In clathrin-mediated endocytosis, adaptor proteins bind to the lipid membrane and recruit clathrin triskelia to the surface. These triskelia assemble a hexagonal cage around the plasma membrane vesicle. **(B)** Binding model of the clathrin module. Included are seven adaptor or accessory proteins (SYP1, EDE1, YAP1801/2, ENT1/2, and SLA2), clathrin heavy chains already assumed to be in trimer form, and clathrin light chains. Five of the adaptor/accessory proteins can bind directly to the lipid membrane. Picture generated with RuleBender.

We emphasize that this minimal model is based on the known concentrations and binding properties of the component proteins, and thus we are not attempting to optimize the model to best describe *in vivo* observations. Furthermore, this kinetic model does not account for biomechanics of the membrane budding or coupling to the cytoskeleton, or molecular structure, which are important features of CME. As we see in our simulations, our vesicles form ~10 times faster than vesicle formation *in vivo*. However, clathrin-coated vesicles (pre-scission) are observed to assemble *in vitro* with minimal components, without the cytoskeleton or any energy sources[154,157]. We thus included in our model all proteins from the larger CME network (Fig 2.1) that directly connect clathrin coat assembly to the membrane

surface, linking the assembly process with the ultimate endocytic goal of transmembrane receptor and cargo uptake. Our model thus represents a useful qualitative framework to assess how stoichiometric balance in clathrin-coat components can impact vesicle formation and thus cargo uptake.

An important feature that our model does capture is the reduction in dimensionality (3D to 2D) which accompanies binding to the membrane surface [158]. Once localized to the membrane via either lipid binding or recruitment by other proteins, proteins are concentrated in units of area^{-1}, with binding constants of $K_D^{2D}=K_D^{3D}/(2\sigma)$, where σ is a lengthscale in the nanometer range[174], as discussed in ref[158]. Transitioning to the membrane can drive dramatic increases in complex formation due to higher effective concentrations of components[158]. In our simulations here, we found that this is a critical factor controlling vesicle formation. Besides this division between the cytoplasm and the membrane surface, there is no other spatial resolution. A full list of model assumptions can be found in Appendix E.

4.2.3 Adaptor proteins are underexpressed and can tune vesicle formation

We first evaluated whether this nine-protein module (Fig 4.2) was significantly balanced. The clathrin heavy chains and light chains are close in expression, as expected since these two have a strong binding affinity ($\sim$ 1nM)[167]. But clathrin was overexpressed compared to its adaptor proteins by over 3-fold. Functionally, a full triskelia has up to six binding sites for adaptor proteins, but only one needs to be bound to localize it to the membrane. Hence, it is not strictly necessary for the adaptor proteins to be balanced. However, we found that when balanced copy

Table 4.1. Parameters for clathrin membrane recruitment model.

Parameter	Description	Value	Notes and References
Vol_CP	Cytoplasm reaction volume	37.2 μm^3	60% of cell volume{159}, for average of haploid cell volume (42 fL) and diploid cell volume (82 fL){160}
SA_PM	Plasma membrane surface area	75.7 μm^2	Assuming a spherical cell, calculated from average of median haploid volume (42 fL) and diploid cell volume (82 fL){160}
σ	Lengthscale conversion between K_a^{2D} and K_a^{3D}	1 nm	{158}
Kd_CHC_CHC	Clathrin heavy chain polymerization	100 μM	Order of. {161}
Kd_CHC_ENT	Clathrin heavy chain binding to ENT1/2	22 μM	Binding of amphiphysin peptide to clathrin box{162}
Kd_CHC_YAP	Clathrin heavy chain binding to YAP1801/2	160 μM	Based on AP180 binding in humans{163}
Kd_EDE_ENT	EDE1 to ENT1/2 binding	12 μM	Binding of EH domain to NPF motif. {164}
Kd_EDE_YAP	EDE1 to YAP1802 binding	0.6 μM	Binding of Eps15 to sAP180 in humans. Kd of 0.5-0.7 μM {165}
Kd_EDE_EDE	EDE1 dimerization	0.127 μM	{166}
Kd_CHC_CLC	Clathrin heavy chain to light chain binding	0.1 nM	Upper limit of binding strength to CHC1 trimers. {167}
Kd_CLC_SLA	Clathrin light chain to SLA2 binding	22 μM	HIP1R (human homolog of SLA2) binds to clathrin cages with Kd in the low nanomolar range{168}, but experiment was not with isolated light chains. We choose to assign same affinity as CHC1 to ENT1/2.{162}
Kd_SLA_SLA	SLA2 dimerization	1 nM	Arbitrarily strong rate chosen. For HIP1R (human homolog) virtually no monomers in vitro.{[169]}
Kd_SYP_SYP	SYP1 dimerization	2.5 μM	Rate based on FCHo2 (human homolog) self-binding.{170}
Kd_SYP_EDE	SYP1 to EDE1 binding	0.227 μM	{166}
Kd_L_ENT	ENT1/2 binding to lipid	0.02 μM	{171}
Kd_L_YAP	YAP1801 binding to lipid	0.3 μM	{171}
Kd_L_SLA	SLA2 binding to lipid	0.2 μM	{171}
Kd_L_SYP	SYP1 binding to lipid	53 μM	F-BAR domain binding to a single PIP2 molecule. {172} Other papers suggest binding additional lipids or binding cargo, so may be stronger.
L_0	Density of lipid PtdIns(4,5)P2 partners	25,292 particles/ μm^2	Estimated from experiments with 3T3/NIH fibroblasts{173}
CHC1_0	Total clathrin heavy chain trimers	6426	19278 heavy chains{2}. Divide by 3.
CLC1_0	Total clathrin light chains	14538	{2}
EDE1_0	EDE1 total proteins	5964	{2}
ENT_0	ENT1/2 total proteins	3075	Sum of ENT1 and ENT2 proteins{2}
YAP1801_0	YAP1801 total proteins	357	{2}
YAP1802_0	YAP1802 total proteins	264	{2}
SLA2_0	SLA2 total proteins	3904	{2}
SYP1_0	SYP1 total proteins	2467	{2}
k_dump	Rate of deletion for a complex of >=100 triskelia	1000 s^{-1}	Arbitrarily high rate chosen
k_recyc	Rate of protein recycling to the cytoplasm	1000 s^{-1}	Arbitrarily high rate chosen

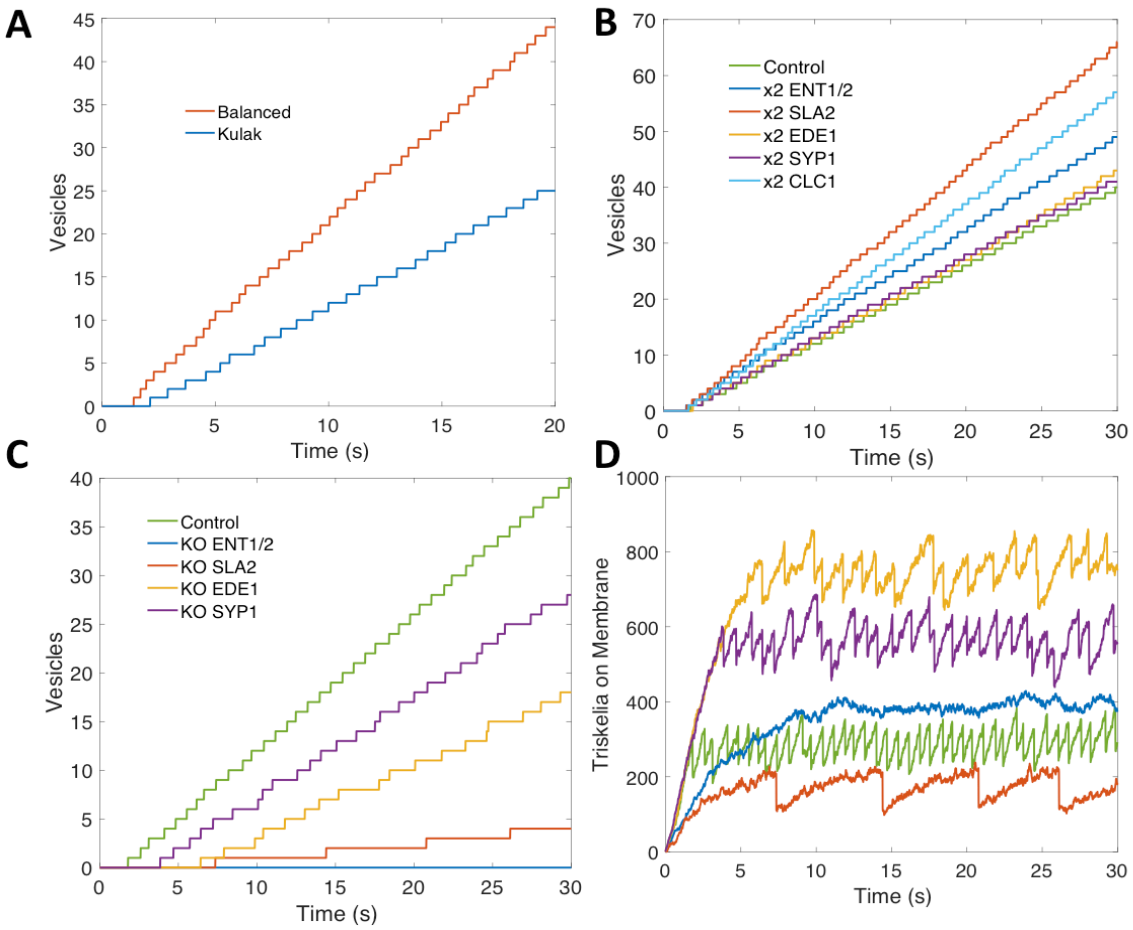


Figure 4.3. Endocytosis is tunable with adaptor proteins. (A) Vesicles were formed faster with balanced copy numbers, indicating that the biological copy numbers are not optimized for maximum vesicle formation. **(B)** Adaptor proteins in the network were underexpressed. Vesicle frequency could be increased by doubling their concentrations. **(C,D)** The system is sensitive to adaptor protein knockouts. Knocking out either SYP1 or ENT1/2 nearly halts vesicle formation. SYP1 and EDE1 appear to have an aggregating effect, allowing vesicles to form with less triskelia on the membrane.

numbers were used instead of observed copy numbers, vesicles formed faster and with fewer components (Fig 4.3a) Thus the biological copy numbers do not appear optimized for maximum vesicle formation, though they are sufficient to drive vesicle formation.

Our model assumes these proteins are well-mixed throughout the cytosol, but cells can spatially regulate proteins, altering the local concentration. We simulate this by altering the expression of the adaptor proteins in our model. Knocking out either SLA2 or ENT1/2 pushes the copy numbers even further out-of-balance, and nearly halts vesicle formation (Fig 4.3c,d). Increasing their expression increases vesicle formation because they are below saturation. Decreasing the other adaptor or scaffold proteins also increases imbalance and has a negative effect on the speed of vesicles, although it is less severe. Clathrin-coat assembly is quite sensitive to these membrane-binding protein concentrations because they not only recruit clathrin to the membrane, but they stabilize the triskelion in 2D, where they can then exploit reduced dimensionality to drive binding[158]. If clathrin polymerized effectively in solution, far fewer adaptor proteins would be needed to link large clathrin-cages to the membrane surface. We speculate that this sensitivity to the membrane-binding adaptor proteins and their observed underexpression could allow the cell to better tune productive vesicle formation to occur only when enough cargo is localized [175]. The adaptor proteins ultimately localize the cargo-bound membrane receptors to clathrin-coated sites, a process called cargo loading[176,177]. By increasing or decreasing the local concentration of adaptors, clathrin recruitment can be halted or sped up. With balanced copy numbers, the process is more stable to perturbations in copy numbers, and therefore less efficiently tuned.

Despite the underexpression of adaptor proteins, we observed a very high adaptor to triskelia ratio in completed vesicles (~19). A single triskelion can bind three SLA2 and three ENT1/2 proteins, which can bind three EDE1 and SYP1

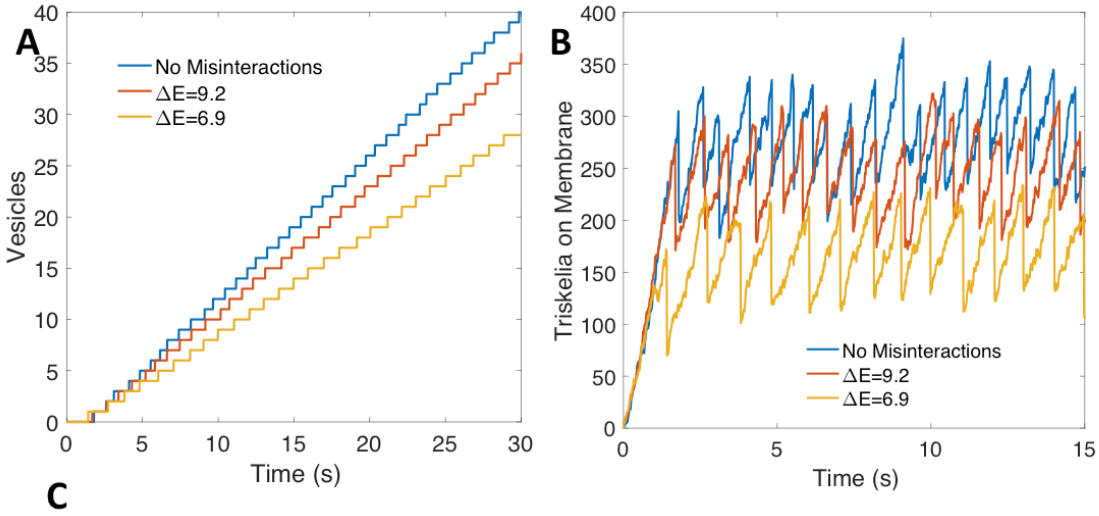
proteins, leading to a seeming saturation of 12 adaptors per triskelion. However, most of these proteins can also dimerize with a strong affinity, allowing them to bind to other complexes of adaptor proteins. Our model lacks steric hindrance that would otherwise prevent this high level of aggregation, but nonetheless there is a clear gap in strength between adaptor protein interactions and clathrin interactions (Table 4.1). The weakness of these clathrin interactions, particularly polymerization (~100 μM)[161], prevent spontaneous cage formation in the cytosol. It is the aggregation of adaptor proteins and localization to the 2D cell membrane that allows cage formation to occur; at least 81% of triskelia were brought to the membrane by adaptor proteins. This suggests another possible reason for overexpression of clathrin: to compensate for lower binding affinity by saturating adaptor proteins.

4.2.4 Misinteractions have a significant impact for strong-binding interactions

To determine the overall influence of misinteractions on vesicle formation, and its dependence on protein binding affinity, we added misinteractions at two different strengths (Methods), with an average ratio of $K_{D,\text{nonspecific}}$ to $K_{D,\text{specific}}$ of 10,000 and 1,000. Despite the weakness of the misinteractions, they decreased the frequency of vesicle formation (Fig 4.4a,b), though this effect was overall less significant than that of copy number alteration (Fig 4.3).

In section 2, we found that strong-binding proteins are more sensitive to stoichiometric balance because they are prone to misinteractions. The strongest binders in the network are the clathrin heavy-chain to light chain interaction (Table

4.1), and they are both more highly expressed relative to the adaptor partners. Misinteractions dramatically increased the number of both heavy and light chains that were not properly assembled into triskelion (~10 fold), because they became trapped in misinteractions (Fig 4.4c). For the weaker binding adaptor proteins, the misinteractions increased non-functional aggregation but to a much lower extent, resulting in about 2-fold increase of adaptor proteins in vesicle complexes. Although this 2-fold increase may seem high given the weakness of the misinteractions, it is driven by the localization of these adaptor proteins on the membrane, which concentrates the proteins and promotes binding between any pair of available binding interfaces. [158]



C

Misinteractions	Adaptors / Triskelia	Excess CHC1[1]	Excess CLC1[1]	ENT1/2	SLA2	SYP1	EDE1	YAP1801/2
None	18.7	29	36	553	366	479	470	3
ΔE=9.2	22.5	71	88	674	438	576	560	2
ΔE=6.9	33.2	208	262	988	668	857	802	6

[1]Excess means not part of a complete triskelia complex. The molecules may be in a partial complex

Figure 4.4. Misinteractions interfere with clathrin recruitment. **(A)** Adding misinteractions to the network decreased vesicle formation and **(B)** interfered with recruitment of triskelia to the membrane. This was caused by aggregates containing too many adaptor proteins, draining them from the cytoplasmic pool. **(C)** Average adaptor proteins in each vesicle. With strong misinteractions, vesicle aggregates contained many adaptors and incomplete triskelia.

Ultimately, misinteractions reduced the frequency of vesicle formation because each vesicle contained a very large aggregate of proteins that drained the cytoplasmic pool of adaptors needed to form new vesicles. The adaptor protein composition is shown in Fig 4.4c. Without misinteractions, vesicles had an average of 18.7 adaptor proteins per full triskelia, whereas strong misintearctions increased the ratio to 33.2. An interesting consequence of misinteractions is that it initially sped up the formation of the first vesicle, due to the large aggregates assembling on the membrane. However, subsequent vesicles were slower to accumulate than without misinteractions. In contrast, without misinteractions, the speed of initial vesicle formation always correlates with the speed of subsequent vesicles formed.

4.3 Discussion

4.3.1 Perfect balance is not observed, even if it would improve the functional outcome of the protein network

We do not expect the cell to perfectly optimize the yield of all of its many assemblies. Each network we have evaluated here is ultimately part of a larger, global cellular network. Perfectly optimizing isolated, local modules does not appear to be a significant pressure for the cell, particularly when a sufficient balance, such as we observe for the vesicle-forming module, maintains functionality. Correlations

in copy numbers are nonetheless often significant relative to randomly assigned copy numbers.

In Chapter 3 we found that copy number imbalance can lead to misinteractions and the features of biological IINs (power-law-like degree distribution, square and hub motifs, sparseness) typically have fewer misinteractions under balanced copy numbers but more misinteractions under imbalance. These networks thus should require more tightly controlled balance to avoid misinteractions. But misinteractions are of course not the only pressure on copy numbers. For multi-protein assembly in an obligate complex (ARP2/3) and in a minimal model of vesicle formation for CME, we found that the functional cost of imbalance was dominated more by its impact on determining specific functional complexes than avoiding misinteractions. Nonetheless, the fact that misinteractions can decrease vesicle formation, by sequestering away adaptor proteins into large aggregates, shows that misinteractions are worse than simply having an excess of free proteins. If this result can be generalized, it may have important implications for mechanistic modeling of biological systems, as misinteractions and other system errors are rarely taken into account.

4.3.2 Observed imbalances in the non-equilibrium vesicle forming module could provide benefits to assembling cargo-selective vesicles

Although the functional effects of copy number balance are usually discussed in the context of number of complete complexes at equilibrium, we have shown that non-equilibrium dynamics can be affected as well. While the clathrin heavy chains and

light chains were balanced with each other, they were overexpressed compared to their adaptor proteins, and this limited the frequency of vesicle formation. Although we found that perfectly balanced copy numbers therefore improved vesicle formation frequency compared to observed copy numbers, we speculate that specific imbalances could still be selected for evolutionarily. There are various possible reasons for this imbalance: the function of endocytosis is cargo uptake, and there is a cargo loading process before endocytosis occurs.[176,177] Hence to maximize function, controlled endocytosis around high-cargo areas of the membrane may be preferably to frequent, spontaneous endocytosis, and the adaptor proteins can serve as an intentional bottleneck in the process. Clathrin, which cannot directly bind to the membrane, may be kept at a high expression in the cytosol so that there is enough triskelia to quickly form a vesicle no matter where the endocytic site occurs. However, the observed underexpression could also result from other adaptor proteins not included in our model, or because clathrin interactions have weaker affinities than interactions between adaptor proteins and must saturate them.

Finally, the predictions of our minimal vesicle-forming model are ultimately limited by the approximations we made to simulate the clathrin coat assembly and vesicle formation. Our model vesicles formed about 10 times faster than is observed *in vivo*. To fully capture the dynamics of this complex process, an ideal model would include all the proteins in our CME network (Fig 2.1), and include both the known biochemistry of binding interactions and the physics and biomechanics of membrane bending and scission. In yeast, the cytoskeleton is needed to help induce membrane budding, after which energy-consuming proteins such as dynamin

scission off the vesicle from the plasma membrane for transport into the cell [177,178]. However, such a modeling approach does not exist, due to the computational limitations of simulating such large complexes and membrane remodeling, and the lack of biochemical data.

Based on the model we did construct, however, there are some more specific limitations. The first is that while rule-based modeling is a convenient way to model complex formation, some theoretical aggregates may be impossible due to steric hindrance. Our model predicted that a vesicle of 100 triskelia could contain ~1900 additional proteins. Assuming each vesicle is a sphere with 100nm diameter, the allowable surface area per adaptor/scaffold protein would only be $\sim 17 nm^2$, which is too small to accommodate the excluded volume of the large, disordered regions of proteins such as ENT1/2[179]. Second, we did not include cooperativity in our model. Molecules localized in the same aggregate do not interact at a faster rate in conventional rule-based modeling. Clathrin triskelia weakly polymerize, as noted above, but the aggregation effect of the adaptor proteins – especially the SYP1/EDE1 complex – localizes triskelia close together, allowing them to bind strongly. Future work could consider effects of cooperativity on assembly, as well as construct more detailed spatial and structural models of the vesicle forming process.

4.4 Methods

4.4.1 ARP2/3 Complex

The model was simulated to equilibrium using a stochastic simulation method (the Gillespie algorithm). Binding interactions were encoded via the rule-based language

BioNetGen and simulated via the Network Free Simulation (NFSim) software [151]. Trimer cooperativity was modeled by increasing the rate of the third reaction if three members of a correct trimer were held together by two reactions. For example, if A is bound to B is bound to C, and a binding between A and C is possible, that reaction rate was set to be arbitrarily high. Reaction rates were arbitrary, but interactions with the core subunit ARC19 were set to be ~10 fold stronger than interactions between periphery subunits, as this increased yield. Yield was measured via the equation

$$Yield = \frac{N_{desired}}{N_{desired} + N_{undesired}} \qquad \text{Eq. 4.1}$$

Where N_{desired} is the number of *proteins* in complete complexes (equal to seven times the number of complex complexes) and $N_{\text{undesired}}$ is the number of proteins in incomplete or misbound complexes. Completely free proteins were ignored.

4.4.2 Simulating clathrin recruitment to the membrane

A subnetwork of nine proteins – clathrin heavy chain (CHC1), clathrin light chain (CLC1), SLA2, ENT1/2, EDE1, SYP1, and YAP1801/2 – was defined based on known binding interactions (Table 4.1). Because the existence of multiple interfaces, allowing noncompetitive binding, results in a large number of possible species we simulated our model using the Network Free Simulator (NFSim)[151]. Binding dissociation constants were obtained from the literature, including for protein-lipid binding. (Table 4.1) For simplicity, the heavy chains were already assumed to be in

trimer form, and ENT1/2 was combined into a single protein as the binding partners were the same.

The cell membrane and the cell cytoplasm function as different compartments with different volumes, but NFSim is not integrated with BioNetGen's compartment language. We bypassed this problem by doubling the number of rules: besides the main rule for each reaction, an additional rule stated that if both proteins are on the cell membrane then the k_{on} rate should be increased according to the membrane volume. Cell membrane 'volume' was determined by multiplying the membrane surface area by a factor 2σ=2 nm to capture the change in binding affinities between 3D and 2D (see Appendix E).

Since our primary goal was to measure clathrin recruitment to the membrane, any complex on the membrane with at least 100 triskelia (a complex of three CHC1 and three CLC1) was considered a "vesicle" and deleted at a high rate k_{dump}. Proteins in the vesicle were then added back to the cytoplasmic pool at a rate k_{recyc}, which was set to be equal to k_{dump} to indicate fast recycling. However, we clarify that even fast recycling is not instantaneous, and that proteins are added back one at a time rather than all at once. Fast vesicle formation thus could still drain the pool of adaptor proteins.

Misinteraction strengths were determined by calculating the geometric mean of the dissociation constants of each interface, as this provided a K_D based on the arithmetic mean of the binding energies.

$$K_{D,mean} = \sqrt[n]{K_{D,1} * K_{D,2} \ldots * K_{D.n}}$$

$$= \sqrt[n]{e^{-\Delta E_1 / K_B T} * e^{-\Delta E_2 / K_B T} \ldots * e^{-\Delta E_n / K_B T}}$$

$$= \sqrt[n]{e^{-(\Delta E_1 + \Delta E_2 + \ldots \Delta E_n) / K_B T}}$$

$$= e^{-(\Delta E_1 + \Delta E_2 + \ldots \Delta E_n) / n K_B T}$$

The K_D of a misinteraction between two interfaces was set to be:

$$f * \sqrt{K_{D,mean,1} * K_{D,mean,2}} \qquad \text{Eq. 4.2}$$

where f=10,000 (weak misinteractions, corresponding to an energy gap of ~9.21 J) or 1,000 (stronger misinteractions, energy gap of ~6.91 J)

Network maps were generated using Cytoscape[150] and RuleBender[7]. Plots were generated in MATLAB. See Appendix E for BioNetGen code for the ARP2/3 complex and the clathrin recruitment model, as well as further notes on the model.

Chapter 5. Conclusions

5.1 Results Summary

In this dissertation I have shown that interface-interaction networks (IINs) have a distinctive topology that can be explained by pressure to maintain high binding specificity: that is, the two constraints of making functional binding strong while making nonfunctional binding weak. By organizing the network into motifs that allow complementary binding (hubs, squares, and pairs), the proteins can optimize their amino acid sequences so as to bind strongly to their intended partners while binding weakly to their unintended partners. When the additional constraint of keeping protein diversity minimized is added, IINs form a scale-free, fragmented structure into distinct modules. These are the same properties observed into two manually-curated biological IINs: that of the clathrin-mediated endocytosis (CME) network in yeast and the ErbB signaling network in humans.

I have also shown that the CME network has statistically significant levels of protein copy-number balance, a strategy cells use to minimize waste and prevent misinteractions. Proteins that participated in transient interactions – such as kinases and phosphatases – were out of balance. The ErbB network – being a signaling network – does not have this level of balance, but shows designed imbalance by increasing protein abundance as one travels downstream through the signaling pathways. I have shown that protein copy numbers and IIN topology can be co-optimized to prevent misinteractions. Under biological IIN properties, misinteraction frequency is lower than in random networks provided that copy

numbers are balanced. But misinteraction frequency is higher under imbalance, suggesting further pressure to maintain copy number balance for strong binding interactions. Increasing overall binding strength may strengthen functional interactions, but also increases misinteraction frequency under imbalance, meaning that there is an upper limit of binding strength for out-of-balance proteins.

Finally, I analyzed some functional consequences of imbalance and misinteractions in the CME network using a dynamic model of vesicle formation. Misinteractions decrease vesicle formation by sequestering adaptor proteins – already underexpressed compared to clathrin – into aggregates. Clathrin heavy chains and light chains – which bind strongly to each other – are balanced. But they are overexpressed compared to adaptor proteins – which bind moderately strongly to each other but weakly to clathrin. Because of the difference in binding affinities, clathrin may be overexpressed in order to saturate the adaptor proteins.

5.2 Medical Relevance

Protein aggregation has been implicated in several neurodegenerative diseases, including Parkinson's, Alzheimer's, and Huntington's. Ciryam et al. implicated "supersaturated" proteins (high abundance, low solubility) as involved in all three diseases[35], as would be expected since these proteins are prone to misinteractions and aggregation. If other aggregation-prone proteins can be identified, then we may be able to predict what gene copy number variations (CNVs) are linked to what diseases. CNVs are known to be a part of several human disease including various cancers and multiple sclerosis[13,14,180]. My results state that strongly binding proteins

are kept in balance to avoid misinteractions, but these "sticky" proteins are prone to misinteractions even if their specificity is high. This means that even doubling their concentration may increase misinteraction frequency, compromising cell function.

Interface-interaction networks can greatly improve the nascent field of network medicine. In fact, the large-scale automatic human IIN discussed in chapter 2 was created specifically to study disease networks[67]. Mutations in protein network modules are not a good predictor of diseases, but mutations in domain-binding network modules are[64]. Hence there is a need to understand the correct structural properties of IINs, but automatically constructed IINs suffer from numerous biases. By comparing the topology of automatically constructed IINs to the topology of the more accurate manually curated IINs, these biases could be corrected in future large-scale studies, hopefully capturing more accurate structural interaction data. The automatically generated IINs also failed to include interactions mediated by short linear motifs. These motifs are a common mediator of both functional and nonfunctional interactions, and thus likely play a role in aggregated-based disease states.

5.3 Future Directions

One unanswered question is whether the network topology determines which proteins are more sensitive to imbalance. We know that for incomplete complexes, overexpression of core proteins is more deleterious than overexpression of periphery proteins[4]. This suggests that "party hubs" (hub proteins with many

interfaces) would be sensitive as well, whereas "date hubs" (hub proteins that bind competitively with many partners, using few interfaces) may or may not be.

This project added two more dimensions to a traditional protein-protein interaction network: the binding interfaces used and the abundances of the proteins. A third dimension would be binding affinities between the partners. The results of Hein et al. state that strongly bound proteins are more likely to be balanced than weakly binding ones[10]. By adding this dimension as an edge property, one can weigh the stoichiometric balancing algorithm we provide in Chapter 3 to prioritize balance for some edges over others. Difference in affinities is also a possible explanation for imbalance, such as the case for clathrin binding to adaptor proteins. However, correct binding affinities are lacking for several protein interactions, and so more data will need to be collected before this can be done.

Our misinteraction toy models only used binary, competitive binding. But recent studies have postulated that keeping complex subunits out-of-balance may be used to decrease the likelihood of misassembly[32]. Toy networks utilizing noncompetitive binding – allowing both misinteractions and misassembly – may be used to confirm whether some network topologies are better at increasing functional complex yield than others.

Appendix A. Supplementary Methods for IIN Sampling

A.1 Modified IIN Sampling Approaches

A.1.1 Self-loop isolation. Our sampling procedure did not initially apply any penalty to highly connected self-binding interfaces, resulting in their random distribution in sampled IINs (Fig S1). By adding a simple penalty against high connectivity for self-binding interfaces, we could reproduce the accurate isolation of these nodes without affecting other network properties.

A.1.2 Unbiased shuffle. We also sampled IINs for a given PPIN but kept the number of interfaces per protein fixed as in the PPIN. The only move was then to allow edges to move between these interfaces. This sampling produced similar results to the unbiased sampling of the full range of IINs (Table S6).

A.1.3 Modified fitness function. We modified our fitness function penalty on the total interfaces to test whether allowing larger fluctuations in the number of interfaces per protein would improve the IIN selectivity for scale-free or random PPINs. For the modified fitness function, absolute number of interfaces was penalized instead of interfaces per protein, such that the μ term of the fitness equation was changed to $e^{\mu(N_{interfaces}-N_{proteins})}$ and μ was lowered to 0.032 to produce a realistic number of interfaces.

A.2 Calculated Properties of Networks

A.2.1 Global clustering coefficient. Given by

$$C_{global,3} = \frac{3N_{triangle}}{N_{open}+N_{triangle}} \qquad \text{Eq A.1}$$

where N_{open} is the number of open triplets and $N_{triangle}$ the number of closed triplets.

A.2.2 Four-node motifs, or tetramers. These were enumerated by finding all four-node subgraphs connected by at least one path, and determining which of six possible architectures each subgraph matched: chain, square, hub, flag, 5-edge, or 6-edge. A single node may belong to more than one subgraph, but a subgraph of four nodes may only be classified as one of the six motifs. The ratios of the amount of each motif to the total number of tetramers were used as a global statistic of the likelihood of each motif. We refer to this as the motif frequency. The three subgraphs with clustering (flags, 5-edge, and 6-edge subgraphs) were grouped into a single frequency due to their rarity.

A.2.3 Fragmentation. The fragmenting or modularity of the network was quantified using the size of the largest component in the network. To normalize, we also calculated the percentage of network interfaces contained in the largest component.

A.3 Quantifying network degree distribution without orphan nodes

To establish the degree distribution of an observed network, we needed to best match that network with networks of the same size but varying degree distributions. To generate the networks for comparison (with N nodes and M edges), we had to modify the algorithm of Goh et al. to prevent orphan nodes (k=0). To summarize, beginning with N nodes, each node is assigned an individual weight of $1^{-\alpha}$, $2^{-\alpha}$, $3^{-\alpha}$ … $N^{-\alpha}$. Edges are then added by selecting two nodes with probabilities equal to the normalized weights. Self-edges were allowed and if an edge already existed then another pair of nodes would be selected. To prevent orphans, we performed this procedure with M-R edges, and used the remaining R edges to connect the orphans back into the network using the same probabilities as above. If there were too many orphans to reconnect, the network was discarded and the procedure rerun. The optimal value of R was defined through a recursive formula that was based on the expected number of orphans produced by the unmodified algorithm. Specifically, we found $R = \lim_{n\to\infty} a_n$ where $a_n = \sum_{i=1}^{N}\left(1 - i^{-P.A.E.} \Big/ \sum_{j=1}^{N} j^{-P.A.E.}\right)^{2(M-a_{n-1})}$ and a_0=0. Without orphans, the sparse networks in particular ($\langle k\rangle \approx 1$) were more similar to one another regardless of the P.A.E., since each node was required to have at least one connection.

A.4 Theoretical properties of IINs constrained to PPINs

The expected number of interfaces for a protein of degree *k* can be calculated from the probability mass function of such a protein having *n* interfaces, where *n* varies from 1 to *k*. We find this distribution is captured by normalizing the Stirling numbers of the second kind:

$$S_k^{(n)} = \frac{1}{n!}\sum_{i=0}^{n}(-1)^i \binom{n}{i}(n-i)^k. \qquad \text{Eq. A.2}$$

The normalization factor is the Bell number, introduced in the main text, that counts the total number of ways to partition the *k* edges into interfaces, $B_k = \sum_{n=0}^{k} S_k^{(n)}$. The expected number of interfaces for a protein of degree k is

$$\langle n \rangle_k = \sum_{n=1}^{k} n S_k^{(n)} / B_k \qquad \text{Eq. A.3}$$

and values for proteins in both manually curated PPINs are reported in Table S1. The expected number of interfaces per IIN is then the sum over all the proteins, and is 200 for the CME PPIN and 411 for the ErbB PPIN (when duplicate edges are included-see Table S1).

The distribution of IIN sizes (in number of interfaces) for a given PPIN is the convolution over all proteins of their Stirling distributions. Each Stirling distribution is narrower and left shifted compared to a Binomial distribution, and their convolution results in an explosion of possible networks centered around the expected interface size. Sparse and dense IINs are then extremely rare.

The total number of IINs for a PPIN is the product of its protein's Bell numbers.

$$Total\ IINs\ possible = \prod_{j=1}^{N_{pro}} B_{\deg(j)} \qquad \text{Eq. A.4}$$

where deg(j) is the degree of protein j. Scale-free PPINs will have significantly more types of IINs possible relative to a random PPIN given the very large Bell numbers of their hub proteins.

To determine the effect of PPIN structure on the degree distribution of IINs, we resorted to computational approaches, and used the unbiased MC sampling ($k_BT=\infty$). Results in Fig. S4 show that most IINs have a random degree distribution, but it is more probable to produce a scale-free IIN from scale-free PPINs than random PPINs.

A.5 Different construction methods for biological IINs

A.5.1 ErbB network. We analyzed two versions of the ErbB network, the original version, where every phosphosite is a separate interface, and a reduced version, where copies of interfaces on a protein with the same specificity for binding partners were represented as a single interface. This reduced the network from 387 interfaces to 303, and from 545 edges to 417, mostly due to the large number of phosphosites per protein that often were all targeted by the same kinase domains. The results of the motif selectivities were the same for both networks, but with the smaller size and smaller number of duplicated edges, the reduced IIN was easier to sample.

A.5.2 Automatically constructed networks. The automatically constructed IINs (Fig. S2) were downloaded from the studies of Wang et al [67] on the human structural interaction network (hSIN), and from the study of Deeds et al [70], whose cytoplasmic yeast SIN network was originally constructed by Kim et al [62]. In both cases domains were assigned to the PPINs using the crystal structures of bound proteins complexes in *i*PFam. Since most protein complexes have not been crystallized, if the proteins in the PPIN contained domains with homologs that interacted, these were assigned as the predicted domain-domain interactions. Either because *i*PFam contains a limited number of linear motif interactions, or linear motifs are not recorded as known domains for specific proteins by PFam, assignments of binding sites such as PRRs and phosphosites as partners were not captured. Also, interfaces were assigned by domain, but the two are not synonymous because protein domains can be large and contain multiple, distinct binding interfaces.

A.5.3 Automatic construction of the CME IIN. Using the S. cerevisiae database on the INstruct website [96], we used the hSIN method of Wang et al to reconstruct the Clathrin-mediated endocytosis IIN. As Fig. S2d shows, only 44 interactions are present, compared to 206 in the manually curated network. Nearly all of the predicted domain-domain interactions are incorrect, or they are assigned to proteins that are not actually observed to bind directly to one another after reading the cited literature. Manual curation, while tedious, allows one to use biochemical data that identifies binding sites, whether a crystal structure exists or not.

Appendix B. CME and ErbB Network Interactions

Table B.1 Clathrin-mediated endocytosis network

206 edges between 195 interfaces

Interface 1	Interface 2	Interface 1 Type	Interface 2 Type
ABP1.0	ACT1.2	Cofilin Domain	Filament, subunit 1, 2
ABP1.1	RVS167.0	PRR	SH3 domain
ABP1.2	AIM21.2	SH3	PRR
ABP1.2	APP1.1	SH3	PRR
ABP1.2	ARK1.1	SH3	PRR
ABP1.2	BSP1.2	SH3	PRR to ABP1
ABP1.2	INP52.1	SH3	PRR
ABP1.2	PRK1.2	SH3	PRR
ABP1.2	SCP1.1	SH3	PRR
ABP1.3	ARP2.6	Acidic Domain	side, front to acidic motifs
ABP1.3	ARP3.5	Acidic Domain	side, front to acidic motifs
ABP1.4	LSB3.1	PRR	SH3 domain
ABP1.5	YSC84.1	PRR	SH3 domain
ABP1.6	SLA1.4	PRR	SH3 domain 1-2
ABP1.6	SLA1.7	PRR	SH3 domain 3
ACT1.0	ACT1.1	Back, Barbed End	Back, Pointed End
ACT1.0	CAP1.0	Back, Barbed End	interface to actin
ACT1.0	CAP2.1	Back, Barbed End	interface to actin
ACT1.0	PFY1.0	Back, Barbed End	interface to actin
ACT1.1	ARP2.4	Back, Pointed End	back, barbed end to actin
ACT1.1	ARP3.4	Back, Pointed End	back, barbed end to actin
ACT1.2	ARC18.1	Filament, subunit 1, 2	interface to actin filament
ACT1.2	ARC35.2	Filament, subunit 1, 2	interface to actin filament
ACT1.2	ARC40.2	Filament, subunit 1, 2	interface to actin filament/acidic motifs
ACT1.2	COF1.0	Filament, subunit 1, 2	cofilin domain
ACT1.2	LAS17.4	Filament, subunit 1, 2	WH2 domain
ACT1.2	MYO3.1	Filament, subunit 1, 2	head domain
ACT1.2	MYO5.3	Filament, subunit 1, 2	head domain
ACT1.2	PAN1.3	Filament, subunit 1, 2	coiled coil domain. Contains WH2 actin binding region 1142-~1190)
ACT1.2	SAC6.0	Filament, subunit 1, 2	CH domain
ACT1.2	SCP1.0	Filament, subunit 1, 2	actin bundling interface (Calpnonin like repeat, not the CH domain)
ACT1.2	SLA2.1	Filament, subunit 1, 2	Talin like domain (ILWEQ) 760-895 followed by coiled-coil

ACT1.2	TWF1.0	Filament, subunit 1, 2	cofilin domain
ACT1.2	VRP1.4	Filament, subunit 1, 2	WH2 domain
ACT1.2	YSC84.2	Filament, subunit 1, 2	#N/A
ACT1.3	AIP1.1	Filament, subunit 4	interface to ACT1, binds simultatneously with COF1
AIM21.0	BBC1.0	PRR	SH3 domain
AIM21.0	YSC84.1	PRR	SH3 domain
AIM21.1	CAP1.3	Speculated to CAP1/2	speculated interface to AIM21
AIM21.1	CAP2.3	Speculated to CAP1/2	speculated interface to AIM21
AIM21.2	RVS167.0	PRR	SH3 domain
AIM3.0	LSB3.1	PRR	SH3 domain
AIM3.0	RVS167.0	PRR	SH3 domain
AIM3.0	YSC84.1	PRR	SH3 domain
AIM3.1	ABP1.2	PRR	SH3
AIM3.2	PRK1.1	Phospho site	kinase domain
AIP1.0	COF1.1	interface to COF1, binds simultaneously with ACT1	interface to AIP1
AKL1.0	AKL1.0	Homodimer interface, probably functional form, would not interfere with kinase activity	Homodimer interface, probably functional form, would not interfere with kinase activity
AKL1.1	EDE1.4	Kinase domain	phospho site
AKL1.1	ENT2.0	Kinase domain	phospho site
AKL1.1	LSB3.2	Kinase domain	phospho site
AKL1.1	SCD5.1	Kinase domain	phospho site
AKL1.1	SLA1.6	Kinase domain	phospho site, probably in C-term repeats region
APL1.0	APM4.1	beta-mu subunit interface	mu-beta subunit interface
APL1.1	APS2.1	beta-sigma subunit interface	sigma-beta subunit interface
APL1.2	APL3.2	beta-alpha subunit interface	alpha-beta subunit interface
APL3.0	APS2.0	alpha-sigma subunit interface	sigma-alpha subunit interface
APL3.1	APM4.0	alpha-mu subunit interface	mu-alpha subunit interface
APM4.2	APS2.2	mu-sigma subunit interface	sigma-mu subunit interface
APP1.0	LSB3.1	PRR	SH3 domain
APP1.0	RVS167.0	PRR	SH3 domain
APP1.0	YSC84.1	PRR	SH3 domain
APP1.1	BBC1.0	PRR	SH3 domain
APP1.1	BZZ1.0	PRR	SH3 domain
APP1.1	MYO5.0	PRR	SH3 domain
APP1.1	RVS167.0	PRR	SH3 domain
ARC15.0	ARC19.2	interface to arc19	interface to arc15
ARC15.1	ARC40.1	interface to arc40	interface to arc15
ARC15.2	ARP2.1	interface to arp2	interface to arc15
ARC18.0	ARP3.2	interface to arp3	interface to arc18
ARC19.0	ARC35.0	interface to arc35	interface to arc19
ARC19.1	ARC40.0	interface to arc40	interface to arc19
ARC19.3	ARP2.3	interface to arp2	interface to arc19
ARC19.4	ARP3.1	interface to arp3	interface to arc19
ARC35.1	ARP3.3	interface to arp3	interface to arc35
ARC40.2	CRN1.2	interface to actin filament/acidic motifs	acidic domain
ARC40.2	LAS17.1	interface to actin	acidic domain

		filament/acidic motifs	
ARC40.2	MYO3.2	interface to actin filament/acidic motifs	acidic domain
ARC40.2	MYO5.2	interface to actin filament/acidic motifs	acidic domain
ARC40.2	PAN1.5	interface to actin filament/acidic motifs	acidic domain
ARK1.0	LSB3.2	kinase domain	phospho site
ARK1.0	PAN1.2	kinase domain	phospho site, LR2 multiple consensus
ARK1.0	SLA1.6	kinase domain	phospho site, probably in C-term repeats region
ARP2.0	LAS17.0	barbed end groove, to C helix	C helix
ARP2.2	ARP3.0	interface to arp3	interface to arp2
ARP2.5	PRK1.0	speculated, to PRK1	interface to arp2
ARP2.6	LAS17.1	side, front to acidic motifs	acidic domain
ARP2.6	MYO3.2	side, front to acidic motifs	acidic domain
ARP2.6	MYO5.2	side, front to acidic motifs	acidic domain
ARP2.6	PAN1.5	side, front to acidic motifs	acidic domain
ARP3.5	CRN1.2	side, front to acidic motifs	acidic domain
ARP3.5	MYO3.2	side, front to acidic motifs	acidic domain
ARP3.5	MYO5.2	side, front to acidic motifs	acidic domain
ARP3.5	PAN1.5	side, front to acidic motifs	acidic domain
ARP3.6	CRN1.3	barbed end groove, to C helix	C helix
BBC1.0	BBC1.1	SH3 domain	PRR
BBC1.0	LAS17.2	SH3 domain	PRR_3
BBC1.0	LAS17.6	SH3 domain	PRR_2
BBC1.1	MYO3.0	PRR	SH3 domain
BBC1.1	MYO5.0	PRR	SH3 domain
BSP1.0	LSB3.1	PRR to LSB3, 4	SH3 domain
BSP1.0	YSC84.1	PRR to LSB3, 4	SH3 domain
BSP1.1	INP52.0	interface to INP52/not PRR	N terminal region
BZZ1.0	MYO5.1	SH3 domain	PRR
BZZ1.0	VRP1.1	SH3 domain	PRR
CAP1.1	CAP2.0	dimer interface to CAP2	dimer interface to CAP1
CAP1.2	TWF1.1	speculated interface to TWF1	speculated interface to CAP1
CAP2.2	TWF1.2	speculated interface to TWF1	speculated interface to CAP2
CHC1.0	CHC1.0	self binding to trimers	self binding to trimers
CHC1.1	ENT1.0	N terminal domain	Clathrin Box
CHC1.1	ENT2.1	N terminal domain	Clathrin Box
CHC1.1	YAP1802.2	N terminal domain	Clathrin Box
CHC1.2	CLC1.0	interface to light chain	interface to heavy chain
CLC1.1	SLA2.4	interface to SLA2	*interface to CLC1, expected non-competing with SLA2.0 partners or homodimerization
CRN1.0	ACT1.2	coiled coil domain	Filament, subunit 1, 2
CRN1.0	CRN1.0	coiled coil domain	coiled coil domain
CRN1.1	ACT1.3	beta propeller domain	Filament, subunit 4
CRN1.2	ARP2.6	acidic domain	side, front to acidic motifs
CRN1.3	ARP2.0	C helix	barbed end groove, to C helix

EDE1.1	ENT1.1	EH domain	NPF motifs
EDE1.1	ENT2.2	EH domain	NPF motifs
EDE1.2	PAL1.0	speculated, to PAL1	speculated interface to EDE1
EDE1.3	EDE1.3	coiled coil domain	coiled coil domain
END3.1	PAN1.1	C terminal region	long repeat 2
ENT1.1	PAN1.0	NPF motifs	EH domain 2
ENT2.0	PRK1.1	phospho site	kinase domain
GTS1.0	YSC84.1	PRR	SH3 domain
GTS1.1	YAP1802.1	speculated interface to YAP1801	speculated to GTS1
LAS17.0	ARP3.6	C helix	barbed end groove, to C helix
LAS17.1	ARP3.5	acidic domain	side, front to acidic motifs
LAS17.2	BZZ1.0	PRR_3	SH3 domain
LAS17.2	LSB3.1	PRR_3	SH3 domain
LAS17.2	MYO3.0	PRR_3	SH3 domain
LAS17.2	MYO5.0	PRR_3	SH3 domain
LAS17.2	RVS167.0	PRR_3	SH3 domain
LAS17.6	LSB3.1	PRR_2	SH3 domain
LAS17.7	BZZ1.0	PRR_0	SH3 domain
LAS17.8	LSB3.1	PRR_1	SH3 domain
LAS17.8	RVS167.0	PRR_1	SH3 domain
LAS17.9	MYO3.0	PRR_4	SH3 domain
LAS17.9	MYO5.0	PRR_4	SH3 domain
LSB3.0	LSB3.0	dimerization interface	dimerization interface
LSB3.1	GTS1.0	SH3 domain	PRR
LSB5.1	LAS17.6	VHS domain (LAS17 also binds residues 40-213)	PRR_2
MYO3.0	PAN1.6	SH3 domain	PRR
MYO5.0	MYO5.1	SH3 domain	PRR
MYO5.0	VRP1.3	SH3 domain	PRR
MYO5.1	MYO3.0	PRR	SH3 domain
PAN1.0	ENT2.2	EH domain 2	NPF motifs
PAN1.0	YAP1801.2	EH domain 2	NPF motifs (5) Cterm
PAN1.2	PRK1.1	phospho site, LR2 multiple consensus	kinase domain
PAN1.3	PAN1.4	coiled coil domain. Contains WH2 actin binding region 1142-~1190)	N terminal long repeat 1
PAN1.6	MYO5.0	PRR	SH3 domain
PRK1.1	ENT1.2	kinase domain	phospho site
PRK1.1	LAS17.10	kinase domain	phospho site
PRK1.1	YAP1801.1	kinase domain	phospho site
RVS167.0	BSP1.0	SH3 domain	PRR to LSB3, 4
RVS167.0	BSP1.3	SH3 domain	PRR to RVS167
RVS167.0	GTS1.0	SH3 domain	PRR
RVS167.0	VRP1.0	SH3 domain	PRR
RVS167.1	RVS161.0	dimerization interface	dimerization interface
RVS167.2	RVS167.2	oligomerization domain	oligomerization domain
SCD5.0	END3.0	region 302-500	N terminal domain
SCD5.0	PAN1.1	region 302-500	long repeat 2
SCD5.1	PRK1.1	phospho site	kinase domain
SLA1.0	CHC1.1	Clathrin Box	N terminal domain
SLA1.0	SLA1.8	Clathrin Box	SHD2 domain
SLA1.1	LSB3.1	PRR	SH3 domain

SLA1.1	RVS167.0	PRR	SH3 domain
SLA1.1	YSC84.1	PRR	SH3 domain
SLA1.2	END3.0	*C terminal repeat TGGAMMP to END3 (not competing with PAN1)	N terminal domain
SLA1.3	LSB5.0	SHD1 domain	NPF motif region sufficient (including GAT domain may strengthen interaction)
SLA1.4	APP1.1	SH3 domain 1-2	PRR
SLA1.4	BSP1.0	SH3 domain 1-2	PRR to LSB3, 4
SLA1.4	INP52.2	SH3 domain 1-2	PRR
SLA1.4	LAS17.2	SH3 domain 1-2	PRR_3
SLA1.4	SYP1.4	SH3 domain 1-2	PRR
SLA1.5	PAN1.4	*C terminal repeat to PAN1 (not competing with END3)	N terminal long repeat 1
SLA1.6	PRK1.1	phospho site, probably in C-term repeats region	kinase domain
SLA1.7	APP1.1	SH3 domain 3	PRR
SLA1.7	LAS17.2	SH3 domain 3	PRR_3
SLA1.8	SLA1.8	SHD2 domain	SHD2 domain
SLA1.9	SLA2.0	Gap1 domain	*coiled coil domain
SLA2.0	PAN1.3	*coiled coil domain	coiled coil domain. Contains WH2 actin binding region 1142-~1190)
SLA2.2	ARK1.0	phospho site	kinase domain
SLA2.2	PRK1.1	phospho site	kinase domain
SLA2.3	SLA2.3	*homodimer interface, Central coil region, not competing	*homodimer interface, Central coil region, not competing
SYP1.0	LAS17.5	interface to LAS17	speculated to SYP1
SYP1.1	LSB3.1	PRR	SH3 domain
SYP1.2	EDE1.0	C terminal region (muHD)	C terminal domain
SYP1.3	SYP1.3	BAR domain	BAR domain
VRP1.1	SLA1.4	PRR	SH3 domain 1-2
VRP1.2	LAS17.3	C terminal region	WH1 domain
VRP1.3	MYO3.0	PRR	SH3 domain
YAP1801.0	GTS1.1	speculated to GTS1	speculated interface to YAP1801
YAP1801.3	CHC1.1	Clathrin Box	N terminal domain
YAP1802.0	EDE1.1	NPF motifs (5)	EH domain
YAP1802.0	PAN1.0	NPF motifs (5)	EH domain 2
YAP1802.3	ARK1.0	phospho site	kinase domain
YAP1802.3	PRK1.1	phospho site	kinase domain
YAP1802.4	END3.1	coiled coil domain	C terminal region
YSC84.0	LSB3.0	dimerization interface	dimerization interface
YSC84.0	YSC84.0	dimerization interface	dimerization interface
YSC84.1	LAS17.2	SH3 domain	PRR_3
YSC84.1	LAS17.6	SH3 domain	PRR_2
YSC84.1	LAS17.8	SH3 domain	PRR_1

Table B.2 ErbB signaling network, full IIN

540 edges between 377 interfaces

Interface 1	Interface 2
G001_ABI1_414_PPPPPVDYEDEE	G040_EPS8_D_007_SH3_1
G002_AKT1_D_030_Pkinase	G011_CASP9_pS183
G002_AKT1_D_030_Pkinase	G011_CASP9_pS196
G002_AKT1_D_030_Pkinase	G021_CREB1_pS133
G002_AKT1_D_030_Pkinase	G043_FOXO1A_pT24
G002_AKT1_D_030_Pkinase	G043_FOXO1A_pS256
G002_AKT1_D_030_Pkinase	G043_FOXO1A_pS319
G002_AKT1_D_030_Pkinase	G045_GAB2_pS159
G002_AKT1_D_030_Pkinase	G048_GRB10_pS428
G002_AKT1_D_030_Pkinase	G126_RAC1_pS71
G002_AKT1_D_030_Pkinase	G127_RAF1_pS259
G002_AKT1_pY315	G160_SRC_D_004_Pkinase_Tyr
G002_AKT1_pY326	G160_SRC_D_004_Pkinase_Tyr
G002_AKT1_Pro_424_427	G160_SRC_D_007_SH3_1
G002_AKT1_pT308	G190_PDPK1_D_030_Pkinase
G002_AKT1_pS473	G190_PDPK1_D_030_Pkinase
G002_AKT1_D_030_Pkinase	G193_BAD_pS75
G002_AKT1_D_030_Pkinase	G193_BAD_pS99
G002_AKT1_D_030_Pkinase	G193_BAD_pS136
G004_APPL1_D_BAR	G125_RAB5A_D_123_Rab
G006_ARAF_D_035_RBD	G057_HRAS_D_042_Ras
G006_ARAF_D_030_Pkinase	G070_MAP2K1_pS218
G006_ARAF_D_030_Pkinase	G070_MAP2K1_pS222
G009_BCAR1_pY115	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY387	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY267	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY234	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY179	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY653	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY287	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY192	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY362	G160_SRC_D_004_Pkinase_Tyr
G009_BCAR1_pY165	G160_SRC_D_004_Pkinase_Tyr
G010_CAMK2A_D_030_Pkinase	G021_CREB1_pS133
G010_CAMK2A_D_030_Pkinase	G150_SMAD2_pS110
G010_CAMK2A_D_030_Pkinase	G150_SMAD2_pS240
G010_CAMK2A_D_030_Pkinase	G150_SMAD2_pS260
G012_CAV1_pY14	G160_SRC_D_004_Pkinase_Tyr
G013_CAV2_pY19	G160_SRC_D_004_Pkinase_Tyr
G014_CBL_D_013_Cbl_N	P_032_EGFR_1063_to_1075
G014_CBL_D_015_Cbl_N3	P_032_EGFR_1063_to_1075
G014_CBL_pY700	G032_EGFR_D_004_Pkinase_Tyr
G014_CBL_pY731	G032_EGFR_D_004_Pkinase_Tyr
G014_CBL_pY774	G032_EGFR_D_004_Pkinase_Tyr
G014_CBL_491_ASLPPVP	G050_GRB2_D_007_SH3_1_D2
G014_CBL_817_SQVPERPPKPFPRRINSY	G146_SH3KBP1_D_007_SH3_1_D1
G014_CBL_817_SQVPERPPKPFPRRINSY	G146_SH3KBP1_D_007_SH3_1_D2

G014_CBL_D_013_Cbl_N	G159_SPRY2_pY55
G014_CBL_D_015_Cbl_N3	G159_SPRY2_pY55
G014_CBL_D_015_Cbl_N3	G159_SPRY2_49_to_61
G014_CBL_D_015_Cbl_N3	G160_SRC_413_to_425
G014_CBL_817_SQVPERPPKPFPRRINSY	G228_ARHGEF7_D_051_SH3_2
G015_CBLB_902_PARPPK	G146_SH3KBP1_D_007_SH3_1_D1
G015_CBLB_899_SQAPARPPKPRPRRTAY	G146_SH3KBP1_D_007_SH3_1_D1
G015_CBLB_899_SQAPARPPKPRPRRTAY	G146_SH3KBP1_D_007_SH3_1_D2
G017_CDC42_D_042_Ras	G092_PAK1_D_043_PBD
G017_CDC42_D_042_Ras	G169_TNK2_D_048_GTPase_binding
G017_CDC42_D_042_Ras	G172_VAV2_D_019_RhoGEF
G017_CDC42_D_042_Ras	G228_ARHGEF7_D_019_RhoGEF
G021_CREB1_pS133	G140_RPS6KA3_D_030_Pkinase_D1
G021_CREB1_pS133	G140_RPS6KA3_D_030_Pkinase_D2
G022_CRK_pY221	G032_EGFR_D_004_Pkinase_Tyr
G022_CRK_D_006_SH2	G032_EGFR_pY1016
G022_CRK_D_006_SH2	G032_EGFR_pY1125
G024_CSK_D_004_Pkinase_Tyr	G160_SRC_pY530
G027_DNM1_751_to_805	G050_GRB2_D_007_SH3_1_D1
G027_DNM1_806_to_851	G050_GRB2_D_007_SH3_1_D1
G027_DNM1_806_to_851	G144_SH3GL2_D_007_SH3_1
G027_DNM1_751_to_805	G144_SH3GL2_D_007_SH3_1
G027_DNM1_pY231	G160_SRC_D_004_Pkinase_Tyr
G027_DNM1_pY597	G160_SRC_D_004_Pkinase_Tyr
G027_DNM1_751_to_805	G160_SRC_D_007_SH3_1
G027_DNM1_806_to_851	G160_SRC_D_007_SH3_1
G029_DUSP1_pS296	G080_MAPK1_D_030_Pkinase
G029_DUSP1_pS323	G080_MAPK1_D_030_Pkinase
G029_DUSP1_pS359	G080_MAPK1_D_030_Pkinase
G029_DUSP1_D_KIM	G080_MAPK1_D_030_Pkinase
G029_DUSP1_D_040_DSPc	G080_MAPK1
G029_DUSP1_D_KIM	G081_MAPK14_D_030_Pkinase
G029_DUSP1_D_040_DSPc	G081_MAPK14
G029_DUSP1_pS364	G082_MAPK3_D_030_Pkinase
G029_DUSP1_D_KIM	G082_MAPK3_D_030_Pkinase
G029_DUSP1_D_040_DSPc	G082_MAPK3
G029_DUSP1_D_KIM	G084_MAPK8_D_030_Pkinase
G029_DUSP1_D_040_DSPc	G084_MAPK8
G031_EGF_D_001_EGF	G032_EGFR_D_002_Recep_L_domain_D1
G031_EGF_D_001_EGF	G032_EGFR_D_002_Recep_L_domain_D2
G032_EGFR_D_004_Pkinase_Tyr	G038_EPS15_pY849
G032_EGFR_D_004_Pkinase_Tyr	G041_ERRFI1_315_374
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY285
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY307
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY406
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY447
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY472
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY619
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY657
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY689
G032_EGFR_pY1092	G044_GAB1_D_MBD
G032_EGFR_pY1110	G044_GAB1_D_MBD
G032_EGFR_pY1138	G050_GRB2_D_006_SH2

G032_EGFR_pS1064	G050_GRB2_D_006_SH2
G032_EGFR_pY1016	G088_NCK1_D_006_SH2
G032_EGFR_pY1016	G099_PIK3R1_D_006_SH2_D1
G032_EGFR_pY1016	G099_PIK3R1_D_006_SH2_D2
G032_EGFR_pY1016	G100_PIK3R2_D_006_SH2_D1
G032_EGFR_D_004_Pkinase_Tyr	G104_PLCG1_pY472
G032_EGFR_D_004_Pkinase_Tyr	G104_PLCG1_pY771
G032_EGFR_D_004_Pkinase_Tyr	G104_PLCG1_pY783
G032_EGFR_pY1016	G104_PLCG1_D_006_SH2_D1
G032_EGFR_pY1016	G104_PLCG1_D_006_SH2_D2
G032_EGFR_pY998	G105_PLCG2_D_006_SH2_D1
G032_EGFR_pY1016	G119_PTPN11_D_012_Y_phosphatase
G032_EGFR_pY1016	G119_PTPN11_D_006_SH2_D1
G032_EGFR_pY1197	G122_PTPN6_D_006_SH2_D1
G032_EGFR_pY1197	G122_PTPN6_D_006_SH2_D2
G032_EGFR_pY1197	G122_PTPN6_D_012_Y_phosphatase
G032_EGFR_D_004_Pkinase_Tyr	G131_RASA1_pY460
G032_EGFR_pY915	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY998	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY1016	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY1125	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY1172	G131_RASA1_D_006_SH2_D1
G032_EGFR_D_004_Pkinase_Tyr	G136_RGS16_pY168
G032_EGFR_D_004_Pkinase_Tyr	G136_RGS16_pY177
G032_EGFR_pY998	G147_SHC1_D_005_PID
G032_EGFR_pY1016	G147_SHC1_D_005_PID
G032_EGFR_pY1110	G147_SHC1_D_005_PID
G032_EGFR_pY1138	G147_SHC1_D_005_PID
G032_EGFR_pY1172	G147_SHC1_D_005_PID
G032_EGFR_pY1192	G147_SHC1_D_005_PID
G032_EGFR_pY869	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY915	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY944	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY998	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY1092	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY1125	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_D_004_Pkinase_Tyr	G161_STAT1_pY701
G032_EGFR_D_004_Pkinase_Tyr	G165_STAT5B_pY699
G032_EGFR_D_004_Pkinase_Tyr	G165_STAT5B_pY725
G032_EGFR_D_004_Pkinase_Tyr	G165_STAT5B_pY740
G032_EGFR_D_004_Pkinase_Tyr	G165_STAT5B_pY743
G032_EGFR	G172_VAV2_D_006_SH2
G034_ELK1_D_D-domain	G080_MAPK1_D_030_Pkinase
G034_ELK1_D_DEF-domain	G080_MAPK1_D_030_Pkinase
G034_ELK1_D_D-domain	G082_MAPK3_D_030_Pkinase
G034_ELK1_D_DEF-domain	G082_MAPK3_D_030_Pkinase
G036_EPN1_496_to_575	G038_EPS15_1_313
G038_EPS15_777_to_789	G050_GRB2_D_007_SH3_1_D1
G038_EPS15_777_to_789	G050_GRB2_D_007_SH3_1_D2
G040_EPS8_D_535_to_821	G156_SOS1
G040_EPS8_D_007_SH3_1	G170_USP6NL
G042_FOS_D_071_bZIP_1	G062_JUN_G062_D_071_bZIP_1
G042_FOS_D_DEF-domain	G080_MAPK1_D_030_Pkinase

G042_FOS_pS21	G080_MAPK1_D_030_Pkinase
G042_FOS_pS32	G080_MAPK1_D_030_Pkinase
G042_FOS_pS42	G080_MAPK1_D_030_Pkinase
G042_FOS_pS70	G080_MAPK1_D_030_Pkinase
G042_FOS_pS113	G080_MAPK1_D_030_Pkinase
G042_FOS_pS374	G080_MAPK1_D_030_Pkinase
G042_FOS_D_DEF-domain	G082_MAPK3_D_030_Pkinase
G042_FOS_pS21	G082_MAPK3_D_030_Pkinase
G042_FOS_pS32	G082_MAPK3_D_030_Pkinase
G042_FOS_pS42	G082_MAPK3_D_030_Pkinase
G042_FOS_pS70	G082_MAPK3_D_030_Pkinase
G042_FOS_pS113	G082_MAPK3_D_030_Pkinase
G042_FOS_pS374	G082_MAPK3_D_030_Pkinase
G042_FOS	G084_MAPK8_D_030_Pkinase
G044_GAB1_337_to_346	G050_GRB2_D_007_SH3_1_D2
G044_GAB1_517_to_522	G050_GRB2_D_007_SH3_1_D2
G044_GAB1_pY447	G099_PIK3R1_D_006_SH2_D1
G044_GAB1_pY472	G099_PIK3R1_D_006_SH2_D1
G044_GAB1_pY589	G099_PIK3R1_D_006_SH2_D1
G044_GAB1_pY447	G099_PIK3R1_D_006_SH2_D2
G044_GAB1_pY472	G099_PIK3R1_D_006_SH2_D2
G044_GAB1_pY589	G099_PIK3R1_D_006_SH2_D2
G044_GAB1_pY657	G119_PTPN11_D_012_Y_phosphatase
G044_GAB1_pY689	G119_PTPN11_D_012_Y_phosphatase
G045_GAB2_502_to_528	G050_GRB2_D_007_SH3_1_D2
G045_GAB2_348_to_367	G050_GRB2_D_007_SH3_1_D1
G045_GAB2_502_to_528	G050_GRB2_D_007_SH3_1_D1
G045_GAB2_pS623	G080_MAPK1_D_030_Pkinase
G045_GAB2_pS623	G082_MAPK3_D_030_Pkinase
G045_GAB2_pY643	G119_PTPN11_D_012_Y_phosphatase
G047_GJA1_pS255	G082_MAPK3_D_030_Pkinase
G047_GJA1_pS279	G082_MAPK3_D_030_Pkinase
G047_GJA1_pS282	G082_MAPK3_D_030_Pkinase
G047_GJA1_pS255	G083_MAPK7_D_030_Pkinase
G047_GJA1_pS282	G083_MAPK7_D_030_Pkinase
G047_GJA1_pS368	G113_PRKCG_D_030_Pkinase
G047_GJA1_pY247	G160_SRC_D_004_Pkinase_Tyr
G047_GJA1_pY265	G160_SRC_D_004_Pkinase_Tyr
G050_GRB2_D_007_SH3_1_D1	G061_JAK2_494_to_506
G050_GRB2_D_007_SH3_1_D2	G061_JAK2_494_to_506
G050_GRB2_D_006_SH2	G104_PLCG1_pY783
G050_GRB2_D_006_SH2	G117_PTK2B_pY881
G050_GRB2_D_006_SH2	G119_PTPN11_pY584
G050_GRB2_D_006_SH2	G147_SHC1_pY427
G050_GRB2_D_007_SH3_1_D1	G156_SOS1_1149_to_1158
G050_GRB2_D_007_SH3_1_D1	G156_SOS1_1133_to_1333
G050_GRB2_D_007_SH3_1_D1	G157_SOS2_1019_to_1032
G050_GRB2_D_007_SH3_1_D2	G157_SOS2_1019_to_1032
G050_GRB2_pY160	G160_SRC_D_004_Pkinase_Tyr
G050_GRB2_D_007_SH3_1_D2	G171_VAV1_D_051_SH3_2
G057_HRAS_D_042_Ras	G098_PIK3CG_D_025_PI3K_rbd
G057_HRAS_D_042_Ras	G098_PIK3CG_D_028_PI3_PI4_kinase
G057_HRAS_D_042_Ras	G098_PIK3CG

G057_HRAS_D_042_Ras	G127_RAF1_D_035_RBD
G057_HRAS_D_042_Ras	G130_RALGDS_D_034_RA
G057_HRAS_D_042_Ras	G131_RASA1_D_011_RasGAP
G057_HRAS_D_042_Ras	G156_SOS1_D_023_RasGEF_N
G057_HRAS_D_042_Ras	G156_SOS1_D_024_RasGEF
G057_HRAS_D_042_Ras	G179_RIN1_D_034_RA
G057_HRAS_D_042_Ras	G187_BRAF_D_035_RBD
G061_JAK2_pY1007	G119_PTPN11_D_012_Y_phosphatase
G061_JAK2_D_004_Pkinase_Tyr_D1	G161_STAT1_pY701
G061_JAK2_D_004_Pkinase_Tyr_D2	G161_STAT1_pY701
G061_JAK2_D_004_Pkinase_Tyr_D1	G164_STAT5A_pY694
G061_JAK2_D_004_Pkinase_Tyr_D2	G164_STAT5A_pY694
G062_JUN_pS73	G084_MAPK8_D_030_Pkinase
G062_JUN_G062_D_D-domain	G084_MAPK8_D_030_Pkinase
G065_KRAS_D_042_Ras	G127_RAF1_D_035_RBD
G065_KRAS_D_042_Ras	G130_RALGDS_D_034_RA
G065_KRAS_D_042_Ras	G156_SOS1_D_023_RasGEF_N
G065_KRAS_D_042_Ras	G156_SOS1_D_024_RasGEF
G065_KRAS_D_042_Ras	G187_BRAF_D_035_RBD
G069_KRT8_pS74	G081_MAPK14_D_030_Pkinase
G070_MAP2K1_pS218	G075_MAP3K1_D_030_Pkinase
G070_MAP2K1_pS222	G075_MAP3K1_D_030_Pkinase
G070_MAP2K1_D_030_Pkinase	G080_MAPK1_pY187
G070_MAP2K1_pT292	G080_MAPK1_D_030_Pkinase
G070_MAP2K1_D_D-site	G080_MAPK1_D_030_Pkinase
G070_MAP2K1_D_030_Pkinase	G082_MAPK3_pT202
G070_MAP2K1_D_030_Pkinase	G082_MAPK3_pY204
G070_MAP2K1_pT292	G127_RAF1_D_030_Pkinase
G070_MAP2K1_pS222	G127_RAF1_D_030_Pkinase
G070_MAP2K1_pT286	G127_RAF1_D_030_Pkinase
G070_MAP2K1_pT386	G127_RAF1_D_030_Pkinase
G070_MAP2K1_pS218	G127_RAF1_D_030_Pkinase
G070_MAP2K1_pS222	G187_BRAF_D_030_Pkinase
G070_MAP2K1_pS218	G187_BRAF_D_030_Pkinase
G070_MAP2K1_D_030_Pkinase	G188_KSR1_D_004_Pkinase_Tyr
G071_MAP2K2_D_030_Pkinase	G080_MAPK1_pT185
G071_MAP2K2_D_030_Pkinase	G080_MAPK1_pY187
G071_MAP2K2_D_D-site	G080_MAPK1_D_030_Pkinase
G071_MAP2K2_D_030_Pkinase	G082_MAPK3_pT202
G071_MAP2K2_D_030_Pkinase	G082_MAPK3_pY204
G071_MAP2K2_pS218	G127_RAF1_D_030_Pkinase
G071_MAP2K2_pS222	G127_RAF1_D_030_Pkinase
G072_MAP2K3_D_030_Pkinase	G081_MAPK14_pT180
G072_MAP2K3_D-site	G081_MAPK14_D_030_Pkinase
G073_MAP2K5_D_036_PB1	G077_MAP3K2_D_036_PB1
G080_MAPK1_D_030_Pkinase	G087_MYC_pT58
G080_MAPK1_D_030_Pkinase	G087_MYC_pS62
G080_MAPK1_D_030_Pkinase	G121_PTPN5_D_KIM
G080_MAPK1_pT185	G121_PTPN5_D_012_Y_phosphatase
G080_MAPK1_pY187	G121_PTPN5_D_012_Y_phosphatase
G080_MAPK1_D_030_Pkinase	G123_PTPRR_pT361
G080_MAPK1_D_030_Pkinase	G123_PTPRR_D_KIM
G080_MAPK1_pT185	G123_PTPRR_D_012_Y_phosphatase

G080_MAPK1_pY187	G123_PTPRR_D_012_Y_phosphatase
G080_MAPK1_D_030_Pkinase	G127_RAF1_pS29
G080_MAPK1_D_030_Pkinase	G127_RAF1_pS289
G080_MAPK1_D_030_Pkinase	G127_RAF1_pS296
G080_MAPK1_D_030_Pkinase	G127_RAF1_pS301
G080_MAPK1_D_030_Pkinase	G127_RAF1_pS642
G080_MAPK1_D_030_Pkinase	G138_RPS6KA1_D_D-domain
G080_MAPK1_D_030_Pkinase	G140_RPS6KA3_pT365
G080_MAPK1_D_030_Pkinase	G140_RPS6KA3_pS369
G080_MAPK1_D_030_Pkinase	G140_RPS6KA3_pT577
G080_MAPK1_D_030_Pkinase	G151_SMAD3_pT179
G080_MAPK1_D_030_Pkinase	G151_SMAD3_pS204
G080_MAPK1_D_030_Pkinase	G151_SMAD3_pS208
G080_MAPK1_D_030_Pkinase	G151_SMAD3_pS213
G080_MAPK1_D_030_Pkinase	G156_SOS1_pS1132
G080_MAPK1_D_030_Pkinase	G156_SOS1_pS1167
G080_MAPK1_D_030_Pkinase	G156_SOS1_pS1178
G080_MAPK1_D_030_Pkinase	G156_SOS1_pS1193
G080_MAPK1_D_030_Pkinase	G156_SOS1_pS1197
G080_MAPK1_D_030_Pkinase	G158_SP1_pT453
G080_MAPK1_D_030_Pkinase	G158_SP1_pT739
G080_MAPK1_D_030_Pkinase	G185_DUSP4_D_KIM
G080_MAPK1_pT185	G185_DUSP4_D_040_DSPc
G080_MAPK1_pY187	G185_DUSP4_D_040_DSPc
G080_MAPK1_D_030_Pkinase	G186_DUSP6_D_KIM
G080_MAPK1	G186_DUSP6_D_040_DSPc
G080_MAPK1_D_030_Pkinase	G186_DUSP6_pS159
G080_MAPK1_D_030_Pkinase	G186_DUSP6_pS197
G081_MAPK14_D_030_Pkinase	G185_DUSP4_D_KIM
G081_MAPK14_pT180	G185_DUSP4_D_040_DSPc
G081_MAPK14_pY182	G185_DUSP4_D_040_DSPc
G082_MAPK3_D_030_Pkinase	G121_PTPN5_D_KIM
G082_MAPK3_pT202	G121_PTPN5_D_012_Y_phosphatase
G082_MAPK3_pY204	G121_PTPN5_D_012_Y_phosphatase
G082_MAPK3_D_030_Pkinase	G123_PTPRR_pT361
G082_MAPK3_D_030_Pkinase	G123_PTPRR_D_KIM
G082_MAPK3_pT202	G123_PTPRR_D_012_Y_phosphatase
G082_MAPK3_pY204	G123_PTPRR_D_012_Y_phosphatase
G082_MAPK3_D_030_Pkinase	G140_RPS6KA3_pT365
G082_MAPK3_D_030_Pkinase	G140_RPS6KA3_pS369
G082_MAPK3_D_030_Pkinase	G140_RPS6KA3_pT577
G082_MAPK3_D_030_Pkinase	G156_SOS1_pS1137
G082_MAPK3_D_030_Pkinase	G156_SOS1_pS1167
G082_MAPK3_D_030_Pkinase	G156_SOS1_pS1178
G082_MAPK3_D_030_Pkinase	G156_SOS1_pS1193
G082_MAPK3_D_030_Pkinase	G156_SOS1_pS1197
G082_MAPK3_D_030_Pkinase	G185_DUSP4_D_KIM
G082_MAPK3_pT202	G185_DUSP4_D_040_DSPc
G082_MAPK3_pY204	G185_DUSP4_D_040_DSPc
G084_MAPK8_D_030_Pkinase	G185_DUSP4_D_KIM
G084_MAPK8_pT183	G185_DUSP4_D_040_DSPc
G084_MAPK8_pY185	G185_DUSP4_D_040_DSPc
G088_NCK1_D_007_SH3_1_D1	G131_RASA1_Pro_135_145

G088_NCK1_D_007_SH3_1_D3	G131_RASA1_Pro_135_145
G088_NCK1_D_007_SH3_1_D1	G174_WASL_304_to_316
G088_NCK1_D_007_SH3_1_D2	G174_WASL_304_to_316
G088_NCK1_D_007_SH3_1_D3	G174_WASL_304_to_316
G092_PAK1_D_183_to_188	G126_RAC1_D_042_Ras
G092_PAK1_D_030_Pkinase	G127_RAF1_pS338
G095_PIK3CA_D_114_PI3K_p85B	G099_PIK3R1_D_inter_SH2
G095_PIK3CA_D_114_PI3K_p85B	G100_PIK3R2_D_inter_SH2
G095_PIK3CA_D_114_PI3K_p85B	G101_PIK3R3_D_inter_SH2
G096_PIK3CB_D_114_PI3K_p85B	G099_PIK3R1_D_inter_SH2
G096_PIK3CB_D_114_PI3K_p85B	G100_PIK3R2_D_inter_SH2
G096_PIK3CB_D_114_PI3K_p85B	G101_PIK3R3_D_inter_SH2
G097_PIK3CD_D_114_PI3K_p85B	G099_PIK3R1_D_inter_SH2
G097_PIK3CD_D_114_PI3K_p85B	G100_PIK3R2_D_inter_SH2
G099_PIK3R1_Pro_80_to_104	G160_SRC_D_007_SH3_1
G099_PIK3R1_Pro_299_318	G160_SRC_D_007_SH3_1
G104_PLCG1_D_007_SH3_1	G156_SOS1_1121_to_1134
G104_PLCG1_pY783	G160_SRC_D_004_Pkinase_Tyr
G115_PRKCZ_pT410	G190_PDPK1_D_030_Pkinase
G117_PTK2B_pY906	G119_PTPN11_D_012_Y_phosphatase
G117_PTK2B_pY402	G160_SRC_D_006_SH2
G117_PTK2B_Pro_852_to_1052	G160_SRC_D_007_SH3_1
G119_PTPN11_D_012_Y_phosphatase	G124_PXN
G119_PTPN11_D_012_Y_phosphatase	G160_SRC_pY530
G119_PTPN11_D_012_Y_phosphatase	G164_STAT5A_pY694
G122_PTPN6_D_012_Y_phosphatase	G160_SRC_pY338
G122_PTPN6_D_012_Y_phosphatase	G160_SRC_pY419
G122_PTPN6_D_012_Y_phosphatase	G160_SRC_pY530
G125_RAB5A_D_123_Rab	G170_USP6NL_D_140_TBC
G126_RAC1_D_042_Ras	G171_VAV1_D_019_RhoGEF
G126_RAC1_D_042_Ras	G172_VAV2_D_019_RhoGEF
G126_RAC1_D_042_Ras	G173_VAV3_D_019_RhoGEF
G126_RAC1_D_042_Ras	G228_ARHGEF7_D_019_RhoGEF
G127_RAF1_pS259	G176_YWHAB_D_143_13-3-3
G127_RAF1_D_030_Pkinase	G187_BRAF_D_030_Pkinase
G127_RAF1_D_030_Pkinase	G188_KSR1_D_004_Pkinase_Tyr
G128_RALB_D_042_Ras	G129_RALBP1_D_RalBD
G128_RALB_D_042_Ras	G130_RALGDS_D_023_RasGEF_N
G128_RALB_D_042_Ras	G130_RALGDS_D_024_RasGEF
G129_RALBP1_D_RalBD	G184_RALA_D_042_Ras
G130_RALGDS_D_023_RasGEF_N	G184_RALA_D_042_Ras
G130_RALGDS_D_024_RasGEF	G184_RALA_D_042_Ras
G147_SHC1_pY350	G160_SRC_D_004_Pkinase_Tyr
G156_SOS1_1021_to_1034	G160_SRC_D_007_SH3_1
G160_SRC_D_004_Pkinase_Tyr	G163_STAT3_pY705
G160_SRC_D_004_Pkinase_Tyr	G190_PDPK1_pY373
G160_SRC_D_004_Pkinase_Tyr	G190_PDPK1_pY485
G171_VAV1_D_019_RhoGEF	G222_RHOA_D_042_Ras
G172_VAV2_D_019_RhoGEF	G222_RHOA_D_042_Ras
G173_VAV3_D_019_RhoGEF	G222_RHOA_D_042_Ras
G176_YWHAB_D_143_13-3-3	G187_BRAF_D_030_Pkinase
G190_PDPK1_D_030_Pkinase	G191_PRKCE
G222_RHOA_D_042_Ras	G228_ARHGEF7_D_019_RhoGEF

G230_MAP2K4	G075_MAP3K1_D_030_Pkinase
G230_MAP2K4	G077_MAP3K2_D_030_Pkinase
G230_MAP2K4	G079_MAP3K4_D_030_Pkinase
G231_MAP2K6	G079_MAP3K4_D_030_Pkinase
G234_P53	G084_MAPK8_D_030_Pkinase
G235_ERBB2_pY1221	G051_GRB7_D_006_SH2
G235_ERBB2_pY1139	G051_GRB7_D_006_SH2
G235_ERBB2_Pkinase_Tyr	G237_ERBB4_Pkinase_Tyr
G230_MAP2K4_D_D-site	G084_MAPK8_D_030_Pkinase
G230_MAP2K4_D_030_Pkinase	G084_MAPK8
G231_MAP2K6_D_D-site	G081_MAPK14_D_030_Pkinase
G231_MAP2K6_D_030_Pkinase	G081_MAPK14
G232_LIMK1_pT508	G092_PAK1_D_030_Pkinase
G234_P53_pS33	G080_MAPK1_D_030_Pkinase
G234_P53_pS33	G081_MAPK14_D_030_Pkinase
G235_ERBB2_pY1221	G100_PIK3R2_D_006_SH2_D1
G235_ERBB2_pY1139	G100_PIK3R2_D_006_SH2_D1
G235_ERBB2_pY1139	G050_GRB2_D_006_SH2
G235_ERBB2_pY1222	G147_SHC1_D_005_PID
G235_ERBB2_pY1221	G147_SHC1_D_005_PID
G001_ABI1_D_007_SH3_1	G156_SOS1_1133_to_1333
G017_CDC42_D_042_Ras	G075_MAP3K1_D_030_Pkinase
G017_CDC42_D_042_Ras	G079_MAP3K4_D_030_Pkinase
G017_CDC42_D_042_Ras	G156_SOS1_D_019_RhoGEF
G017_CDC42_D_042_Ras	G229_MAP3K11_D_CRIB
G017_CDC42_D_042_Ras	G233_ARHGAP3_D_029_RhoGAP
G050_GRB2_D_007_SH3_1_D1	G092_PAK1_D_Pro_rich
G050_GRB2_D_007_SH3_1_D2	G092_PAK1_D_Pro_rich
G065_KRAS_D_042_Ras	G131_RASA1_D_011_RasGAP
G065_KRAS_D_042_Ras	G179_RIN1_D_034_RA
G071_MAP2K2_pS218	G075_MAP3K1_D_030_Pkinase
G071_MAP2K2_pS222	G075_MAP3K1_D_030_Pkinase
G072_MAP2K3_pS218	G229_MAP3K11_D_004_Pkinase_Tyr
G072_MAP2K3_pS222	G229_MAP3K11_D_004_Pkinase_Tyr
G074_MAP2K7	G075_MAP3K1_D_030_Pkinase
G074_MAP2K7_D_D-site	G084_MAPK8_D_030_Pkinase
G074_MAP2K7_D_030_Pkinase	G084_MAPK8
G075_MAP3K1_D_030_Pkinase	G126_RAC1_D_042_Ras
G079_MAP3K4_D_030_Pkinase	G126_RAC1_D_042_Ras
G081_MAPK14_D_030_Pkinase	G230_MAP2K4_D_D-site
G082_MAPK3_D_030_Pkinase	G087_MYC_pT58
G082_MAPK3_D_030_Pkinase	G087_MYC_pS62
G091_NRAS_D_042_Ras	G127_RAF1_D_035_RBD
G091_NRAS_D_042_Ras	G131_RASA1_D_011_RasGAP
G091_NRAS_D_042_Ras	G156_SOS1_D_023_RasGEF_N
G091_NRAS_D_042_Ras	G156_SOS1_D_024_RasGEF
G091_NRAS_D_042_Ras	G179_RIN1_D_034_RA
G092_PAK1_D_030_Pkinase	G193_BAD_pS112
G092_PAK1_D_030_Pkinase	G193_BAD_pS136
G098_PIK3CG_D_114_PI3K_p85B	G099_PIK3R1_D_inter_SH2
G106_PLD1_pT147	G111_PRKCA_D_030_Pkinase
G106_PLD1_pS2	G111_PRKCA_D_030_Pkinase
G106_PLD1_pS561	G111_PRKCA_D_030_Pkinase

G106_PLD1	G112_PRKCB1_D_030_Pkinase
G106_PLD1	G113_PRKCG_D_030_Pkinase
G106_PLD1	G114_PRKCI_D_030_Pkinase
G106_PLD1	G115_PRKCZ_D_030_Pkinase
G106_PLD1	G191_PRKCE_D_030_Pkinase
G107_PLD2	G111_PRKCA_D_030_Pkinase
G107_PLD2	G112_PRKCB1_D_030_Pkinase
G107_PLD2	G113_PRKCG_D_030_Pkinase
G107_PLD2	G114_PRKCI_D_030_Pkinase
G107_PLD2	G115_PRKCZ_D_030_Pkinase
G107_PLD2	G191_PRKCE_D_030_Pkinase
G125_RAB5A_D_123_Rab	G179_RIN1_D_033_VPS9
G126_RAC1_D_042_Ras	G156_SOS1_D_019_RhoGEF
G126_RAC1_D_042_Ras	G233_ARHGAP3_D_029_RhoGAP
G126_RAC1_D_042_Ras	G229_MAP3K11_D_CRIB
G140_RPS6KA3_D_030_Pkinase_D1	G156_SOS1_pS1137
G140_RPS6KA3_D_030_Pkinase_D1	G156_SOS1_pS1167
G140_RPS6KA3_D_030_Pkinase_D1	G156_SOS1_pS1178
G140_RPS6KA3_D_030_Pkinase_D1	G156_SOS1_pS1193
G140_RPS6KA3_D_030_Pkinase_D1	G156_SOS1_pS1197
G140_RPS6KA3_pT365	G190_PDPK1_D_030_Pkinase
G140_RPS6KA3_pS369	G190_PDPK1_D_030_Pkinase
G140_RPS6KA3_pT577	G190_PDPK1_D_030_Pkinase
G160_SRC_D_004_Pkinase_Tyr	G161_STAT1_pY701
G229_MAP3K11_D_004_Pkinase_Tyr	G230_MAP2K4
G229_MAP3K11_D_004_Pkinase_Tyr	G231_MAP2K6_pS207
G235_ERBB2	G122_PTPN6_D_006_SH2_D1
G235_ERBB2	G122_PTPN6_D_006_SH2_D2
G235_ERBB2_pY1139	G119_PTPN11_D_006_SH2_D1
G235_ERBB2_pY1139	G119_PTPN11_D_006_SH2_D2
G235_ERBB2_pY1222	G131_RASA1_D_006_SH2_D1
G235_ERBB2_pY1221	G131_RASA1_D_006_SH2_D1
G235_ERBB2_pY1139	G131_RASA1_D_006_SH2_D1
G235_ERBB2_pY1023	G131_RASA1_D_006_SH2_D1
G235_ERBB2_pY1221	G172_VAV2_D_006_SH2
G235_ERBB2_pY1139	G172_VAV2_D_006_SH2
G235_ERBB2	G044_GAB1_D_MBD
G235_ERBB2_pY1222	G104_PLCG1_D_006_SH2_D1
G235_ERBB2_pY1222	G104_PLCG1_D_006_SH2_D2
G235_ERBB2_pY1221	G104_PLCG1_D_006_SH2_D1
G235_ERBB2_pY1221	G104_PLCG1_D_006_SH2_D2
G235_ERBB2_pY1139	G104_PLCG1_D_006_SH2_D2
G235_ERBB2_pY1023	G104_PLCG1_D_006_SH2_D1
G235_ERBB2_pY1139	G105_PLCG2_D_006_SH2_D1
G235_ERBB2_pY1139	G105_PLCG2_D_006_SH2_D2
G235_ERBB2_pY1222	G099_PIK3R1_D_006_SH2_D1
G235_ERBB2_pY1222	G099_PIK3R1_D_006_SH2_D2
G235_ERBB2_pY1221	G099_PIK3R1_D_006_SH2_D1
G235_ERBB2_pY1221	G099_PIK3R1_D_006_SH2_D2
G235_ERBB2_pY1139	G099_PIK3R1_D_006_SH2_D1
G235_ERBB2_pY1139	G099_PIK3R1_D_006_SH2_D2
G235_ERBB2_pY1023	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3	G122_PTPN6_D_006_SH2_D1

G236_ERBB3	G122_PTPN6_D_006_SH2_D2
G236_ERBB3	G119_PTPN11_D_006_SH2_D1
G236_ERBB3	G119_PTPN11_D_006_SH2_D2
G236_ERBB3_pY1289	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1276	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1262	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1222	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1197	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY868	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1276	G172_VAV2_D_006_SH2
G236_ERBB3	G050_GRB2_D_006_SH2
G236_ERBB3	G044_GAB1_D_MBD
G236_ERBB3_pY1289	G104_PLCG1_D_006_SH2_D2
G236_ERBB3_pY1289	G104_PLCG1_D_006_SH2_D1
G236_ERBB3_pY1276	G104_PLCG1_D_006_SH2_D1
G236_ERBB3_pY1276	G104_PLCG1_D_006_SH2_D2
G236_ERBB3	G105_PLCG2_D_006_SH2_D1
G236_ERBB3	G105_PLCG2_D_006_SH2_D2
G236_ERBB3_pY1289	G147_SHC1_D_006_SH2
G236_ERBB3_pY1276	G147_SHC1_D_006_SH2
G236_ERBB3_pY1289	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1289	G099_PIK3R1_D_006_SH2_D2
G236_ERBB3_pY1276	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1276	G099_PIK3R1_D_006_SH2_D2
G236_ERBB3_pY1260	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1260	G099_PIK3R1_D_006_SH2_D2
G236_ERBB3_pY1222	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1222	G099_PIK3R1_D_006_SH2_D2
G236_ERBB3_pY1197	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1197	G099_PIK3R1_D_006_SH2_D2
G236_ERBB3_pY1054	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1054	G099_PIK3R1_D_006_SH2_D2
G237_ERBB4_pY1284	G147_SHC1_D_005_PID
G237_ERBB4_pY1188	G147_SHC1_D_005_PID
G237_ERBB4	G099_PIK3R1_D_006_SH2_D1
G237_ERBB4	G099_PIK3R1_D_006_SH2_D2
G237_ERBB4	G100_PIK3R2_D_006_SH2_D1
G236_ERBB3	G100_PIK3R2_D_006_SH2_D2
G237_ERBB4	G101_PIK3R3_D_006_SH2_D1
G237_ERBB4	G101_PIK3R3_D_006_SH2_D2
G235_ERBB2_pY1222	G101_PIK3R3_D_006_SH2_D1
G235_ERBB2_pY1221	G101_PIK3R3_D_006_SH2_D2
G235_ERBB2_pY1139	G101_PIK3R3_D_006_SH2_D1
G235_ERBB2_pY1139	G101_PIK3R3_D_006_SH2_D2
G235_ERBB2_pY1023	G101_PIK3R3_D_006_SH2_D1
G235_ERBB2_pY1023	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1289	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1289	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1276	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1276	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1260	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1260	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1224	G101_PIK3R3_D_006_SH2_D1

G236_ERBB3_pY1222	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1222	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1197	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1197	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1054	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1054	G101_PIK3R3_D_006_SH2_D2
G065_KRAS_D_042_Ras	G098_PIK3CG_D_025_PI3K_rbd
G065_KRAS_D_042_Ras	G098_PIK3CG_D_028_PI3_PI4_kinase
G065_KRAS_D_042_Ras	G098_PIK3CG
G091_NRAS_D_042_Ras	G098_PIK3CG_D_025_PI3K_rbd
G091_NRAS_D_042_Ras	G098_PIK3CG_D_028_PI3_PI4_kinase
G091_NRAS_D_042_Ras	G098_PIK3CG
G098_PIK3CG_D_114_PI3K_p85B	G100_PIK3R2_D_inter_SH2
G098_PIK3CG_D_114_PI3K_p85B	G101_PIK3R3_D_inter_SH2
G127_RAF1_D_030_Pkinase	G127_RAF1_D_030_Pkinase
G187_BRAF_D_030_Pkinase	G187_BRAF_D_030_Pkinase

Table B.3 ErbB signaling network, reduced IIN

415 edges between 297 interfaces

Interface 1	Interface 2
G001_ABI1_414_PPPPPVDYEDEE	G040_EPS8_D_007_SH3_1
G002_AKT1_D_030_Pkinase	G021_CREB1_pS133
G002_AKT1_D_030_Pkinase	G127_RAF1_pS259
G002_AKT1_D_030_Pkinase	G011_CASP9_pS183
G002_AKT1_D_030_Pkinase	G043_FOXO1A_pT24
G002_AKT1_D_030_Pkinase	G045_GAB2_pS159
G002_AKT1_D_030_Pkinase	G048_GRB10_pS428
G002_AKT1_D_030_Pkinase	G126_RAC1_pS71
G002_AKT1_D_030_Pkinase	G193_BAD_pS75
G002_AKT1_D_030_Pkinase	G193_BAD_pS136
G002_AKT1_pY315	G160_SRC_D_004_Pkinase_Tyr
G002_AKT1_Pro_424_427	G160_SRC_D_007_SH3_1
G002_AKT1_pT308	G190_PDPK1_D_030_Pkinase
G004_APPL1_D_BAR	G125_RAB5A_D_123_Rab
G006_ARAF_D_035_RBD	G057_HRAS_D_042_Ras
G006_ARAF_D_030_Pkinase	G070_MAP2K1_pS218
G009_BCAR1_pY115	G160_SRC_D_004_Pkinase_Tyr
G010_CAMK2A_D_030_Pkinase	G021_CREB1_pS133
G010_CAMK2A_D_030_Pkinase	G150_SMAD2_pS110
G012_CAV1_pY14	G160_SRC_D_004_Pkinase_Tyr
G013_CAV2_pY19	G160_SRC_D_004_Pkinase_Tyr
G014_CBL_D_013_Cbl_N	P_032_EGFR_1063_to_1075
G014_CBL_D_013_Cbl_N	G159_SPRY2_pY55
G014_CBL_D_015_Cbl_N3	P_032_EGFR_1063_to_1075
G014_CBL_D_015_Cbl_N3	G159_SPRY2_pY55
G014_CBL_D_015_Cbl_N3	G159_SPRY2_49_to_61
G014_CBL_D_015_Cbl_N3	G160_SRC_413_to_425
G014_CBL_pY700	G032_EGFR_D_004_Pkinase_Tyr
G014_CBL_491_ASLPPVP	G050_GRB2_D_007_SH3_1_D2

G014_CBL_817_SQVPERPPKPFPRRINSY	G146_SH3KBP1_D_007_SH3_1_D1
G014_CBL_817_SQVPERPPKPFPRRINSY	G146_SH3KBP1_D_007_SH3_1_D2
G014_CBL_817_SQVPERPPKPFPRRINSY	G228_ARHGEF7_D_051_SH3_2
G015_CBLB_902_PARPPK	G146_SH3KBP1_D_007_SH3_1_D1
G015_CBLB_899_SQAPARPPKPRPRRTAY	G146_SH3KBP1_D_007_SH3_1_D1
G015_CBLB_899_SQAPARPPKPRPRRTAY	G146_SH3KBP1_D_007_SH3_1_D2
G017_CDC42_D_042_Ras	G172_VAV2_D_019_RhoGEF
G017_CDC42_D_042_Ras	G075_MAP3K1_D_030_Pkinase
G017_CDC42_D_042_Ras	G079_MAP3K4_D_030_Pkinase
G017_CDC42_D_042_Ras	G092_PAK1_D_043_PBD
G017_CDC42_D_042_Ras	G169_TNK2_D_048_GTPase_binding
G017_CDC42_D_042_Ras	G228_ARHGEF7_D_019_RhoGEF
G017_CDC42_D_042_Ras	G156_SOS1_D_019_RhoGEF
G017_CDC42_D_042_Ras	G229_MAP3K11_D_CRIB
G017_CDC42_D_042_Ras	G233_ARHGAP3_D_029_RhoGAP
G021_CREB1_pS133	G140_RPS6KA3_D_030_Pkinase_D1
G021_CREB1_pS133	G140_RPS6KA3_D_030_Pkinase_D2
G022_CRK_pY221	G032_EGFR_D_004_Pkinase_Tyr
G022_CRK_D_006_SH2	G032_EGFR_pY1016
G022_CRK_D_006_SH2	G032_EGFR_pY1125
G024_CSK_D_004_Pkinase_Tyr	G160_SRC_pY530
G027_DNM1_806_to_851	G050_GRB2_D_007_SH3_1_D1
G027_DNM1_806_to_851	G160_SRC_D_007_SH3_1
G027_DNM1_806_to_851	G144_SH3GL2_D_007_SH3_1
G027_DNM1_pY231	G160_SRC_D_004_Pkinase_Tyr
G029_DUSP1_pS296	G080_MAPK1_D_030_Pkinase
G029_DUSP1_D_KIM	G080_MAPK1_D_030_Pkinase
G029_DUSP1_D_KIM	G081_MAPK14_D_030_Pkinase
G029_DUSP1_D_KIM	G082_MAPK3_D_030_Pkinase
G029_DUSP1_D_KIM	G084_MAPK8_D_030_Pkinase
G029_DUSP1_D_040_DSPc	G080_MAPK1
G029_DUSP1_D_040_DSPc	G081_MAPK14
G029_DUSP1_D_040_DSPc	G082_MAPK3
G029_DUSP1_D_040_DSPc	G084_MAPK8
G029_DUSP1_pS364	G082_MAPK3_D_030_Pkinase
G031_EGF_D_001_EGF	G032_EGFR_D_002_Recep_L_domain_D1
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY447
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY657
G032_EGFR_D_004_Pkinase_Tyr	G104_PLCG1_pY783
G032_EGFR_D_004_Pkinase_Tyr	G038_EPS15_pY849
G032_EGFR_D_004_Pkinase_Tyr	G041_ERRFI1_315_374
G032_EGFR_D_004_Pkinase_Tyr	G044_GAB1_pY285
G032_EGFR_D_004_Pkinase_Tyr	G104_PLCG1_pY472
G032_EGFR_D_004_Pkinase_Tyr	G131_RASA1_pY460
G032_EGFR_D_004_Pkinase_Tyr	G136_RGS16_pY168
G032_EGFR_D_004_Pkinase_Tyr	G161_STAT1_pY701
G032_EGFR_D_004_Pkinase_Tyr	G165_STAT5B_pY699
G032_EGFR_pY1092	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY1092	G044_GAB1_D_MBD
G032_EGFR_pY1110	G044_GAB1_D_MBD
G032_EGFR_pY1110	G147_SHC1_D_005_PID
G032_EGFR_pY1138	G050_GRB2_D_006_SH2
G032_EGFR_pY1138	G147_SHC1_D_005_PID

G032_EGFR_pS1064	G050_GRB2_D_006_SH2
G032_EGFR_pY1016	G119_PTPN11_D_012_Y_phosphatase
G032_EGFR_pY1016	G088_NCK1_D_006_SH2
G032_EGFR_pY1016	G099_PIK3R1_D_006_SH2_D1
G032_EGFR_pY1016	G099_PIK3R1_D_006_SH2_D2
G032_EGFR_pY1016	G100_PIK3R2_D_006_SH2_D1
G032_EGFR_pY1016	G104_PLCG1_D_006_SH2_D1
G032_EGFR_pY1016	G104_PLCG1_D_006_SH2_D2
G032_EGFR_pY1016	G119_PTPN11_D_006_SH2_D1
G032_EGFR_pY1016	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY1016	G147_SHC1_D_005_PID
G032_EGFR_pY998	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY998	G105_PLCG2_D_006_SH2_D1
G032_EGFR_pY998	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY998	G147_SHC1_D_005_PID
G032_EGFR_pY1197	G122_PTPN6_D_012_Y_phosphatase
G032_EGFR_pY1197	G122_PTPN6_D_006_SH2_D1
G032_EGFR_pY915	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY915	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY1125	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR_pY1125	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY1172	G131_RASA1_D_006_SH2_D1
G032_EGFR_pY1172	G147_SHC1_D_005_PID
G032_EGFR_pY1192	G147_SHC1_D_005_PID
G032_EGFR_pY869	G160_SRC_D_004_Pkinase_Tyr
G032_EGFR	G172_VAV2_D_006_SH2
G034_ELK1_D_D-domain	G080_MAPK1_D_030_Pkinase
G034_ELK1_D_D-domain	G082_MAPK3_D_030_Pkinase
G034_ELK1_D_DEF-domain	G080_MAPK1_D_030_Pkinase
G034_ELK1_D_DEF-domain	G082_MAPK3_D_030_Pkinase
G036_EPN1_496_to_575	G038_EPS15_1_313
G038_EPS15_777_to_789	G050_GRB2_D_007_SH3_1_D1
G038_EPS15_777_to_789	G050_GRB2_D_007_SH3_1_D2
G040_EPS8_D_535_to_821	G156_SOS1
G040_EPS8_D_007_SH3_1	G170_USP6NL
G042_FOS_D_071_bZIP_1	G062_JUN_G062_D_071_bZIP_1
G042_FOS_D_DEF-domain	G080_MAPK1_D_030_Pkinase
G042_FOS_D_DEF-domain	G082_MAPK3_D_030_Pkinase
G042_FOS	G084_MAPK8_D_030_Pkinase
G044_GAB1_337_to_346	G050_GRB2_D_007_SH3_1_D2
G044_GAB1_pY447	G099_PIK3R1_D_006_SH2_D1
G044_GAB1_pY447	G099_PIK3R1_D_006_SH2_D2
G044_GAB1_pY589	G099_PIK3R1_D_006_SH2_D1
G044_GAB1_pY589	G099_PIK3R1_D_006_SH2_D2
G044_GAB1_pY657	G119_PTPN11_D_012_Y_phosphatase
G045_GAB2_348_to_367	G050_GRB2_D_007_SH3_1_D1
G045_GAB2_502_to_528	G050_GRB2_D_007_SH3_1_D1
G045_GAB2_502_to_528	G050_GRB2_D_007_SH3_1_D2
G045_GAB2_pS623	G080_MAPK1_D_030_Pkinase
G045_GAB2_pS623	G082_MAPK3_D_030_Pkinase
G045_GAB2_pY643	G119_PTPN11_D_012_Y_phosphatase
G047_GJA1_pS255	G082_MAPK3_D_030_Pkinase
G047_GJA1_pS255	G083_MAPK7_D_030_Pkinase

G047_GJA1_pS279	G082_MAPK3_D_030_Pkinase
G047_GJA1_pS368	G113_PRKCG_D_030_Pkinase
G047_GJA1_pY247	G160_SRC_D_004_Pkinase_Tyr
G050_GRB2_D_007_SH3_1_D1	G061_JAK2_494_to_506
G050_GRB2_D_007_SH3_1_D1	G156_SOS1_1149_to_1158
G050_GRB2_D_007_SH3_1_D1	G156_SOS1_1133_to_1333
G050_GRB2_D_007_SH3_1_D1	G157_SOS2_1019_to_1032
G050_GRB2_D_007_SH3_1_D1	G092_PAK1_D_Pro_rich
G050_GRB2_D_007_SH3_1_D2	G061_JAK2_494_to_506
G050_GRB2_D_007_SH3_1_D2	G157_SOS2_1019_to_1032
G050_GRB2_D_007_SH3_1_D2	G171_VAV1_D_051_SH3_2
G050_GRB2_D_007_SH3_1_D2	G092_PAK1_D_Pro_rich
G050_GRB2_D_006_SH2	G104_PLCG1_pY783
G050_GRB2_D_006_SH2	G235_ERBB2_pY1139
G050_GRB2_D_006_SH2	G236_ERBB3
G050_GRB2_D_006_SH2	G117_PTK2B_pY881
G050_GRB2_D_006_SH2	G119_PTPN11_pY584
G050_GRB2_D_006_SH2	G147_SHC1_pY427
G050_GRB2_pY160	G160_SRC_D_004_Pkinase_Tyr
G057_HRAS_D_042_Ras	G098_PIK3CG_D_025_PI3K_rbd
G057_HRAS_D_042_Ras	G098_PIK3CG_D_028_PI3_PI4_kinase
G057_HRAS_D_042_Ras	G098_PIK3CG
G057_HRAS_D_042_Ras	G127_RAF1_D_035_RBD
G057_HRAS_D_042_Ras	G130_RALGDS_D_034_RA
G057_HRAS_D_042_Ras	G131_RASA1_D_011_RasGAP
G057_HRAS_D_042_Ras	G156_SOS1_D_023_RasGEF_N
G057_HRAS_D_042_Ras	G179_RIN1_D_034_RA
G057_HRAS_D_042_Ras	G187_BRAF_D_035_RBD
G061_JAK2_pY1007	G119_PTPN11_D_012_Y_phosphatase
G061_JAK2_D_004_Pkinase_Tyr_D1	G161_STAT1_pY701
G061_JAK2_D_004_Pkinase_Tyr_D1	G164_STAT5A_pY694
G062_JUN_pS73	G084_MAPK8_D_030_Pkinase
G062_JUN_G062_D_D-domain	G084_MAPK8_D_030_Pkinase
G065_KRAS_D_042_Ras	G098_PIK3CG_D_025_PI3K_rbd
G065_KRAS_D_042_Ras	G098_PIK3CG_D_028_PI3_PI4_kinase
G065_KRAS_D_042_Ras	G098_PIK3CG
G065_KRAS_D_042_Ras	G127_RAF1_D_035_RBD
G065_KRAS_D_042_Ras	G130_RALGDS_D_034_RA
G065_KRAS_D_042_Ras	G131_RASA1_D_011_RasGAP
G065_KRAS_D_042_Ras	G156_SOS1_D_023_RasGEF_N
G065_KRAS_D_042_Ras	G179_RIN1_D_034_RA
G065_KRAS_D_042_Ras	G187_BRAF_D_035_RBD
G069_KRT8_pS74	G081_MAPK14_D_030_Pkinase
G070_MAP2K1_pS218	G127_RAF1_D_030_Pkinase
G070_MAP2K1_pS218	G075_MAP3K1_D_030_Pkinase
G070_MAP2K1_pS218	G187_BRAF_D_030_Pkinase
G070_MAP2K1_D_030_Pkinase	G080_MAPK1_pY187
G070_MAP2K1_D_030_Pkinase	G082_MAPK3_pT202
G070_MAP2K1_D_030_Pkinase	G188_KSR1_D_004_Pkinase_Tyr
G070_MAP2K1_pT292	G080_MAPK1_D_030_Pkinase
G070_MAP2K1_pT292	G127_RAF1_D_030_Pkinase
G070_MAP2K1_D_D-site	G080_MAPK1_D_030_Pkinase
G070_MAP2K1_pT286	G127_RAF1_D_030_Pkinase

G071_MAP2K2_D_030_Pkinase	G080_MAPK1_pT185
G071_MAP2K2_D_030_Pkinase	G080_MAPK1_pY187
G071_MAP2K2_D_030_Pkinase	G082_MAPK3_pT202
G071_MAP2K2_D_D-site	G080_MAPK1_D_030_Pkinase
G071_MAP2K2_pS218	G127_RAF1_D_030_Pkinase
G071_MAP2K2_pS218	G075_MAP3K1_D_030_Pkinase
G072_MAP2K3_D_030_Pkinase	G081_MAPK14_pT180
G072_MAP2K3_D-site	G081_MAPK14_D_030_Pkinase
G073_MAP2K5_D_036_PB1	G077_MAP3K2_D_036_PB1
G080_MAPK1_D_030_Pkinase	G234_P53_pS33
G080_MAPK1_D_030_Pkinase	G140_RPS6KA3_pT365
G080_MAPK1_D_030_Pkinase	G087_MYC_pT58
G080_MAPK1_D_030_Pkinase	G121_PTPN5_D_KIM
G080_MAPK1_D_030_Pkinase	G123_PTPRR_pT361
G080_MAPK1_D_030_Pkinase	G123_PTPRR_D_KIM
G080_MAPK1_D_030_Pkinase	G127_RAF1_pS29
G080_MAPK1_D_030_Pkinase	G138_RPS6KA1_D_D-domain
G080_MAPK1_D_030_Pkinase	G151_SMAD3_pT179
G080_MAPK1_D_030_Pkinase	G156_SOS1_pS1132
G080_MAPK1_D_030_Pkinase	G156_SOS1_pS1167
G080_MAPK1_D_030_Pkinase	G158_SP1_pT453
G080_MAPK1_D_030_Pkinase	G185_DUSP4_D_KIM
G080_MAPK1_D_030_Pkinase	G186_DUSP6_D_KIM
G080_MAPK1_pT185	G121_PTPN5_D_012_Y_phosphatase
G080_MAPK1_pT185	G123_PTPRR_D_012_Y_phosphatase
G080_MAPK1_pT185	G185_DUSP4_D_040_DSPc
G080_MAPK1_pY187	G121_PTPN5_D_012_Y_phosphatase
G080_MAPK1_pY187	G123_PTPRR_D_012_Y_phosphatase
G080_MAPK1_pY187	G185_DUSP4_D_040_DSPc
G080_MAPK1	G186_DUSP6_D_040_DSPc
G081_MAPK14_D_030_Pkinase	G230_MAP2K4_D_D-site
G081_MAPK14_D_030_Pkinase	G231_MAP2K6_D_D-site
G081_MAPK14_D_030_Pkinase	G234_P53_pS33
G081_MAPK14_D_030_Pkinase	G185_DUSP4_D_KIM
G081_MAPK14_pT180	G185_DUSP4_D_040_DSPc
G081_MAPK14_pY182	G185_DUSP4_D_040_DSPc
G082_MAPK3_D_030_Pkinase	G140_RPS6KA3_pT365
G082_MAPK3_D_030_Pkinase	G087_MYC_pT58
G082_MAPK3_D_030_Pkinase	G121_PTPN5_D_KIM
G082_MAPK3_D_030_Pkinase	G123_PTPRR_pT361
G082_MAPK3_D_030_Pkinase	G123_PTPRR_D_KIM
G082_MAPK3_D_030_Pkinase	G156_SOS1_pS1167
G082_MAPK3_D_030_Pkinase	G185_DUSP4_D_KIM
G082_MAPK3_D_030_Pkinase	G156_SOS1_pS1137
G082_MAPK3_pT202	G121_PTPN5_D_012_Y_phosphatase
G082_MAPK3_pT202	G123_PTPRR_D_012_Y_phosphatase
G082_MAPK3_pT202	G185_DUSP4_D_040_DSPc
G084_MAPK8_D_030_Pkinase	G234_P53
G084_MAPK8_D_030_Pkinase	G230_MAP2K4_D_D-site
G084_MAPK8_D_030_Pkinase	G074_MAP2K7_D_D-site
G084_MAPK8_D_030_Pkinase	G185_DUSP4_D_KIM
G084_MAPK8_pT183	G185_DUSP4_D_040_DSPc
G088_NCK1_D_007_SH3_1_D1	G131_RASA1_Pro_135_145

G088_NCK1_D_007_SH3_1_D1	G174_WASL_304_to_316
G088_NCK1_D_007_SH3_1_D2	G174_WASL_304_to_316
G092_PAK1_D_183_to_188	G126_RAC1_D_042_Ras
G092_PAK1_D_030_Pkinase	G232_LIMK1_pT508
G092_PAK1_D_030_Pkinase	G193_BAD_pS136
G092_PAK1_D_030_Pkinase	G127_RAF1_pS338
G092_PAK1_D_030_Pkinase	G193_BAD_pS112
G095_PIK3CA_D_114_PI3K_p85B	G099_PIK3R1_D_inter_SH2
G095_PIK3CA_D_114_PI3K_p85B	G100_PIK3R2_D_inter_SH2
G095_PIK3CA_D_114_PI3K_p85B	G101_PIK3R3_D_inter_SH2
G096_PIK3CB_D_114_PI3K_p85B	G099_PIK3R1_D_inter_SH2
G096_PIK3CB_D_114_PI3K_p85B	G100_PIK3R2_D_inter_SH2
G096_PIK3CB_D_114_PI3K_p85B	G101_PIK3R3_D_inter_SH2
G097_PIK3CD_D_114_PI3K_p85B	G099_PIK3R1_D_inter_SH2
G097_PIK3CD_D_114_PI3K_p85B	G100_PIK3R2_D_inter_SH2
G099_PIK3R1_Pro_80_to_104	G160_SRC_D_007_SH3_1
G104_PLCG1_D_007_SH3_1	G156_SOS1_1121_to_1134
G104_PLCG1_pY783	G160_SRC_D_004_Pkinase_Tyr
G115_PRKCZ_pT410	G190_PDPK1_D_030_Pkinase
G117_PTK2B_pY906	G119_PTPN11_D_012_Y_phosphatase
G117_PTK2B_pY402	G160_SRC_D_006_SH2
G117_PTK2B_Pro_852_to_1052	G160_SRC_D_007_SH3_1
G119_PTPN11_D_012_Y_phosphatase	G160_SRC_pY530
G119_PTPN11_D_012_Y_phosphatase	G164_STAT5A_pY694
G119_PTPN11_D_012_Y_phosphatase	G124_PXN
G122_PTPN6_D_012_Y_phosphatase	G160_SRC_pY530
G122_PTPN6_D_012_Y_phosphatase	G160_SRC_pY338
G125_RAB5A_D_123_Rab	G170_USP6NL_D_140_TBC
G125_RAB5A_D_123_Rab	G179_RIN1_D_033_VPS9
G126_RAC1_D_042_Ras	G171_VAV1_D_019_RhoGEF
G126_RAC1_D_042_Ras	G172_VAV2_D_019_RhoGEF
G126_RAC1_D_042_Ras	G173_VAV3_D_019_RhoGEF
G126_RAC1_D_042_Ras	G075_MAP3K1_D_030_Pkinase
G126_RAC1_D_042_Ras	G079_MAP3K4_D_030_Pkinase
G126_RAC1_D_042_Ras	G228_ARHGEF7_D_019_RhoGEF
G126_RAC1_D_042_Ras	G156_SOS1_D_019_RhoGEF
G126_RAC1_D_042_Ras	G229_MAP3K11_D_CRIB
G126_RAC1_D_042_Ras	G233_ARHGAP3_D_029_RhoGAP
G127_RAF1_pS259	G176_YWHAB_D_143_13-3-3
G127_RAF1_D_030_Pkinase	G127_RAF1_D_030_Pkinase
G127_RAF1_D_030_Pkinase	G187_BRAF_D_030_Pkinase
G127_RAF1_D_030_Pkinase	G188_KSR1_D_004_Pkinase_Tyr
G128_RALB_D_042_Ras	G129_RALBP1_D_RalBD
G128_RALB_D_042_Ras	G130_RALGDS_D_023_RasGEF_N
G129_RALBP1_D_RalBD	G184_RALA_D_042_Ras
G130_RALGDS_D_023_RasGEF_N	G184_RALA_D_042_Ras
G147_SHC1_pY350	G160_SRC_D_004_Pkinase_Tyr
G156_SOS1_1021_to_1034	G160_SRC_D_007_SH3_1
G160_SRC_D_004_Pkinase_Tyr	G161_STAT1_pY701
G160_SRC_D_004_Pkinase_Tyr	G163_STAT3_pY705
G160_SRC_D_004_Pkinase_Tyr	G190_PDPK1_pY373
G171_VAV1_D_019_RhoGEF	G222_RHOA_D_042_Ras
G172_VAV2_D_019_RhoGEF	G222_RHOA_D_042_Ras

G173_VAV3_D_019_RhoGEF	G222_RHOA_D_042_Ras
G176_YWHAB_D_143_13-3-3	G187_BRAF_D_030_Pkinase
G190_PDPK1_D_030_Pkinase	G140_RPS6KA3_pT365
G190_PDPK1_D_030_Pkinase	G191_PRKCE
G222_RHOA_D_042_Ras	G228_ARHGEF7_D_019_RhoGEF
G230_MAP2K4	G075_MAP3K1_D_030_Pkinase
G230_MAP2K4	G079_MAP3K4_D_030_Pkinase
G230_MAP2K4	G229_MAP3K11_D_004_Pkinase_Tyr
G230_MAP2K4	G077_MAP3K2_D_030_Pkinase
G231_MAP2K6	G079_MAP3K4_D_030_Pkinase
G235_ERBB2_pY1221	G099_PIK3R1_D_006_SH2_D1
G235_ERBB2_pY1221	G099_PIK3R1_D_006_SH2_D2
G235_ERBB2_pY1221	G100_PIK3R2_D_006_SH2_D1
G235_ERBB2_pY1221	G104_PLCG1_D_006_SH2_D1
G235_ERBB2_pY1221	G104_PLCG1_D_006_SH2_D2
G235_ERBB2_pY1221	G131_RASA1_D_006_SH2_D1
G235_ERBB2_pY1221	G147_SHC1_D_005_PID
G235_ERBB2_pY1221	G172_VAV2_D_006_SH2
G235_ERBB2_pY1221	G051_GRB7_D_006_SH2
G235_ERBB2_pY1221	G101_PIK3R3_D_006_SH2_D2
G235_ERBB2_Pkinase_Tyr	G237_ERBB4_Pkinase_Tyr
G230_MAP2K4_D_030_Pkinase	G084_MAPK8
G231_MAP2K6_D_030_Pkinase	G081_MAPK14
G235_ERBB2_pY1139	G099_PIK3R1_D_006_SH2_D1
G235_ERBB2_pY1139	G099_PIK3R1_D_006_SH2_D2
G235_ERBB2_pY1139	G100_PIK3R2_D_006_SH2_D1
G235_ERBB2_pY1139	G104_PLCG1_D_006_SH2_D2
G235_ERBB2_pY1139	G105_PLCG2_D_006_SH2_D1
G235_ERBB2_pY1139	G119_PTPN11_D_006_SH2_D1
G235_ERBB2_pY1139	G131_RASA1_D_006_SH2_D1
G235_ERBB2_pY1139	G172_VAV2_D_006_SH2
G235_ERBB2_pY1139	G051_GRB7_D_006_SH2
G235_ERBB2_pY1139	G119_PTPN11_D_006_SH2_D2
G235_ERBB2_pY1139	G105_PLCG2_D_006_SH2_D2
G235_ERBB2_pY1139	G101_PIK3R3_D_006_SH2_D1
G235_ERBB2_pY1139	G101_PIK3R3_D_006_SH2_D2
G235_ERBB2_pY1222	G099_PIK3R1_D_006_SH2_D1
G235_ERBB2_pY1222	G099_PIK3R1_D_006_SH2_D2
G235_ERBB2_pY1222	G104_PLCG1_D_006_SH2_D1
G235_ERBB2_pY1222	G104_PLCG1_D_006_SH2_D2
G235_ERBB2_pY1222	G131_RASA1_D_006_SH2_D1
G235_ERBB2_pY1222	G147_SHC1_D_005_PID
G235_ERBB2_pY1222	G101_PIK3R3_D_006_SH2_D1
G001_ABI1_D_007_SH3_1	G156_SOS1_1133_to_1333
G072_MAP2K3_pS218	G229_MAP3K11_D_004_Pkinase_Tyr
G074_MAP2K7	G075_MAP3K1_D_030_Pkinase
G074_MAP2K7_D_030_Pkinase	G084_MAPK8
G091_NRAS_D_042_Ras	G098_PIK3CG_D_025_PI3K_rbd
G091_NRAS_D_042_Ras	G098_PIK3CG_D_028_PI3_PI4_kinase
G091_NRAS_D_042_Ras	G098_PIK3CG
G091_NRAS_D_042_Ras	G127_RAF1_D_035_RBD
G091_NRAS_D_042_Ras	G131_RASA1_D_011_RasGAP
G091_NRAS_D_042_Ras	G156_SOS1_D_023_RasGEF_N

G091_NRAS_D_042_Ras	G179_RIN1_D_034_RA
G098_PIK3CG_D_114_PI3K_p85B	G099_PIK3R1_D_inter_SH2
G098_PIK3CG_D_114_PI3K_p85B	G100_PIK3R2_D_inter_SH2
G098_PIK3CG_D_114_PI3K_p85B	G101_PIK3R3_D_inter_SH2
G106_PLD1_pT147	G111_PRKCA_D_030_Pkinase
G106_PLD1	G113_PRKCG_D_030_Pkinase
G106_PLD1	G112_PRKCB1_D_030_Pkinase
G106_PLD1	G114_PRKCI_D_030_Pkinase
G106_PLD1	G115_PRKCZ_D_030_Pkinase
G106_PLD1	G191_PRKCE_D_030_Pkinase
G107_PLD2	G113_PRKCG_D_030_Pkinase
G107_PLD2	G111_PRKCA_D_030_Pkinase
G107_PLD2	G112_PRKCB1_D_030_Pkinase
G107_PLD2	G114_PRKCI_D_030_Pkinase
G107_PLD2	G115_PRKCZ_D_030_Pkinase
G107_PLD2	G191_PRKCE_D_030_Pkinase
G140_RPS6KA3_D_030_Pkinase_D1	G156_SOS1_pS1167
G140_RPS6KA3_D_030_Pkinase_D1	G156_SOS1_pS1137
G229_MAP3K11_D_004_Pkinase_Tyr	G231_MAP2K6_pS207
G235_ERBB2	G044_GAB1_D_MBD
G235_ERBB2	G122_PTPN6_D_006_SH2_D1
G235_ERBB2_pY1023	G099_PIK3R1_D_006_SH2_D1
G235_ERBB2_pY1023	G104_PLCG1_D_006_SH2_D1
G235_ERBB2_pY1023	G131_RASA1_D_006_SH2_D1
G235_ERBB2_pY1023	G101_PIK3R3_D_006_SH2_D1
G235_ERBB2_pY1023	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3	G044_GAB1_D_MBD
G236_ERBB3	G105_PLCG2_D_006_SH2_D1
G236_ERBB3	G119_PTPN11_D_006_SH2_D1
G236_ERBB3	G122_PTPN6_D_006_SH2_D1
G236_ERBB3	G119_PTPN11_D_006_SH2_D2
G236_ERBB3	G105_PLCG2_D_006_SH2_D2
G236_ERBB3	G100_PIK3R2_D_006_SH2_D2
G236_ERBB3_pY1289	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1289	G099_PIK3R1_D_006_SH2_D2
G236_ERBB3_pY1289	G104_PLCG1_D_006_SH2_D1
G236_ERBB3_pY1289	G104_PLCG1_D_006_SH2_D2
G236_ERBB3_pY1289	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1289	G147_SHC1_D_006_SH2
G236_ERBB3_pY1289	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1289	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1276	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1276	G099_PIK3R1_D_006_SH2_D2
G236_ERBB3_pY1276	G104_PLCG1_D_006_SH2_D1
G236_ERBB3_pY1276	G104_PLCG1_D_006_SH2_D2
G236_ERBB3_pY1276	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1276	G172_VAV2_D_006_SH2
G236_ERBB3_pY1276	G147_SHC1_D_006_SH2
G236_ERBB3_pY1276	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1276	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1262	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1222	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1222	G099_PIK3R1_D_006_SH2_D2

G236_ERBB3_pY1222	G131_RASA1_D_006_SH2_D1
G236_ERBB3_pY1222	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1222	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1260	G099_PIK3R1_D_006_SH2_D1
G236_ERBB3_pY1260	G099_PIK3R1_D_006_SH2_D2
G236_ERBB3_pY1260	G101_PIK3R3_D_006_SH2_D1
G236_ERBB3_pY1260	G101_PIK3R3_D_006_SH2_D2
G237_ERBB4_pY1284	G147_SHC1_D_005_PID
G237_ERBB4	G099_PIK3R1_D_006_SH2_D1
G237_ERBB4	G099_PIK3R1_D_006_SH2_D2
G237_ERBB4	G100_PIK3R2_D_006_SH2_D1
G237_ERBB4	G101_PIK3R3_D_006_SH2_D1
G237_ERBB4	G101_PIK3R3_D_006_SH2_D2
G236_ERBB3_pY1224	G101_PIK3R3_D_006_SH2_D1
G187_BRAF_D_030_Pkinase	G187_BRAF_D_030_Pkinase

Appendix C. Additional Figures and Data

C.1 Additional figures for Chapter 2

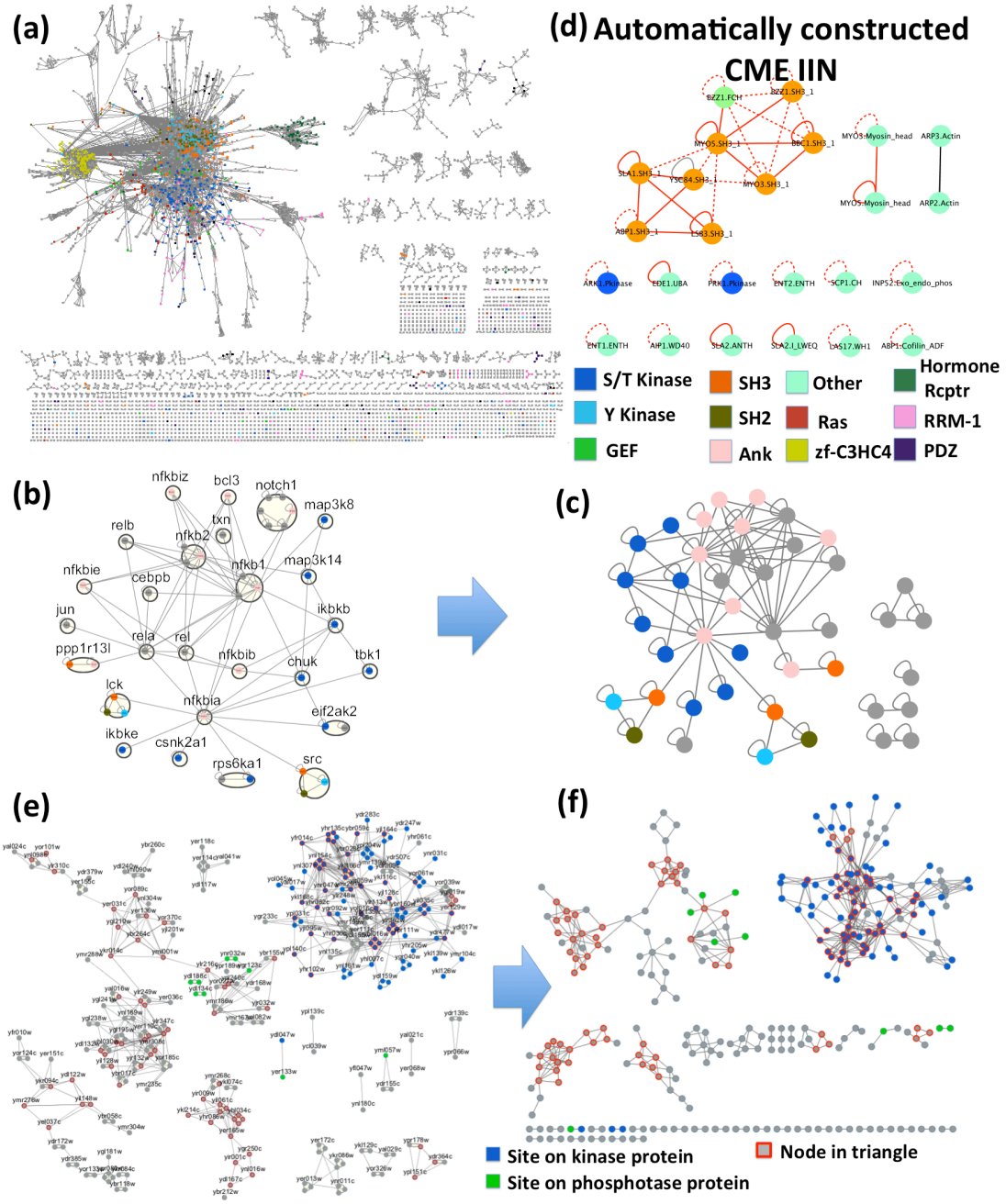
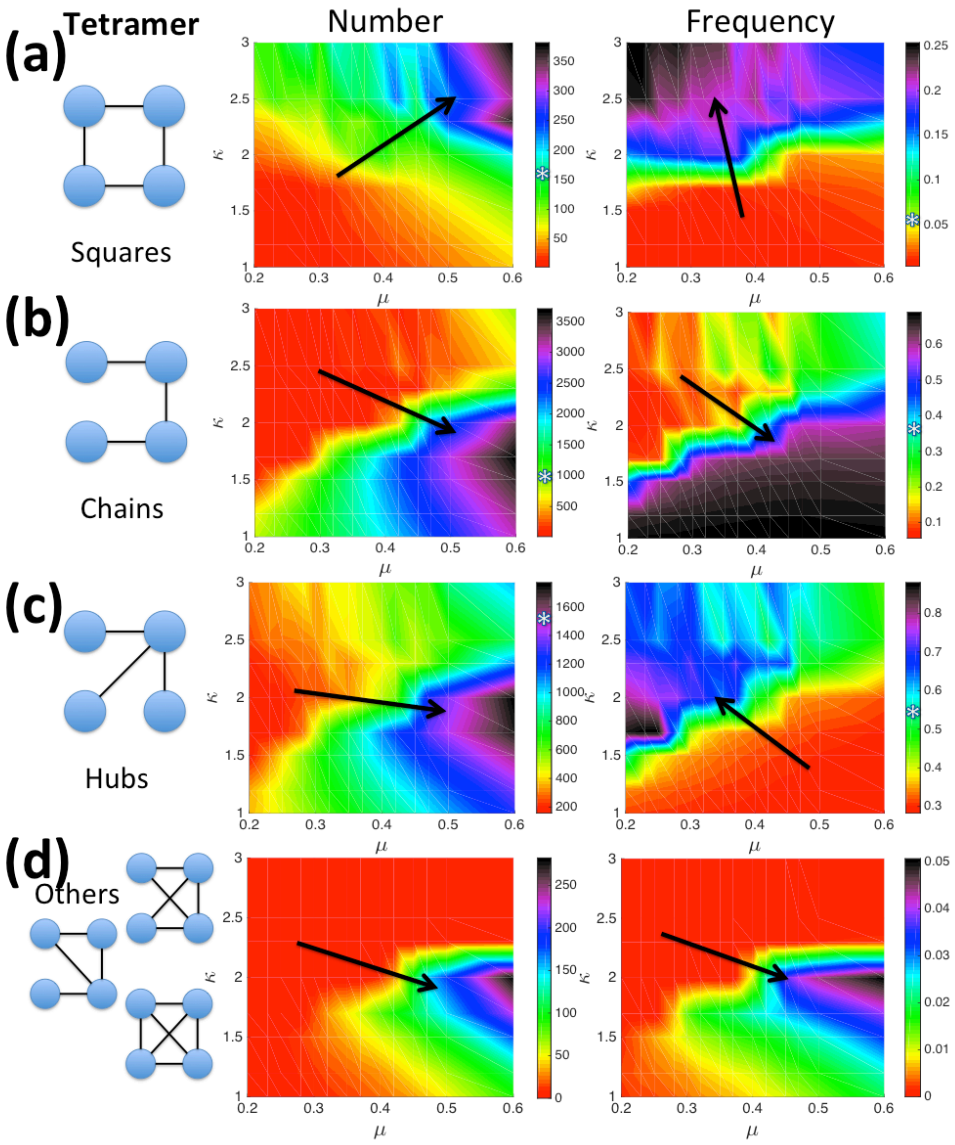


Figure C.1. Automatically constructed IINs differ from manually curated networks. (a) The IIN from the human structural interaction network [67]. Although it is fragmented, it is only so at a level equivalent to the protein network (Table 1). Hence the IIN does not create many distinct interface modules like the manually curated networks. (**b)** Site graph and isolated IIN **(c)** for a small portion of the human structural interaction network of (A). **(d)** Yeast CME network reconstructed with the *S. cerevisiae* INstruct database [96]. The IIN is clearly much smaller than the manually curated Yeast CME network of Fig. 1C, and many of the assigned interactions from the Instruct database do not occur in the manually curated network. Black edges indicate correct domain interactions. Red edges are for incorrect domain interactions, but correct PPIs. Dashed red edges indicate interactions we removed because they weren't in the reference literature or were found to not be direct. Gray edges are interactions for domains we didn't define. **(e, f)** Yeast structural interaction network [62] with only the cytoplasmic proteins [70]. The IIN **(f)** is again only fragmented at the same level as the PPIN. Interface types are not annotated in the published data, so interfaces on kinase and phosphatase proteins are highlighted. There are no self-loops and many triangles.



(e)

β	μ	κ	ω	Effects
ON	ON	ON	OFF	More edges, more interfaces (<k>≈2), more squares (edges close chains)
ON	ON	OFF	ON	Lower PAE (0.3-0.4), less squares and more chains, giant component with 80-90% of network
ON	OFF	ON	ON	Lower PAE (0.4-0.5), more interfaces, less squares. Fragmented network of pairs and small star hubs
OFF	ON	ON	ON	More clustering/triangles, especially when starting from dense extreme. PAE unaffected.
OFF	ON	OFF	ON	Moves towards dense extreme (one interface per protein)
ON	OFF	OFF	ON	No clustering, binomial network (PAE=0-0.2), <k>≈2 (network size constrained by strength of ω), few squares

Figure C.2. Fitness function parameters determine number and frequency of four-node motifs in sampled IINs. The number and frequency of **(a)** square motifs **(b)** chain motifs **(c)** hub motifs and **(d)** the three remaining tetramer types, all from IINs sampled with a fitness function where the parameters κ and μ are varied. Arrows indicate direction of increase of motif frequencies. Results from Monte Carlo sampling performed with $k_BT=1$ on the CME PPIN (Fig. 1a). The other two parameters not shown on axes were set to $\beta=4$ and $\omega=0.1$ for these simulations. The last three tetramer types in **(d)** ('Others') include clustering – penalized by β – which only occur as μ is increased since this drives the IIN closer to the PPIN in structure. The white stars indicate the statistics of the real CME IIN, which contains no clustering. **(e)** Each of the four parameters in the fitness function (κ, μ, β, ω) control structural aspects of the sampled IIN structures (Methods). By turning off each parameter, we illustrate how the networks respond with fewer biasing forces on their structural elements. Without any control of interfaces (μ is off), interfaces are more abundant and the network is relatively disconnected, whereas without any square bias (κ is off), squares are uncommon, chains are not penalized, and therefore the network does not fragment, resulting in a giant component.

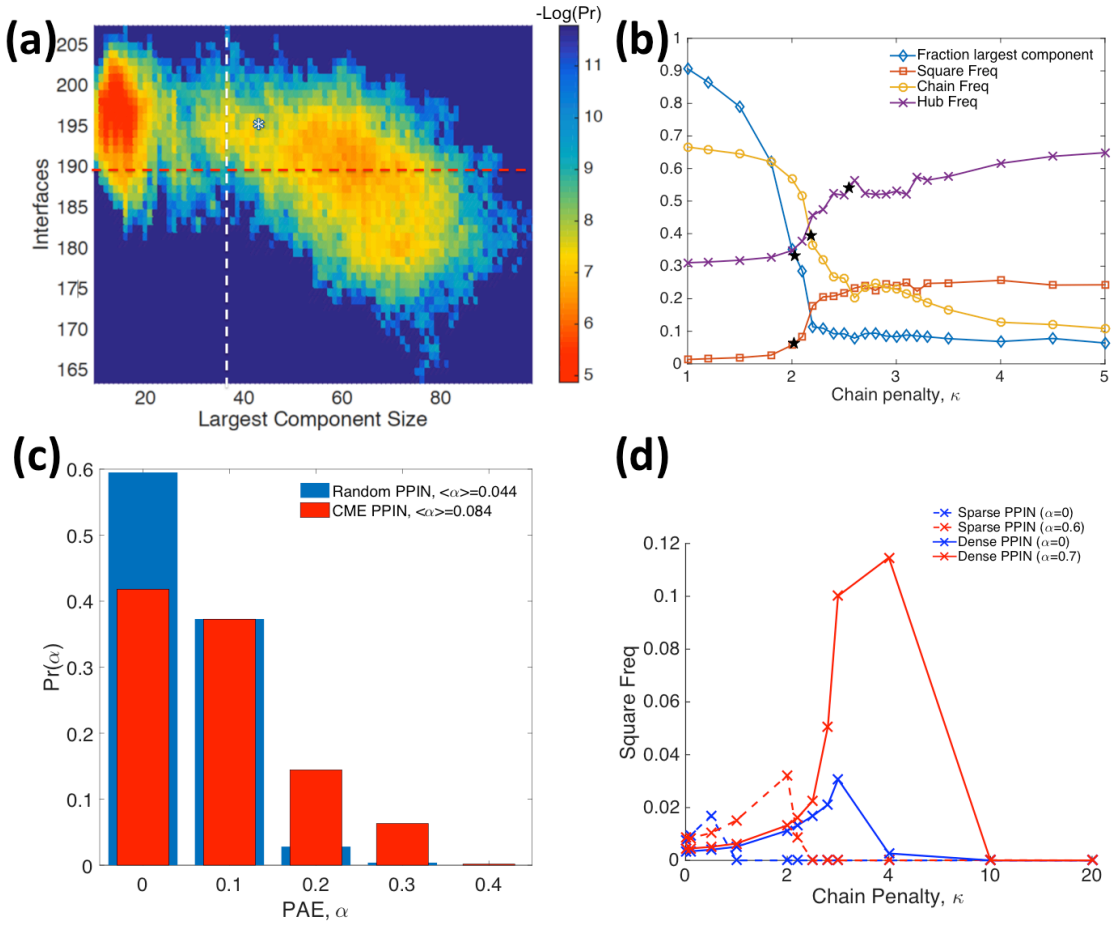


Figure C.3. IIN properties vary as the structure of the PPIN varies. **(a)** Probability distribution of IINs sampled with a fixed parameter set (the optimal parameters) as a function of the reaction coordinates of largest component size and interface number shows two basins. IINs were sampled for the CME PPIN, and the white star indicates the statistics of the actual CME IIN. The white line divides the networks 50-50. The sampled networks had to pass a threshold (red line) to transition left. **(b)** Effects of κ on fragmentation and tetramer frequency in IINs sampled from the ErbB PPIN. Black stars indicate observed values. Other parameters used were: β=4, μ=0.5, ω=0.025, k_BT=1. **(c)** The distribution of PAEs with unbiased sampling ($k_BT=\infty$) is broader for scale-free like (red bars) PPINs, meaning scale-free like IINs are more common. **(d)** For both sparse and dense PPINs, the scale-free like version (red curves) produced a higher square frequency over nearly all κ values. Other parameters were β=4, μ=0.45, ω=0.1, and k_BT=1.

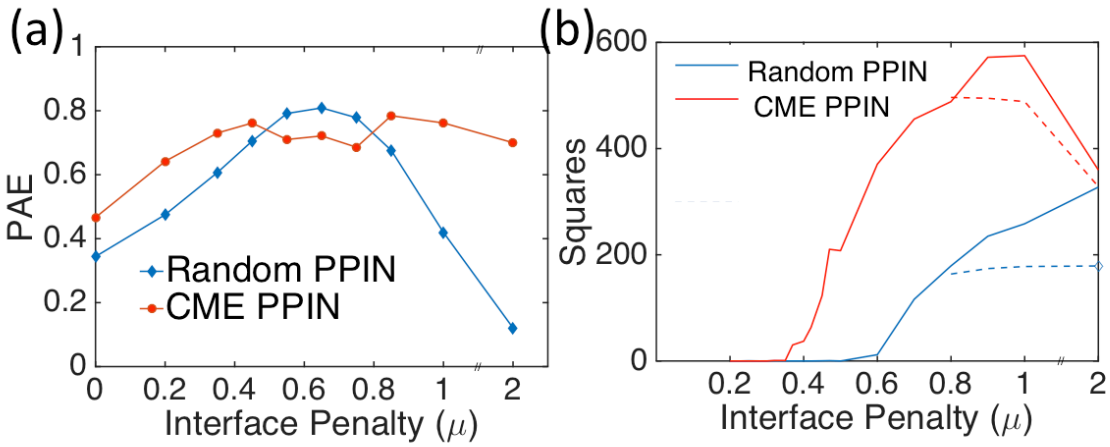
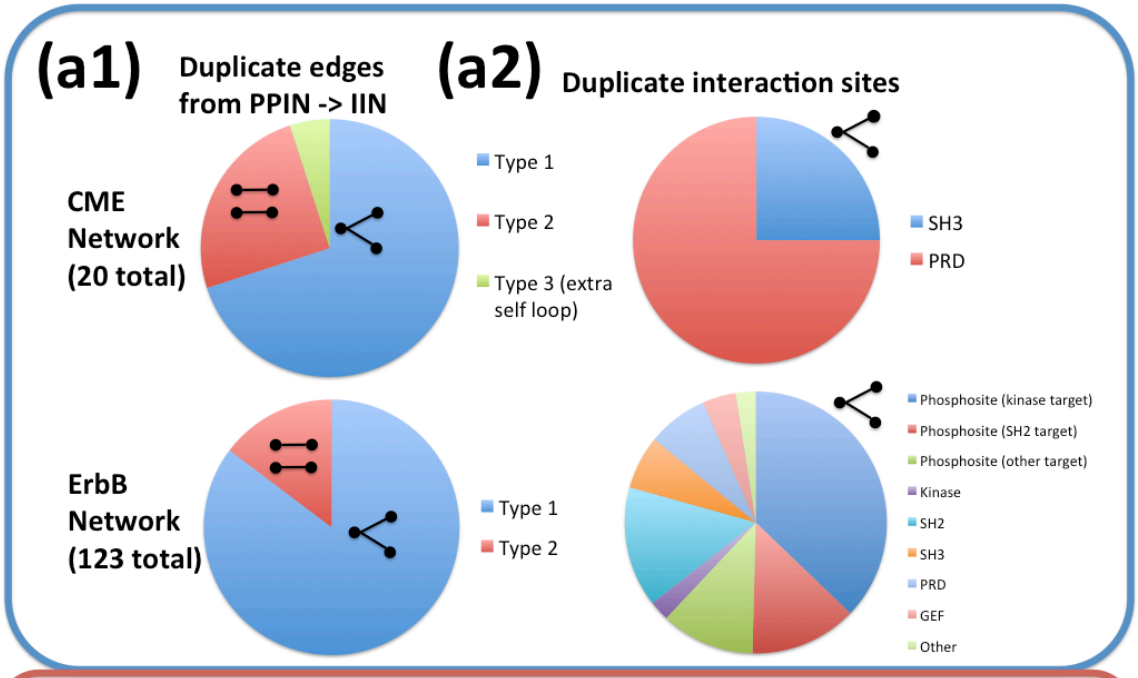
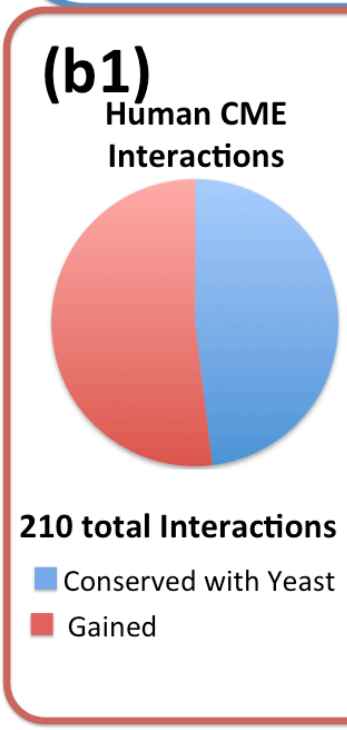


Figure C.4. Random PPINs have more constraints in selecting fit IINs.
(a) By varying the parameters of the fitness function, we verified that random PPINs (blue curves) are more limited than scale-free like PPINs (red curves) for producing sampled IINs with large PAEs. Large PAEs indicate hub *interfaces* are present. **(b)** Random PPINs also limit the frequency of square motifs in their IINs. Squares appear readily in the scale-free like PPIN (red curve) thanks to the presence of hub proteins which produce more tetramers in the PPIN that can become squares in the IIN. Edge duplication is one mechanism to produce additional squares (solid lines vs no edge duplication in dashed lines), usually by closing chains into squares.





(b2)	Human proteins	Gained inter-actions	% alignment with yeast	Sequence conser-vation
AP2 complex	AP2B1, AP2A2, AP2A1, AP2M1, AP2S1	38	AP2B1: 64%	32%
Kinases	BMP2K, AAK	9	All: ~30%	38%
SH3 containing proteins	DBNL, FNBP1L, PACSIN2, SH3YL1, MYO1E, MYO1F, BIN1, SH3GL2, SH3KBP1, ITSN1	14	-	-
PRD containing proteins	WAS, WASL, WIPF1, WIPF2, SH3KBP1, ITSN1	9	-	-
Actin Capping	CAPZA	20	94%	32%
Clathrin light chain	CLTA/B	18	~56%	19%
Synapto-janin	SYNJ1	7	50%	35%

Figure C.5. Network rewiring from human to yeast CME networks is dominated by a few proteins and numerous PPIs are duplicated in the IINs due to repeated domain copies.
(a) Many PPIs in the CME and ErbB networks produce multiple edges in their IINs due mostly to repeated copies of the same domain type. **(a1)** These extra binding modes between protein pairs where the two edges share an interface (type I) outnumber modes involving separate interfaces for each edge (type II). **(a2)** Of the type I extra binding modes, about 75% result from multiple copies of unstructured binding sites (e.g. PRRs, phosphosites). **(b)** CME interactions of Human functional homologs are compared to the Yeast interactome. **(b1)** About half the human interactions are conserved in yeast as well. **(b2)** Gained interactions were not most prominent in SH3 containing proteins, but were most heavily centered in the AP-2 complex. The AP-2 complex acquires a critical beta-appendage domain not present in yeast that acts as a hub interface in metazoans, binding multiple types of linear motifs [102] and clathrin. Both the actin capping protein and the clathrin light chains do not appear to make structural changes, but the low sequence conservation could drive acquisition of new partners to surface patches. Both lack canonically recognized binding domains.

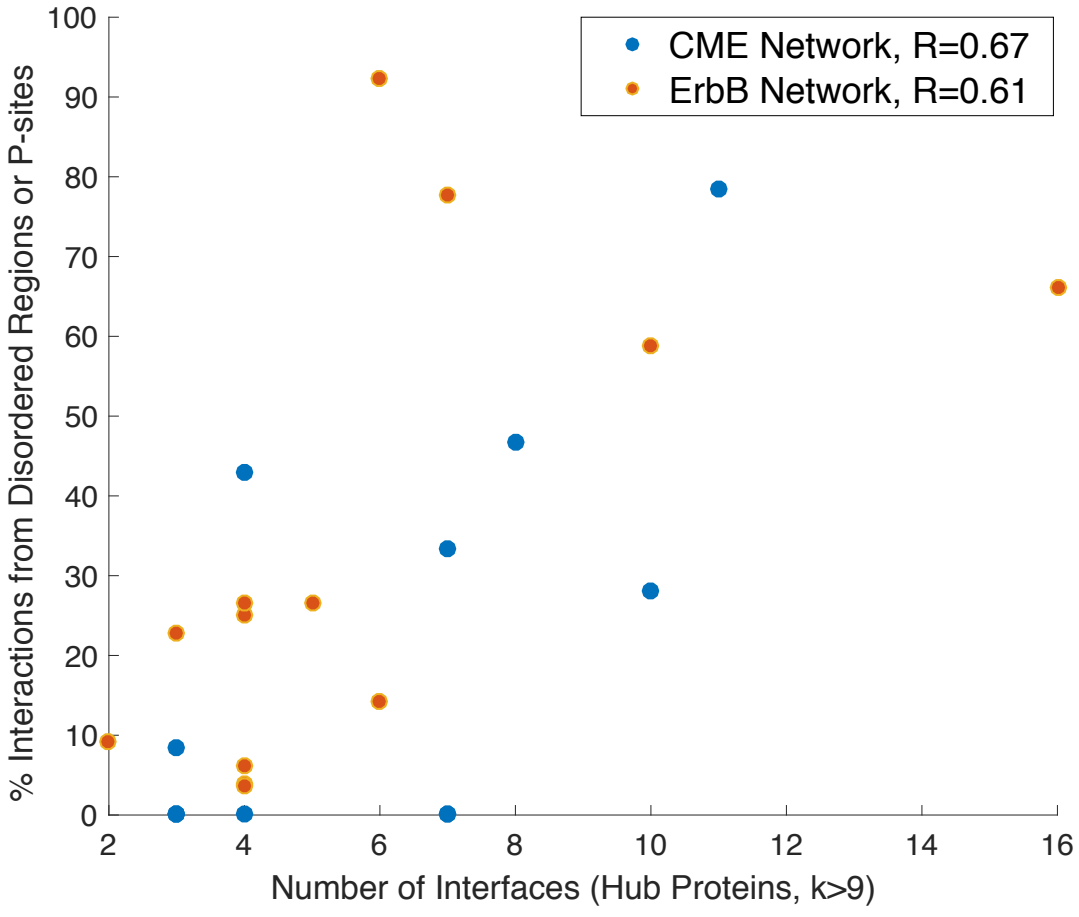


Figure C.6. Hubs mediate interactions through disordered regions. Hub proteins in both the CME and ErbB networks lie on a spectrum between having multiple disordered interfaces, or having a small number of versatile structured interfaces that bind to disordered regions on other proteins. As the plot shows, the more interfaces a hub has, the more likely it is that they mediate most of their interactions through disordered regions. Note that for the ErbB network, we counted all phosphosites as disordered, although this may not be the case in reality. For the CME network, disordered regions included proline rich regions (PRRs), acidic domains, and clathrin boxes [181].

C.2 Additional Data for Chapter 2

Table C.1. Statistics of best individual IINs for random vs scale-free like PPINs.

	Original Fitness Function		Modified Fitness Function	Original Fitness Function	
	CME PPIN	Random PPIN	Random PPIN	ErbB PPIN	Random (ErbB) PPIN
Interfaces	174	215	191	290	356
Edges	202	190	201	364	278
Pref. Attach. Exp.	0.8	0.7	0.7	1	1
Largest Component (%)	9.8%	5.1%	32%	11%	2.3%
C_{Global}	0	0	0.033	0	0
Tetramers	1534	463	2915	9423	234
Square	0.17	0.056	0.017	0.21	0.051
Chain	0.18	0.13	0.48	0.27	0.013
Hub	0.65	0.82	0.23	0.52	0.94
Other	0	0	0.28	0	0
Fitness Penalty (modified function)	74.62	357.9	251.5	---	---
Fitness Penalty (original function)	279.1	366.9	517.5	548.8	564.9

Table C.2. Residue conservation analysis for Human ErbB and Yeast CME proteins

	More Conserved than Average	Score (0 is average, <0 is conserved)
All CME and ErbB Domains/Interfaces		
Hub Interfaces (55)	89%	-0.42±0.36
Non-hubs (371)	70%	-0.22±0.6
PRRs[a] (48)	41%	0.11±0.63
Residues not in domains (174 proteins)	15%	0.3±0.4
Non-hub interfaces		
Bind to hubs (212)	64%	-0.18±0.7
Do not bind hubs (159)	77%	-0.3±0.45

[a] Proline rich regions

Table C.3: P-values for number of interfaces on CME proteins given number of connections.

Protein	Edges in IIN	Interfaces in PPIN	Expected Interfaces (Eq. S7) (Edges Column B)	p-value (Eq. S8)	Bell # (Edges Column B)
ABP1	16	7	6.9	0.72	10480142147
ACT1	24	4	9.4	0.000026	4.45959E+17
AIM21	6	3	3.3	1	203
AIM3	5	3	2.9	1	52
AIP1	2	2	1.5	1	2
AKL1	6	2	3.3	0.24	203
APL1	3	3	2	0.4	5
APL3	3	3	2	0.4	5
APM4	3	3	2	0.4	5
APP1	10	2	4.9	0.0048	115975
APS2	3	3	2	0.4	5
ARC15	3	3	2	0.4	5
ARC18	2	2	1.5	1	2

ARC19	5	5	2.9	0.038	52
ARC35	3	3	2	0.4	5
ARC40	8	3	4.1	0.34	4140
ARK1	6	2	3.3	0.24	203
ARP2	13	7	5.9	1	27644437
ARP3	13	7	5.9	1	27644437
BBC1	7	2	3.7	0.098	877
BSP1	7	4	3.7	1	877
BZZ1	5	1	2.9	0.038	52
CAP1	4	4	2.5	0.13	15
CAP2	4	4	2.5	0.13	15
CHC1	7	3	3.7	0.6	877
CLC1	2	2	1.5	1	2
COF1	2	2	1.5	1	2
CRN1	8	4	4.1	1	4140
EDE1	7	5	3.7	0.6	877
END3	4	2	2.5	1	15
ENT1	4	3	2.5	1	15
ENT2	5	3	2.9	1	52
GTS1	5	2	2.9	0.52	52
INP52	3	3	2	0.4	5
LAS17	28	11	11	0.74	6.16054E+21
LSB3	14	3	6.2	0.0042	190899322
LSB5	2	2	1.5	1	2
MYO3	10	3	4.9	0.092	115975
MYO5	13	4	5.9	0.1	27644437
PAL1	1	1	1	1	1
PAN1	17	7	7.2	0.61	82864869804
PFY1	1	1	1	1	1
PRK1	12	3	5.6	0.021	4213597
RVS161	1	1	1	1	1
RVS167	14	3	6.2	0.0042	190899322
SAC6	1	1	1	1	1
SCD5	4	2	2.5	1	15
SCP1	2	2	1.5	1	2
SLA1	23	10	9.1	0.84	4.4152E+16
SLA2	7	5	3.7	0.6	877
SYP1	5	5	2.9	0.038	52
TWF1	3	3	2	0.4	5
VRP1	7	5	3.7	0.6	877
YAP1801	4	4	2.5	0.13	15
YAP1802	7	5	3.7	0.6	877
YSC84/LSB4	13	3	5.9	0.0097	27644437

C.3 Additional Figures for Chapter 3

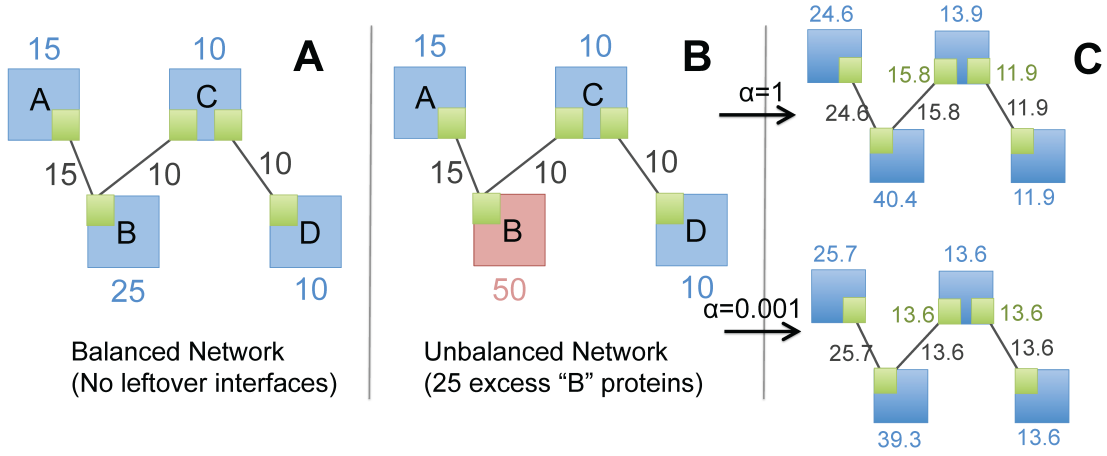


Figure C.7. Balanced vs Unbalanced Network. **A)** This left network has just enough proteins (teal numbers) to form the desired number of complexes (gray numbers). **(B)** This network has an excess of "B" proteins and is thus not balanced. **(C)** Two balanced solutions from our balancing algorithm. The top solution, which uses α=1 (equal weight on proximity to C_0 and equalizing interfaces on the same protein) gave a solution where the two interfaces on protein C were not equal. The bottom solution, which used α=0.001 (prioritizes equal interfaces) gave a solution where the two interfaces on C had the same copy numbers. We note that in many networks such a balanced solution is not possible; see Figure 3.8 for an example.

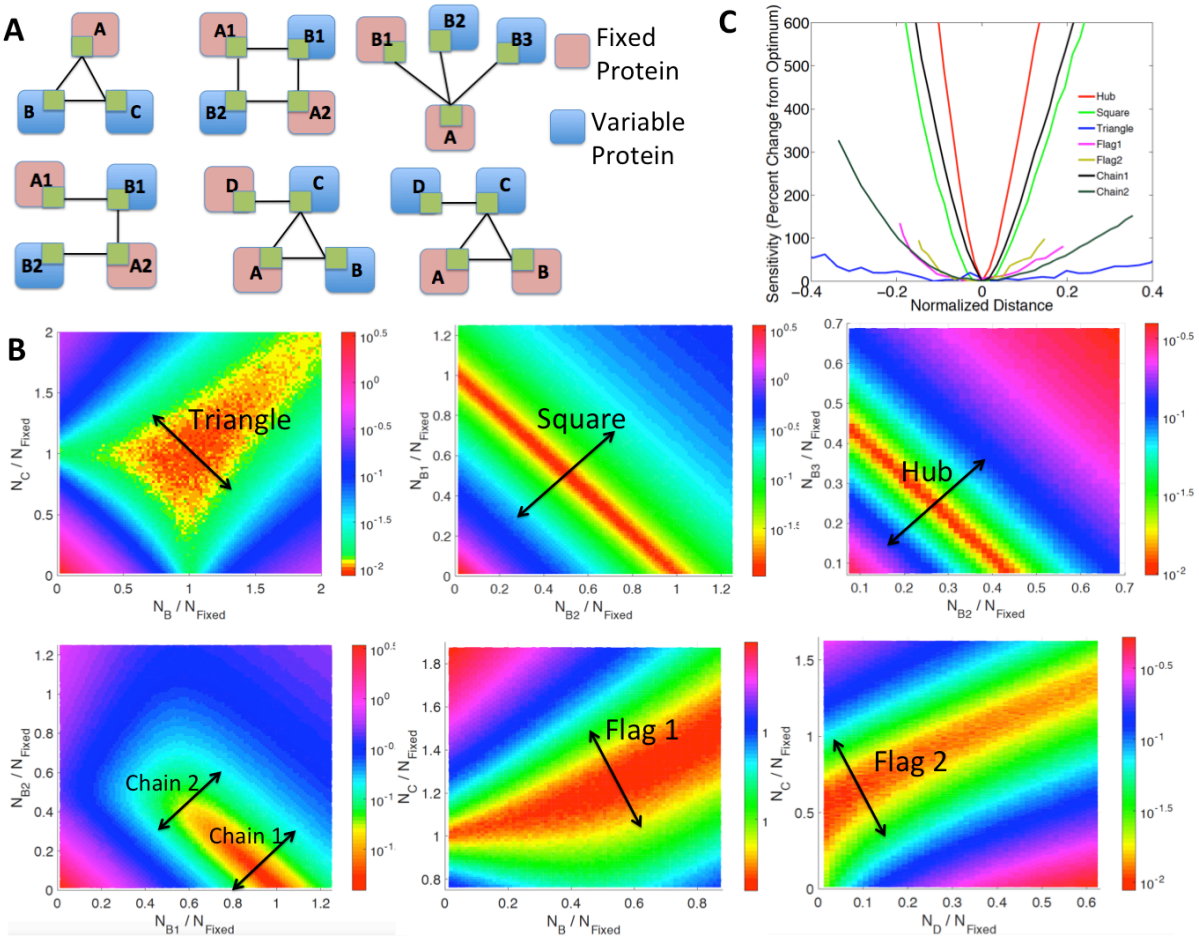


Figure C.8. Misinteraction frequency in the small network motifs. (A) Small networks used to construct the surface plots. For all simulations, two proteins had variable concentrations (blue) while the others had fixed concentrations (pink). **(B)** Surface plots of misinteraction frequency (color bar-Eq 1 main text). Misinteraction frequency is measured as $N_{\text{nonspecific}}$ / (N_{specific} + N_{free}); that is, number of nonspecific complexes divided by all other species; at steady-state as described in the main text. Each plot corresponds to each respective network in **A**. The X and Y-axes are the concentrations of the variable proteins divided by the total concentrations of the fixed proteins. The black line is the principal component, which was used as an axis to measure the sensitivity of misinteractions as one moved away from a local minimum. For the chain we used two arbitrary local minima because the absolute minimum was when B2=0, a trivial solution. For the flag network we used two different sets of fixed and variable proteins because the surface plots were asymmetric. **(C)** The sensitivity of each network to misinteraction frequency as the protein concentrations moved away from an optimum (local minimum). Sensitivity is measured as percent change from the optimal (lowest) misinteraction frequency.

Appendix D. Protein Abundances for CME and ErbB Networks

Table D.1

CME Network (Yeast)

All copy numbers from Kulak, 2014[2]

Protein	Copy Number
ABP1	30791
ACT1	117202
AIM21	1978
AIM3	220
AIP1	9739
AKL1	1307
APL1	319
APL3	426
APM4	298
APP1	70
APS2	78
ARC15	8217
ARC18	15195
ARC19	1583
ARC35	8495
ARC40	5066
ARK1	346
ARP2	11254
ARP3	8098
BBC1	6204
BSP1	458
BZZ1	2631
CAP1	10000
CAP2	5445
CHC1	19278
CLC1	14538
COF1	201066
CRN1	6509
EDE1	5964
END3	4196
ENT1	1750
ENT2	1325
GTS1	2007
INP52	393

ErbB Network (Humans)

Copy numbers pulled from five different studies[2,10,146-148]

Protein	Copy Number	Source
ABI1	67031	Kulak, 2014
AKT1	26277	Kulak, 2014
APPL1	99337	Kulak, 2014
ARAF	50606	Kulak, 2014
BCAR1	29286	Kulak, 2014
CAMK2A	#N/A	
CASP9	63095	Kulak, 2014
CAV1	28070	Kulak, 2014
CAV2	5593	Kulak, 2014
CBL	30423	Kulak, 2014
CBLB	1227	Kulak, 2014
CDC42	1010502	Kulak, 2014
CREB1	8461	Kulak, 2014
CRK	180641	Kulak, 2014
CSK	198129	Kulak, 2014
DNM1	113	Kulak, 2014
DUSP1	2459	Hein, 2015
EGF	#N/A	
EGFR	92675	Kulak, 2014
ELK1	2074	Kulak, 2014
EPN1	315	Kulak, 2014
EPS15	72354	Kulak, 2014
EPS8	1472	Kulak, 2014
ERRFI1	923	Kulak, 2014
FOS	1845	Hein, 2015
FOXO1A	59816	Nagaraj, 2011
GAB1	1531	Kulak, 2014
GAB2	1707	Kulak, 2014
GJA1	69294	Nagaraj, 2011
GRB10	24164	Kulak, 2014
GRB2	628016	Kulak, 2014
GRB7	460	Kulak, 2014
HRAS	131924	Kulak, 2014
JAK2	387	Kulak, 2014

LAS17	165	
LSB3	2910	
LSB5	787	
MYO3	1841	
MYO5	2219	
PAL1	1230	
PAN1	3357	
PFY1	29465	
PRK1	177	
RVS161	13654	
RVS167	6020	
SAC6	30402	
SCD5	46	
SCP1	2456	
SLA1	2964	
SLA2	3904	
SYP1	2467	
TWF1	4975	
VRP1	473	
YAP1801	357	
YAP1802	264	
YSC84	1833	
JUN	4110	Beck, 2011
KRAS	146936	Hein, 2015
KRT8	10993552	Hein, 2015
MAP2K1	170516	Kulak, 2014
MAP2K2	325020	Kulak, 2014
MAP2K3	244496	Kulak, 2014
MAP2K5	838	Hein, 2015
MAP2K7	89412	Kulak, 2014
MAP3K1	392	Kulak, 2014
MAP3K2	12568	Kulak, 2014
MAP3K4	96	Kulak, 2014
MAPK1	544669	Kulak, 2014
MAPK14	242167	Kulak, 2014
MAPK3	58855	Kulak, 2014
MAPK7	1060	Kulak, 2014
MAPK8	31107	Kulak, 2014
MYC	27128	Kulak, 2014
NCK1	49266	Kulak, 2014
NRAS	32806	Kulak, 2014
PAK1	7259	Kulak, 2014
PIK3CA	5982	Kulak, 2014
PIK3CB	1787	Kulak, 2014
PIK3CD	158	Kulak, 2014
PIK3CG	#N/A	
PIK3R1	448	Kulak, 2014
PIK3R2	16715	Kulak, 2014
PIK3R3	2060	Kulak, 2014
PLCG1	69601	Hein, 2015
PLCG2	5957	Hein, 2015
PLD1	3278	Kulak, 2014
PLD2	6520	Kulak, 2014
PRKCA	27673	Kulak, 2014
PRKCB1	#N/A	
PRKCG	121760	Nagaraj, 2011
PRKCI	72034	Kulak, 2014
PRKCZ	3801	Kulak, 2014
PTK2B	5423	Kulak, 2014
PTPN11	299893	Kulak, 2014
PTPN5	#N/A	
PTPN6	4198	Kulak, 2014
PTPRR	10058	Kulak, 2014
PXN	94096	Kulak, 2014
RAB5A	68671	Kulak, 2014
RAC1	1675011	Kulak, 2014
RAF1	35757	Kulak, 2014
RALB	27800	Kulak, 2014
RALBP1	11288	Kulak, 2014
RALGDS	#N/A	

RASA1	26455	Kulak, 2014
RGS16	#N/A	
RPS6KA1	400730	Kulak, 2014
RPS6KA3	112170	Kulak, 2014
SH3GL2	4333	Kulak, 2014
SH3KBP1	89628	Kulak, 2014
SHC1	111966	Kulak, 2014
SMAD2	139990	Kulak, 2014
SMAD3	3425	Kulak, 2014
SOS1	7565	Kulak, 2014
SOS2	144	Kulak, 2014
SP1	248	Kulak, 2014
SPRY2	#N/A	
SRC	1220	Kulak, 2014
STAT1	201741	Kulak, 2014
STAT3	241307	Kulak, 2014
STAT5A	135833	Kulak, 2014
STAT5B	31869	Kulak, 2014
TNK2	646	Kulak, 2014
USP6NL	1818	Kulak, 2014
VAV1	59	Kulak, 2014
VAV2	106531	Kulak, 2014
VAV3	167890	Nagaraj, 2011
WASL	93628	Kulak, 2014
YWHAB	3298759	Kulak, 2014
RIN1	65438	Kulak, 2014
RALA	589879	Kulak, 2014
DUSP4	#N/A	
DUSP6	#N/A	
BRAF	785	Kulak, 2014
KSR1	954	Kulak, 2014
PDPK1	24665	Kulak, 2014
PRKCE	768	Hein, 2015
BAD	22271	Kulak, 2014
RHOA	358430	Kulak, 2014
ARHGEF7	100897	Kulak, 2014
MAP3K11	513	Kulak, 2014
MAP2K4	72580	Kulak, 2014
MAP2K6	77223	Kulak, 2014
LIMK1	4809	Kulak, 2014
ARHGAP3	#N/A	
P53	4220	Beck, 2011
ERBB2	835	Kulak, 2014
ERBB3	1823	Wisniewski, 2014
ERBB4	#N/A	

Appendix E. Further Notes and Code for BioNetGen Models

E.1 Vesicle Forming Model Notes

The purpose of simulating the vesicle forming module was to determine the relative effects of copy number imbalance on clathrin recruitment and vesicle formation at the cell membrane. Our model utilized nine base species: the clathrin heavy chain (CHC1) in trimer form, clathrin light chain (CLC1), the lipid PtdIns(4,3)P_2, and the adaptor proteins ENT1/2, SLA2, YAP1801, YAP1802, SYP1, and EDE1. ENT1 and ENT2 were combined into one specie with a concentration defined by the sum of their individual copy numbers because their partners and binding affinities were equivalent. Parameters with notes and references are shown in Table 4.1. Steps such as triggering by cargo or the role of actin in vesicle invagination and scission were not captured.

E.1.1 Spatial Compartments. Spatial resolution was not characterized. Instead, the cytosol and membrane were treated as two separate compartments. The solution volume and the membrane surface area were defined based on the known dimensions of the yeast cell. For species and reactions that occurred in the membrane compartment, the concentrations and binding affinities are all converted from units of um^{-2} to standard solution units L^{-1} based on a single length-scale conversion, σ. The 'volume' of the membrane compartment is then given by Surface

Area (SA)* 2σ. This parameter σ is not an arbitrary definition of the membrane depth. Rather, it is most accurately represented as the conversion between the equilibrium constants in 3D versus those in 2D, such that $K_a^{2D}=K_a^{3D}/(2\sigma)$. As we detailed in a recent study[158], σ is thus a thermodynamic property of each binding pair, and will depend on the molecular properties of the proteins involved. Although rarely measured, both experiment and theory[174,182,183] quantify σ on the nanometer lengthscale. In the NFsim simulations, all protein pairs will bind based on the same value of σ, which we set to 1 nm. The reduced dimensionality of the membrane surface, which is here captured in the effective 'volume' of the membrane compartment, creates higher concentrations of proteins on the membrane that then promotes binding. Binding events where one protein is in the cytosol and one protein on the membrane requires a search within the solution compartment for the membrane bound protein. Thus, these reactions are based on the true 3D cytosolic volume to define the K_a for the reaction.

E.1.2 Vesicle formation reaction . Vesicle formation was treated as instantaneous once an aggregate of 100 triskelia was formed. A triskelia in our model was required to contain a heavy chain trimer bound to 3 light chains, otherwise it was not counted toward the sum. The aggregate was deleted at a rapid rate of 1000 s^-1 (NFSim does not allow a true instantaneous rule, but treats every rule as a type of reaction). Proteins were then added back to the cytosol at a rate of 1000 s^-1, until the total number of proteins equaled the starting number. In reality, it takes time for both the vesicle to scission -- with the help of the cell cytoskeleton -- and further

time for the clathrin cage to be broken up and the proteins returned to the cytosol. But because our focus was on clathrin recruitment as it controls vesicle formation, and because parameters for protein recycling time are unavailable in the literature, we choose to make these rates uniformly rapid.

E.1.3 Binding affinities. Binding rates for human homologs were used when the binding rates for yeast proteins were not available. In the case of SYP1, a binding rate between a generic F-BAR domain and a PtdIns(4,3)P_2 molecule was used, but this assumes that the domain only binds one PtdIns(4,3)P_2 molecule when it may bind two. SYP1 may also bind to other types of lipids or to transmembrane "cargo" receptors. Thus the true binding rate may be larger. But we found that increasing binding rate by a factor of 100 did not increase vesicle formation, because the limiting step was EDE1 binding to SYP1.

E.1.4 Structural assumptions. Steric hindrance may weaken or even prevent binding depending on the composition of the aggregate. For example, even though one SLA2 protein can bind one clathrin light chain, the composition of a SLA2 - triskelia complex may be less than 3 to 1 due to steric inhibition. But without further structural information, this steric inhibition is impossible to characterize. Similarly, EDE1 has three EH domains, but it is unclear how many partners it may bind at once. We provided EDE1 with two EH binding sites in our model.

E.2 BNGL Code for ARP2/3 Complex Model

```
begin model
begin parameters
      NA 6.022e23 #mol^-1
      sigma 0.002 #um
      vol_CP 3.2e-15 #volume L
      sa_PM 2.1715 #um^2 Assuming a sphere

      #Binding parameters
      kon 0.1
      kon2 1
      kon3 1
      kon4 1000
      kon_arc18_arp3 kon
      kon_arp2_arc15 kon
      kon_arc15_arc40 kon
      kon_arc15_arc19 kon
      kon_arc19_arc35 kon
      kon_arc19_arc40 kon
      kon_arc19_arp2 kon
      kon_arc19_arp3 kon
      kon_arc35_arp3 kon
      kon_arp2_arp3 kon
      koff 1
end parameters

begin molecule types
      ARC35(a,b)
      ARC18(a)
      ARC15(a,b,c)
      ARC19(a,b,c,d,f)
      ARC40(a,b)
      ARP2(a,b,c)
      ARP3(a,b,c,d)
end molecule types

begin seed species
      ARC35(a,b) 8495
      ARC18(a) 15195
      ARC15(a,b,c) 8217
      ARC19(a,b,c,d,f) 1583
      ARC40(a,b) 5066
```

```
      ARP2(a,b,c) 11254
      ARP3(a,b,c,d) 8098
end seed species

begin observables

      Molecules FreeARC35 ARC35(a,b)
      Molecules FreeARC18 ARC18(a)
      Molecules FreeARC15 ARC15(a,b,c)
      Molecules FreeARC19 ARC19(a,b,c,d,f)
      Molecules FreeARC40 ARC40(a,b)
      Molecules FreeARP2 ARP2(a,b,c)
      Molecules FreeARP3 ARP3(a,b,c,d)
      Species FullComplex
ARC19(a!1,b!2,c!3,d!4,f!5).ARC18(a!6).ARC35(a!5,b!7).ARC15(a!1,b!
8,c!9).ARC40(a!2,b!8).ARP2(a!9,b!3,c!10).ARP3(a!6,b!4,c!7,d!10)
      Species TwoARC19 ARC19().ARC19()
      Species ThreeARC19 ARC19().ARC19().ARC19()
      Species FourARC19 ARC19().ARC19().ARC19().ARC19()

      Molecules ARC_19 ARC19()

end observables

begin functions
      BooleanFunc(x) = if(ARC_19(x)<15,kon,0)
      BooleanFunc2(x) = if(ARC_19(x)<15,kon2,0)
end functions

begin reaction rules
      %x:ARC15(a)+ARC19(a)<->ARC15(a!1).ARC19(a!1)
            BooleanFunc2(x), koff
      %x:ARC15(b)+ARC40(b)<->ARC15(b!1).ARC40(b!1)
            BooleanFunc(x), koff
      %x:ARC15(c)+ARP2(a)<->ARC15(c!1).ARP2(a!1) BooleanFunc(x),
            koff
      ARC18(a)+ARP3(a)<->ARC18(a!1).ARP3(a!1) kon3, koff
      ARC19(f)+%x:ARC35(a)<->ARC19(f!1).ARC35(a!1)
            BooleanFunc2(x), koff
      ARC19(b)+%x:ARC40(a)<->ARC19(b!1).ARC40(a!1)
            BooleanFunc2(x), koff
      ARC19(c)+%x:ARP2(b)<->ARC19(c!1).ARP2(b!1) BooleanFunc2(x),
            koff
      ARC19(d)+%x:ARP3(b)<->ARC19(d!1).ARP3(b!1) BooleanFunc2(x),
```

```
        koff
    %x:ARC35(b)+ARP3(c)<->ARC35(b!1).ARP3(c!1) BooleanFunc(x),
        koff
    %x:ARP2(c)+ARP3(d)<->ARP2(c!1).ARP3(d!1) BooleanFunc(x),
        koff

    #Need to add a new set of rules for trimer binding. If two
        members of the trimer are
    # bound, the third binding event should be much more likely
    #Will need 12 rules in total, 3 for each possible trimer

    ARC19(b!1,a).ARC40(a!1,b!2).ARC15(b!2,a) ->
        ARC19(b!1,a!3).ARC40(a!1,b!2).ARC15(b!2,a!3) kon4
    ARC19(b!1,a!3).ARC40(a!1,b).ARC15(b,a!3) ->
        ARC19(b!1,a!3).ARC40(a!1,b!2).ARC15(b!2,a!3) kon4
    ARC19(b,a!3).ARC40(a,b!2).ARC15(b!2,a!3) ->
        ARC19(b!1,a!3).ARC40(a!1,b!2).ARC15(b!2,a!3) kon4

    ARC19(a!1,c).ARP2(a!3,b).ARC15(a!1,c!3) ->
        ARC19(a!1,c!2).ARP2(a!3,b!2).ARC15(a!1,c!3) kon4
    ARC19(a,c!2).ARP2(a!3,b!2).ARC15(a,c!3) ->
        ARC19(a!1,c!2).ARP2(a!3,b!2).ARC15(a!1,c!3) kon4
    ARC19(a!1,c!2).ARP2(a,b!2).ARC15(a!1,c) ->
        ARC19(a!1,c!2).ARP2(a!3,b!2).ARC15(a!1,c!3) kon4

    ARC19(c!1,d!2).ARP2(b!1,c).ARP3(b!2,d) ->
        ARC19(c!1,d!2).ARP2(b!1,c!3).ARP3(b!2,d!3) kon4
    ARC19(c!1,d).ARP2(b!1,c!3).ARP3(b,d!3) ->
        ARC19(c!1,d!2).ARP2(b!1,c!3).ARP3(b!2,d!3) kon4
    ARC19(c,d!2).ARP2(b,c!3).ARP3(b!2,d!3) ->
        ARC19(c!1,d!2).ARP2(b!1,c!3).ARP3(b!2,d!3) kon4

    ARC19(d!1,f!2).ARP3(b!1,c).ARC35(a!2,b) ->
        ARC19(d!1,f!2).ARP3(b!1,c!3).ARC35(a!2,b!3) kon4
    ARC19(d!1,f).ARP3(b!1,c!3).ARC35(a,b!3) ->
        ARC19(d!1,f!2).ARP3(b!1,c!3).ARC35(a!2,b!3) kon4
    ARC19(d,f!2).ARP3(b,c!3).ARC35(a!2,b!3) ->
        ARC19(d!1,f!2).ARP3(b!1,c!3).ARC35(a!2,b!3) kon4

end reaction rules
end model

simulate_nf({suffix=>"nf",t_end=>0.01,n_steps=>1000});
```

E.3 BNGL Code for Vesicle Forming Model

```
begin model
begin parameters

      #Volumes
      NA 6.022e23 #mol^-1
      sigma 0.002 #um
      vol_CP 37.2e-15 #volume L
      sa_PM 75.7 #um^2
      vol_PM sa_PM*sigma*1e-15 #volume L
      #Binding parameters (Kd in units of microMolar. Convert to
           molar, then get kon rate, then divide by NA)
      koff 1
      kdump 1000 #Rate of destruction for Clathrin complexes of a
           certain size
      kc 1000#200 #Recycling rate for adding molecules back into
           the pool

      kon_chc_chc (1/(100*1e-6))/NA #Liter/second
      kon_chc_ent (1/(22*1e-6))/NA
      kon_chc_yap (1/(160*1e-6))/NA
      kon_ede_ent (1/(12*1e-6))/NA
      kon_ede_yap (1/(0.6*1e-6))/NA
      kon_ede_ede (1/(0.127*1e-6))/NA
      kon_chc_clc (1/(0.0001*1e-6))/NA
      kon_clc_sla2 (1/(22*1e-6))/NA
      kon_sla2_sla2 (1/(0.001*1e-6))/NA
      kon_syp_syp (1/(2.5*1e-6))/NA
      kon_syp_ede (1/(0.227*1e-6))/NA

      kon_l_ent (1/(0.02*1e-6))/NA
      kon_l_yap (1/(0.3*1e-6))/NA
      kon_l_sla2 (1/(0.2*1e-6))/NA
      kon_l_syp (1/(53*1e-6))/NA

      #Rate constants for cytoplasm
      kon_chc_chc_cy kon_chc_chc/vol_CP #1/second
      kon_chc_ent_cy kon_chc_ent/vol_CP
      kon_chc_yap_cy kon_chc_yap/vol_CP
      kon_ede_ent_cy kon_ede_ent/vol_CP
      kon_ede_yap_cy kon_ede_yap/vol_CP
      kon_ede_ede_cy kon_ede_ede/vol_CP
```

```
        kon_chc_clc_cy kon_chc_clc/vol_CP
        kon_clc_sla2_cy kon_clc_sla2/vol_CP
        kon_sla2_sla2_cy kon_sla2_sla2/vol_CP
        kon_syp_syp_cy kon_syp_syp/vol_CP
        kon_syp_ede_cy kon_syp_ede/vol_CP

        kon_l_ent_cy kon_l_ent/vol_CP
        kon_l_yap_cy kon_l_yap/vol_CP
        kon_l_sla2_cy kon_l_sla2/vol_CP
        kon_l_syp_cy kon_l_syp/vol_CP

        #Rate increase for lipid membranes

        kon_chc_chc_pm kon_chc_chc/vol_PM - kon_chc_chc_cy
        kon_chc_ent_pm kon_chc_ent/vol_PM - kon_chc_ent_cy
        kon_chc_yap_pm kon_chc_yap/vol_PM - kon_chc_yap_cy
        kon_ede_ent_pm kon_ede_ent/vol_PM - kon_ede_ent_cy
        kon_ede_yap_pm kon_ede_yap/vol_PM - kon_ede_yap_cy
        kon_ede_ede_pm kon_ede_ede/vol_PM - kon_ede_ede_cy
        kon_chc_clc_pm kon_chc_clc/vol_PM - kon_chc_clc_cy
        kon_clc_sla2_pm kon_clc_sla2/vol_PM - kon_clc_sla2_cy
        kon_sla2_sla2_pm kon_sla2_sla2/vol_PM - kon_sla2_sla2_cy
        kon_syp_syp_pm kon_syp_syp/vol_PM - kon_syp_syp_cy
        kon_syp_ede_pm kon_syp_ede/vol_PM - kon_syp_ede_cy

        kon_l_ent_pm kon_l_ent/vol_PM - kon_l_ent_cy
        kon_l_yap_pm kon_l_yap/vol_PM - kon_l_yap_cy
        kon_l_sla2_pm kon_l_sla2/vol_PM - kon_l_sla2_cy
        kon_l_syp_pm kon_l_syp/vol_PM - kon_l_syp_cy

        #Initial Copy Numbers
        CHC1_0 6426 #19278/3
        CLC1_0 14538
        EDE1_0 5964
        ENT1_0 1750
        ENT2_0 1325
        YAP1801_0 357
        YAP1802_0 264
        SLA2_0 3904
        SYP1_0 2467
        L_0 2.5292e4*sa_PM # particles/um^2 -> particles

end parameters
begin molecule types
```

```
      CHC1(a1,a2,a3,b1,b2,b3,c1,c2,c3)
      CLC1(a,b)
      EDE1(a,a,b,c)
      ENT(a,b,lb)
      YAP1801(a,lb)
      YAP1802(a,b)
      SLA2(a,b,lb)
      SYP1(a,b,lb)
      L(p)
      Ve()
end molecule types

begin seed species
      CHC1(a1,a2,a3,b1,b2,b3,c1,c2,c3) CHC1_0
      CLC1(a,b)  CLC1_0
      EDE1(a,a,b,c)    EDE1_0
      ENT(a,b,lb)      ENT1_0 + ENT2_0
      YAP1801(a,lb)    YAP1801_0
      YAP1802(a,b)     YAP1802_0
      SLA2(a,b,lb) SLA2_0
      SYP1(a,b,lb) SYP1_0
      L(p)  L_0
      Ve() 0
end seed species

begin observables
      Molecules ClathrinLip CHC1().L()
      Molecules Clathrin CHC1()
      Molecules CHCself CHC1().CHC1()
      Molecules FreeCHC CHC1(a1,a2,a3,b1,b2,b3)
      Molecules Yap_CHC YAP1801(a!+).CHC1()
      Molecules ENT_CHC ENT(a!+).CHC1()
      Molecules ClathrinLight CLC1()
      Molecules EDE_1 EDE1()
      Molecules ENT ENT()
      Molecules SLA_2 SLA2()
      Molecules SYP_1 SYP1()
      Molecules ENT_L ENT(lb!1).L(p!1)
      Molecules YAP_L YAP1801(lb!1).L(p!1)
      Molecules SLA_L SLA2(lb!1).L(p!1)
      Molecules SLA_CHC SLA2(a!+).CHC1()
      Molecules YAP_1801 YAP1801()
      Molecules YAP_1802 YAP1802()
      Molecules Lipids L()
```

```
      Molecules EDEdub EDE1(a!+,a!+)
      Molecules EDEsin EDE1(a!+)
      Molecules Vesicles Ve()
      Molecules Triskelia
            CHC1(c1!1,c2!2,c3!3).CLC1(a!1).CLC1(a!2).CLC1(a!3)
      Molecules TL
            CHC1(c1!1,c2!2,c3!3).CLC1(a!1).CLC1(a!2).CLC1(a!3).L()
end observables

begin functions
#Need a function that causes a clathrin "dump" back
# into the cytosol after a clathrin cage reaches a certain size.
      BooleanFunc(x) = if(Triskelia(x)>=100,kdump,0)

      CreateL() = if(Lipids<L_0,kc,0)
      CreateCHC() = if(Clathrin<CHC1_0,kc/2,0) #Assuming CHC to
adaptor ratio is roughly 2 to 1
      CreateCLC() = if(ClathrinLight<CLC1_0,kc/2,0)
      CreateENT() = if(ENT<ENT1_0+ENT2_0,kc,0)
      CreateEDE1() = if(EDE_1<EDE1_0,kc,0)
      CreateYAP1801() = if(YAP_1801<YAP1801_0,kc,0)
      CreateYAP1802() = if(YAP_1802<YAP1802_0,kc,0)
      CreateSLA2() = if(SLA_2<SLA2_0,kc,0)
      CreateSYP1() = if(SYP_1<SYP1_0,kc,0)

end functions


begin reaction rules
      CHC1(a1) + CHC1(a1) <-> CHC1(a1!1).CHC1(a1!1)
            kon_chc_chc_cy, koff
      CHC1(a1).L() + CHC1(a1).L() ->
            CHC1(a1!1).L().CHC1(a1!1).L() kon_chc_chc_pm
      CHC1(a2) + CHC1(a2) <-> CHC1(a2!1).CHC1(a2!1)
            kon_chc_chc_cy, koff
      CHC1(a2).L() + CHC1(a2).L() ->
            CHC1(a2!1).L().CHC1(a2!1).L() kon_chc_chc_pm
      CHC1(a3) + CHC1(a3) <-> CHC1(a3!1).CHC1(a3!1)
            kon_chc_chc_cy, koff
      CHC1(a3).L() + CHC1(a3).L() ->
            CHC1(a3!1).L().CHC1(a3!1).L() kon_chc_chc_pm
      CHC1(a1) + CHC1(a2) <-> CHC1(a1!1).CHC1(a2!1)
            kon_chc_chc_cy, koff
      CHC1(a1).L() + CHC1(a2).L() ->
```

```
      CHC1(a1!1).L().CHC1(a2!1).L() kon_chc_chc_pm
CHC1(a2) + CHC1(a3) <-> CHC1(a2!1).CHC1(a3!1)
      kon_chc_chc_cy, koff
CHC1(a2).L() + CHC1(a3).L() ->
      CHC1(a2!1).L().CHC1(a3!1).L() kon_chc_chc_pm
CHC1(a1) + CHC1(a3) <-> CHC1(a1!1).CHC1(a3!1)
      kon_chc_chc_cy, koff
CHC1(a1).L() + CHC1(a3).L() ->
      CHC1(a1!1).L().CHC1(a3!1).L() kon_chc_chc_pm

CHC1(b1) + ENT(a) <-> CHC1(b1!1).ENT(a!1) kon_chc_ent_cy,
      koff
CHC1(b1).L() + ENT(a).L() -> CHC1(b1!1).L().ENT(a!1).L()
      kon_chc_ent_pm
CHC1(b1) + YAP1801(a) <-> CHC1(b1!1).YAP1801(a!1)
      kon_chc_yap_cy, koff
CHC1(b1).L() + YAP1801(a).L() ->
      CHC1(b1!1).L().YAP1801(a!1).L() kon_chc_yap_pm
CHC1(b1) + YAP1802(b) <-> CHC1(b1!1).YAP1802(b!1)
      kon_chc_yap_cy, koff
CHC1(b1).L() + YAP1802(b).L() ->
      CHC1(b1!1).L().YAP1802(b!1).L() kon_chc_yap_pm

CHC1(b2) + ENT(a) <-> CHC1(b2!1).ENT(a!1) kon_chc_ent_cy,
      koff
CHC1(b2).L() + ENT(a).L() -> CHC1(b2!1).L().ENT(a!1).L()
      kon_chc_ent_pm
CHC1(b2) + YAP1801(a) <-> CHC1(b2!1).YAP1801(a!1)
      kon_chc_yap_cy, koff
CHC1(b2).L() + YAP1801(a).L() ->
      CHC1(b2!1).L().YAP1801(a!1).L() kon_chc_yap_pm
CHC1(b2) + YAP1802(b) <-> CHC1(b2!1).YAP1802(b!1)
      kon_chc_yap_cy, koff
CHC1(b2).L() + YAP1802(b).L() ->
      CHC1(b2!1).L().YAP1802(b!1).L() kon_chc_yap_pm

CHC1(b3) + ENT(a) <-> CHC1(b3!1).ENT(a!1) kon_chc_ent_cy,
      koff
CHC1(b3).L() + ENT(a).L() -> CHC1(b3!1).L().ENT(a!1).L()
      kon_chc_ent_pm
CHC1(b3) + YAP1801(a) <-> CHC1(b3!1).YAP1801(a!1)
      kon_chc_yap_cy, koff
CHC1(b3).L() + YAP1801(a).L() ->
      CHC1(b3!1).L().YAP1801(a!1).L() kon_chc_yap_pm
```

```
CHC1(b3) + YAP1802(b) <-> CHC1(b3!1).YAP1802(b!1)
     kon_chc_yap_cy, koff
CHC1(b3).L() + YAP1802(b).L() ->
     CHC1(b3!1).L().YAP1802(b!1).L() kon_chc_yap_pm

CHC1(c1) + CLC1(a) <-> CHC1(c1!1).CLC1(a!1) kon_chc_clc_cy,
     koff
CHC1(c1).L() + CLC1(a).L() -> CHC1(c1!1).L().CLC1(a!1).L()
     kon_chc_clc_pm
CHC1(c2) + CLC1(a) <-> CHC1(c2!1).CLC1(a!1) kon_chc_clc_cy,
     koff
CHC1(c2).L() + CLC1(a).L() -> CHC1(c2!1).L().CLC1(a!1).L()
     kon_chc_clc_pm
CHC1(c3) + CLC1(a) <-> CHC1(c3!1).CLC1(a!1) kon_chc_clc_cy,
     koff
CHC1(c3).L() + CLC1(a).L() -> CHC1(c3!1).L().CLC1(a!1).L()
     kon_chc_clc_pm
EDE1(a) + ENT(b) <-> EDE1(a!1).ENT(b!1) kon_ede_ent_cy,
     koff
EDE1(a).L() + ENT(b).L() -> EDE1(a!1).L().ENT(b!1).L()
     kon_ede_ent_pm
EDE1(a) + YAP1802(a) <-> EDE1(a!1).YAP1802(a!1)
     kon_ede_yap_cy, koff
EDE1(a).L() + YAP1802(a).L() ->
     EDE1(a!1).L().YAP1802(a!1).L() kon_ede_yap_pm
EDE1(b) + EDE1(b) <-> EDE1(b!1).EDE1(b!1) kon_ede_ede_cy,
     koff
EDE1(b).L() + EDE1(b).L() -> EDE1(b!1).L().EDE1(b!1).L()
     kon_ede_ede_pm
L(p) + ENT(lb) <-> L(p!1).ENT(lb!1) kon_l_ent_cy, koff
L(p) + ENT(lb).L() -> L(p!1).ENT(lb!1).L() kon_l_ent_pm
L(p) + YAP1801(lb) <-> L(p!1).YAP1801(lb!1) kon_l_yap_cy,
     koff
L(p) + YAP1801(lb).L() -> L(p!1).YAP1801(lb!1).L()
     kon_l_yap_pm

CLC1(b) + SLA2(a) <-> CLC1(b!1).SLA2(a!1) kon_clc_sla2_cy,
     koff
CLC1(b).L() + SLA2(a).L() -> CLC1(b!1).L().SLA2(a!1).L()
     kon_clc_sla2_pm
SLA2(b) + SLA2(b) <-> SLA2(b!1).SLA2(b!1) kon_sla2_sla2_cy,
     koff
SLA2(b).L() + SLA2(b).L() -> SLA2(b!1).L().SLA2(b!1).L()
     kon_sla2_sla2_pm
```

```
      L(p) + SLA2(lb) <-> L(p!1).SLA2(lb!1) kon_l_sla2_cy, koff
      L(p) + SLA2(lb).L() -> L(p!1).SLA2(lb!1).L() kon_l_sla2_pm

      SYP1(b) + SYP1(b) <-> SYP1(b!1).SYP1(b!1) kon_syp_syp_cy,
           koff
      SYP1(b).L() + SYP1(b).L() -> SYP1(b!1).L().SYP1(b!1).L()
           kon_syp_syp_pm
      SYP1(a) + EDE1(c) <-> SYP1(a!1).EDE1(c!1) kon_syp_ede_cy,
           koff
      SYP1(a).L() + EDE1(c).L() -> SYP1(a!1).L().EDE1(c!1).L()
           kon_syp_ede_pm

      L(p) + SYP1(lb) <-> L(p!1).SYP1(lb!1) kon_l_syp_cy, koff
      L(p) + SYP1(lb).L() -> L(p!1).SYP1(lb!1).L() kon_l_syp_pm

      #Need to "dump" everything back into cytosol:
      # 1) Delete the complex
      # 2) Generate new molecules until the total amount reaches
           the original "total"
      %c::L() -> Ve() BooleanFunc(c)

      0 -> CHC1(a1,a2,a3,b1,b2,b3,c1,c2,c3) CreateCHC()
      0 -> CLC1(a,b)+CLC1(a,b)+CLC1(a,b) CreateCLC() #Create 3
           times as much to balance with CHC
      0 -> EDE1(a,a,b,c) CreateEDE1()
      0 -> ENT(a,b,lb) CreateENT()
      0 -> YAP1801(a,lb) CreateYAP1801()
      0 -> YAP1802(a,b) CreateYAP1802()
      0 -> L(p)+L(p)+L(p)+L(p) CreateL() #Faster recycling since
           four kinds of proteins bind it
      0 -> SLA2(a,b,lb) CreateSLA2()
      0 -> SYP1(a,b,lb) CreateSYP1()

end reaction rules
end model

simulate_nf({suffix=>"nf",t_end=>20,n_steps=>5000,param=>"-gml
      2000000"});
```

Glossary of Acronyms and Common Terms

DBH: Dosage-balance hypothesis. States that protein concentrations should be balanced according to the stoichiometry of the protein complexes they form. Dosage balance is also referred to as stoichiometric balance.

Dosage sensitivity: The observed deleterious effect of overexpressing certain genes in a cell.

CME: Clathrin-mediated endocytosis, a process by which cells take in external matter by using clathrin proteins to form a hexagonal cage which alters the curvature of the cell membrane.

CNV: Copy-number variation, usually of a gene.

CSD: Chi-square distance. See Eq. 3.5.

Degree: The number of connections of a node in a network

Domain: A structured region of a protein. Large proteins have a modular design where conserved domains are connected by strands of disordered regions.

Hub (protein): A protein that contains a high number of connections compared to other proteins in the network. Statistics to identify hubs vary.

Date hub: A protein with many connections but few interfaces.

Party hub: A protein with many connections spread across many interfaces.

IIN: Interface-interaction network. A map of the binding partners of interfaces on proteins.

Interface: A region of a protein used to interact with another protein. Also called a binding site. May be located on a domain or a disordered region.

IPH: Interaction-promiscuity hypothesis. States that the primary cause of dosage sensitivity is an increase in misinteractions.

JSD: Jensen-Shannon distance. See Eq. 3.6.

Misinteraction: A weak protein-protein interaction not selected for by evolution.

Monte Carlo: A computational method that relies on repeated random sampling. The algorithm in this thesis uses transition probabilities to sample possible states.

Motif: A pattern that appears more often than one would expect at random.
Network motifs refer to subgraphs usually 3-4 nodes in size.
Linear motifs refer to amino acid sequence patterns on disordered regions of proteins, usually used to bind to structured regions of other proteins

PPIN: Protein-protein interaction network, a map of the binding partners of proteins

PRR: Proline rich region. A common linear motif that binds to an SH3 domain. Also referred to (inaccurately) as Proline rich domain (PRD) in the literature.

Quadratic Programming: Algorithm used to minimize a quadratic objective function subject to linear constraints.

Rule-based modeling: Dynamic modeling method where one defines rules that proteins use to bind, rather than manually enumerating all possible complexes.

Scale-free network: A network where the number of connections of each node (degree) fits a power-law distribution $P(k) \propto k^{-\gamma}$. So-called because the distribution properties of $P(k)$ are independent of the scale of k, as $P(ak)=a^{-\gamma}P(k)$

Site Graph: PPIN map displaying the interfaces that proteins use to bind. Also called a contact map.

SLiM: Short Linear Motif (see Motif)

SH3: Src-homology 3 domain, a common domain structure that binds to a proline-rich region.

Bibliography

1 Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737-741, doi:10.1038/nature02046 (2003).

2 Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods* **11**, 319-324, doi:10.1038/nmeth.2834 (2014).

3 Veitia, R. A., Bottani, S. & Birchler, J. A. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends in genetics : TIG* **24**, 390-397, doi:10.1016/j.tig.2008.05.005 (2008).

4 Veitia, R. A. & Potier, M. C. Gene dosage imbalances: action, reaction, and models. *Trends Biochem Sci* **40**, 309-317, doi:10.1016/j.tibs.2015.03.011 (2015).

5 Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci U S A* **109**, 14746-14753, doi:10.1073/pnas.1207726109 (2012).

6 Satina, S., Blakeslee, A. F. & Avery, A. G. Balanced and unbalanced haploids in Datura. *J. Hered.* **28**, 192-202 (1937).

7 Smith, A. M., Xu, W., Sun, Y., Faeder, J. R. & Marai, G. E. RuleBender: integrated modeling, simulation and visualization for rule-based intracellular biochemistry. *BMC Bioinformatics* **13 Suppl 8**, S3, doi:10.1186/1471-2105-13-S8-S3 (2012).

8 Sopko, R. *et al.* Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* **21**, 319-330, doi:10.1016/j.molcel.2005.12.011 (2006).

9 Makanae, K., Kintaka, R., Makino, T., Kitano, H. & Moriya, H. Identification of dosage-sensitive genes in Saccharomyces cerevisiae using the genetic tug-of-war method. *Genome Res* **23**, 300-311, doi:10.1101/gr.146662.112 (2013).

10 Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712-723, doi:10.1016/j.cell.2015.09.053 (2015).

11 Zhou, J., Lemos, B., Dopman, E. B. & Hartl, D. L. Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster. *Genome Biol Evol* **3**, 1014-1024, doi:10.1093/gbe/evr023 (2011).

12 Deng, X. *et al.* Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster. *Nat Genet* **43**, 1179-1185, doi:10.1038/ng.948 (2011).

13 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451-481, doi:10.1146/annurev.genom.9.081307.164217 (2009).

14 McElroy, J. P. *et al.* Copy number variation in pediatric multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)* **19**, 1014-1021, doi:10.1177/1352458512469696 (2013).

15 Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* **64**, 3060-3071 (2004).

16 Oberdorf, R. & Kortemme, T. Complex topology rather than complex membership is a determinant of protein dosage sensitivity. *Molecular systems biology* **5**, 253, doi:10.1038/msb.2009.9 (2009).

17 Levchenko, A., Bruck, J. & Sternberg, P. W. Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. *Proc Natl Acad Sci U S A* **97**, 5818-5823 (2000).

18 Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533-538, doi:10.1126/science.1188308 (2010).

19 Georgiev, P., Chlamydas, S. & Akhtar, A. Drosophila dosage compensation: males are from Mars, females are from Venus. *Fly (Austin)* **5**, 147-154 (2011).

20 Goldberg, A. L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**, 895-899, doi:10.1038/nature02263 (2003).

21 Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194-197, doi:10.1038/nature01771 (2003).

22 Yang, J., Lusk, R. & Li, W.-H. Organismal complexity, protein complexity, and gene duplicability. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15661-15665, doi:10.1073/pnas.2536672100 (2003).

23 Veitia, R. A. On gene dosage balance in protein complexes: a comment on Semple JI, Vavouri T, Lehner B. A simple principle concerning the robustness of protein complex activity to changes in gene expression. *BMC Syst Biol* **3**, 16, doi:10.1186/1752-0509-3-16 (2009).

24 Semple, J. I., Vavouri, T. & Lehner, B. A simple principle concerning the robustness of protein complex activity to changes in gene expression. *BMC Syst Biol* **2**, 1, doi:10.1186/1752-0509-2-1 (2008).

25 Moriya, H. Quantitative nature of overexpression experiments. *Mol Biol Cell* **26**, 3932-3939, doi:10.1091/mbc.E15-07-0512 (2015).

26 Kiel, C., Verschueren, E., Yang, J.-S. & Serrano, L. Integration of protein abundance and structure data reveals competition in the ErbB signaling network. *Science signaling* **6**, ra109, doi:10.1126/scisignal.2004560 (2013).

27 Tomala, K. & Korona, R. Evaluating the fitness cost of protein expression in Saccharomyces cerevisiae. *Genome Biol Evol* **5**, 2051-2060, doi:10.1093/gbe/evt154 (2013).

28 Vavouri, T., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198-208, doi:10.1016/j.cell.2009.04.029 (2009).

29 Davey, N. E. *et al.* Attributes of short linear motifs. *Mol Biosyst* **8**, 268-281, doi:10.1039/c1mb05231d (2012).

30 Gsponer, J., Futschik, M. E., Teichmann, S. A. & Babu, M. M. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* **322**, 1365-1368, doi:10.1126/science.1163581 (2008).

31 Matalon, O., Horovitz, A. & Levy, E. D. Different subunits belonging to the same protein complex often exhibit discordant expression levels and evolutionary properties. *Curr Opin Struct Biol* **26**, 113-120, doi:10.1016/j.sbi.2014.06.001 (2014).

32 Murugan, A., Zou, J. & Brenner, M. P. Undesired usage and the robust self-assembly of heterogeneous structures. *Nat Commun* **6**, 6203, doi:10.1038/ncomms7203 (2015).

33 Olzscha, H. *et al.* Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell* **144**, 67-78, doi:10.1016/j.cell.2010.11.050 (2011).

34 Stefani, M. & Dobson, C. M. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J Mol Med (Berl)* **81**, 678-699, doi:10.1007/s00109-003-0464-5 (2003).

35 Ciryam, P., Tartaglia, G. G., Morimoto, R. I., Dobson, C. M. & Vendruscolo, M. Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. *Cell reports* **5**, 781-790, doi:10.1016/j.celrep.2013.09.043 (2013).

36 Zhang, J., Maslov, S. & Shakhnovich, E. I. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Molecular systems biology* **4**, 210, doi:10.1038/msb.2008.48 (2008).

37 Schreiber, G. & Keating, A. E. Protein binding specificity versus promiscuity. *Curr Opin Struct Biol* **21**, 50-61, doi:10.1016/j.sbi.2010.10.002 (2011).

38 Li, Y., Zhang, X. & Cao, D. The role of shape complementarity in the protein-protein interactions. *Sci Rep* **3**, 3271, doi:10.1038/srep03271 (2013).

39 Johnson, M. E. & Hummer, G. Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 603-608, doi:10.1073/pnas.1010954108 (2011).

40 Zarrinpar, A., Park, S.-H. & Lim, W. A. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676-680, doi:10.1038/nature02178 (2003).

41 Ellis, R. J. & Minton, A. P. Cell biology: join the crowd. *Nature* **425**, 27-28, doi:10.1038/425027a (2003).

42 Ellis, R. J. Macromolecular crowding: obvious but underappreciated. *Trends Biochem Sci* **26**, 597-604 (2001).

43 Zhou, H. X., Rivas, G. & Minton, A. P. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu Rev Biophys* **37**, 375-397, doi:10.1146/annurev.biophys.37.032807.125817 (2008).

44 Munishkina, L. A., Cooper, E. M., Uversky, V. N. & Fink, A. L. The effect of macromolecular crowding on protein aggregation and amyloid fibril formation. *J Mol Recognit* **17**, 456-464, doi:10.1002/jmr.699 (2004).

45 Coquel, A. S. *et al.* Localization of protein aggregation in Escherichia coli is governed by diffusion and nucleoid macromolecular crowding effect. *PLoS Comput Biol* **9**, e1003038, doi:10.1371/journal.pcbi.1003038 (2013).

46 Eisen, H. N. & Chakraborty, A. K. Evolving concepts of specificity in immune reactions. *Proc Natl Acad Sci U S A* **107**, 22373-22380, doi:10.1073/pnas.1012051108 (2010).

47 DeLano, W. L., Ultsch, M. H., de Vos, A. M. & Wells, J. A. Convergent solutions to binding at a protein-protein interface. *Science* **287**, 1279-1283 (2000).

48 Tzeng, S. R. & Kalodimos, C. G. Protein dynamics and allostery: an NMR view. *Curr Opin Struct Biol* **21**, 62-67, doi:10.1016/j.sbi.2010.10.007 (2011).

49 Gsponer, J. & Babu, M. M. Cellular strategies for regulating functional and nonfunctional protein aggregation. *Cell Rep* **2**, 1425-1437, doi:10.1016/j.celrep.2012.09.036 (2012).

50 Levy, E. D., De, S. & Teichmann, S. A. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 20461-20466, doi:10.1073/pnas.1209312109 (2012).

51 Heo, M., Maslov, S. & Shakhnovich, E. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc Natl Acad Sci U S A* **108**, 4258-4263, doi:10.1073/pnas.1009392108 (2011).

52 Rocha, E. P. & Danchin, A. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* **21**, 108-116, doi:10.1093/molbev/msh004 (2004).

53 Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927-931 (2001).

54 Yang, J. R., Liao, B. Y., Zhuang, S. M. & Zhang, J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci U S A* **109**, E831-840, doi:10.1073/pnas.1117408109 (2012).

55 Levy, E. D., Landry, C. R. & Michnick, S. W. How perfect can protein interactomes be? *Sci Signal* **2**, pe11, doi:10.1126/scisignal.260pe11 (2009).

56 Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics* **5**, 101-113, doi:10.1038/nrg1272 (2004).

57 Barabasi, A. L. Scale-free networks: a decade and beyond. *Science* **325**, 412-413, doi:10.1126/science.1173299 (2009).

58 Alon, U. *An introduction to systems biology : design principles of biological circuits*. (Chapman & Hall/CRC, 2007).

59 Yeger-Lotem, E. *et al.* Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A* **101**, 5934-5939, doi:10.1073/pnas.0306752101 (2004).

60 Ryall, K. A. *et al.* Network reconstruction and systems analysis of cardiac myocyte hypertrophy signaling. *The Journal of biological chemistry* **287**, 42259-42268, doi:10.1074/jbc.M112.382937 (2012).

61 Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* **100**, 12123-12128, doi:10.1073/pnas.2032324100 (2003).

62 Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938-1941, doi:10.1126/science.1136174 (2006).

63 Ryall, K. A. & Tan, A. C. Systems biology approaches for advancing the discovery of effective drug combinations. *J Cheminform* **7**, 7, doi:10.1186/s13321-015-0055-9 (2015).

64 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12**, 56-68, doi:10.1038/nrg2918 (2011).

65 Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**, 64-68, doi:10.1038/ng881 (2002).

66 Johnson, M. E. & Hummer, G. Interface-resolved network of protein-protein interactions. *PLoS computational biology* **9**, e1003065, doi:10.1371/journal.pcbi.1003065 (2013).

67 Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology* **30**, 159-164, doi:10.1038/nbt.2106 (2012).

68 Watkins, A. M. & Arora, P. S. Structure-based inhibition of protein-protein interactions. *Eur J Med Chem* **94**, 480-488, doi:10.1016/j.ejmech.2014.09.047 (2015).

69 Kumar, A., Butler, B. M., Kumar, S. & Ozkan, S. B. Integration of structural dynamics and molecular evolution via protein interaction networks: a new

era in genomic medicine. *Curr Opin Struct Biol* **35**, 135-142, doi:10.1016/j.sbi.2015.11.002 (2015).

70 Deeds, E. J., Krivine, J., Feret, J., Danos, V. & Fontana, W. Combinatorial complexity and compositional drift in protein interaction networks. *PLoS One* **7**, e32032, doi:10.1371/journal.pone.0032032 (2012).

71 Johnson, M. E. & Hummer, G. Evolutionary pressure on the topology of protein interface interaction networks. *The journal of physical chemistry. B* **117**, 13098-13106, doi:10.1021/jp402944e (2013).

72 Holland, D. O., Shapiro, B. H., Xue, P. & Johnson, M. E. Protein-protein binding selectivity and network topology constrain global and local properties of interface binding networks. *Sci Rep* **7**, 5631, doi:10.1038/s41598-017-05686-2 (2017).

73 Mosca, R., Ceol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat Methods* **10**, 47-53, doi:10.1038/nmeth.2289 (2013).

74 Acuner Ozbabacan, S. E., Gursoy, A., Nussinov, R. & Keskin, O. The structural pathway of interleukin 1 (IL-1) initiated signaling reveals mechanisms of oncogenic mutations and SNPs in inflammation and cancer. *PLoS Comput Biol* **10**, e1003470, doi:10.1371/journal.pcbi.1003470 (2014).

75 Weatheritt, R. J., Luck, K., Petsalaki, E., Davey, N. E. & Gibson, T. J. The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics* **28**, 976-982, doi:10.1093/bioinformatics/bts072 (2012).

76 Bjorkholm, P. & Sonnhammer, E. L. Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics* **25**, 3020-3025, doi:10.1093/bioinformatics/btp522 (2009).

77 Chylek, L. A. *et al.* Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems. *Wiley Interdiscip Rev Syst Biol Med* **6**, 13-36, doi:10.1002/wsbm.1245 (2014).

78 Park, J., Lee, D. S., Christakis, N. A. & Barabasi, A. L. The impact of cellular networks on disease comorbidity. *Mol Syst Biol* **5**, 262, doi:10.1038/msb.2009.16 (2009).

79 Ozbabacan, S. E. A., Gursoy, A., Nussinov, R. & Keskin, O. The Structural Pathway of Interleukin 1 (IL-1) Initiated Signaling Reveals Mechanisms of Oncogenic Mutations and SNPs in Inflammation and Cancer. *Plos Comput Biol* **10**, doi:10.1371/journal.pcbi.1003470 (2014).

80 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).

81 Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Protein Sci* **86**, 2 9 1-2 9 37, doi:10.1002/cpps.20 (2016).

82 Szilagyi, A. & Zhang, Y. Template-based structure modeling of protein-protein interactions. *Curr Opin Struct Biol* **24**, 10-23, doi:10.1016/j.sbi.2013.11.005 (2014).

83 Stein, A., Mosca, R. & Aloy, P. Three-dimensional modeling of protein interactions and complexes is going 'omics. *Curr Opin Struct Biol* **21**, 200-208, doi:10.1016/j.sbi.2011.01.005 (2011).

84 Dinkel, H. *et al.* ELM 2016-data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res* **44**, D294-D300, doi:10.1093/nar/gkv1291 (2016).

85 van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589-6631, doi:10.1021/cr400525m (2014).

86 Johnson, M. E. & Hummer, G. Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 603-608, doi:10.1073/pnas.1010954108 (2011).

87 Heo, M. Y., Maslov, S. & Shakhnovich, E. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 4258-4263, doi:10.1073/pnas.1009392108 (2011).

88 Yang, J.-R., Liao, B.-Y., Zhuang, S.-M. & Zhang, J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E831-840, doi:10.1073/pnas.1117408109 (2012).

89 Gastner, M. T. & Newman, M. E. Optimal design of spatial distribution networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **74**, 016117, doi:10.1103/PhysRevE.74.016117 (2006).

90 Levy, E. D. & Pereira-Leal, J. B. Evolution and dynamics of protein interactions and networks. *Curr Opin Struc Biol* **18**, 349-357, doi:10.1016/j.sbi.2008.03.003 (2008).

91 Yook, S. H., Oltvai, Z. N. & Barabasi, A. L. Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928-942, doi:10.1002/pmic.200300636 (2004).

92 Scott McShan, R. W. The implications of hub-and-spoke routing for airline costs and competitiveness. *Logistics and Transportation Review* **25**, 209-230 (1989).

93 Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378-382, doi:Doi 10.1038/35019019 (2000).

94 Orengo, C. A. & Thornton, J. M. Protein families and their evolution-a structural perspective. *Annu Rev Biochem* **74**, 867-900, doi:10.1146/annurev.biochem.74.082803.133029 (2005).

95 Beltrao, P. & Serrano, L. Specificity and evolvability in eukaryotic protein interaction networks. *Plos Comput Biol* **3**, 258-267, doi:ARTN e25 10.1371/journal.pcbi.0030025 (2007).

96 Meyer, M. J., Das, J., Wang, X. J. & Yu, H. Y. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* **29**, 1577-1579, doi:10.1093/bioinformatics/btt181 (2013).

97 Kaneko, T., Li, L. & Li, S. S. C. The SH3 domain - a family of versatile peptide- and protein-recognition module. *Front Biosci* **13**, 4938-4952, doi:10.2741/3053 (2008).

98 Rushworth, L. K., Hindley, A. D., O'Neill, E. & Kolch, W. Regulation and role of Raf-1/B-Raf heterodimerization. *Mol Cell Biol* **26**, 2262-2272, doi:10.1128/MCB.26.6.2262-2272.2006 (2006).

99 Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432, doi:Doi 10.1093/Bioinformatics/Btq675 (2011).

100 Xin, X. F. *et al.* SH3 interactome conserves general function over specific form. *Molecular systems biology* **9**, doi:ARTN 652

10.1038/msb.2013.9 (2013).

101 Weinberg, J. & Drubin, D. G. Clathrin-mediated endocytosis in budding yeast. *Trends Cell Biol* **22**, 1-13, doi:10.1016/j.tcb.2011.09.001 (2012).

102 Schmid, E. M. *et al.* Role of the AP2 beta-appendage hub in recruiting partners for clathrin-coated vesicle assembly. *Plos Biol* **4**, 1532-1548, doi:ARTN e262

10.1371/journal.pbio.0040262 (2006).

103 Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. Evolutionary rate in the protein interaction network. *Science* **296**, 750-752, doi:DOI 10.1126/science.1068696 (2002).

104 Manna, B., Bhattacharya, T., Kahali, B. & Ghosh, T. C. Evolutionary constraints on hub and non-hub proteins in human protein interaction network: Insight

from protein connectivity and intrinsic disorder. *Gene* **434**, 50-55, doi:10.1016/j.gene.2008.12.013 (2009).

105 Haynes, C. *et al.* Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *Plos Comput Biol* **2**, 890-901, doi:ARTN e100 10.1371/journal.pcbi.0020100 (2006).

106 Brown, C. J. *et al.* Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* **55**, 104-110, doi:10.1007/s00239-001-2309-6 (2002).

107 Batada, N. N. *et al.* Stratus not altocumulus: A new view of the yeast protein interaction network. *Plos Biol* **4**, 1720-1731, doi:ARTN e317 10.1371/journal.pbio.0040317 (2006).

108 Batada, N. N. *et al.* Still stratus not altocumulus: Further evidence against the date/party hub distinction. *Plos Biol* **5**, 1202-1206, doi:ARTN e154 10.1371/journal.pbio.0050154 (2007).

109 Drummond, D. A. & Wilke, C. O. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* **10**, 715-724, doi:10.1038/nrg2662 (2009).

110 Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Reviews Genetics* **7**, 337-348, doi:10.1038/nrg1838 (2006).

111 Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A. Modeling of Protein Interaction Networks. *Complexus* **1**, 38-44 (2003).

112 Eisenberg, E. & Levanon, E. Y. Preferential attachment in the protein network evolution. *Physical Review Letters* **91**, doi:ARTN 138701 10.1103/PhysRevLett.91.138701 (2003).

113 Kim, W. K. & Marcotte, E. M. Age-Dependent Evolution of the Yeast Protein Interaction Network Suggests a Limited Role of Gene Duplication and Divergence. *Plos Comput Biol* **4**, doi:ARTN e1000232 10.1371/journal.pcbi.1000232 (2008).

114 Wagner, A. How the global structure of protein interaction networks evolves. *P Roy Soc B-Biol Sci* **270**, 457-466, doi:10.1098/rspb.2002.2269 (2003).

115 Middendorf, M., Ziv, E. & Wiggins, C. H. Inferring network mechanisms: The Drosophila melanogaster protein interaction network. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3192-3197, doi:10.1073/pnas.0409515102 (2005).

116 Navlakha, S. & Kingsford, C. Network archaeology: uncovering ancient networks from present-day interactions. *PLoS Comput Biol* **7**, e1001119, doi:10.1371/journal.pcbi.1001119 (2011).

117 Cotton, J. A. & Page, R. D. Rates and patterns of gene duplication and loss in the human genome. *Proc Biol Sci* **272**, 277-283, doi:10.1098/rspb.2004.2969 (2005).

118 Pan, D. & Zhang, L. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. *Genome Biol* **8**, R158, doi:10.1186/gb-2007-8-8-r158 (2007).

119 Katju, V. & Bergthorsson, U. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet* **4**, 273, doi:10.3389/fgene.2013.00273 (2013).

120 Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D complex: A structural classification of protein complexes. *Plos Comput Biol* **2**, 1395-1406, doi:ARTN e155
10.1371/journal.pcbi.0020155 (2006).

121 Ispolatov, I., Yuryev, A., Mazo, I. & Maslov, S. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res* **33**, 3629-3635, doi:10.1093/nar/gki678 (2005).

122 Matthews, L. R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* **11**, 2120-2126, doi:DOI 10.1101/gr.205301 (2001).

123 Madan Babu, M., Teichmann, S. A. & Aravind, L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* **358**, 614-633, doi:10.1016/j.jmb.2006.02.019 (2006).

124 Schmid, E. M. & McMahon, H. T. Integrating molecular and network biology to decode endocytosis. *Nature* **448**, 883-888 (2007).

125 Ti, S. C., Jurgenson, C. T., Nolen, B. J. & Pollard, T. D. Structural and biochemical characterization of two binding sites for nucleation-promoting factor WASp-VCA on Arp2/3 complex. *Proc Natl Acad Sci U S A* **108**, E463-471, doi:10.1073/pnas.1100125108 (2011).

126 Caldarelli, G., Pastor-Satorras, R. & Vespignani, A. Structure of cycles and local ordering in complex networks. *Eur Phys J B* **38**, 183-186 (2004).

127 Goh, K. I., Kahng, B. & Kim, D. Universal behavior of load distribution in scale-free networks. *Phys Rev Lett* **87**, 278701, doi:10.1103/PhysRevLett.87.278701 (2001).

128 Newman, M. E., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys* **64**, 026118, doi:10.1103/PhysRevE.64.026118 (2001).

129 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066 (2002).

130 Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344-350, doi:10.1093/nar/gkw408 (2016).

131 Camacho, C. *et al.* BLAST+: architecture and applications. *Bmc Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

132 Chica, C., Labarga, A., Gould, C. M., Lopez, R. & Gibson, T. J. A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *Bmc Bioinformatics* **9**, doi:Artn 229

10.1186/1471-2105-9-229 (2008).

133 Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**, D286-D293, doi:10.1093/nar/gkv1248 (2016).

134 Pessia, E., Makino, T., Bailly-Bechet, M., McLysaght, A. & Marais, G. A. Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proc Natl Acad Sci U S A* **109**, 5346-5351, doi:10.1073/pnas.1116763109 (2012).

135 Reichard, P. Ribonucleotide reductases: substrate specificity by allostery. *Biochem Biophys Res Commun* **396**, 19-23, doi:10.1016/j.bbrc.2010.02.108 (2010).

136 Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183-1186, doi:10.1126/science.1070919 (2002).

137 Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat Genet* **38**, 636-643, doi:10.1038/ng1807 (2006).

138 Chen, Q. & Pollard, T. D. Actin filament severing by cofilin dismantles actin patches and produces mother filaments for new patches. *Curr Biol* **23**, 1154-1162, doi:10.1016/j.cub.2013.05.005 (2013).

139 Bravo-Cordero, J. J., Magalhaes, M. A., Eddy, R. J., Hodgson, L. & Condeelis, J. Functions of cofilin in cell locomotion and invasion. *Nat Rev Mol Cell Biol* **14**, 405-415, doi:10.1038/nrm3609 (2013).

140 Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **81**, 2340-2361, doi:10.1021/j100540a008 (1977).

141 Newman, J. R. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840-846, doi:10.1038/nature04785 (2006).

142 Chong, Y. T. *et al.* Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell* **161**, 1413-1424, doi:10.1016/j.cell.2015.04.051 (2015).

143 Crivat, G. & Taraska, J. W. Imaging proteins inside cells with fluorescent tags. *Trends Biotechnol* **30**, 8-16, doi:10.1016/j.tibtech.2011.08.002 (2012).

144 Landry, J. J. *et al.* The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda)* **3**, 1213-1224, doi:10.1534/g3.113.005777 (2013).

145 Gertz, E. M. & Wright, S. J. Object-oriented software for quadratic programming. *ACM Transactions on Mathematical Software* **29**, 58-81 (2003).

146 Wisniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics* **13**, 3497-3506, doi:10.1074/mcp.M113.037309 (2014).

147 Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549, doi:10.1038/msb.2011.82 (2011).

148 Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* **7**, 548, doi:10.1038/msb.2011.81 (2011).

149 Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. *Ieee T Inform Theory* **49**, 1858-1860, doi:10.1109/Tit.2003.813506 (2003).

150 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

151 Sneddon, M. W., Faeder, J. R. & Emonet, T. Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. *Nat Methods* **8**, 177-183, doi:10.1038/nmeth.1546 (2011).

152 Motley, A., Bright, N. A., Seaman, M. N. & Robinson, M. S. Clathrin-mediated endocytosis in AP-2-depleted cells. *The Journal of cell biology* **162**, 909-918, doi:10.1083/jcb.200305145 (2003).

153 Jost, M., Simpson, F., Kavran, J. M., Lemmon, M. A. & Schmid, S. L. Phosphatidylinositol-4,5-bisphosphate is required for endocytic coated vesicle formation. *Curr Biol* **8**, 1399-1402 (1998).

154 Dannhauser, P. N. & Ungewickell, E. J. Reconstitution of clathrin-coated bud and vesicle formation with minimal components. *Nat Cell Biol* **14**, 634-639, doi:10.1038/ncb2478 (2012).

155 McMahon, H. T. & Boucrot, E. Molecular mechanism and physiological functions of clathrin-mediated endocytosis. *Nat Rev Mol Cell Biol* **12**, 517-533, doi:10.1038/nrm3151 (2011).

156 Mishra, S. K. *et al.* Disabled-2 exhibits the properties of a cargo-selective endocytic clathrin adaptor. *The EMBO journal* **21**, 4915-4926 (2002).

157 Kelly, B. T. *et al.* Clathrin adaptors. AP2 controls clathrin polymerization with a membrane-activated switch. *Science* **345**, 459-463, doi:10.1126/science.1254836 (2014).

158 Yogurtcu, O. N. & Johnson, M. E. Cytosolic proteins can exploit membrane localization to trigger functional assembly. *bioRxiv* (2017).

159 Alberts, B. *Molecular biology of the cell*. Sixth edition. edn, (Garland Science, Taylor and Francis Group, 2015).

160 Jorgensen, P., Nishikawa, J. L., Breitkreutz, B. J. & Tyers, M. Systematic identification of pathways that couple cell growth and division in yeast. *Science* **297**, 395-400, doi:10.1126/science.1070850 (2002).

161 Wakeham, D. E., Chen, C. Y., Greene, B., Hwang, P. K. & Brodsky, F. M. Clathrin self-assembly involves coordinated weak interactions favorable for cellular regulation. *EMBO J* **22**, 4980-4990, doi:10.1093/emboj/cdg511 (2003).

162 Miele, A. E., Watson, P. J., Evans, P. R., Traub, L. M. & Owen, D. J. Two distinct interaction motifs in amphiphysin bind two independent sites on the clathrin terminal domain beta-propeller. *Nat Struct Mol Biol* **11**, 242-248, doi:10.1038/nsmb736 (2004).

163 Zhuo, Y. *et al.* Dynamic interactions between clathrin and locally structured elements in a disordered protein mediate clathrin lattice assembly. *J Mol Biol* **404**, 274-290, doi:10.1016/j.jmb.2010.09.044 (2010).

164 de Beer, T. *et al.* Molecular mechanism of NPF recognition by EH domains. *Nat Struct Biol* **7**, 1018-1022, doi:10.1038/80924 (2000).

165 Morgan, J. R., Prasad, K., Jin, S., Augustine, G. J. & Lafer, E. M. Eps15 homology domain-NPF motif interactions regulate clathrin coat assembly during synaptic vesicle recycling. *J Biol Chem* **278**, 33583-33592, doi:10.1074/jbc.M304346200 (2003).

166 Boeke, D. *et al.* Quantification of cytosolic interactions identifies Ede1 oligomers as key organizers of endocytosis. *Mol Syst Biol* **10**, 756, doi:10.15252/msb.20145422 (2014).

167 Winkler, F. K. & Stanley, K. K. Clathrin heavy chain, light chain interactions. *EMBO J* **2**, 1393-1400 (1983).

168 Engqvist-Goldstein, A. E. *et al.* The actin-binding protein Hip1R associates with clathrin during early stages of endocytosis and promotes clathrin assembly in vitro. *J Cell Biol* **154**, 1209-1223, doi:10.1083/jcb.200106089 (2001).

169 Wilbur, J. D. *et al.* Actin binding by Hip1 (huntingtin-interacting protein 1) and Hip1R (Hip1-related protein) is regulated by clathrin light chain. *J Biol Chem* **283**, 32870-32879, doi:10.1074/jbc.M802863200 (2008).

170 Henne, W. M. *et al.* Structure and analysis of FCHo2 F-BAR domain: a dimerizing and membrane recruitment module that effects membrane curvature. *Structure* **15**, 839-852, doi:10.1016/j.str.2007.05.002 (2007).

171 Stahelin, R. V. *et al.* Contrasting membrane interaction mechanisms of AP180 N-terminal homology (ANTH) and epsin N-terminal homology (ENTH) domains. *J Biol Chem* **278**, 28993-28999, doi:10.1074/jbc.M302865200 (2003).

172 Moravcevic, K. *et al.* Comparison of Saccharomyces cerevisiae F-BAR domain structures reveals a conserved inositol phosphate binding site. *Structure* **23**, 352-363, doi:10.1016/j.str.2014.12.009 (2015).

173 Yoon, Y., Lee, P. J., Kurilova, S. & Cho, W. In situ quantitative imaging of cellular lipids using molecular sensors. *Nat Chem* **3**, 868-874, doi:10.1038/nchem.1163 (2011).

174 Wu, Y., Vendome, J., Shapiro, L., Ben-Shaul, A. & Honig, B. Transforming binding affinities from three dimensions to two with application to cadherin clustering. *Nature* **475**, 510-513, doi:10.1038/nature10183 (2011).

175 Loerke, D. *et al.* Cargo and dynamin regulate clathrin-coated pit maturation. *PLoS Biol* **7**, e57, doi:10.1371/journal.pbio.1000057 (2009).

176 Weinberg, J. & Drubin, D. G. Clathrin-mediated endocytosis in budding yeast. *Trends Cell Biol* **22**, 1-13, doi:10.1016/j.tcb.2011.09.001 (2012).

177 Boettner, D. R., Chi, R. J. & Lemmon, S. K. Lessons from yeast for clathrin-mediated endocytosis. *Nat Cell Biol* **14**, 2-10, doi:10.1038/ncb2403 (2011).

178 Lu, R., Drubin, D. G. & Sun, Y. Clathrin-mediated endocytosis in budding yeast at a glance. *J Cell Sci* **129**, 1531-1536, doi:10.1242/jcs.182303 (2016).

179 Busch, D. J. *et al.* Intrinsically disordered proteins drive membrane curvature. *Nat Commun* **6**, 7875, doi:10.1038/ncomms8875 (2015).

180 Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712, doi:10.1038/nature08516 (2010).

181 Dafforn, T. R. & Smith, C. J. Natively unfolded domains in endocytosis: hooks, lines and linkers. *EMBO Rep* **5**, 1046-1052, doi:10.1038/sj.embor.7400276 (2004).

182 Hu, J., Lipowsky, R. & Weikl, T. R. Binding constants of membrane-anchored receptors and ligands depend strongly on the nanoscale roughness of

membranes. *Proc Natl Acad Sci U S A* **110**, 15283-15288, doi:10.1073/pnas.1305766110 (2013).

183 Weikl, T. R., Hu, J., Xu, G. K. & Lipowsky, R. Binding equilibrium and kinetics of membrane-anchored receptors and ligands in cell adhesion: Insights from computational model systems and theory. *Cell Adh Migr* **10**, 576-589, doi:10.1080/19336918.2016.1180487 (2016).

Biography

David O. Holland was born and raised in Northern Virginia. He completed his B.S. in Biomedical Engineering with a minor in Applied Mathematics at the University Virginia, working with Dr. Jeffrey Saucerman to model the β2 adrenergic pathway in heart myocytes. He began his PhD in Biomedical Engineering at the Johns Hopkins University in September of 2011, and joined Dr. Margaret Johnson's lab in late 2013. At Hopkins he designed and taught two courses: an introduction to systems biology and an introduction to network science. He has volunteered with the School of Medicine Science Outreach Program (SOP), which helps introduce science to elementary-level school children.