Structure-informed theoretical modeling defines principles governing avidity in bivalent protein interactions

Reagan Portelance[1], Anqi Wu[1], Alekhya Kandoor[1], and Kristen M. Naegle[1]✉

[1]Department of Biomedical Engineering and the Department for Genome Sciences, University of Virginia, Charlottesville, Virginia, United States of America.

In signaling cascades, where domain-motif interactions tend to interact with relatively low affinity (allowing for reversibility), signaling proteins often encode multiple domains or motifs. This presents the possibility for avidity where multivalent binding drastically increases the interaction strength and duration. However, given the large combinatorial space, predicting and validating multivalent interactions that interact with avidity is a challenge. Here, we integrate mechanistic modeling, structure-based analysis, and experimental approaches as a framework for defining the conditions under which avidity plays a role. We explore the tandem SH2 domain family of interactions with bisphosphorylated partners as a multivalent archetype, which encompasses key secondary messengers in tyrosine kinase signaling networks. While certain multivalent interactions have been shown to be necessary in immune receptor recruitment of partners, bivalent recruitment of tandem SH2 domains more broadly is poorly understood. Theoretical modeling suggests that maximum avidity occurs with closely spaced or flexibly linked phosphotyrosine sites, combined with moderate monovalent affinities – exactly around the innate range of SH2 domain affinity. Surprisingly, despite sequence diversity, structure-based analysis showed relatively conserved three-dimensional spacing between SH2 domains across all tandem SH2 families, which we corroborate experimentally, suggesting evolutionary optimization for avidity interactions. The combination of structure-based analysis of domain spacing with available monovalent experimental data appears to be sufficiently accurate to predict and rank order high affinity interactions of tandem SH2 domain recruitment to the EGFR C-terminal tail. Using these principles we extended bivalent predictions into the full phosphoproteome space and structural parameterization of other partners of SH2 domain binding, providing resources and methods for more rapid expansion of bivalent analysis. These approaches lay the groundwork for larger utility in multivalent prediction and testing to help better understand protein interactions that drive cell signaling.

SH2 domains | protein interactions | computational biology | phosphotyrosine | avidity | bivalent | biolayer interferometry | BLI

Correspondence: ***kmn4mj@virginia.edu***

Introduction

Avidity is the property that defines the increase in effective binding strength observed in interactions that involve multiple monovalent components. This occurs due to biophysical constraints where the initial interaction locks the unbound interface into a region of space constrained by their shared linker, thus increasing its "effective concentration" and, as a result, its binding driving force (1). It is a well-documented phenomenon in antibody binding, where multiple antibody epitopes boost affinity (2), which scientists have exploited for antibody-drug conjugates to increase therapeutic impact (3, 4). Cell signaling also relies on multivalent interactions to drastically increase sensitivity and specificity in network connections as we can see in a few well-documented cases involving tandem SH2 proteins that can engage with two phosphotyrosine (pTyr) sites simultaneously (5, 6). The necessity of bivalent recruitment is well documented in immune receptor-based signaling, specifically through the patterning of two pTyr sequences, such as ZAP70 recruitment to CD3ζ chain of the T-cell receptor (7, 8), SYK recruitment to Fc receptors for immunoglobin E (FcϵRI), and PTPN6/PTPN11 (SHP1/SHP2) recruitment to the programmed cell death protein 1 (PD-1) and B- and T- lymphocyte attenuator (BTLA) (9, 10). In these cases it has been shown that protein-protein interactions are high affinity, thanks to the bivalent recruitment of the tandem SH2 domain proteins and that biophysical constraints of the linkers guide that recruitment (11).

In addition to tyrosine phosphatases PTPN6/PTPN11 and tyrosine kinases ZAP70/SYK, three additional families (PI3K, PLCγ, and RASA1) covering ten total human proteins also contain tandem SH2 domains. These are key proteins that transition signaling from receptor-proximal phosphotyrosine signaling to downstream lipid signaling (PI3K and PLCγ) and serine/threonine signaling (RASA1). Multisite phosphorylation is also more broadly observed on other receptor tyrosine kinases (RTKs) and on downstream targets such as GAB1, which gets robustly phosphorylated to act as a docking protein for multiple tandem SH2 proteins such as PTPN11, PI3K, and RASA1, allowing for simultaneous regulation of different signaling networks (12). Beyond tandem SH2 domain containing proteins, approximately half of the SH2 domain family of proteins contain two protein interaction modules (e.g. SH3-SH2 and PTB-SH2) (13) and multidomain protein interaction modules are found broadly across the proteome. Clearly, there is strong potential for multivalency in key signaling interactions across a wide range of signaling pathways, and we are likely missing critical information by ignoring the impact of avidity on these systems. However, we currently lack the ability to predict when and how avidity will shape multivalent interactions.

One challenge to broadly identifying multivalent avid-

ity interactions across the proteome is the complication of modeling multivalent interactions in a framework that captures the biophysics of the constrained interaction and the many species that are formed (such as an N-terminal SH2 domain bound to both pTyr sites and bivalent species bound in different orientations). Errington et al. developed an ordinary differential equation (ODE) model that accounts for the biophysical parameters of the linker constraint, using this to scale the driving force of the second interaction in the complex once the first has formed (14), which we adapted here for the purpose of exploring the theoretical regimes that define asymmetric binding.

An additional challenge in the field of tandem SH2 domain interactions is the experimental validation of a theoretical model, especially for interactions where the bisphosphorylated partners might be far apart and separated by a highly flexible linker – such as in receptor tails (e.g. EGFR). Peptide synthesis of phosphorylated proteins becomes both costly and complex, with limitations on total size and number of pTyr residues that can be incorporated (15). We developed a synthetic toolkit that can produce large amounts of multiply-phosphorylated protein, including sites across the entirety of the EGFR C-terminal tail in a single protein construct (16). Using both of these advances, computational and experimental, we set out to test the theoretical modeling results. Using structure-based analysis, we also explored how to inform the linker biophysics for more accurate models, finding that, despite high sequence length variability, tandem SH2 domains fall into two groups of different linker distances and despite the differences, they both can drive interactions with high avidity to similar pTyr partners. Based on the findings, we then sought to generalize identifying high affinity bivalent interactions involving pairs of pTyr sites and human tandem SH2 domain proteins using available quantitative and high-throughput data, which recovers known and new protein interactions. Given the importance of parameterizing biophysical properties of linkers and domain sizes, we also undertook a comprehensive evaluation of other partners involved in multivalency with SH2 domains in tyrosine kinase signaling, identifying other generalized conservation that may guide high affinity interactions that determine signaling. All together, the framework presented here, and the findings in the tandem SH2 domain family, offer an approach for tackling the high complexity of identifying and testing protein interactions with emergent properties that significantly differ from their monovalent parts.

Results

Establishing a computational model of tandem SH2 binding. Errington et al.'s ordinary differential equation (ODE) model simulates a surface plasmon resonance (SPR) experiment between a multivalent receptor-ligand pair, tracking all possible binding configurations over time – 15 of which exist in a bivalent system (14) (Fig. S1). Key parameters in the model include the monovalent k_{on} and k_{off} rate constants, species concentrations, and linker parameters. Linker parameters include the sequence length and protein flexibility – as determined by the persistence length (l_p) – which are used to calculate the "effective concentration" of an unbound species following initial binding of the first domain, thus altering the driving force of the subsequent interaction and producing avidity. We adapted the model to allow monovalent affinity parameterization of all possible SH2-pTyr interactions and to improve control over the persistence length parameter. We modeled dynamic behavior under different SH2 domain concentrations, extracting the fraction bound to both SH2 domains (in either configuration) at equilibrium to estimate the bivalent dissociation constant ($K_{D,B}$). We additionally simulated the estimation of monovalent domain interactions with the SH2 ligand (by setting the other domain affinity to 1M and effectively removing that interaction). We refer to these as $K_{D,N}$ and $K_{D,C}$ and calculate avidity as the inverse of the bivalent dissociation constant ($K_{A,B}$) divided by the sum of the inverse of the monovalent affinities ($K_{A,N}$ and $K_{A,C}$) (Equation 1):

$$Avidity = \frac{K_{A,B}}{K_{A,N} + K_{A,C}} \quad \textbf{(1)}$$

Simulations recapitulate avidity in PTPN11:GAB1 interaction. To test the ability of this modeling approach to be suitably parameterized from available data and to predict avidity, we selected an interesting cytosolic interaction that is known to occur bivalently and depend on avidity – PTPN11 binding to Y627/Y659 bisphosphorylated GAB1 (17, 18). The linker separating the PTPN11 SH2 domains is 9 amino acids long, and the GAB1 pTyr sites are separated by 31 amino acids. Since SH2 domains interact strongly with both the pTyr site and its six surrounding residues (-2 to +4), these are assumed to be encompassed in the pTyr binding motif, and therefore not included in the linker. Thus, the linker length of GAB1 is set at 25 amino acids, and all subsequent simulations similarly remove these six residues from the pTyr linker. Distance across the PTPN11 linker was measured using PDB structures as 17.3Å, allowing us to adjust the linker length to be 5.75 a.a. to better represent that distance given an estimate of roughly 3Å per amino acid. However, no PDB structures exist that span the GAB1 C-terminal tail, so the 25 amino acid linker was retained. The persistence length of the SH2 linker was set to 30Å as has been measured in interdomain spanning protein linkers (19). The pTyr linker was set to 4Å to match previously measured persistence lengths of intrinsically disordered proteins (20, 21) (Fig. 1A). Estimates of monovalent K_D values were taken from a deep re-analysis of fluorescence polarization experiments (22, 23) by Ronan et al. (24). The N-SH2 domain of PTPN11 is estimated to bind to site Y627 with a monovalent K_D of 287 nM whereas the C-SH2 domain used in the studies was shown to be nonfunctional making it impossible to draw a conclusion on binding affinity. We know that the C-SH2 domain is capable of binding to Y659 (17) so we modeled a monovalent K_D of 20 μM which is the minimal detectable binding affinity based on the experimental limits of the studies (22, 23). For simplicity, we assumed no binding of the C-SH2 domain to site Y627 to prevent interference with the N-terminal interaction. Rate

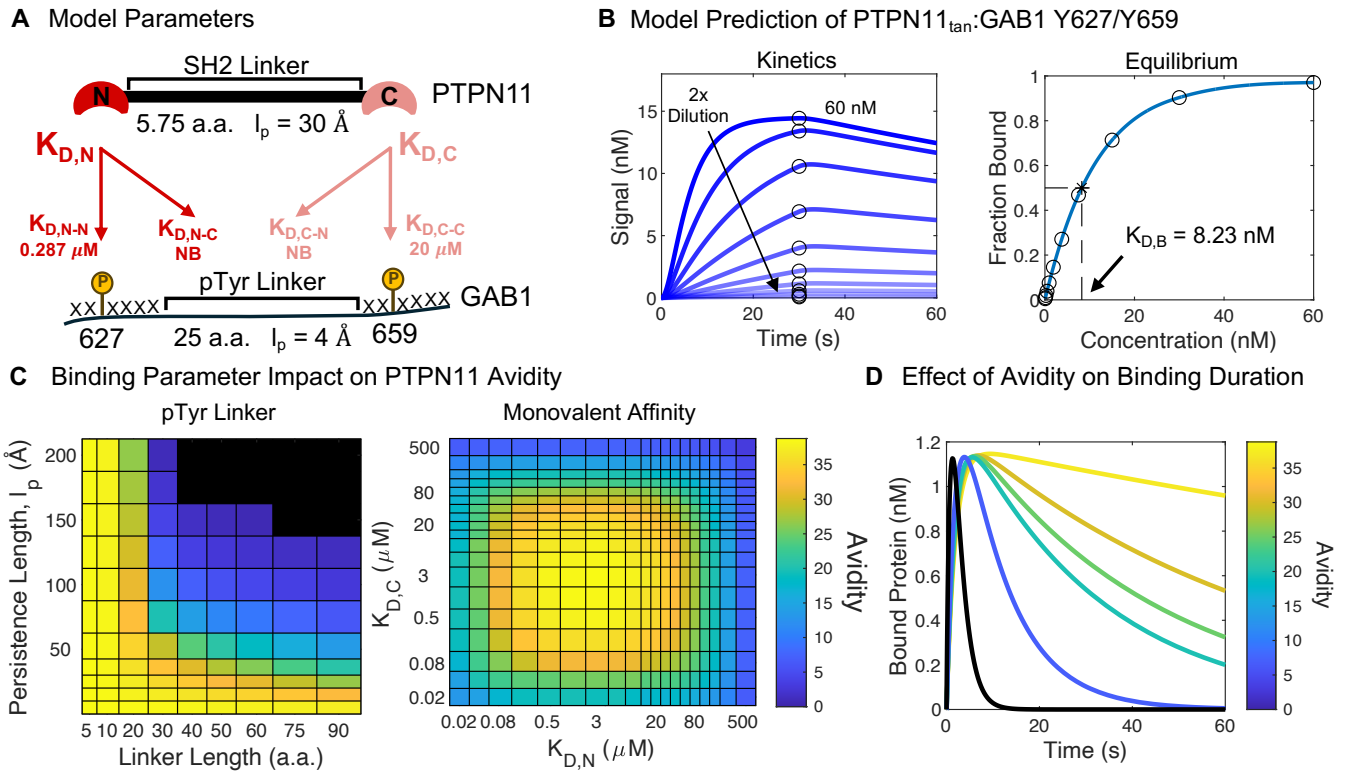

Fig. 1. Computational simulations to determine theoretical effects of interaction parameters on avidity. **A)** The generalized ODE model considers linker parameters for both partners and the monovalent affinities of all pairs. An interaction that is not expected to occur is listed as a non-binder (NB). This toy model is annotated with parameters specifically used to model tandem SH2 domain binding of PTPN11 to doubly phosphorylated Y627/Y659 of GAB1, based on literature and sequence values. **B)** A serial dilution simulation of surface plasmon resonance (SPR) experiments of $PTPN11_{tan}$ and GAB1 was performed using the ODE model. Equilibrium concentrations of bound protein were used to generate a concentration effect curve and estimate the $K_{D,B}$. **C)** We used the model to independently explore the impact of ligand linker biophysics and affinity of the monovalent species. Here, the PTPN11 tandem SH2 domain parameters were used, and we either fixed monovalent affinities with changing linker parameters, or linker parameters were fixed with changing monovalent affinities. The heat map color indicates the avidity measured under each condition with black indicating an inability to achieve bivalent binding. **D)** The kinetics of transient binding (interaction does not reach equilibrium) of the SH2 domain are plotted here under different model parameters that yield varying levels of avidity. The black curve showcases the behavior of a monovalent species that experiences no avidity.

constants were set with a constant k_{off} of 1 s^{-1} - approximating data found in PTPN11 and PIK3R1 SH2 binding studies (25, 26) - adjusting k_{on} for the desired K_D. Although specific rates would be important to dynamic models, here we are specifically interested in using equilibrium behavior to understand overall binding affinity between a tandem SH2 domain and a bisphosphorylated partner, so holding k_{off} constant ensures consistency and provides further confidence in comparisons made across proteins.

Simulations were performed with a series of eleven PTPN11 concentrations, and the resulting concentration effect curve provided a predicted $K_{D,B}$ of 8.23 nM and avidity of 35, meaning the presence of both SH2 domains makes the binding affinity of the interaction 35-fold stronger than the individual monomers (Fig. 1B). In short, the capacity for tandem SH2 domain interaction with GAB1 converts a moderate affinity interaction that could at best be 287 nM to a strong interaction of 8.23 nM, which has significant implications for the degree and duration of the PTPN11:GAB1 interaction and suggests that the modeling approach informed by available monovalent interaction data successfully recapitulates a known bivalent interaction property.

The role of linkers on shaping avidity. Having demonstrated our ability to correctly model avidity within the PTPN11-GAB1 interaction, we wanted to explore the impact of different parameters of the model. We first looked at different linker behaviors, since this parameter drives the biophysical constraints of the interaction. Keeping the model linker parameters for PTPN11 and fixing $K_{D,N}$ and $K_{D,C}$ at 4 μM, we simulated a series of receptor linker and persistence lengths while calculating avidity under each condition (Fig. 1C). We observed maximum avidity occurs in conditions where the pTyr linker is relatively short or when it is highly flexible. As the distance between the pTyr sites increases and the linker becomes more rigid, avidity decreases until bivalent binding is not possible. Across the theoretical range of linkers, it turns out that the maximum avidity attainable is closely approximated by that of the PTPN11:GAB1 interaction, suggesting that partnership is fairly optimized. It should be noted that the model cannot identify occlusion effects – the case where SH2 domains engaged with two pTyr would physically overlap for very short pTyr spacing. To account for this effect, we set a lower limit on the pTyr separation modeled, since below this separation it would be expected that the two pTyr interactions could not occur and avidity would worsen and move towards monovalent competitive interactions. For the case of SH2 engagement, we hypothesized this minimum spacing would be close to 12 amino acids separating the two pTyr sites (a 6

amino acid linker; Fig 1A), based on the diameter of the SH2 domain as measured across crystal structures, but in reality it would be slightly smaller due to the SH2 domains not being perfect spheres. In support of this, the smallest bisphosphorylated ligands where we could find evidence of tandem SH2 engagement are the 10 amino acid ITAM sequences in the TCR CD3 chains which engage with ZAP70 through avidity (7, 8). Therefore, we did not model any linker below a separation of 10 amino acids, which correlates to a linker length of 4Å.

The role of monovalent affinities on shaping avidity. The other key parameter that determines the driving force behind a tandem SH2 interaction is the monovalent affinities of each SH2 domain. To test the impact of monovalent K_D on avidity we fixed the linker parameters to that of PTPN11 and GAB1 and then independently tested monovalent affinities (1C). The strongest avidities occurred when each monovalent affinity was moderately high. As either $K_{D,N}$ or $K_{D,C}$ became too large or too small, avidity decreased. The behavior in these two regimes is quite different, however, from a protein-protein interaction standpoint. Under high monovalent affinities, interactions are still high, and there is diminishing benefit or need for a second recruitment site. On the opposite end of the avidity drop-off, under low enough monovalent affinities, the partners are not interacting and there is no initial binding event that seeds the biophysical constraint that leads to a secondary binding event. Theoretically, the range where avidity is strongest occurs in the typical range of human SH2 domain binding (0.1 to 10 μM) (27).

The primary model of avidity is each domain binding with near exclusivity to one cognate partner on the other protein with little to know cross-reactivity, such as in the PTPN11-GAB1 model. To explore how cross-reactivity between partners impacts avidity, we altered the relative contributions of each potential domain/pTyr interaction towards the overall $K_{D,N}$ and $K_{D,C}$ by increasing the degree of cross-reactivity for both domains. Monovalent affinity remained constant, with $K_{D,N-N}$ and $K_{D,C-C}$ decreasing to compensate for the increases in $K_{D,N-C}$ and $K_{D,C-N}$ (Fig. S2). The maximum avidity decreases with cross-reactivity, though remains in the greater than 15-fold range, when both domains have reasonably strong affinity (2.5μM) for the ligand, except under extreme conditions where both domains are competing for the same single pTyr site with near exclusivity. However, moving to a model where there is one strong recruitment domain (2.5μM) and one weaker domain (25μM) results in a significantly larger reduction in avidity with the introduction of cross-reactivity. Hence, the characteristics of high avidity include strong preferences for separate domain specificity with minimal cross-reactivity.

The effects of avidity on binding duration. Signaling proteins bind in a transient manner, meaning they bind, unbind, and typically re-bind a number of times before fully dissociating and moving away within the cell (28). Since we have established that avidity within tandem SH2 proteins likely occurs due to an increased effective concentration of unbound domain upon initial binding of the first SH2 domain, it stands to reason that a multivalent protein in the process of unbinding from a target where one domain is still bound would experience the same increased driving force. In this manner, we hypothesize that protein interactions that experience stronger avidity will re-bind with a stronger affinity, thus extending the amount of time two proteins are engaged. To test this hypothesis, the PTPN11:GAB1 interaction was used as a baseline, and linker and monovalent K_D parameters were altered to achieve different avidities, including a simulated monovalent condition where no avidity occurs. It is worth highlighting that our selection of a constrained k_{off} of 1 s^{-1} determines the specific time scale of this simulation and the x-axis is an arbitrary unit connected to this constant. The different conditions were then used to simulate a transient binding interaction where there is no longer a constant influx of PTPN11 to maintain a steady concentration and drive the system to equilibrium, and instead the proteins are allowed to bind and fully unbind from one another (Fig. 1D). As hypothesized, the conditions that represented a larger avidity also resulted in the proteins remaining bound for a significantly longer period of time.

Establishing an experimental model of tandem SH2 domain binding. Although the adapted ODE modeling approach appears to describe a known bivalent interaction well, experimental validation of model outputs is necessary, especially since there may often be uncertainty in the parameterization of monovalent affinities and linker biophysics. Therefore, we set out to establish a complementary experimental validation. We selected biolayer interferometry (BLI) for measuring binding kinetics between an SH2 domain and a phosphorylated peptide partner immobilized to a streptavidin probe. However, synthesizing multiply phosphorylated peptides is challenging and expensive, especially at the length necessary to cover physiologically relevant partners. This limitation was overcome by utilizing our lab's SISA-KiT system which involves co-expressing a protein fused to a p40 polyproline sequence with a constitutively active tyrosine kinase fused to an ABL SH3 domain (Fig. 2A). The affinity between the ABL SH3 domain and the p40 sequence localizes the kinase to our protein of interest, driving phosphorylation and helping us achieve multi-site phosphorylation. This approach has the benefit of cheaply and easily producing large batches of phosphorylated proteins, although it produces a mixture of peptide species, including nonphosphorylated, singly phosphorylated, and doubly phosphorylated partners. In some ways, this experimental system is a strength – it captures the complexity of ligand presentation in cells, where it is unlikely that tandem SH2 domains interact with a pure population of doubly phosphorylated partners. We set out to establish the relevancy of this experimental system and to test model predictions of tandem SH2 domain binding.

Experimentally testing PTPN11:GAB1 interaction. To evaluate the general experimental approach of recombinant production of bisphosphorylated proteins using SiSA-KiT, we

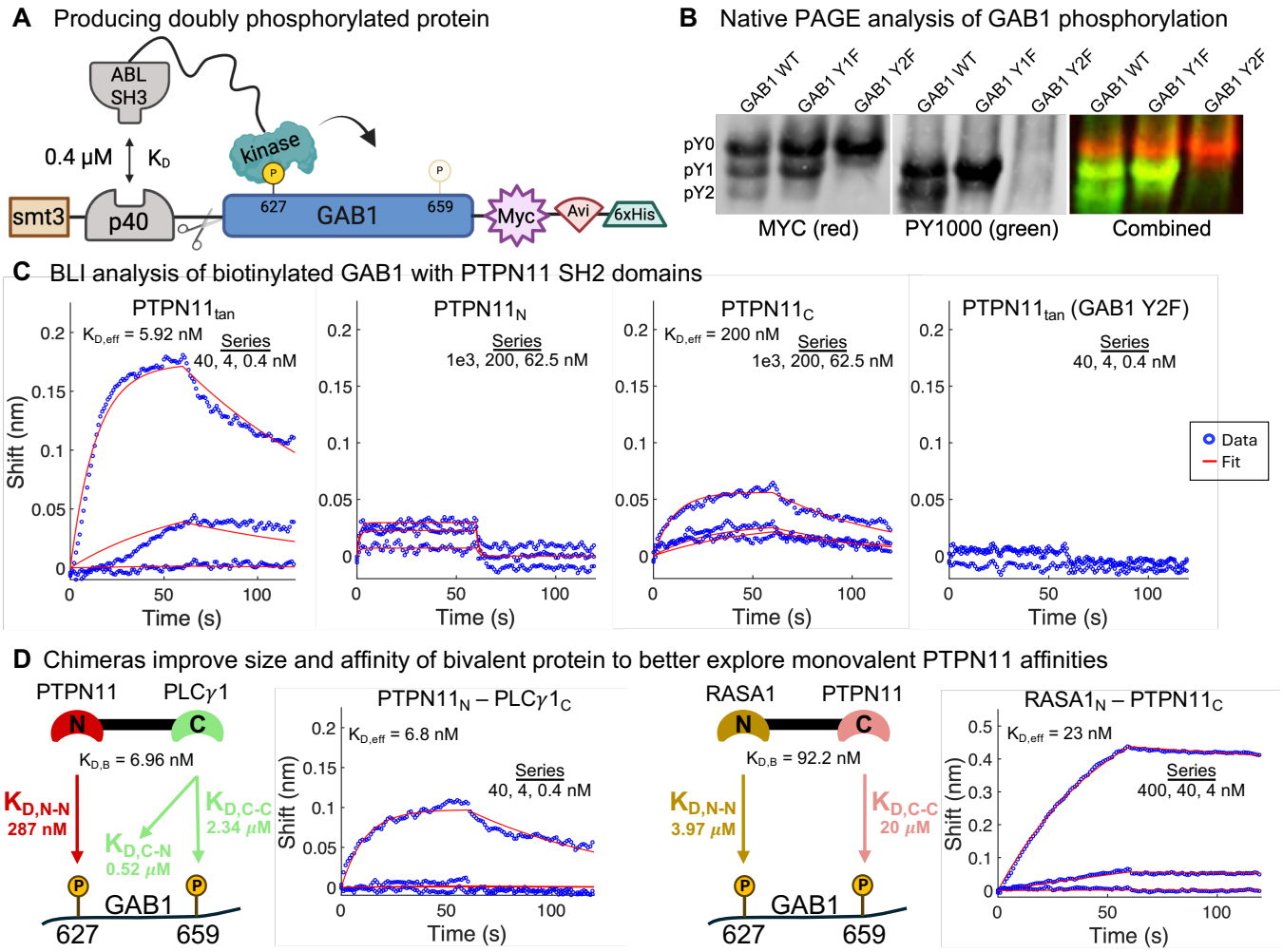

Fig. 2. Experimental testing of the PTPN11:GAB1 interaction. **A)** SISA-KiT was used to phosphorylate both pTyr sites within the GAB1 region expressed using SRC kinase catalytic domain. The secondary interaction used in SISA-KiT to enhance phosphorylation of the substrate is the ABL SH3 domain with the p40 polyproline sequence (APTYSPPPPP). **B)** The degree of phosphorylation on purified GAB1 was evaluated by running native PAGE, immunoblotting with MYC, and confirming by pan-specific phosphoantibody. GAB1 WT refers to an unmutated GAB1 region, Y1F is a Y627F mutant, and Y2F is a Y627F/Y659F negative control. **C)** Purified GAB1 proteins co-expressed with kinase were biotinylated using BirA ligase and streptavidin BLI tips were used to measure the binding kinetics with SH2 domain. Shown here are the background corrected on- and off-kinetics and the fit of the BLI data with a 1:1 kinetic model. We adjusted the serial dilution series to match the predicted affinity, and the specific concentrations of SH2 domain in the experiment are indicated. Where fit was sufficiently reliable in the experiment, we report an effective affinity ($K_{D,eff}$). **D)** Chimeras were produced by swapping out the $PTPN11_N$ and $PTPN11_C$ SH2 domains with $RASA1_N$ and $PLC\gamma1_C$, respectively. Simulations of these chimeras interacting with GAB1 were performed in the model, and BLI experiments were performed to calculate $K_{D,eff}$ for each new species.

selected the positive control GAB1 and PTPN11 interaction. We expressed the approximately 100 amino acid long span (residues 590 to 694) of the GAB1 C-terminus, where Y627/Y659 were the only tyrosines, along with a non-phosphorylatable control (Y2F) and an additional form with only Y659 (Y1F) and phosphorylated them using the SRC kinase. To test the yield of phosphorylation, we used native PAGE to resolve the phosphorylated forms, estimating 20% of our GAB1 protein was doubly phosphorylated and 30% of the protein was singly phosphorylated (Fig. 2B). These constructs were engineered to include Avitag™, which we used to biotinylate for the purposes of using BLI to measure interactions with GAB1 species. For BLI experiments, we tested three serial dilutions of SH2 domain protein, using the Y2F control to test for the phosphospecific binding of the interaction, and performed a global fit to measure binding affinity. The range of serial dilutions were adjusted throughout the experiments to better match the affinity of each interaction since binding saturation leads to nonlinearities and the inability to accurately extract on and off rates (Fig. S4). Like other experimental kinetic or thermodynamic analyses, BLI signal represents the mixture of all configurations of the SH2 domain protein bound to the ligand and hence we refer to the measured affinity from BLI as an "effective" affinity ($K_{D,eff}$).

We hypothesized that if the degree of doubly phosphorylated protein was sufficient to induce bivalent binding, then the tandem SH2 domain binding would be significantly stronger than the monovalent SH2 domain binding. Hence, for this experimental test we also tested the monovalent interactions of the N-terminal and C-terminal PTPN11 domains with the phosphorylated GAB1, allowing us to calculate the avidity effect of the interaction (Fig. 2C, S3). By presenting the same phospholigand species mixture to individual N- or C-terminal domains, the measurement reflects the combined $K_{D,N}$ and $K_{D,C}$ interaction with both phosphorylation sites and can be compared directly to the tandem SH2 domain in-

teraction. Throughout this work, we report the lowest $K_{D,eff}$ of duplicate experiments, except in cases of poor fit, as determined in prior work (24).

Consistent with the literature values, individual PTPN11 domains bound weakly to GAB1, whereas the tandem PTPN11 domains bound strongly with an estimated $K_{D,eff}$ of 5.92 nM, 34-fold stronger than the highest monovalent affinity, almost exactly matching the avidity predicted by the model. The bivalent affinity also closely matched values found using ITC experiments, lending additional confidence to this BLI approach (18). $PTPN11_C$ bound fairly well with a $K_{D,eff}$ of 200 nM, greatly exceeding our parameterization of 20 μM. $PTPN11_N$ exhibited detectable binding, but the combination of low signal and rapid on-rate made the global fit unreliable, resulting in $K_{D,eff}$ estimates spanning six orders of magnitude for replicates with similar binding curves. Therefore, we were unable to report a $K_{D,eff}$ for this domain. However, the bivalent affinity demonstrates strongly that both domains are functionally intact in the tandem domain species, and the prediction of the bivalent interaction was 8.23 nM, which is very similar to the experimentally measured 5.92 nM interaction. Together, these suggest that, as a set of interactions, the model affinities and the linker parameters were well calibrated to predict the overall increase in binding affinity in tandem interactions, which is two orders of magnitude better than the best monovalent affinity. Importantly, these results demonstrate that the experimental approach of producing large phosphorylated species, coupled with BLI, is capable of measuring high avidity interactions to test model predictions.

SH2 chimeras to interrogate model parameters. To overcome the signal limitations of testing low affinity SH2 monomers with BLI, we developed a new approach for evaluating the discrepancy in monovalent affinities of each PTPN11 domain. We selected two alternate tandem SH2 domain proteins, RASA1 and PLCγ1, which also had previously characterized binding affinities with GAB1 Y627 and Y659 (24). We made two domain "swaps" that would allow us to isolate and interrogate each monovalent PTPN11 domain under bivalent conditions that are better for BLI – increasing protein size and boosting affinity within the optimum range of the instrument (Fig. 2D). The $PTPN11_N$ - PLC$\gamma1_C$ chimera replaces the PTPN11 C-SH2 domain that was lacking binding affinity estimates in the literature with a new C-terminal domain with previously measurable affinity to both GAB1 sites. The $RASA1_N$ - $PTPN11_C$ chimera replaces the immeasurable PTPN11 N-terminal domain with an interaction previously characterized with GAB1 Y627 as 3.97 μM. We simulated and experimentally tested these chimeras, keeping the originally proposed PTPN11 domain interaction parameters (Fig. 1A). For the $PTPN11_N$ - PLC$\gamma1_C$ chimera, the model and BLI experiments were in strong agreement with a $K_{D,B}$ of 6.96 nM and a $K_{D,eff}$ of 6.8 nM, suggesting that our parameterization of the PTPN11 N-SH2 domain is fairly accurate, and the monomer is binding with an affinity of roughly 300 nM. However, for the $RASA1_N$ - $PTPN11_C$ chimera, the model and experimental results differed by a wide margin. The model provided a $K_{D,B}$ prediction of 92.2 nM whereas BLI yielded a $K_{D,eff}$ of 23 nM. This disagreement suggests that the parameterization of the PTPN11 C-SH2 affinity of 20 μM was inaccurate, and the monomer is capable of binding with a stronger affinity as supported by the $PTPN11_C$ BLI results. This is not surprising given the 20 μM parameter was arbitrarily chosen due to limitations in literature data.

These results highlight the importance of having accurate experimental data to parameterize the ODE model. However, the strong agreement between the model and experimental results of the $PTPN11_{tan}$:GAB1 interaction, despite the uncertainty surrounding $PTPN11_C$ demonstrates the ability of the model to handle errors in one domain if the remaining monovalent interaction is relatively strong and can drive avidity. Furthermore, these results introduce the idea that certain pTyr sites within the human proteome that are not expected to interact with any SH2 domain-containing protein, based on available monovalent binding experiments, may actually be a critical binding site to enhance bivalent binding of a tandem SH2 domain partner, changing both amount and duration of the interaction.

Exploring generalizability of tandem SH2 domain binding. Having established that the computational approach can successfully predict a known high affinity bivalent interaction with PTPN11-GAB1 and the ability to confirm experimentally, we set out to explore the generalizability of this approach to more broadly predict tandem SH2 domain recruitment. First, we asked if we could develop predictions for each family member of the tandem SH2 domains, which vary in the SH2 domain linker region. Secondly, we ask if we can scan a set of different pTyr pairs along a receptor tyrosine kinase C-terminal tail and identify the preferred binding partners of an SH2 domain.

Expanding to all tandem SH2 domain family members. The ten tandem SH2 proteins within the human proteome are spread across five structurally homologous groups, each exhibiting a unique SH2 linker. PTPN11 and PLCγ1 contain a short, flexible SH2 linker of 9/10 amino acids in length whereas ZAP70 and PIK3R1 contain a much longer linker of 60/195 amino acids which is rich in alpha helices, making it significantly more rigid (the persistence length of alpha helices is typically between 150 – 200Å (29, 30)). Finally, RASA1's SH2 domains are separated by an SH3 domain, providing moderate rigidity to accompany its length of 78 amino acids (31). When measuring the linker distance for these proteins using PDB structure estimates, the differences in size across proteins are noticeably smaller, but there still exists a modest variability – most notably between PTPN11 and ZAP70 which exhibit an average linker distance of 17.3 and 40.9 A, respectively (Fig. 3A). We were curious whether the difference in SH2 linker parameters between ZAP70 and PTPN11 would also correspond to differences in preferred pTyr linker parameters, so we simulated a pTyr linker sweep experiment for ZAP70 (Fig 3B). Somewhat surprisingly, the results of this simulation match quite well to the same experiment performed for PTPN11 (Fig 1C), with the only real dif-

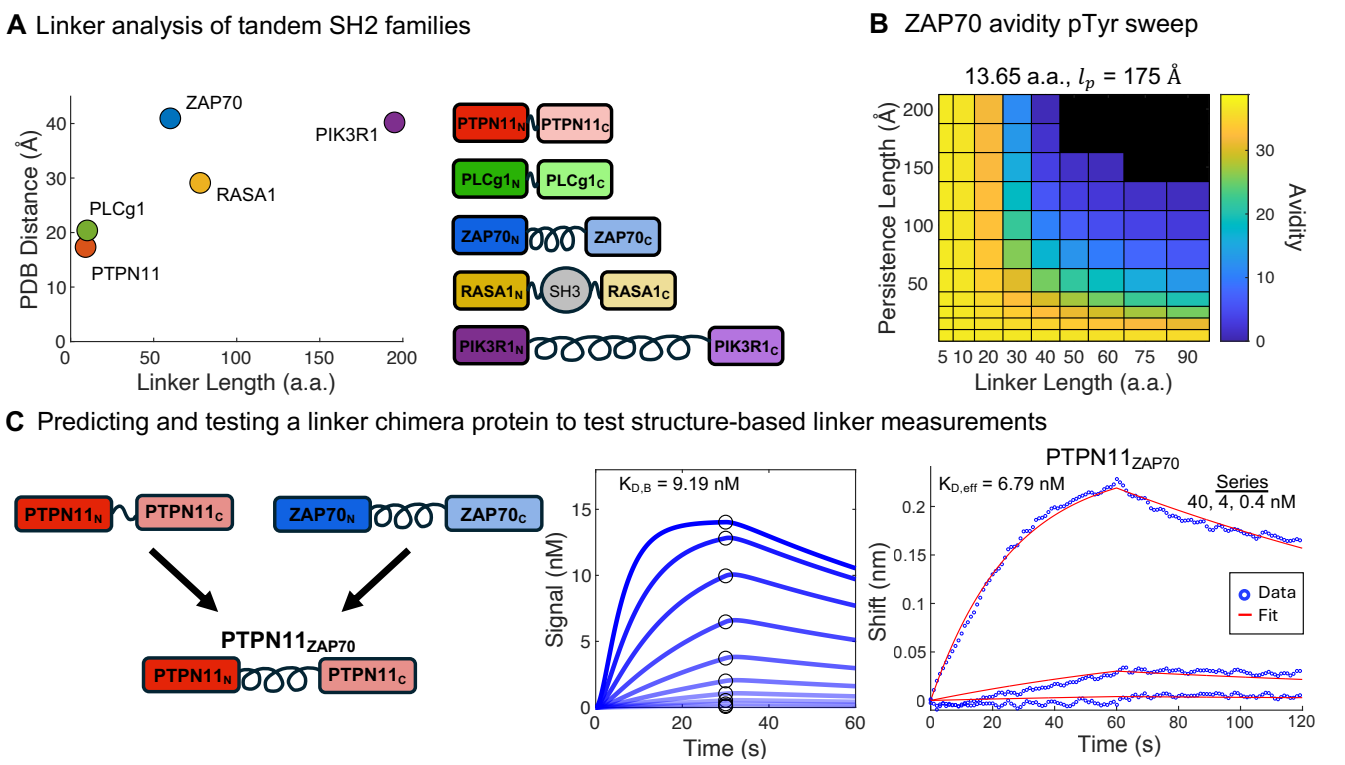

Fig. 3. Evaluating linkers between all tandem SH2 domains across human SH2 families. **A)** The structure-based end-to-end distance versus the primary sequence length for each tandem SH2 protein family is shown. Structure-based distance is averaged across available structures for that protein. **B)** We conducted a pTyr linker sweep simulation where ZAP70 linker parameters were held constant for the SH2 linker and pTyr linker parameters were varied, with avidity being calculated under each condition. Setting the structure estimate linker length at 13.65 results in a linker distance that closely matches 40.9 Å. **C)** We tested the hypothesis that structure-based distance is similar between family members, despite high sequence variability by replacing the short PTPN11 linker region between PTPN11 SH2 domains with the large, helical region of ZAP70. We modeled and experimentally measured the effective affinity of the $\mathrm{PTPN11}_{\mathrm{ZAP70}}$ chimera with doubly phosphorylated GAB1 Y627/659. Experimentally measured chimera affinity (6.79nM) is very close to that of the PTPN11 binding with GAB1 (5.92nM), suggesting that the model hypothesis of similar binding characteristics across varying SH2 linker parameters is accurate.

ference being a modestly improved avidity response to rigid pTyr linkers of a slightly greater length.

To confirm that the variability across linkers causes minimal changes to binding characteristics of the tandem proteins, we created a linker "swap" by replacing the PTPN11 SH2 domain linker with the ZAP70 linker (Fig 3C). The ($\mathrm{PTPN11}_{\mathrm{ZAP70}}$) chimera is predicted to bind with high avidity on a similar scale to the original PTPN11 wild type protein (9.19 nM compared to 8.23 nM of PTPN11). Experimentally, the $\mathrm{PTPN11}_{\mathrm{ZAP70}}$ bound to GAB1 with an effective affinity of 6.79 nM, closely matching the model estimate and, as expected, being only slightly higher than the original $\mathrm{PTPN11}_{\mathrm{tan}}$ $\mathrm{K}_{\mathrm{D,eff}}$ of 5.92 nM (Fig. S5).

Given the stark differences between the flexibility of the SH2 linkers as well as their amino acid separation, it is surprising how small of an impact it has on avidity. Structurally, the longer linkers of proteins such as PIK3R1 are folded in a way that minimizes their end-to-end distance. PIK3R1's linker is only 2.3x as long as PTPN11's linker based on physical distance despite containing 22x as many amino acids, and nearly exactly matches the distance across ZAP70's linker despite containing more than 3x the number of amino acids. This allows PIK3R1 and ZAP70 to engage just as strongly with short pTyr linkers as proteins such as PTPN11 and PLCg1. These results suggest that all tandem SH2 proteins evolved to engage with strong avidity to the same subsets of pTyr pairs.

Ranking EGFR phosphotyrosine pair recruitment of PLCγ1. We have demonstrated the utility of the ODE model to investigate the PTPN11:GAB1 interaction, including chimeras and their impacts. Next, we wanted to test whether the framework could translate well to a new set of more complex interactions. We elected to study recruitment of PLCγ1 to pTyr pairs along the EGFR C-terminal tail, where bivalent interactions are not yet well characterized. The EGFR C-tail contains nine pTyr sites, and both SH2 domains of PLCγ1 have a measured monovalent affinity towards at least four of these sites (Fig. 4A). However, studies on the recruitment of PLCγ1 only focus on a subset of pTyr sites - completely ignoring the presence of Y998 and Y1138 - and only compare the binding affinity of individual sites (32, 33). The precise manner of bivalent engagement to EGFR is unknown, leaving a major gap in our understanding of recruitment strength and configuration which may influence the interactions between EGFR and other signaling proteins. We used the ODE model to predict the $\mathrm{K}_{\mathrm{D,B}}$ of each potential pTyr pair and identify the rank order of preferred binding states of this multifaceted interaction. However, for two pairs spanning almost the entire length of the tail (998/1197 and 1016/1197), the estimation of the linker biophysics was unsolvable in our computing environment and memory and were not considered. All

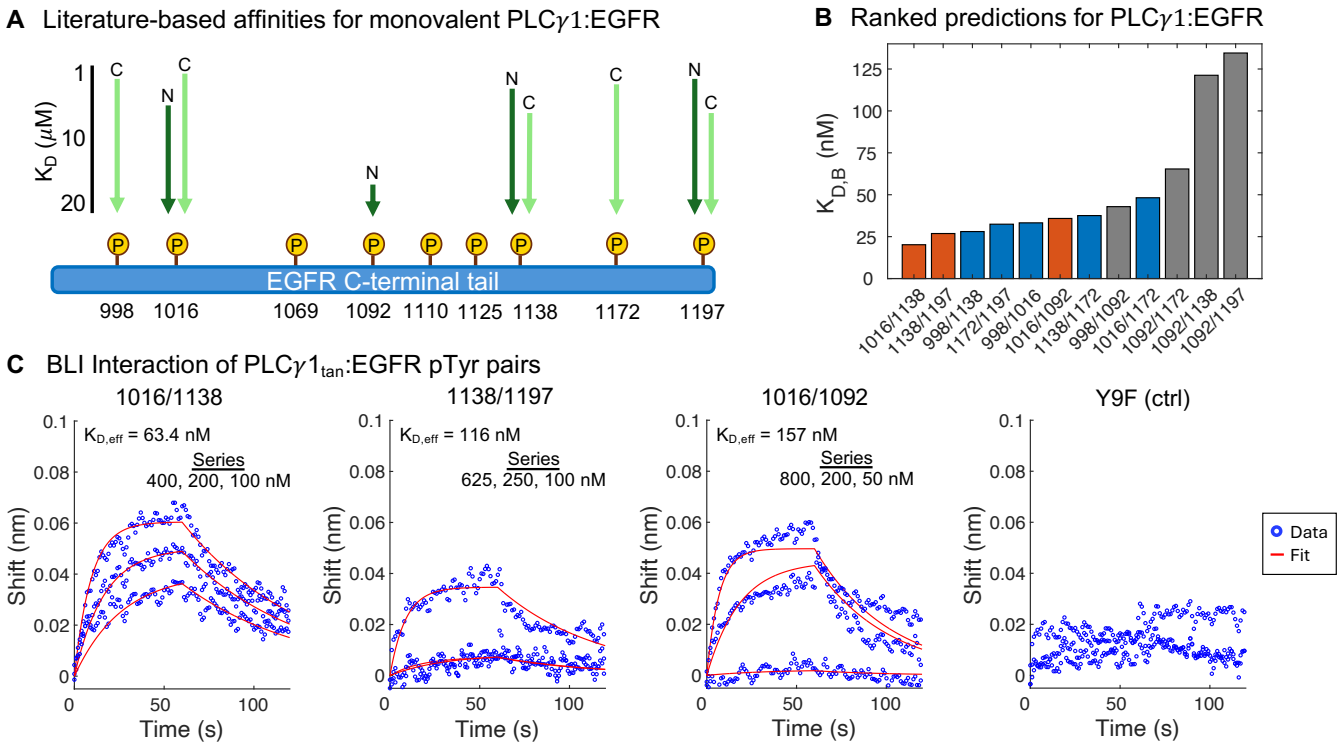

Fig. 4. Predicting and testing bivalent recruitment of PLCγ1 to EGFR C-terminal tail phosphotyrosine pairs. **A)** Literature based monovalent affinities of PLCγ1 with pTyr sites across the EGFR C-terminal tail (24), where N is N-terminal SH2 domain affinity and C is C-terminal SH2 domain affinity, where the array is proportional to affinity. **B)** Using literature affinity values, structure-based SH2 domain distance, and estimated biophysical distances for the disordered tail, we predicted the bivalent binding $K_{D,B}$ of PLCγ1 to EGFR pTyr pairs. Orange and gray bars indicate pairs experimentally tested (orange where interactions were successfully tested in our experimental system). **C)** The experimental data for Y7F pTyr pairs in orange in panel C and a Y9F control by BLI with the effective affinity for each interaction.

other remaining pair predictions were ranked, based on bivalent avidity (Fig. 4B), suggesting a broad range of effective affinities and some with very high avidity.

We selected seven pTyr pairs spanning a range of bivalent affinity predictions for experimental validation. Here, we compared Y7F EGFR C-tail constructs containing two tyrosines to a Y9F control (no tyrosines). We found that native PAGE was unable to sufficiently separate the phosphospecies of these large proteins, so we also used a Phos-Tag approach (34), where we were able to achieve separation of some species of phosphorylation. We used phosphospecific antibodies to the EGFR sites to confirm that we had successfully produced phosphorylation of both target sites in the kinase conditions selected (SFig. S8), although precise ratios of dual phosphorylation could not be calculated. However, the strong sub-micromolar binding demonstrated through BLI suggests that sufficient doubly-phosphorylated EGFR protein was presented.

Three pTyr pairs predicted to have the highest binding affinity produced strong BLI responses and the predicted rank was recapitulated by experiments –1016/1138 was strongest ($K_{D,\mathrm{eff}}$ of 63.4 nM), followed by 1138/1197 ($K_{D,\mathrm{eff}}$ of 116 nM) and 1016/1092 ($K_{D,\mathrm{eff}}$ of 157 nM) (Fig. 4C, SFig. S9). Unfortunately, the other four EGFR pairs tested did not achieve an adequate fit across any replicates, so their $K_{D,\mathrm{eff}}$ cannot be reported (SFig. S10). Poor binding behavior in the BLI system might be attributed to insufficient bisphosphorylated receptor or a limitation of the maximum concentration of PLC$\gamma 1_{\mathrm{tan}}$, both of which would be problematic for these lower affinity interactions along the spectrum of predicted affinity. Regardless, the three strongest pairs that were measurable in our system matched the model-predicted rank order, suggesting that the model can be used to effectively estimate preferred binding partners within a complex system and is useful in wider contexts beyond the PTPN11:GAB1 interaction.

While the relative affinity matches the predicted order, it should be noted that the exact $K_{D,\mathrm{eff}}$ value for each EGFR pair differs from its predicted $K_{D,B}$ by a wider margin than observed in the PTPN11:GAB1 interaction – with systematically stronger model predicted affinities. Our model assumed similar flexibility of the EGFR C-tail as the GAB1 linker (a persistence length of 4Å), but it is possible that the C-tail achieves a higher rigidity following phosphorylation which could account for the reduced affinity. Furthermore, while we were able to utilize a structure-estimated linker length for PLCγ1, no PDB files exist that span the full C-terminal tail of EGFR, so we had to rely on simpler estimations of pTyr linker length that may introduce error into the system. Although estimating the specific persistence length of disordered regions like receptor tyrosine kinase tails may lead to some general errors, it is encouraging that despite this the approach is able to identify affinity interactions that emerge from bivalent cooperation and their relative rank. Importantly, these results demonstrate that tandem SH2 domain recruitment may be a much more generalized phenomenon in tyrosine kinase signaling than was previously anticipated and that these bivalent events might span regions much larger

than the spacing found in the classic immune ITAM-type pTyr patterns.

Generalizing prediction capabilities in tyrosine kinase signaling. The combination of theoretical modeling, structure-based parameterization, and experimental approaches used here highlighted key properties that govern high affinity tandem SH2 domain recruitment. To be more broadly extensible to predicting high affinity interactions though, the ability to identify such interactions without the need for highly detailed monovalent interaction data would be a great benefit, so we set out to compare how monovalent quantitative data and less characterized monovalent interaction data would perform for such purposes. Also, given the importance of SH2 in partnership with other domains on proteins to drive interactions, such as in partnership with SH3 domains (35, 36) or to guide tyrosine kinase targets (37), we set out to evaluate the linker and domain diameter relationships to evaluate the general conservation of biophysical properties. Here, we present new resources that identifies possible high affinity interactions and the biophysical parameterization for extending beyond tandem SH2 domain recruitment.

Constructing a bivalent pTyr interaction resource. Using theoretical modeling and available monovalent affinities, we were able to predict and measure high affinity protein-interactions composed of bivalent pTyr recruitment of tandem SH2 domains. We next wished to explore if we could generalize these principles to identify other high affinity protein-protein interactions across the proteome as a result of tandem SH2 domain recruitment, despite limited monovalent data. For this, we first constructed a bivalent pTyr resource, guided by the principles identified from theoretical modeling – finding all pairs of pTyr sites in the human proteome that are far enough apart to allow for both SH2 domains to bind (10 amino acids) and which are likely separated by disordered segments (we excluded sites found in or separated by domains). This resource consists of 45,678 pairs of phosphotyrosine sites in the nonredundant human proteome from 3,861 proteins, meaning 18.9% of human proteins have at least one pair of pTyr sites. We observed a range of protein-level distribution of number of pairs, with 19% of proteins having between 1 and 3 pairs of pTyr, but some proteins having many, e.g. 28 that have more than 100 pTyr pairs, which includes catenin delta-1 (CTNND1) and IRS2.

This list of possible bivalent partners for tandem SH2 domain families is the first step in building a resource for predicting high affinity interactions, which we then supplemented with binding data to identify specific protein-protein recruitment capacity. We attempted to identify tandem SH2 recruitment capacity across the bivalent pTyr reference by using PSSM scoring from degenerate libraries (38). However, consistent with what others have found (39, 40), we failed to find physiologically relevant predictions from this approach, suggesting that PSSM scoring is insufficient for broad scale prediction of SH2 domain recruitment. In lieu of the ability to predict interactions from PSSMs, we next gathered resources that might allow annotation of a larger set of pTyr pairs with their possible tandem SH2 domain partners. For this, we used quantitative data from fluorescence polarization measurements, expanding into the rest of the dataset we had used to parameterize GAB1:PTPN11 and EGFR:PLCγ1 interactions (24), and incorporating any interaction measured at 10μM or higher. However, given the relatively low-throughput nature of these types of experiments (covering 140 pTyr peptides across 7 proteins, yielding 406 pTyr pairs), we also tested if qualitative data from co-immunoprecipitation experiments (41) could be used, where we hypothesized that interactions that persist through washing and detection represent relatively high affinity interactions. Supplementary Data S4 contains all bivalent pTyr pairs in the human proteome and the subsets of pairs that could be annotated with interaction data from the Ronan et al. FP data or the Martyn et al. IP data. Given the highly quantitative and SH2-extensive nature of the FP data, we tracked and noted a partnership between a pair of sites and a tandem SH2 domain if at least one interaction was found and tracked the most likely configuration, twisted or inline, based on the highest sum of affinity values. For Martyn data, where they performed IP with both wild type domains and superbinders – SH2 domains engineered to deepen the pocket, which has been shown to increase affinity without substantially altering specificity (13, 41, 42) – we tracked and noted whether the partnership was seen in one or the other type of IP – indicating interactions seen in wild type domain pull downs as HA for "high affinity" and in superbinder pull downs as LA for "low affinity". Using the Martyn pulldown data, we are able to annotate 622 human proteins, or 45% of the human pTyr bivalent proteome with information about PIK3R1/R2 and PTPN6/11 interactions and with Ronan we annotated 406 pTyr pairs across the majority of the tandem SH2 domain family.

Next, we wished to understand if these resource types could accurately recover known bivalent partnerships and relevant protein-protein interactions that result from the accumulated evidence of many possible interaction sites across a pair of proteins. Both resources successfully identify GAB1 recruitment of PTPN11 to Y627/Y659, and IP data also captures that the corollary sites on GAB2 recruit PTPN11. If we sum all pairs of interactions between a protein and a tandem SH2 domain protein, we recover known interactions and new hypotheses. In the FP data, PTPN11 is rarely recruited to ERBB family members (only recruited to ERBB2), but is highly recruited to GAB1 (69% of possible pairs recruit PTPN11), consistent with the studies that indicate PTPN11 recruitment to the receptors is through the adaptor protein (43). Additionally, although PTPN11 is key to MET signaling (44), it has not yet been shown if that is through direct interactions, but the protein-level data suggests that 75% of MET pTyr pairs might recruit PTPN11. The FP data can provide insight into the method of binding – we observed strong PLCG1 recruitment to EGFR, consistent with our experiments, and interestingly at a ratio indicating it prefers to bind in a twisted configuration (5:1 twisted to inline).

Using the overall fraction of pairs that can recruit a tandem SH2 protein, it becomes clear that PIK3R1 recruitment to EGFR is poor, but much higher for ERBB3, consistent with known interactions and the differential distribution of YXXM recruitment motifs (45, 46). In the IP data, which covers fewer SH2 domains but has far more total pTyr partners, IRS2 and GAB2 have the highest number of overall high affinity interactions, recovering their role in recruiting PIK3R1, PIK3R2, and PTPN11, and - in the case of GAB2 - PTPN6. PTPN6 interaction data only occurs in the superbinder IP and the highest number of possible pairwise interactions with PTPN6 is CTNND1, a protein that is known specifically to recruit PTPN6 (47), which suggests both the wild type and superbinder interaction data is capable of recovering physiologically-relevant interactions. Overall, this approach to identify biophysically possible pTyr pairs, coupled with available monovalent binding data appears to be useful at identifying both pairwise and full protein level interactions that guide signaling outcomes and suggests the ability to more rapidly identify potential high affinity interactions driven by tandem pTyr engagement is possible without the necessity to perform highly quantitative and controlled assays.

Biophysical analysis of other SH2 domain partnerships. Based on our findings that SH2 domains retained relatively similar physical distances, matched well to bivalent avidity across the entire family, despite higher primary structure diversity, we set out to evaluate other linkers involving SH2 domains. Using CoDIAC (13) we comprehensively extracted the domain diameters and end-to-end linker distances between SH2 domains and SH3, tyrosine kinase, and tyrosine phosphatase domains. We summarized the average values of all of these partners, for use in modeling efforts (Supplementary Data S3), and visualized the relative spacing and sizes of all partners (Fig. S6). Interestingly, the size of these linkers are within a similar range as the SH2-SH2 linkers, with only the SH2-Kinase linkers of SYK/ZAP70 exceeding the maximum SH2-SH2 separation (Fig. S7), suggesting that these multivalent interactions also have the potential to be altered through avidity effects. Additionally, given domain diameters establish occlusion zones, we compared the linker distances, relative to their lengths, and the average combined diameter of a pair (Fig S7). The SH3-SH2 linkers are the smallest overall, which may suggest slightly different idealized polyproline-pTyr separation on cognate partners than what we observed for SH2-SH2 by modeling, even across highly diverse family members. Interestingly, across all tyrosine kinases we have data for, there is a conserved smaller separation between the adaptor domains, than with the kinase catalytic domain. Where SH3-SH2 physical separation is highly conserved, we observed much higher diversity when SH3 domains follow the SH2 domain, suggesting higher diversity in possible partner spacing between the pTyr and the polyproline, or possibly more frequently involved in trans-interactions. Overall, the analysis identifies interesting constraints that could be used to identify partner bivalent motifs and, like SH2-SH2 domains, several well conserved properties which could simplify the identification of partnerships.

Discussion

A guiding principle in the field of phosphotyrosine signaling is that the relatively low affinity of SH2 domains (on the order of μM) is important to the reversibility of signaling – allowing the dissociation of signaling complexes and phosphatase access (48–50). This study, along with existing knowledge in the immune receptor system, indicates that tandem SH2 domain containing proteins can be recruited with a strikingly higher affinity (on the order of nM). Both total amount and dynamics shape cellular outcomes (51, 52), and increasing this by 30- to 100-fold has profound impacts on protein interactions and downstream consequences. Notably, we found that the normal range of affinity for an SH2 domain (0.1 - 10 μM) matches well to the range of monovalent K_D values that maximize avidity, providing a new perspective on SH2 binding affinity. Perhaps the relatively low affinity of SH2 domains evolved because it optimizes the advantage of multivalency within the tandem SH2 proteins to allow for these dramatic improvements in affinity. Additionally, we found it interesting that the tandem SH2 domain families appear to have evolved relatively similar geometries in 3D structure that set them up exceedingly well for high avidity interactions and were surprised to find that the EGFR C-terminal tail allowed for high avidity interactions, despite significant sequence separation between sites – an experimental finding that would not have been possible via peptide synthesis methods. Given the breadth of signaling proteins covered by the tandem SH2 domain family, there is a possibility that almost all phosphotyrosine signaling is governed to some degree by tandem SH2 domain interaction avidity. Additionally, if effective affinities are really in the nM range, then active forms of reversibility may be necessary and might explain findings in the regulation of SH2 domain linker regions (11) and on or near binding interface residues of the SH2 domain (13). Finally, the modeling approach also highlights the importance of particular mutation effects – where loss of the necessary seed sequence interaction might be more important and predictive of dysregulated interactions, then impacts on the lower affinity partner.

Given the high degree of multidomain interaction units throughout the proteome, it is likely that broad swaths of molecular biology research should consider avidity, yet the combinatorial complexity of these systems have created barriers to such avidity incorporation. Combining mechanistic modeling and structure informed parameterization presents a broader framework for identifying protein interactions with high effective affinities due to multivalency, and it provides theoretical constraints that would help identify those interactions that experience avidity. Even with limited information about monovalent species or details of linker biophysics, modeling helps to bound the reasonable space of whether high affinity interactions might occur and rank-order binding partner preference, even if not able to perfectly predict effective affinity. Additionally, the synthetic toolkit for driving tyrosine phosphorylation on proteins can easily be ex-

tended to cover alternate post-translational modifications and mixtures of bivalency (e.g. polyproline and a pTyr for SH3-SH2 testing and possibly acetylation recruitment of tandem bromoodomains). It appears that these constraints can then be used to scan the proteome for possible cognate partners, using less quantitative approaches (such as immunoprecipitation) to identify the seed interaction that provides the biophysical driving force of high affinity interactions. Conservation of biophysical parameters across families, such as SH3-SH2 and SH2-SH2, suggests that the problem might be reduced for the entire family to the same patterning on cognate partners, simplifying the modeling and prediction processes. Of course, even though modeling and experiments of these two part systems suggest important findings for signaling consequences, it does not yet consider competition, cellular localization constraints, or other dynamic processes, such as conformational dynamics that regulate different states of multidomain proteins, that are also at play in physiological signaling systems. Despite this, the framework presented here offer an approach for tackling the high complexity of identifying and testing protein interactions with emergent properties that significantly differ from their monovalent parts, which is an important step in modeling and dissecting complex interactions that govern cell physiology and considering more broadly how to intervene in dysregulated signaling cascades in disease.

Materials and Methods

Plasmid cloning and mutagenesis. We used plasmid backbones as described in Ryan et al. (16) to clone GAB1 and EGFR into the toolkit substrate backbones (sVd with Avitag™) for co-expression and phosphorylation with a kVh-kinase (also described in Ryan et al.). The Avitag™ originally came from the pGEX-SH2A-SH2 plasmid which was a gift from Bruce Mayer (Addgene plasmid # 46481; http://n2t.net/addgene:46481; RRID:Addgene_46481) (53). SH2 domains were cloned into a pGEX backbone, with an N-terminal GST fusion. We used either restriction/ligation cloning or In-Fusion cloning for engineering of backbone vectors, insertion of tags and targeting, and all kinase and substrate insertions. We used Snapgene to design primers, and sequenced verified by Sanger sequencing. Sequencing was performed to get full coverage of insert. Some constructs were whole plasmid sequenced to verify the entire plasmid. For cloning, we used NEB Phusion polymerase (part no. M0530S) and included extensions for cloning into the target backbones using NEB enzymes. We used Agilent's QuickChange mutagenesis kit (part no 210518) to perform point mutations. We used carbenicillin at 100μg per mL for selection of ampicillin plasmids (substrate and GST plasmids) and kanamycin at 50μg per mL for selection of kanamycin plasmids (kinase plasmids). Subcloning Efficiency DH5α Competent Cells: ThermoFisher Scientific Cat. 18265017 were used for DNA propagation. The sources for SH2 domains were: PTPN11 – from pBABE-PTPN11 which was a gift from Dr. Matt Lazzara (54); RASA1 – pDONR223_RASA1_WT was a gift from Jesse Boehm & William Hahn & David Root (Addgene plasmid # 81778 ; http://n2t.net/addgene:81778 ; RRID:Addgene_81778); ZAP70 – pDONR223-ZAP70 was a gift from William Hahn & David Root (Addgene plasmid # 23887 ; http://n2t.net/addgene:23887 ; RRID:Addgene_23887); PLCγ1 – pGEX PLCg1(NC)-SH2 was a gift from Bruce Mayer (Addgene plasmid # 46471 ; http://n2t.net/addgene:46471 ; RRID:Addgene_46471). GAB1 – from pCDNA3.1-HA-GAB1 which was a gift from Dr. Matt Lazzara (55). EGFR C-terminal tail was gifted to us by Linda Pike in both Y9F and Y8F formats (56). We selected the appropriate Y8F plasmids to mutate a second tyrosine to phenylalanine to create the Y7F plasmids.

Protein expression and purification.

Protein induction and bacterial lysis. **SH2 Domains** Plasmids were transformed into BL21 gold DE3 (Agilent part no. 230132) and an isolated colony was used to inoculate a 5 mL LB, 5 μL carbenicillin culture which grew overnight at 37°C. 2.5 mLs of this culture was used to inoculate a 100 mL LB, 100 μL carbenicillin culture which shook at 37°C until it reached the desired OD600 (0.7-1.0). The culture was then induced using 0.5 mM IPTG and grown overnight in an 18°C shaker. The culture was spun down at 8500 rpm for 10 minutes and the pellet was frozen at -20°C until lysis.

Phosphorylated Proteins Plasmids were co-transformed along with a kinase plasmid into BL21 gold DE3 (Agilent part no. 230132) and an isolated colony was used to inoculate a 5 mL LB, 5μL carbenicillin/kanamycin culture which grew overnight at 37°C. 2.5 mLs of this culture was used to inoculate a 100 mL LB, 100μL carbenicillin/kanamycin culture which shook at 37°C until it reached the desired OD600 (0.7-1.0). The culture was then induced using 0.5 mM IPTG and grown overnight in an 18°C shaker. The culture was spun down at 8500 rpm for 10 minutes and resuspended in 100 mL LB, 100 uL carbenicillin/kanamycin, and 0.2% L-arabinose. The culture was grown for another 4 hrs in a 37°C shaker and then spun down at 8500 rpm for 10 minutes and the pellet was frozen at -20°C until lysis.

Lysis We resuspended bacterial cell pellets in a 10:1 ratio of final growth volume to lysis buffer. Lysis buffer was 50mM Tris-HCl at pH7.0 and 150mM NaCl, supplemented with protease inhibitors (EMD Millipore 539137), phosphatase inhibitors (Millipore Sigma P5726), and PMSF (1:1000). For small volume lysis of 1mL to 5mL, we used bead beating, adding 400μL of beads to each 1mL of resuspended cell pellet and bead beating for 2 minutes. For larger lysis volumes, we sonicated resuspended pellets for 5 minutes with 40% pulse sequence. Following lysis, we clarified by centrifugation at 13,000 rpm for 10 minutes at 4°C.

Protein purification. **Nickel/6xHis Purification**. We used Genscript Ni-NTA MagBeads (cat no. L00295), using a magnetic tube stand for isolating beads during decanting. Prior to protein binding, we equilibrated beads in lysis buffer (50mM Tris-HCl, 300mM NaCl (pH 8.0), 0.5% TX-100, 80 mM imidazole). Beads were isolated by magnet and decanted.

We added equilibration buffer again, this time in addition to 1mL of clarified lysate and incubated at 4°C for at least 1hr up to overnight, while rotating. Isolated beads were washed in a 20-fold excess (relative to bead volume) of wash buffer (50mM Tris-HCl, 300mM NaCl, 40mM Imidazole, 0.5% Triton-X (pH 8.0)) a total of three times, incubating in the wash buffer for 5 minutes each time. PreScission protease (GenScript cat no. Z02799) was used to cleave the N-terminal solubility tag and p40 sequence. The beads were resuspended in 50 mM Tris-HCl, 150 mM NaCl, 3 mM DTT at 4 times the resin bead volume and incubated overnight at 4°C with 1 μL of PreScission. Protein was eluted using 2 times the resin bed volume with elution buffer (50mM Tris-HCl, 300mM NaCl, 500mM Imidazole pH 8.0). We repeated elution two total times. We dialyzed protein following manufacturer directions for the Pierce Slide-A-Lyzer mini dialysis with a 3.5kDa cutoff into 50mM Tris-HCl, 150mM NaCl at pH. 7.8. Proteins were aliquoted and frozen at -80°C.

GST Purification We used Pierce Glutathione Agarose (part no. 16100) and followed the manufacturer's base protocol for GST-based purification. Beads were equilibrated in buffer (50mM Tris, 150mM NaCl) and protein capture was performed for one hour to overnight at 4°C. Beads were washed (three times) with a 10-fold excess volume of equilibration buffer, with centrifugation to isolate the beads. We used PreScission protease-based elution according to manufacturers directions (GenScript GST-PreScission protease part no. Z02799), incubating protease with beads resuspended in 50 mM Tris-HCl, 150 mM NaCl, 3 mM DTT overnight at 4°C. PreScission cleavage was performed to ensure no dimerization of the GST domain (49).

Biotinylation. Following nickel- purification and dialysis, phosphorylated substrates were biotinylated using the BirA500 kit (Avidity EC 6.3.4.15). We reacted 10nmol of substrate with 2.5μg of BirA ligase for between 30 minutes and 5 hours at 30°C depending on the protein concentration. Proteins were dialyzed following biotinylation to remove excess biotin. Biotinylation was frequently confirmed through 680IRdye-conjugated streptavidin, incubated for one hour at 4°C (at 1:1500), prior to final washes and LICOR-based imaging.

Western-based analysis. Samples were reduced in Laemmli loading buffer (Boston BioProducts) and boiled for 10 minutes at 95°C, unless they contained imidazole, where we boiled at 70°C for 10 minutes to avoid breaking protein bonds. We performed standard SDS-PAGE analysis using purchased precast NuPAGE gels from Invitrogen, typically using 4-12% gradients or single 10% gels, running in a MOPS SDS running buffer (Novex Invitrogen). Transfer to nitrocellulose membranes occurred in a buffer (Novex Invitrogen) with 20% methanol (membranes were pre-wetted in transfer buffer). Pre-stained ladder from LI-COR (Chameleon Duo) was used along with IRDye-680 and -800 secondary antibodies for infrared scanning. We used Intercept Blocking Buffer (LI-COR), diluted 1:1 in Tris Buffered Saline (TBS) for blocking membranes (one hour at room temperature or overnight at 4°C) and preparing primary and secondary antibodies. Following primary incubation (1 hour at room temperature or overnight at 4°C) and secondary incubation (1 hour at room temperature at 1:10,000, we washed membranes in a TBS-T (1% tween solution) or TBS solution. Antibody stripping was done with 0.2N NaOH for 30 minutes, as needed for efficient stripping. Membranes were scanned on a LI-COR Odyssey system.

Antibodies We used the following antibodies for verification of expression and phosphorylation: MYC antibody Mouse Bio X Cell BE0238 at 1:3000 dilution; Rabbit PY1000 antibody CST 8954S at 1:1500 dilution; EGFR pY998 Rabbit CST 2641 at 1:1000 dilution; EGFR pY992/EGFR pY1016 Rabbit CST 2235 at 1:1000 dilution; ERBB2 pY1196 (with cross-reactivity of EGFR pY1138 and pY1016) Rabbit CST 6942 at 1:1000 dilution; EGFR pY1068/EGFR pY1092 Rabbit CST 3777 at 1:1000 dilution; EGFR pY1148/EGFR pY1172 Rabbit CST 4404 at 1:1000 dilution; EGFR pY1173/EGFR pY 1197 Rabbit CST 4407 at 1:1000 dilution. We used the following secondary antibodies from LI-COR Biosciences (at 1:10,000 dilution): Donkey Anti-Mouse IgG 680RD, Donkey Anti-Rabbit IgG 800CW; Donkey Anti-Rabbit igG 680RD; Donkey Anti-Mouse IgG 800CW.

Native PAGE: Phosphorylation efficiency of phosphoproteins was evaluated by running protein samples by native PAGE to allow for separation of protein species with varying levels of phosphorylation. A 9% acrylamide separating gel was poured with 1.8 mLs 30% acrylamide, 1.5 mLs 1.5 M Tris-HCl, 600 μLs 10% APS, and 6 μL TEMED. A stacking gel was poured using 533 μL 30% acrylamide, 1 mL 1.5 M Tris-HCl, 40 μL 10% APS, and 8 μL TEMED. Gels were cast using the Bio-Rad Mini-PROTEAN® Tetra Handcast System. 1L of 1x Running Buffer (250mM Tris-HCl and 1.92M Glycine) was used, and gels were run at 4°C at 90V until separation was achieved. Samples were prepared in loading dye: 40% Glycerol, 248 mM Tris-HCl, 0.02% Bromophenol Blue.

Protein concentration estimation: We used PAGE-based separation and Coomassie staining (ThermoScientific Pierce Coomassie Brilliant Blue G250 part no 20279), followed by 680-infrared detection by LICOR for estimating protein concentration compared to a BSA standard curve.

Biolayer Interferometry (BLI). Experiments were run on the Gator® Pilot BLI instrument using Streptavidin (SA) probes (Gator SKU 160002), MAX plates (Gator SKU 130062) for probe loading, and black flat plates (Gator SKU 130150) for sample loading. A sampling rate of 10 Hz was used, and a temperature of 30°C was maintained within the instrument. 250 uL of buffer (50 mM Tris-HCl, 150 mM NaCl, 1% Bovine Serum Albumin, filter sterilized) was loaded into the MAX plate to allow for equilibration of the probes, and a 10-minute, 400 rpm equilibration step was performed at the beginning of each experiment. 180 uL of sample was loaded into the black flat plate and held at a tilt throughout the experiment. Experiments were run with a 30 second baseline in buffer, 120 second load of biotinylated,

phosphorylated protein, 30 second baseline in buffer, 60 second association of SH2 domain, and 60 second dissociation in buffer. The load wells contain equal concentrations of phosphorylated protein, whereas the association wells contain a serial dilution of SH2 domain within the first three wells of the column. The final well contains buffer to act as a reference. Plates were shaken throughout the experiment with PTPN11:GAB1 experiments being shaken at 1000 rpm and PLCγ1:EGFR experiments being shaken at 400 rpm.

BLI Analysis. Kinetic analysis was performed within the Gator One® software (version is 2.16.6.0130). Data was aligned on the Y-axis at the association step, and an inter-step correction was applied along with Savitzky-Golay filtering to reduce noise. The reference well is subtracted from the other three wells to account for non-binding artifacts within the signal. A 1:1 global binding model is fit to both the association and dissociation curves of each set of three serial dilutions with an unlinked Rmax to quantify the dissociation constant, K_D. BLI figures contained within this paper uses data processed in the manner described above, and both the data and binding fit curves have been re-plotted in MATLAB®. The data is subsampled to include every 10th data point. All BLI data is provided on Figshare at 10.6084/m9.figshare.31141453, which was not subsampled but did undergo the other pre-processing steps.

ODE Modeling of tandem SH2 domain interactions. The original code was written by Dr. Wesley Errington in the Sarkar lab at the University of Minnesota (14) and shared by compressed folder by Dr. Sarkar. The code was adapted to include greater flexibility within the parameters. Most notably, additional k_{on} and k_{off} parameters were incorporated into the differential equations to allow for different affinities by each SH2 domain for each pTyr site, and persistence length was made universally adjustable instead of binning all protein interactions into two possible groups, flexible and rigid. The output of the original code is a series of plots mapping the binding behavior of the proteins under different initial concentrations of SH2 domain. Fifteen binding configurations are possible within a bivalent system, but for the purposes of this paper, only subsets were summed together and plotted as a single metric of binding. When analyzing tandem SH2 domains, the two bivalently bound configurations (inline and twisted) were isolated, whereas configurations detailing single domain binding were utilized when simulating monovalent binding. Additional code was written to extract the equilibrium concentration of each of these plots and feed that into a concentration effect curve plot. This operates by plotting the fraction of bound species at equilibrium – calculated by dividing the equilibrium concentration by the starting concentration of phosphorylated protein (R0) – against the starting concentration of SH2 domain (L0). Within the model, the concentration of SH2 domain is assumed to remain constant throughout the simulation. The value of L0 that corresponds to 50% of the phosphorylated protein being bound at equilibrium is the dissociation constant (K_D). Results provided were generated using Matlab R2024a or R2024b. All parameters have been provided in Data S1. We used the University of Virginia's High Performance Computing environment Rivanna for model predictions. All Matlab code is provided in our repository at https://github.com/NaegleLab/BivalentModeling_tanSH2.

Estimation of diameters from structure. For simulations provided in this paper, estimates of the molecular weight of SH2 domains and pTyr sites were calculated using the following website https://web.expasy.org/compute_pi/. For pTyr sites, the sequence spanning two sites N-terminal of the pTyr site and four sites C-terminal was used to estimate molecular weight. Sites Y627 and Y659 within Gab1 were calculated and used as an estimate for all pTyr pairs.

SH2 diameters We identified PDB entries containing the domain of interest (SH2) and their corresponding domain boundaries using CoDIAC (13). We included wild-type and mutant structures only if the mutations resided well outside the domain regions. To estimate domain diameters, we assumed each domain adopts an approximately spherical shape. Based on this assumption, the diameter was defined as the longest distance between any two atoms within the domain. To compute this, we measured all pairwise distances between backbone atoms (N and C) within the defined domain boundaries and recorded the maximum distance as the domain's diameter for multiple PDB structures available for the same domain. Since the ODE code assumed the same size for both entities, and since we saw relatively small differences between the N and C terminal diameters, we averaged all diameters for both domains across all available structures. All diameter calculations are provided in Data S2. Code for extraction of domain diameters is provided at https://github.com/NaegleLab/BivalentModeling_tanSH2.

Estimation of linkers from structure. **pTyr linkers** Linker length was estimated as the number of amino acids between the pTyr sites or SH2 domains, unless a PDB file was available for the protein in which case linker length was estimated using the same PDB-extraction based approach as SH2 domains.

SH2 domain linkers Structures exist for all tandem SH2 domain family members and so we used those to estimate the linker lengths. For SH2-SH2 linker used for model simulations within the paper, we used UniProt defined SH2 domain boundaries to define the linker (as the first amino acid after the N-terminal SH2 domain and up to the amino acid before the start of the C-terminal SH2 domain). Then for all available structures, we calculated the average distance between all atoms of the two amino acids farthest apart from each other on the linker region (i.e. the residues immediately adjacent to the domains). All linker calculations are provided in Data S2. Code for extraction of linker distances is provided at https://github.com/NaegleLab/BivalentModeling_tanSH2.

SH3, PTP, and Kinase linkers For the expanded analysis of SH2-SH2, SH2-SH3, SH2-PTP, and SH2-Kinase linkers, we first identified all available structures containing do-

mains of interest using CoDIAC (13). For each structure, domain diameters and inter-domain linker distances were computed as described previously. For this analysis, all distance measurements were calculated using the backbone reference point 'CA' atom. Domain and linker regions were defined using InterPro annotated domain boundaries. The primary analysis focused on adjacent domain pairs connected by linkers of defined amino acid lengths. In addition, we included the non-adjacent N-SH2 and C-SH2 domain pair in RASA1, where an SH3 domain lies between the two SH2 domains. Structures bound to inhibitors were excluded, as inhibitor binding often stabilizes inactive conformations that differ substantially from native states. Structures containing unmodeled regions within either domain or at residues used to define linker boundaries were also excluded. For each domain-domain pair, we assessed the consistency of linker distances and domain diameters across multiple structures. Structures exhibiting outlier behavior relative to the group were excluded from this analysis. For example, the SH2-Kinase linker distance in SRC structure (PDB: 1Y57) was at least 20Å longer than in the other SRC structures (Fig. S11B) and was therefore excluded. Similarly, several HCK structures (PDB: 2C0T, 2C0O, 2C0I) that displayed unusually large SH2 and SH3 domain diameters were excluded (Fig. S11A). Overall, SH2-Kinase domain-linker measurements from all HCK structures showed poor agreement with those from its close homologs SRC and FES (Fig. S11C). Consequently, all the SH2-Kinase measurements from HCK were excluded, whereas SH3-SH2 measurements from HCK that were comparable to those of SRC and FES were retained (Fig. S11C). For all the structures included in this analysis, domain diameters and linker distances are provided in Data S3. This includes both a comprehensive list of structures retrieved from CODIAC as well as a filtered list that was used in the final analysis and a summary of unique domain-domain pairs with average domain diameters and linker distances. All code and data sources are available at Figshare at 10.6084/m9.figshare.31141453.

Estimation of monovalent KD values. Estimates of monovalent K_D were taken from Ronan et al. meta-analysis of fluorescence polarization experiments (24). All k_{off} values were set to 1 s^{-1}, and k_{on} values were adjusted to match the estimated K_D.

Bivalent pTyr resource generation. We used ProteomeScoutAPI v3.0, with data updated as of January 2026, using the indicated non-redundant reference proteome as determined by Uniprot to identify all pairwise pTyr pairs in the unique human proteome (20,405 protein records). We excluded PTMs that occur within InterPro domains. We then built a pairwise pTyr resource, excluding any pairs that were less than 10 amino acids apart and any pairs that had a defined InterPro domain between them. Next, we took the two binding resource datasets, Martyn et a. (41) and Ronan et al. (24) and used the Uniprot ID and the indicated sites to map those into our reference phosphoproteome and indicated in the reference bivalent if an interaction was detected between a tandem SH2 domain and one of the sites in the pair. For Ronan et al. data, we mapped interactions if the monovalent K_D was less than 10μM. We excluded SYK and ZAP70 mapping since the majority of those data were indicated as non-functional. Additionally, only the PTPN6-C domain was available in the experiment and half of the interactions tested also were indicated as non-functional SH2 domain and we excluded the mapping of these interactions. For Martyn et al., we mapped each dataset that was pulled down by a tandem SH2 domain into our reference proteome, using the stimulated form (pervanadate treated) of the dataset from the publication. If it was identified in a pull down with a wild type protein, we indicated this as HA for high affinity and if by a superbinder form it was indicated as LA or low affinity, which may also be due to low expression the Jurkat cells. The full resource, including all pTyr pairs, annotated Ronan data, annotated Martyn data, and summarized protein-protein interaction data is provided in Supplementary Data S4 and the code used to construct, annotate, and summarize along with primary data resources is provided on Figshare at 10.6084/m9.figshare.31141453.

OPAL scoring of bivalent pTyr reference. We attempted to use degenerate library-based scoring to annotate recruitment. Based on structural analysis from Kandoor et al. (13), which identified consistent -2 to +4 interactions, we scored sequences using just the -2 to +4 interaction matrix and we removed from the reference pTyr proteome any N- or C-terminal interactions that were shorter than the span provided and then we performed PSSM matrix multiplication and summation, as per (38), to score each peptide using matrices as published in Yaron-Barir et al. (57). We tested both a percentile cutoff and a score cutoff to indicate likely interactions on the real reference phosphoproteome and a random proteome, where we randomly sorted the 20 amino acids surrounding the central pTyr residue and rescored. We assessed individual and cumulative interactions across all possible scoring approaches and failed to see physiological interactions, with random phosphoproteomes frequently scoring better and having no distinguishable difference in their ability to identify protein-protein level interactions.

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Numbers R35GM138127 and T32GM008715 (to Reagan Portelance). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank Dr. Casim Sarkar and Dr. Wesley Errington for sharing the Matlab code and giving guidance on model adaptations. We would additionally like to thank Devon Semoy and Dr. Neel Shah for their helpful guidance in native PAGE for the separation of phosphospecies. The authors acknowledge Research Computing at The University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication. URL: https://rc.virginia.edu

Bibliography

1. Ravi S. Kane. Thermodynamics of multivalent interactions: influence of the linker. *Langmuir : the ACS journal of surfaces and colloids*, 26(11):8636–8640, June 2010. ISSN 1520-5827 0743-7463. doi: 10.1021/la9047193. Place: United States.
2. Georges Vauquelin and Steven J. Charlton. Exploring avidity: understanding the potential gains in functional affinity and target residence time of bivalent and heterobivalent ligands. *British journal of pharmacology*, 168(8):1771–1785, April 2013. ISSN 1476-5381 0007-1188. doi: 10.1111/bph.12106. Place: England.
3. Natalia Nuñez-Prado, Marta Compte, Seandean Harwood, Ana Álvarez Méndez, Simon Lykkemark, Laura Sanz, and Luis Álvarez Vallina. The coming of age of engineered mul-

tivalent antibodies. *Drug discovery today*, 20(5):588–594, May 2015. ISSN 1878-5832 1359-6446. doi: 10.1016/j.drudis.2015.02.013. Place: England.

4. Angel M. Cuesta, Noelia Sainz-Pastor, Jaume Bonet, Baldomero Oliva, and Luis Alvarez-Vallina. Multivalent antibodies: when design surpasses evolution. *Trends in biotechnology*, 28(7):355–362, July 2010. ISSN 1879-3096 0167-7799. doi: 10.1016/j.tibtech.2010.03.007. Place: England.
5. Elizabeth A. Ottinger, Martyn C. Botfield, and Steven E. Shoelson. Tandem SH2 Domains Confer High Specificity in Tyrosine Kinase Signaling*. *Journal of Biological Chemistry*, 273 (2):729–735, January 1998. ISSN 0021-9258. doi: 10.1074/jbc.273.2.729.
6. Yunxin Deng, Artem K. Efremov, and Jie Yan. Modulating binding affinity, specificity, and configurations by multivalent interactions. *Biophysical journal*, 121(10):1868–1880, May 2022. ISSN 1542-0086 0006-3495. doi: 10.1016/j.bpj.2022.04.017. Place: United States.
7. N. Isakov, R. L. Wange, W. H. Burgess, J. D. Watts, R. Aebersold, and L. E. Samelson. ZAP-70 binding specificity to T cell receptor tyrosine-based activation motifs: the tandem SH2 domains of ZAP-70 bind distinct tyrosine-based activation motifs with varying affinity. *The Journal of experimental medicine*, 181(1):375–380, January 1995. ISSN 0022-1007 1540-9538. doi: 10.1084/jem.181.1.375. Place: United States.
8. J. Sloan-Lancaster, J. Presley, J. Ellenberg, T. Yamazaki, J. Lippincott-Schwartz, and L. E. Samelson. ZAP-70 association with T cell receptor zeta (TCRzeta): fluorescence imaging of dynamic changes upon cellular stimulation. *The Journal of cell biology*, 143(3):613–624, November 1998. ISSN 0021-9525 1540-8140. doi: 10.1083/jcb.143.3.613. Place: United States.
9. Xiaozheng Xu, Takeya Masubuchi, Qixu Cai, Yunlong Zhao, and Enfu Hui. Molecular features underlying differential SHP1/SHP2 binding of immune checkpoint receptors. *eLife*, 10:e74276, November 2021. ISSN 2050-084X. doi: 10.7554/eLife.74276. Publisher: eLife Sciences Publications, Ltd.
10. Nikolaos Patsoukis, Jonathan S. Duke-Cohan, Apoorvi Chaudhri, Halil-Ibrahim Aksoylar, Qi Wang, Asia Council, Anders Berg, Gordon J. Freeman, and Vassiliki A. Boussiotis. Interaction of SHP-2 SH2 domains with PD-1 ITSM induces PD-1 dimerization and SHP-2 activation. *Communications Biology*, 3(1):128, March 2020. ISSN 2399-3642. doi: 10.1038/s42003-020-0845-0.
11. Timothy Travers, William K. Kanagy, Rachael A. Mansbach, Elton Jhamba, Cedric Cleyrat, Byron Goldstein, Diane S. Lidke, Bridget S. Wilson, and S. Gnanakaran. Combinatorial diversity of Syk recruitment driven by its multivalent engagement with FcεRIγ. *Molecular Biology of the Cell*, 30(17):2331–2347, August 2019. ISSN 1059-1524. doi: 10.1091/mbc.E18-11-0722. Publisher: American Society for Cell Biology (mboc).
12. Anatoly Kiyatkin, Edita Aksamitiene, Nick I. Markevich, Nikolay M. Borisov, Jan B. Hoek, and Boris N. Kholodenko. Scaffolding protein Grb2-associated binder 1 sustains epidermal growth factor-induced mitogenic and survival signaling by multiple positive feedback loops. *The Journal of biological chemistry*, 281(29):19925–19938, July 2006. ISSN 0021-9258 1083-351X. doi: 10.1074/jbc.M600482200. Place: United States.
13. Alekhya Kandoor, Gabrielle Martinez, Julianna M. Hitchcock, Savannah Angel, Logan Campbell, Saqib Rizvi, and Kristen M. Naegle. CoDIAC: A comprehensive approach for interaction analysis reveals novel insights into SH2 domain function and regulation. *bioRxiv*, page 2024.07.18.604100, January 2025. doi: 10.1101/2024.07.18.604100.
14. Wesley J. Errington, Bence Bruncsics, and Casim A. Sarkar. Mechanisms of noncanonical binding dynamics in multivalent protein–protein interactions. *Proceedings of the National Academy of Sciences*, 116(51):25659–25667, December 2019. doi: 10.1073/pnas.1902909116. Publisher: Proceedings of the National Academy of Sciences.
15. Tim Bilbrough, Emanuele Piemontese, and Oliver Seitz. Dissecting the role of protein phosphorylation: A chemical biology toolbox. *Chemical Society Reviews*, 51(13):5691–5730, July 2022. ISSN 1460-4744. doi: 10.1039/D1CS00991E.
16. Margaret M. Ryan, Reagan Portelance, Graham F. Newman, Gabrielle Martinez, Swathi Shekharan, Anqi Wu, Savannah Angel, Katherine E. Schaberg, Petra Gilmore, Robert Sprung, Reid Townsend, and Kristen M. Naegle. A signaling inspired synthetic toolkit for efficient production of tyrosine phosphorylated proteins. *bioRxiv*, page 2024.12.22.629992, January 2025. doi: 10.1101/2024.12.22.629992.
17. Jess M. Cunnick, Lin Mei, Craig A. Doupnik, and Jie Wu. Phosphotyrosines 627 and 659 of Gab1 Constitute a Bisphosphoryl Tyrosine-based Activation Motif (BTAM) Conferring Binding and Activation of SHP2 *. *Journal of Biological Chemistry*, 276(26):24380–24387, June 2001. ISSN 0021-9258. doi: 10.1074/jbc.M010275200. Publisher: Elsevier.
18. Tobias Gruber, Marc Lewitzky, Lisa Machner, Ulrich Weininger, Stephan M. Feller, and Jochen Balbach. Macromolecular Crowding Induces a Binding Competent Transient Structure in Intrinsically Disordered Gab1. *Journal of Molecular Biology*, 434(5):167407, March 2022. ISSN 0022-2836. doi: 10.1016/j.jmb.2021.167407.
19. Huan-Xiang Zhou. Quantitative Relation between Intermolecular and Intramolecular Binding of Pro-Rich Peptides to SH3 Domains. *Biophysical Journal*, 91(9):3170–3181, November 2006. ISSN 0006-3495. doi: 10.1529/biophysj.106.090258.
20. Alexander F. Chin, Dmitri Toptygin, W. Austin Elam, Travis P. Schrank, and Vincent J. Hilser. Phosphorylation Increases Persistence Length and End-to-End Distance of a Segment of Tau Protein. *Biophysical Journal*, 110(2):362–371, January 2016. ISSN 0006-3495. doi: 10.1016/j.bpj.2015.12.013.
21. Guillaume Stirnemann, David Giganti, Julio M. Fernandez, and B. J. Berne. Elasticity, structure, and relaxation of extended proteins under force. *Proceedings of the National Academy of Sciences*, 110(10):3847–3852, March 2013. doi: 10.1073/pnas.1300596110. Publisher: Proceedings of the National Academy of Sciences.
22. Ronald J. Hause, Kin K. Leung, John L. Barkinge, Mark F. Ciaccio, Chih-pin Chuu, and Richard B. Jones. Comprehensive Binary Interaction Mapping of SH2 Domains via Fluorescence Polarization Reveals Novel Functional Diversification of ErbB Receptors. *PLoS ONE*, 7(9):e44471, September 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0044471.
23. Kin K Leung, Ronald J Hause, John L Barkinge, Mark F Ciaccio, Chih-Pin Chuu, and Richard B Jones. Enhanced prediction of Src homology 2 (SH2) domain binding potentials using a fluorescence polarization-derived c-Met, c-Kit, ErbB, and androgen receptor interactome. *Molecular & cellular proteomics : MCP*, 13(7):1705–23, July 2014. ISSN 1535-9484. doi: 10.1074/mcp.M113.034876.
24. Tom Ronan, Roman Garnett, and Kristen M. Naegle. New analysis pipeline for high-throughput domain–peptide affinity experiments improves SH2 interaction data. *Journal of Biological Chemistry*, 295(32):11346–11363, August 2020. ISSN 0021-9258. doi: 10.1074/jbc.RA120.012503.
25. Livia Pagano, Valeria Pennacchietti, Francesca Malagrinò, Mariana Di Felice, Julian Toso, Elena Puglisi, Stefano Gianni, and Angelo Toto. Folding and Binding Kinetics of the Tandem of SH2 Domains from SHP2. *International journal of molecular sciences*, 25(12), June 2024. ISSN 1422-0067. doi: 10.3390/ijms25126566. Place: Switzerland.
26. Lorenzo Visconti, Francesca Malagrinò, Livia Pagano, and Angelo Toto. Understanding the Mechanism of Recognition of Gab2 by the N-SH2 Domain of SHP2. *Life (Basel, Switzerland)*, 10(6), June 2020. ISSN 2075-1729. doi: 10.3390/life10060085. Place: Switzerland.
27. Tomonori Kaneko, Rakesh Joshi, Stephan M. Feller, and Shawn SC Li. Phosphotyrosine recognition domains: the typical, the atypical and the versatile. *Cell Communication and Signaling*, 10(1):32, November 2012. ISSN 1478-811X. doi: 10.1186/1478-811X-10-32.
28. Georges Vauquelin and Steven J. Charlton. Long-lasting target binding and rebinding as mechanisms to prolong in vivo drug action. *British journal of pharmacology*, 161(3):488–508, October 2010. ISSN 1476-5381 0007-1188. doi: 10.1111/j.1476-5381.2010.00936.x. Place: England.
29. Sivaraj Sivaramakrishnan and James A. Spudich. Systematic control of protein interaction using a modular ER/K α-helix linker. *Proceedings of the National Academy of Sciences*, 108 (51):20467–20472, December 2011. doi: 10.1073/pnas.1116066108. Publisher: Proceedings of the National Academy of Sciences.
30. Kathryn G. Wilcox, Marlee E. Dingle, Ankit Saha, Michael J. A. Hore, and Svetlana Morozova. Persistence length of α-helical poly-l-lysine. *Soft Matter*, 18(35):6550–6560, 2022. ISSN 1744-683X. doi: 10.1039/D2SM00921H. Publisher: The Royal Society of Chemistry.
31. Magnus Kjaergaard. Estimation of Effective Concentrations Enforced by Complex Linker Architectures from Conformational Ensembles. *Biochemistry*, 61(3):171–182, February 2022. ISSN 0006-2960. doi: 10.1021/acs.biochem.1c00737. Publisher: American Chemical Society.
32. D. Rotin, B. Margolis, M. Mohammadi, R. J. Daly, G. Daum, N. Li, E. H. Fischer, W. H. Burgess, A. Ullrich, and J. Schlessinger. SH2 domains prevent tyrosine dephosphorylation of the EGF receptor: identification of Tyr992 as the high-affinity binding site for SH2 domains of phospholipase C gamma. *The EMBO journal*, 11(2):559–567, February 1992. ISSN 0261-4189 1460-2075. doi: 10.1002/j.1460-2075.1992.tb05087.x. Place: England.
33. Ansuman Chattopadhyay, Manuela Vecchi, Qun-sheng Ji, Raymond Mernaugh, and Graham Carpenter. The Role of Individual SH2 Domains in Mediating Association of Phospholipase C-γ1 with the Activated EGF Receptor*. *Journal of Biological Chemistry*, 274(37): 26091–26097, September 1999. ISSN 0021-9258. doi: 10.1074/jbc.274.37.26091.
34. Eiji Kinoshita, Emiko Kinoshita-kikuta, and Tohru Koike. Phos-tag SDS-PAGE systems for phosphorylation profiling of proteins with a wide range of molecular masses under neutral pH conditions. *Proteomics*, 12:192–202, 2012. doi: 10.1002/pmic.201100524.
35. Shalini S. Yadav and W. Todd Miller. Cooperative activation of Src family kinases by SH3 and SH2 ligands. *Cancer letters*, 257(1):116–123, November 2007. ISSN 0304-3835. doi: 10.1016/j.canlet.2007.07.012.
36. Stéphane Richard, Dongyin Yu, Kendall J. Blumer, Derek Hausladen, Michael W. Olszowy, Patricia A. Connelly, and Andrey S. Shaw. Association of p62, a Multifunctional SH2- and SH3-Domain-Binding Protein, with src Family Tyrosine Kinases, Grb2, and Phospholipase Cγ-1. *Molecular and Cellular Biology*, 15(1):186–197, January 1995. ISSN null. doi: 10.1128/MCB.15.1.186.
37. B J Mayer, P K Jackson, R a Van Etten, and D Baltimore. Point mutations in the abl SH2 domain coordinately impair phosphotyrosine binding in vitro and transforming activity in vivo. *Molecular and cellular biology*, 12(2):609–18, 1992. ISSN 0270-7306. doi: 10.1128/MCB.12.2.609.Updated.
38. Lei Li, Chenggang Wu, Haiming Huang, Kaizhong Zhang, Jacob Gan, and Shawn S-C Li. Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. *Nucleic acids research*, 36(10):3263–73, June 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn161.
39. Kousik Kundu, Fabrizio Costa, Michael Huber, Michael Reth, and Rolf Backofen. Semi-supervised prediction of SH2-peptide interactions from imbalanced high-throughput data. *PloS one*, 8(5):e62732, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0062732.
40. Dejan Gagoski, H. Tomas Rube, Chaitanya Rastogi, Lucas A. N. Melo, Xiaoting Li, Rashmi Voleti, Neel H. Shah, and Harmen J. Bussemaker. Accurate affinity models for SH2 domains from peptide binding assays and free-energy regression. *Protein Science*, 34(11):e70317, 2025. ISSN 1469-896X. doi: 10.1002/pro.70317.
41. Gregory D. Martyn, Gianluca Veggiani, Ulrike Kusebauch, Seamus R. Morrone, Bradley P. Yates, Alex U. Singer, Jiefei Tong, Noah Manczyk, Gerald Gish, Zhi Sun, Igor Kurinov, Frank Sicheri, Michael F. Moran, Robert L. Moritz, and Sachdev S. Sidhu. Engineered SH2 Domains for Targeted Phosphoproteomics. *ACS Chemical Biology*, 17(6):1472–1484, June 2022. ISSN 15548937. doi: 10.1021/acschembio.2c00051.
42. Tomonori Kaneko, Haiming Huang, Xuan Cao, Xing Li, Chengjun Li, Courtney Voss, Sachdev S. Sidhu, and Shawn S. C. Li. Superbinder SH2 Domains Act as Antagonists of Cell Signaling. *Science Signaling*, 5(243):ra68–ra68, September 2012. doi: 10.1126/scisignal.2003021.
43. Armelle Yart, Muriel Laffargue, Patrick Mayeux, Stany Chretien, Christine Peres, Nicholas Tonks, Serge Roche, Bernard Payrastre, Hugues Chap, and Patrick Raynal. A Critical Role for Phosphoinositide 3-Kinase Upstream of Gab1 and SHP2 in the Activation of Ras and Mitogen-activated Protein Kinases by Epidermal Growth Factor *. *Journal of Biological Chemistry*, 276(12):8856–8864, March 2001. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M006966200.
44. Christiane R. Maroun, Monica A. Naujokas, Marina Holgado-Madruga, Albert J. Wong, and Morag Park. The Tyrosine Phosphatase SHP-2 Is Required for Sustained Activation of Extracellular Signal-Regulated Kinase and Epithelial Morphogenesis Downstream from the Met Receptor Tyrosine Kinase. *Molecular and Cellular Biology*, 20(22):8513–8525, November 2000. ISSN null. doi: 10.1128/MCB.20.22.8513-8525.2000.
45. Paolo Fedi, Jacalyn H. Pierce, Pier Paolo Di Fiore, and Matthias H. Kraus. Efficient coupling with phosphatidylinositol 3-kinase, but not phospholipase CϒΥ or GTPase-activating protein, distinguishes ErbB-3 signaling from that of other ErbB/EGFR family members. *Molecular*

and Cellular Biology, 14(1):492–500, January 1994. ISSN 1098-5549. doi: 10.1128/mcb.14.1.492-500.1994.

46. Nathan J. Hellyer, Myong-Soo Kim, and John G. Koland. Heregulin-dependent Activation of Phosphoinositide 3-Kinase and Akt via the ErbB2/ErbB3 Co-receptor. *Journal of Biological Chemistry*, 276(45):42153–42161, November 2001. ISSN 00219258. doi: 10.1074/jbc.M102079200.
47. H. Keilhack, U. Hellman, J. van Hengel, F. van Roy, J. Godovac-Zimmermann, and F. D. Böhmer. The protein-tyrosine phosphatase SHP-1 binds to and dephosphorylates p120 catenin. *The Journal of Biological Chemistry*, 275(34):26376–26384, August 2000. ISSN 0021-9258. doi: 10.1074/jbc.M001315200.
48. Tomonori Kaneko, Haiming Huang, Xuan Cao, Xing Li, Chengjun Li, Courtney Voss, Sachdev S. Sidhu, and Shawn S. C. Li. Superbinder SH2 Domains Act as Antagonists of Cell Signaling. *Science Signaling*, 5(243):ra68–ra68, September 2012. doi: 10.1126/scisignal.2003021. Publisher: American Association for the Advancement of Science.
49. J E Ladbury, M a Lemmon, M Zhou, J Green, M C Botfield, and J Schlessinger. Measurement of the binding of tyrosyl phosphopeptides to SH2 domains: A reappraisal. *Proceedings of the National Academy of Sciences of the United States of America*, 92(8):3199–203, May 1995. ISSN 0027-8424.
50. John E. Ladbury and Stefan T. Arold. *Energetics of Src Homology Domain Interactions in Receptor Tyrosine Kinase-Mediated Signaling*, volume 488. Elsevier Inc., 2011. ISBN 978-0-12-381268-1. doi: 10.1016/B978-0-12-381268-1.00007-0.
51. T. C. Grammer and J. Blenis. Evidence for MEK-independent pathways regulating the prolonged activation of the ERK-MAP kinases. *Oncogene*, 14(14):1635–1642, April 1997. ISSN 0950-9232. doi: 10.1038/sj.onc.1201000. Place: England.
52. Silvia D. M. Santos, Peter J. Verveer, and Philippe I. H. Bastiaens. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nature Cell Biology*, 9(3):324–330, March 2007. ISSN 1476-4679. doi: 10.1038/ncb1543.
53. Kazuya Machida, Christopher M. Thompson, Kevin Dierck, Karl Jablonowski, Satu Kärkkäinen, Bernard Liu, Haimin Zhang, Piers D. Nash, Debra K. Newman, Peter Nollau, Tony Pawson, G. Herma Renkema, Kalle Saksela, Martin R. Schiller, Dong-Guk Shin, and Bruce J. Mayer. High-throughput phosphotyrosine profiling using SH2 domains. *Molecular cell*, 26 (6):899–915, June 2007. ISSN 1097-2765. doi: 10.1016/j.molcel.2007.05.031. Place: United States.
54. Janine M. Buonato, Ingrid S. Lan, and Matthew J. Lazzara. EGF augments TGFβ-induced epithelial-mesenchymal transition by promoting SHP2 binding to GAB1. *Journal of Cell Science*, 128(21):3898–3909, January 2015. ISSN 1477-9137. doi: 10.1242/jcs.169599.
55. Christopher M. Furcht, Janine M. Buonato, and Matthew J. Lazzara. EGFR-activated Src family kinases maintain GAB1-SHP2 complexes distal from EGFR. *Science Signaling*, 8 (376):1–13, 2015. ISSN 19379145. doi: 10.1126/scisignal.2005697.
56. Kamaldeep Gill, Jennifer L. Macdonald-Obermann, and Linda J. Pike. Epidermal growth factor receptors containing a single tyrosine in their C-terminal tail bind different effector molecules and are signaling-competent. *Journal of Biological Chemistry*, 292(50):20744–20755, December 2017. ISSN 0021-9258. doi: 10.1074/jbc.M117.802553.
57. Tomer M. Yaron-Barir, Brian A. Joughin, Emily M. Huntsman, Alexander Kerelsky, Daniel M. Cizin, Benjamin M. Cohen, Amit Regev, Junho Song, Neil Vasan, Ting-Yu Lin, Jose M. Orozco, Christina Schoenherr, Cari Sagum, Mark T. Bedford, R. Max Wynn, Shih-Chia Tso, David T. Chuang, Lei Li, Shawn S.-C. Li, Pau Creixell, Konstantin Krismer, Mina Takegami, Harin Lee, Bin Zhang, Jingyi Lu, Ian Cossentino, Sean D. Landry, Mohamed Uduman, John Blenis, Olivier Elemento, Margaret C. Frame, Peter V. Hornbeck, Lewis C. Cantley, Benjamin E. Turk, Michael B. Yaffe, and Jared L. Johnson. The intrinsic substrate specificity of the human tyrosine kinome. *Nature*, 629:1174–11811–8, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07407-y.

Supplementary Materials for Portelance, Wu, Kandoor, and Naegle

This PDF file includes:. Figures S1 to S8
Captions for Data S1 to S4

Supplementary Figures.

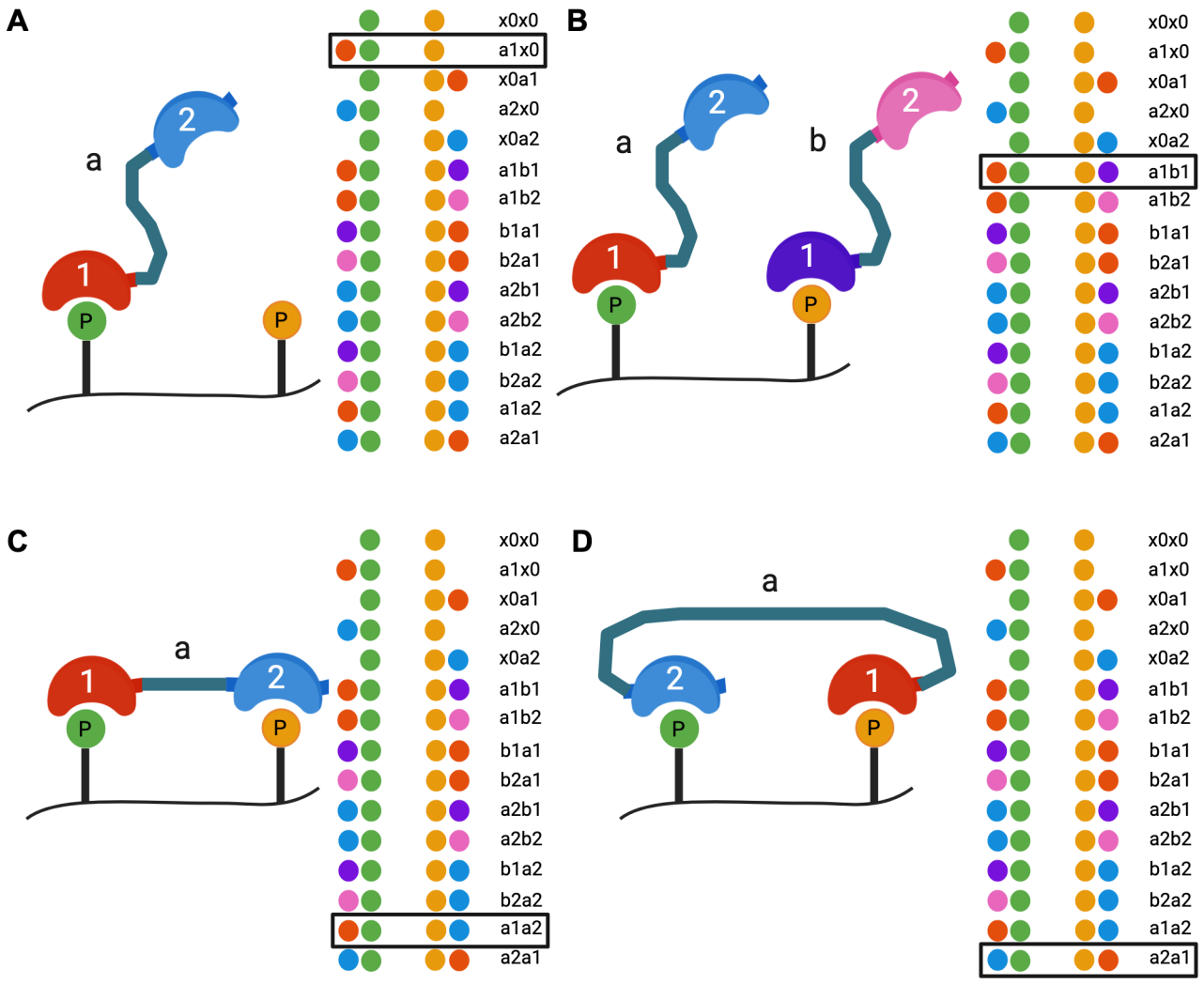

Fig. S1. Bivalent binding states. There are 15 possible binding states for a bivalent system. In addition to the unbound state this includes (**A**) four monovalently bound states, (**B**) eight bivalently bound states spanning across two tandem SH2 proteins, and two versions of a bivalently bound state - one (**C**) inline and one (**D**) twisted.

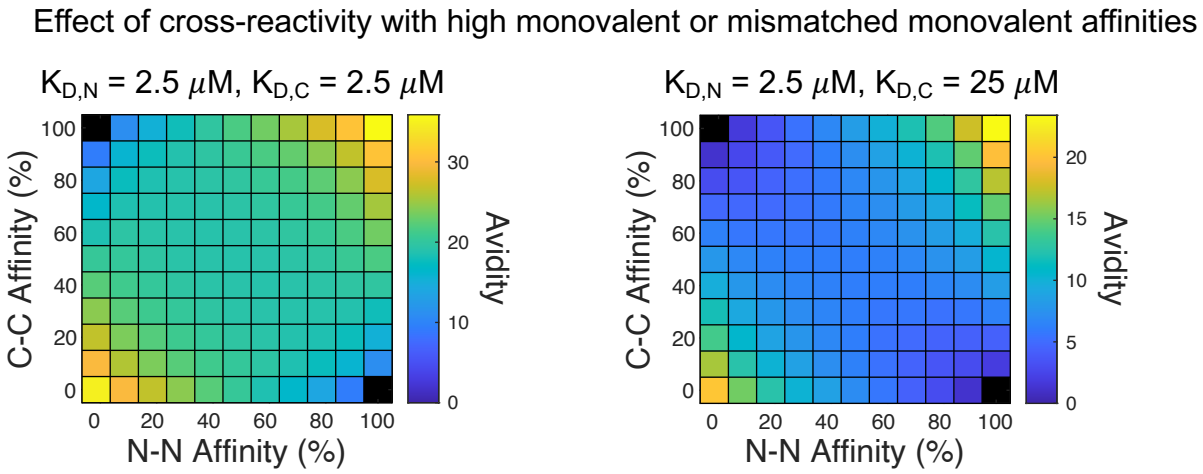

Fig. S2. Cross-reactivity negatively impacts bivalent interactions. In order to test for the impact of cross-reactivity we simulated two interactions with fixed monovalent affinity – either as both high affinity (left) or as one high affinity and one low affinity interaction (right). We then changed the cross-binding interaction affinity from N-SH2 to C-pTyr and C-SH2 to N-pTyr, increasing it from no cross-reactivity up to 100% cross-reactivity while decreasing the N-SH2 to N-pTyr and C-SH2 to C-pTyr affinity accordingly to maintain a constant $K_{D,N}$ and $K_{D,C}$.

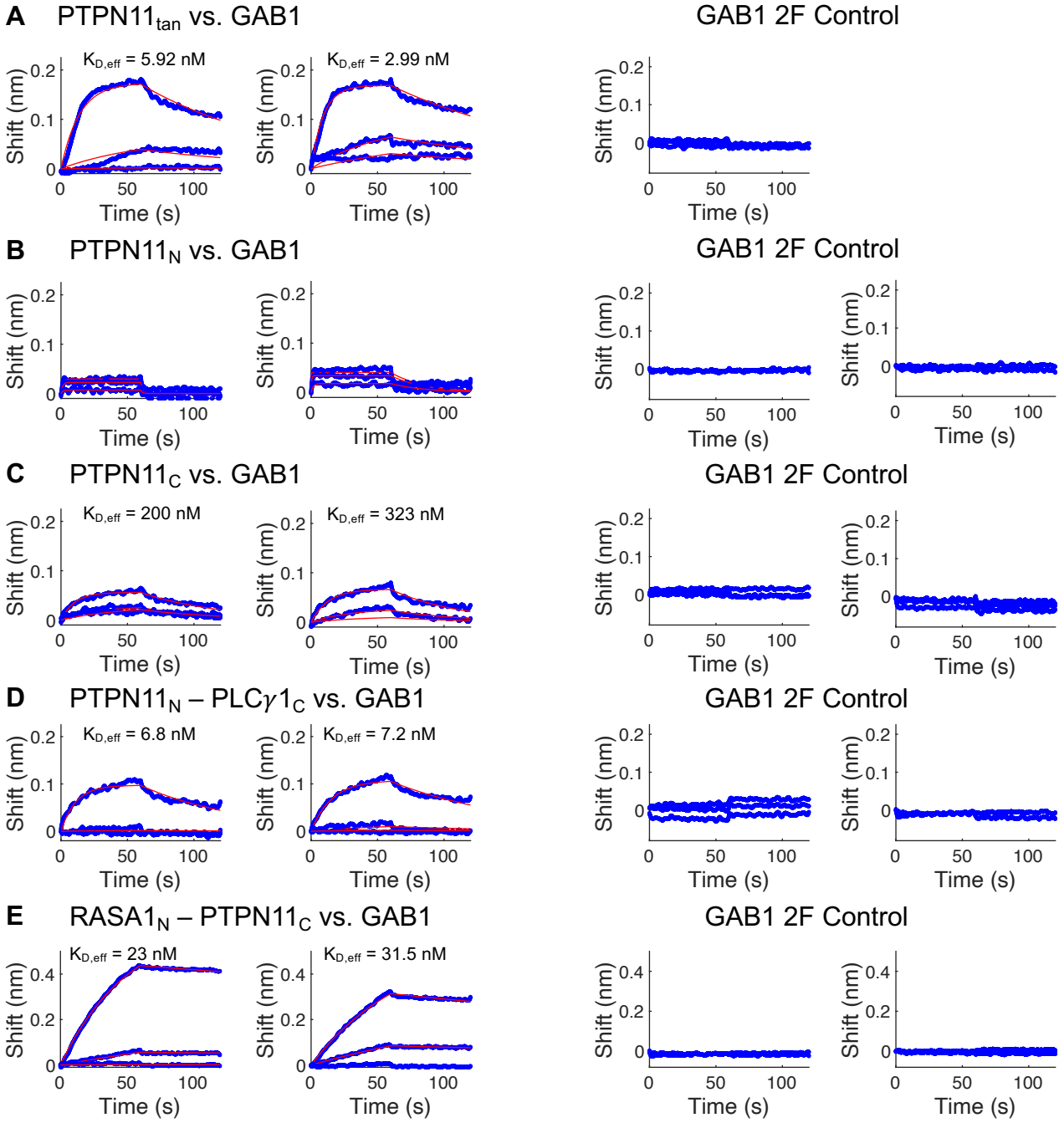

Fig. S3. Replicates and Y2F controls for BLI experiments of PTPN11 interacting with GAB1. GAB1 refers to the same preparation of phosphorylated GAB1 C-terminal tail spanning Y627 and Y659. GAB1 2F refers to the negative control with both tyrosines mutated to phenylalanine. The replicate data and the effective K_D are given for (**A**) Tandem SH2 domains of PTPN11 (40, 4, 0.4 nM) (**B**) The N-terminal SH2 domain of PTPN11 (1000, 250, 62.5 nM) (**C**) The C-terminal SH2 domain of PTPN11 (1000, 250, 62.5 nM) (**D**) The chimera of PTPN11 N-terminal and PLCγ1 C-terminal domain (40, 4, 0.4 nM), and (**E**) The chimera of RASA1 N-terminal domain and the PTPN11 C-terminal domain (400, 40, 4 nM (left) and 200, 50, 12.5 nM (right)).

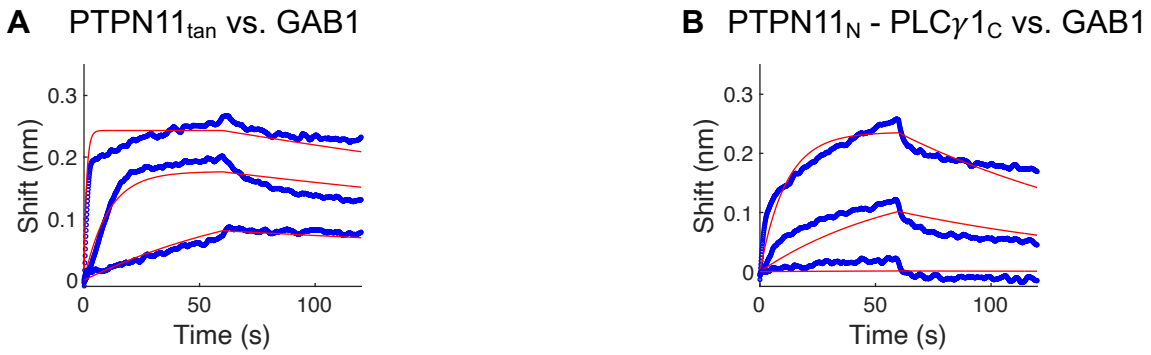

Fig. S4. Examples of serial dilution range that produces nonlinearities. Examples of (**A**) PTPN11$_{\mathrm{tan}}$ and (**B**) PTPN11$_{\mathrm{N}}$-PLCγ1$_{\mathrm{C}}$ vs. GAB1 BLI experiments with too high SH2 concentrations (400, 40, 4 nM) resulting in poor fit.

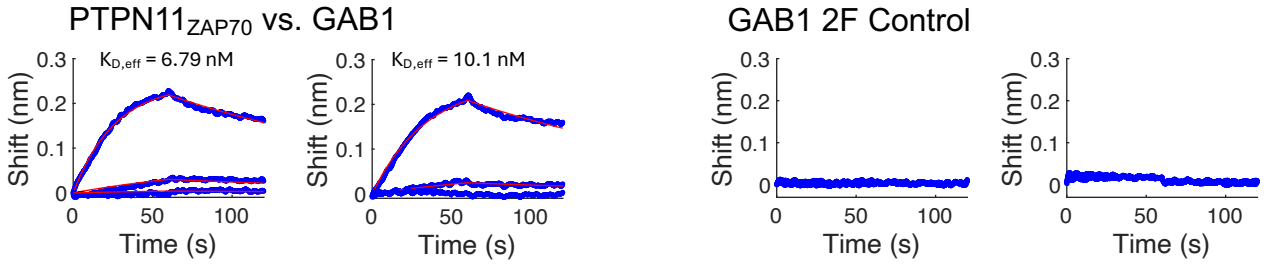

Fig. S5. Replicates and Y2F controls for BLI experiments of PTPN11$_{\mathrm{ZAP70}}$ binding with GAB1. GAB1 refers to the same preparation of phosphorylated GAB1 C-terminal tail spanning Y627 and Y659. GAB1 2F refers to the negative control with both tyrosines mutated to phenylalanine. The replicate data and the effective K_{D} are given for tandem SH2 domains of PTPN11$_{\mathrm{ZAP70}}$ (40, 4, 0.4 nM).

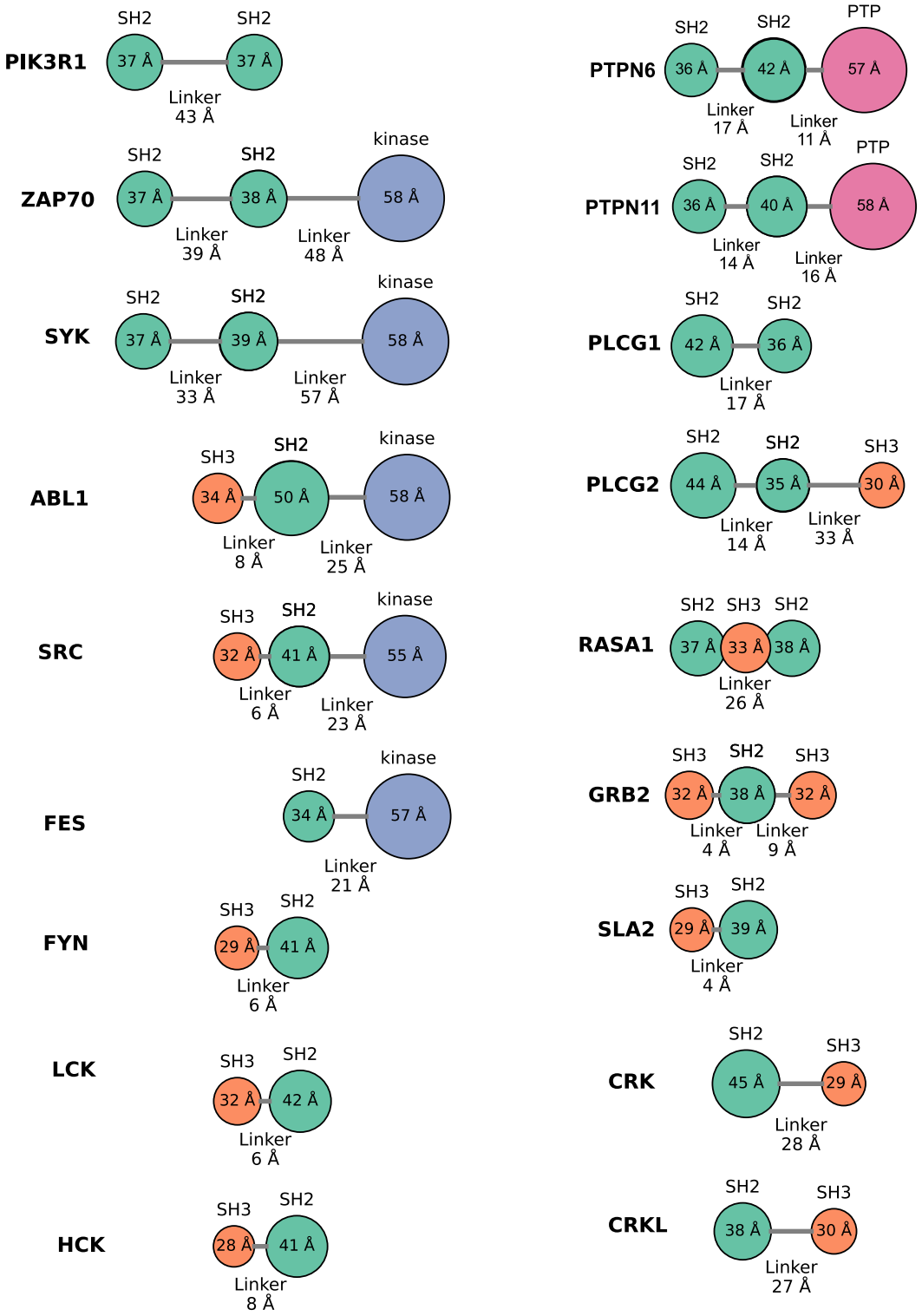

Fig. S6. Domain architecture of SH2 containing proteins. The domain architecture of proteins used to measure the linker distance between SH2, SH3, PTP, and Kinase domains is provided along with the size of each domain and distance across each linker. Domains are only included if a linker distance could be measured.

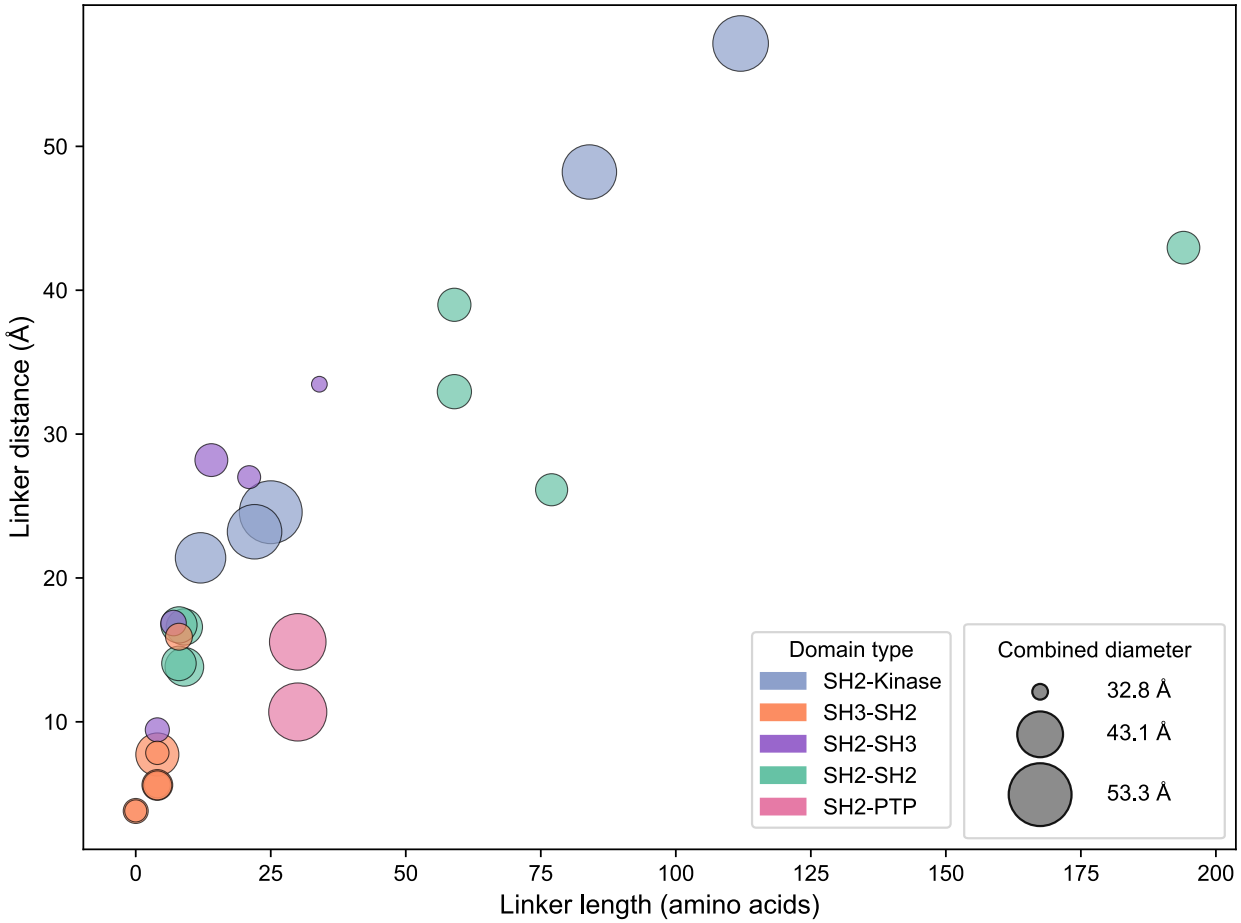

Fig. S7. Linker size between SH2, SH3, PTP, and Kinase domains. All linker distances between SH2, SH3, PTP, and Kinase domains in the human proteome were measured if PDB structures were available. The average distance for each protein was plotted against the number of amino acids comprising the linker. The size of the dot represents the possible occlusion size of the linker calculated by summing the radii of each domain bookending the linker.

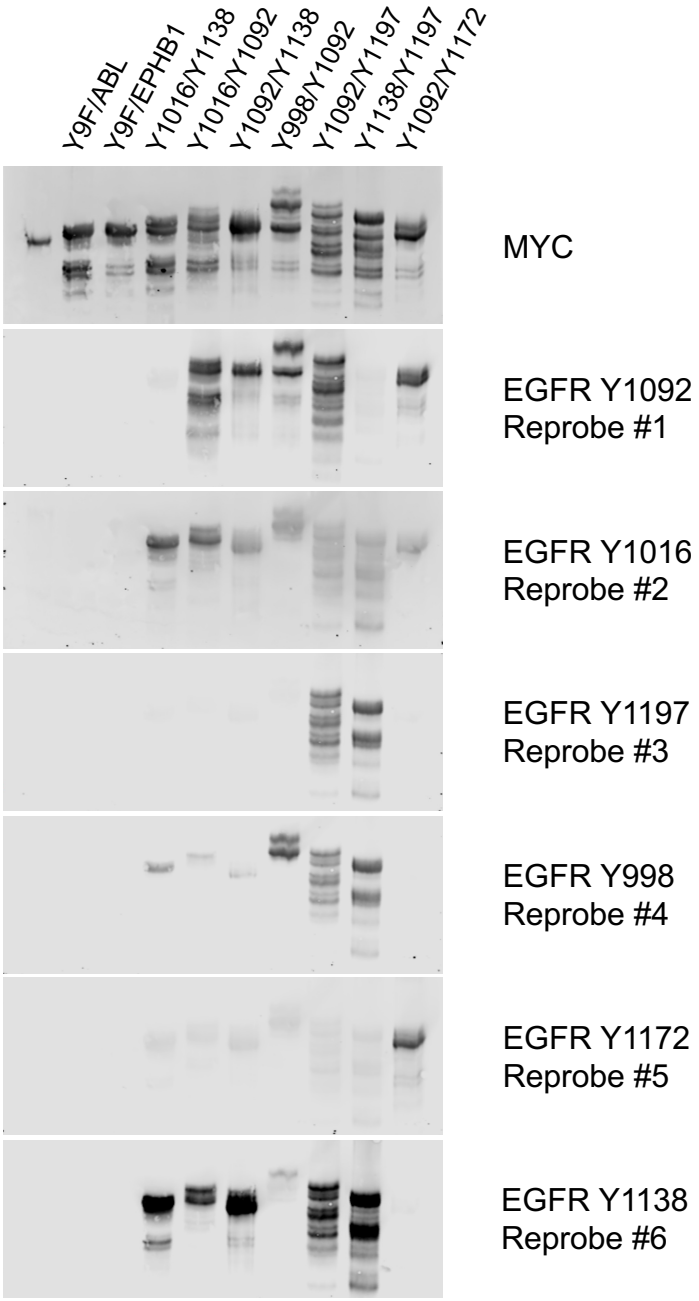

Fig. S8. Validation of double phosphorylation on EGFR Y7F constructs. All EGFR Y7F pairs were run on a Zn-Phostag gel and co-probed with anti-MYC antibody and site-specific antibodies for EGFR pTyr sites 998, 1016, 1092, 1138, 1172, and 1197. The anti-ERBB2 Y1196 antibody was used for site 1138 due to homology between the sites, but the antibody proved to be nonspecific and bound broadly across the pairs. The 998 antibody also demonstrated some nonspecificity, and there was some incomplete stripping during the 1016 probe. Clear separation of phosphospecies were seen in some pairs such as Y998/Y1092 while other pairs such as Y1016/Y1138 showed little separation. It is difficult to conclusively determine the extent of double phosphorylation on any individual pair due to variation in migration, but we did show that both EGFR sites exhibited some degree of phosphorylation in each pair. Whether this is a mixture of singly-phosphorylated species or a properly doubly-phosphorylated species remains unclear.

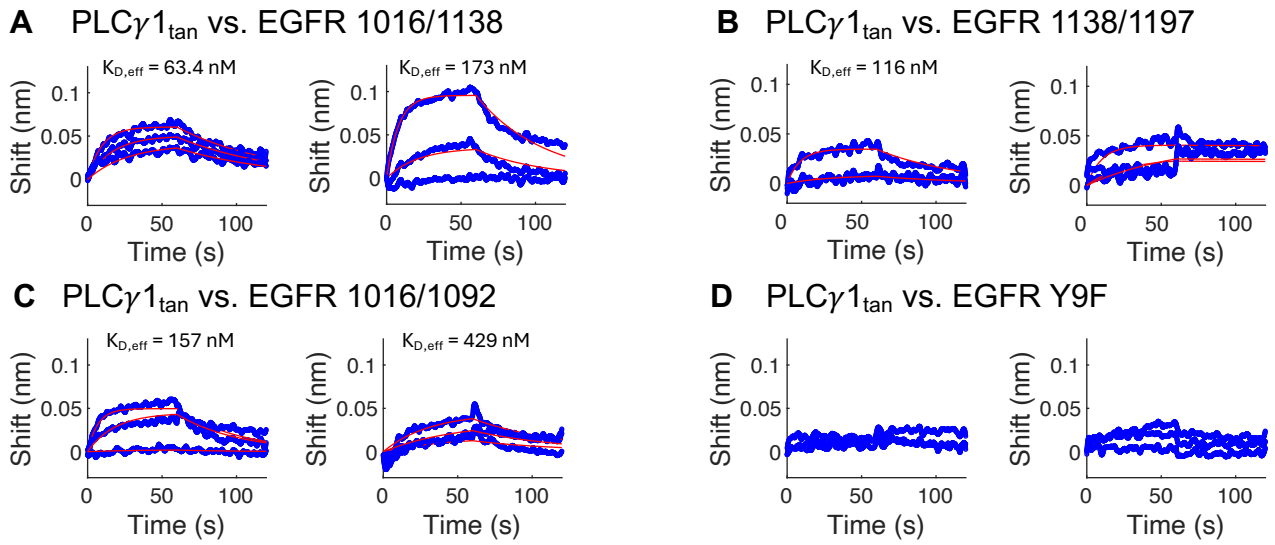

Fig. S9. Replicates and Y9F controls for EGFR BLI experiments. EGFR refers to the different preparations of phosphorylated EGFR C-terminal tail with unique pairs of pTyr sites available and the remaining seven pTyr sites mutated to a phenylalanine. EGFR Y9F refers to the negative control with all tyrosines mutated to phenylalanine. The replicate data and the effective K_D are given for (**A**) EGFR 1016/1138 (400, 200, 100 nM (left) and 800, 200, 50 nM (right)), (**B**) EGFR 1138/1197 (625, 250, 100 nM (left) and 500, 200, 80 nM (right)) (**C**) EGFR 1016/1092 (800, 200, 50 nM (left) and 400, 200, 100 nM (right)), and (**D**) EGFR Y9F (800, 200, 50 nM).

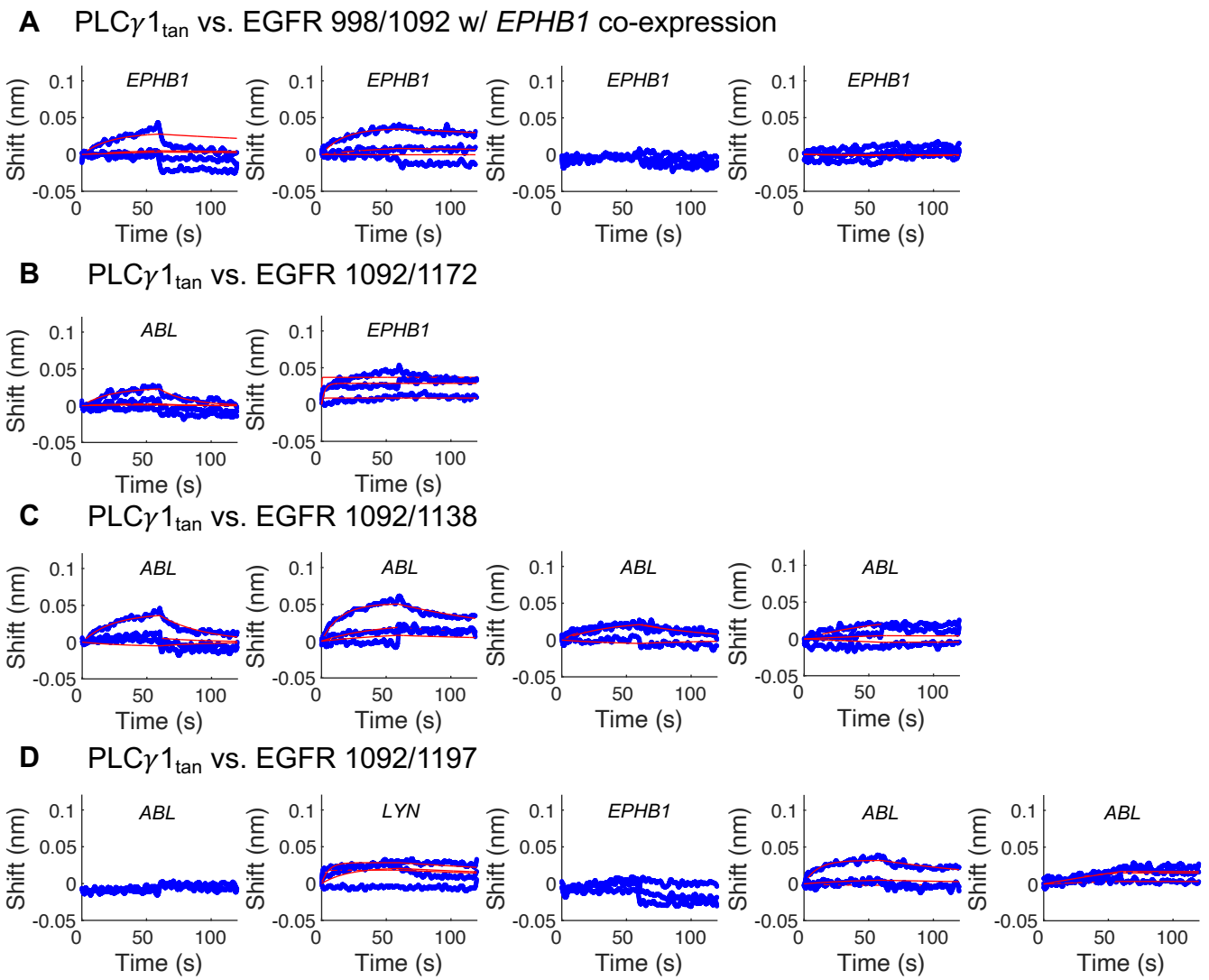

Fig. S10. EGFR Y7F Pairs with Poor BLI Fit. In addition to EGFR Y7F pairs 1016/1138, 1138/1197, and 1092/1197, four additional pairs were produced and tested against PLC$\gamma 1_{\mathrm{tan}}$ using BLI. These four pairs - (**A**) 998/1092, (**B**) 1092/1172, (**C**) 1092/1138, and (**D**) 1092/1197 - did not achieve an adequate fit across any replicate tested so a $K_{\mathrm{D,eff}}$ could not be reported. For certain pairs, multiple kinases co-expressions were tested, and the kinase used for each replicate is listed.

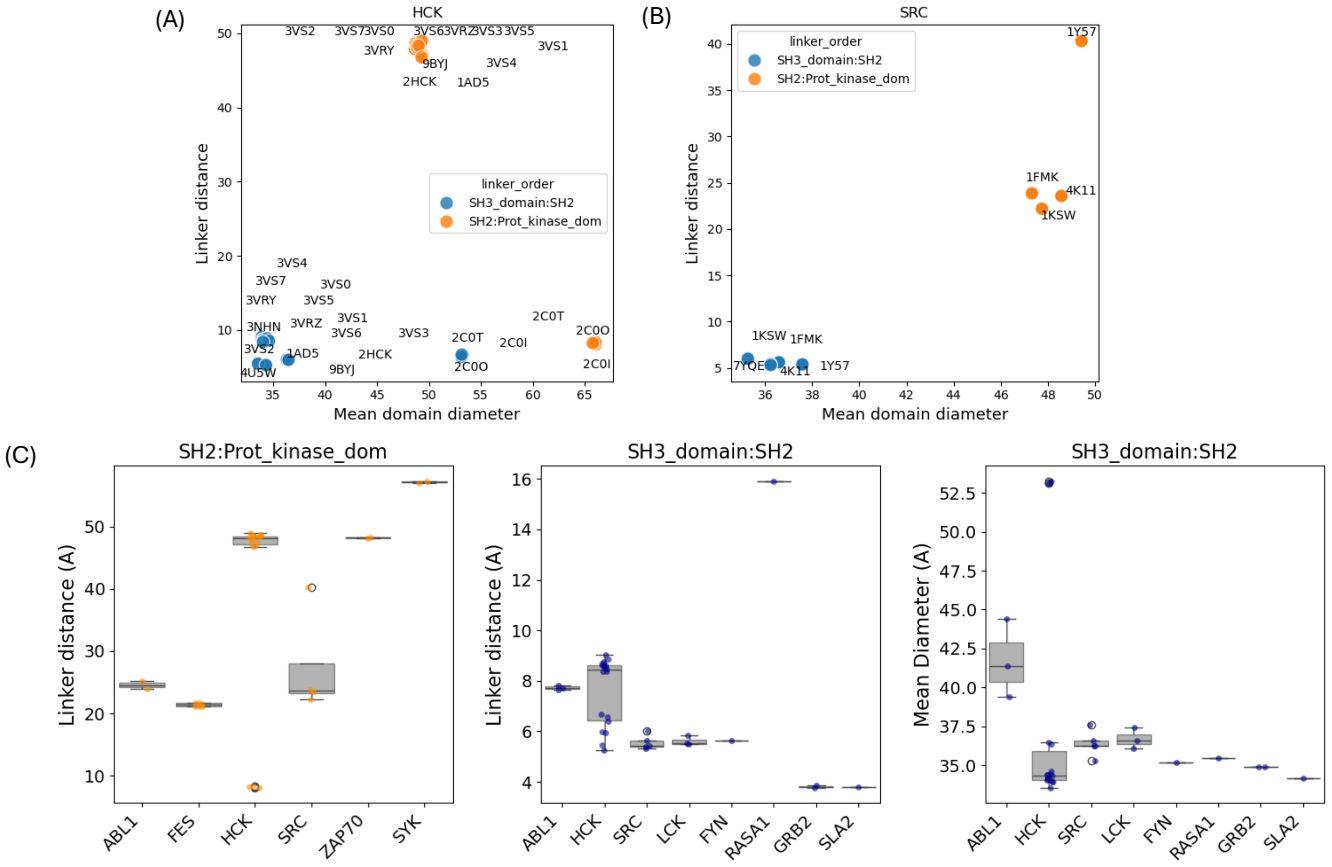

Fig. S11. Assessing outlier structures within each domain-domain linker group. Scatter plots showing the mean domain diameter of the two domains (x-axis) versus the linker distance (y-axis) for (**A**) HCK and (**B**) SRC structures. Each datapoint corresponds to an individual PDB structure and is annotated with its PDB identifier. (**C**) Box plots summarizing linker distances and mean domain diameters for domain-domain linker pairs with multiple available structures, grouped by gene.

Supplementary Data.

Supplementary Data S1: Model parameters used for all SH2 domain predictions. The model parameters used (in parameters.m) and the changes in EffC_Calculator.m used for linker parameters are given for each and every prediction provided in this work.

Supplementary Data S2: Structure-based extraction of tandem SH2 domain linkers and SH2 domain diameters. Sheet 1 of this Excel workbook contains all PDB's analyzed for linker distances according to methods, and the average value derived that was used in modeling. Sheet 2 contains the analysis of all SH2 domain diameters in structures used and a final diameter used in the model that is the average of all values.

Supplementary Data S3: Structure-based extraction of all SH2-relevant linkers. This workbook contains the extraction of all related PDB files that had an SH2 domain with coverage of a second, or more, domain of interest (phosphatase catalytic domain, kinase catalytic domain, or SH3 domain). PDB_filtered indicates the final structures used, eliminating structures with an inhibitor bound, or other outlier issue as indicated in Methods. Domain_linker_summary sheet is the final average linker distance and domain diameters for all pairs of domains containing an SH2 domain across the PDB. Sizes are given in angstroms, and AA indicates length in number of amino acids.

Supplementary Data S4: Bivalent pTyr resource. This workbook contains the full human bivalent pTyr pairs according to methods provided. Sheets labeled Ronan bivalent pTyr and Martyn bivalent pTyr are the reference pairs that had annotated data in the Martyn (41) or Ronan (24) resources. The pair-level data was summarized in Martyn PPI and Ronan PPI sheets – indicating protein-protein interaction level data as accumulated form the pairs of sites.