CoDIAC: A comprehensive approach for interaction analysis reveals novel insights into SH2 domain function and regulation

Alekhya Kandoor[1], Gabrielle Martinez[1], Julianna M. Hitchcock[1], Savannah Angel[1], Logan Campbell[1], Saqib Rizvi[1], and Kristen M. Naegle[1,✉]

[1]Department of Biomedical Engineering and the Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, United States of America.

Protein domains are conserved structural and functional units and are the functional building blocks of proteins. Evolutionary expansion means that domain families are often represented by many members in a species, which are found in various configurations with other domains, which have evolved new specificity for interacting partners. Here, we develop a structure-based interface analysis to comprehensively map domain interfaces from available experimental and predicted structures, including interfaces with other macromolecules and intraprotein interfaces (such as might exist between domains in a protein). We hypothesized that a comprehensive approach to contact mapping of domains could yield new insights. Specifically, we use it to gain information about how domains selectivity interact with ligands, whether domain-domain interfaces of repeated domain partnerships are conserved across diverse proteins, and identify regions of conserved post-translational modifications, using relationship to interaction interfaces as a method to hypothesize the effect of post-translational modifications (and mutations). We applied this approach to the human SH2 domain family, an extensive modular unit that is the foundation of phosphotyrosine-mediated signaling, where we identified a novel approach to understanding the binding selectivity of SH2 domains and evidence that there is coordinated and conserved regulation of multiple SH2 domain binding interfaces by tyrosine and serine/threonine phosphorylation and acetylation, suggesting that multiple signaling systems can regulate protein activity and SH2 domain interactions in a regulated manner. We provide the extensive features of the human SH2 domain family and this modular approach, as an open source Python package for COmprehensive Domain Interface Analysis of Contacts (CoDIAC).

structural biology | SH2 domains | protein interactions | computational biology
Correspondence: ***kmn4mj@virginia.edu***

Introduction

Protein domains, conserved protein sequences that share a common fold and function, are key building blocks for proteins. Recently, it was predicted that the human proteome is composed of 47,576 domains (1), which vary broadly in size and function. For example, domain functions may include binding (e.g. binding to other proteins, RNA, DNA, or lipids) or enzymatic ability (e.g. regulating phosphorylation, proteolysis, DNA repair, or RNA splicing). The recombination of different domains into new protein architectures results in tailoring of existing function (e.g. the addition of a binding domain with an enzyme might enhance localization and enzyme capabilities) or creating entirely new functions (e.g. the recombination of domains that had not existed before). As duplication of domains proceeds across evolution, in addition to shuffling into new protein architectures, the domain sequences continue to evolve, creating new specificity for their targets, enabling new functionality from both the domain-containing protein and new or different sequences (2). Although we learn a great deal about a protein from its individual domain components, there is still an incomplete understanding of domain interaction interfaces and how they drive interactions with other molecules or regulate overall protein function.

Domains are the targets of a vast array of post-translational modifications (3), very few of which have an annotated function. In prior work, we hypothesized that conserved post-translational modifications (PTMs) – PTMs occurring in the same structural position of a domain family across many domains – likely share a common effect across the domain family. Applying this to the heavily modified RRM domain family, we found a conserved tyrosine phosphorylation in the RNA binding motif of the RRM domain (the region that interacts with RNA), suggesting that there may be global regulation of RNA interactions by phosphotyrosine signaling (4). However, our approach lacked the ability to directly connect PTMs comprehensively to the interaction interfaces across domains. Understanding residue-level contacts, such as with ligands or with other possible interaction interfaces (e.g. domain-domain interfaces), might help hypothesize the functional effects of PTMs across the RRM domain and other domains. We were inspired by the work of the Protein Contact Atlas (5), which extracted non-covalent interactions from X-ray structures, creating easy to use adjacency files that list the interaction types between every pair of residues in an experimental structure. This platform generates an exhaustive set of tools to visualize residue-based interaction networks in the context of the secondary elements that are present within the protein structure entities. However, the difficulty in connecting structure with its reference sequence and defined domains within the adjacency files, the absence of a programmatic interface for interacting with the service, and the lack of coverage for NMR, EM, and Al-

phaFold predicted structures (6) by the Protein Contact Atlas led us to develop a modular, Python-based package that can more easily identify relevant structures of interest, define domain boundaries within structures, and integrate contact analysis with post-translational modifications and mutations. All aspects of this pipeline use flat text formats, allowing for easy human- and machine-based interpretation and we use Jalview-based annotation approaches, allowing for integrated visualization. Most of the modules developed within this pipeline are usable as individual components and also in an integrated manner for systematic and a comprehensive analysis of domain families.

Given the centrality of domains to defining and regulating proteins, we believe this pipeline will help enable a broad range of research questions. In this work, we harness this pipeline for the analysis of SH2 domains – a "reader" domain that is central to all of phosphotyrosine (pTyr) mediated signaling. The SH2 domain is interesting, because it is extensive (119 domains in the human proteome on 109 proteins), it co-occurs in diverse protein architectures (e.g alone, with other reader domains, and with a broad range of enzymatic domains), it is a member of a broad class of globular domains that interact with flexible protein ligands (termed domain-motif interactions), and it is extensively post-translationally modified. Although it is well-studied and has decent coverage in the PDB experimentally, structures still only cover a fraction of the "SH2ome" (even lower coverage for structures complexed with ligand) and although much has been learned about some of the determinants that drive ligand binding specificity and binding, including the development of the SH2 domain as a reagent for enrichment of pTyr sites for mass spectrometry pipelines (7), there is still an overall lack of understanding of which SH2 domains ultimately interact with which of the 46,000 pTyr sites in the human proteome (8). Additional, a better understanding of the SH2 domain family as a whole might offer improved methods for increasing the SH2 domain affinity for reagents, while also engineering new specificities for the possibility of developing inhibitors for dysregulated tyrosine kinase signaling (9, 10). Finally, although one structural position of tyrosine phosphorylation has a known function in regulating SH2 domain interactions in SRC family kinases (11–13), relatively little is known about the extensive PTMs on the SH2 domain beyond that position.

Here, we share the details of our generalized contact mapping and domain-centric analysis pipelines, termed COmprehensive Domain Interface Analysis of Contacts (CoDIAC), and the analysis performed using the comprehensive set of contact maps extracted for the human SH2 domain family. We develop an approach to systematically project contacts, including harnessing AlphaFold predictions, to infer contact maps for SH2 domain family members lacking structural data. We recover known insights about the SH2 domain along with new insights. For example, through contact mapping of ligand interfaces, we find that the overall fraction of bonds formed between the SH2 domain and a ligand position is directly correlated with the specificity conferred at that ligand position in selectivity. Additionally, we find that domain-domain pairings that commonly reoccur in the proteome (e.g. the SH3-SH2 combination) have a conserved interaction interface, but generally only when the entire protein domain architecture is conserved. Along with some domain additions or differences causing a rearrangement of a shared interface, we also find that that partial protein structures do not always reflect the interaction interfaces that are natively found in full protein structures. Using conserved structural analysis of PTMs and binding interfaces, we find evidence that both domain-domain and domain-ligand interfaces are extensively modified by serine/threonine phosphorylation, tyrosine phosphorylation, and acetylation, suggesting that the SH2 domain "reader" function might be highly regulated by multiple signaling systems. Finally, the ability to extrapolate across a larger part of the SH2 domain family and consider the multifaceted interfaces of SH2 domains brings new insights into the possible role of clinically-relevant mutations.

Results

CoDIAC: A comprehensive domain-centric pipeline for analysis of contacts. CoDIAC is a python-based package that integrates several resources, allowing for the identification and contact map extraction of all available experimental structures from the PDB (14) and predictions from AlphaFold (6) that structurally cover a domain of interest, defined by the domain's InterPro accession (15). Figure 1A depicts an overview of the CoDIAC pipeline, which starts with building an annotated reference set of proteins and structures, extracts the contact maps, and integrates with other features to produce resource files that can be analyzed, or used for visualization in Jalview (16). CoDIAC uses UniProt (17) as the primary protein reference and pairwise sequence alignment from Biopython (18) in order to align structural sequences with the UniProt reference sequences, for the purpose of identifying domain boundaries and creating features all referenced to a common sequence. In addition to the structurally extracted contact maps, CoDIAC can translate other protein annotations onto the reference sequences (again by pairwise alignment), and we have directly integrated the extraction and translation of post-translational modifications from ProtomeScout (3) and PhosphoSitePlus (19), as well as variants extracted from databases such as OMIM (20) and gnomAD missense mutations (21) that are represented with a clinically significant ClinVar score (22). CoDIAC is available at `https://github.com/NaegleLab/CoDIAC`.

In addition to using the entirety of the CoDIAC pipeline for comprehensive domain-centric analysis, as we do here for the human SH2 domain family, modular components of the CoDIAC pipeline can be used independently for different scientific pursuits. For example, the first part of the pipeline produces highly descriptive flat text files that contain sequences, gene names, and domains (both UniProt annotations and InterPro annotations) for all proteins of any domain of interest, or for all proteins of interest. Hence this could be used to build an understanding of domains across evolution (e.g. by pulling this information from different species of interest), or

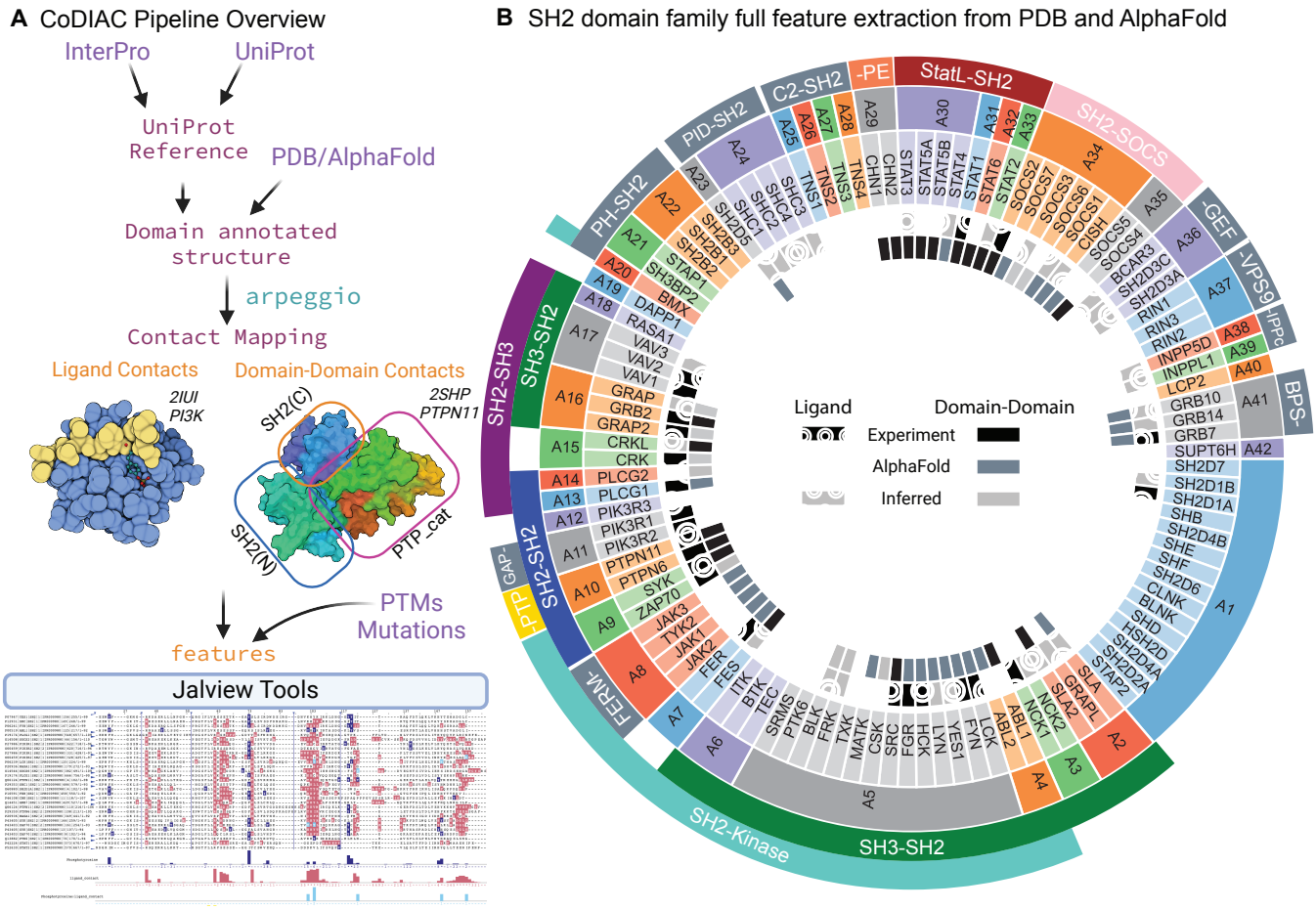

Fig. 1. **A)** Flowchart outlining the key modules of the CoDIAC pipeline that can identify relevant domain containing proteins (agnostic of species), associated PDB and AlphaFold structures, annotate controlled domains from those resources for the extraction of domain interactions between proteins (ligand) and within proteins (domain-domain), and orient PTMs, mutations, and contact interfaces onto a common reference and Jalview-based visualization. **B)** An overview of all data and features extracted for the human SH2 domain family using CoDIAC. The SH2 domain family has been sorted and grouped by shared architectures (A1 for example is a group that contain only the SH2 domain). The immediately adjacent domains to the SH2 domain are labeled on the outermost tracks (e.g. A5 group is an architecture that consists of SH3-SH2-Kinase) and are in color if experimental structures covering the pair of domains exists. The feature generation and inference performed for the SH2 domain family are highlighted in the inner two circles. The innermost circle represents the genes for which we have experimental (black), AlphaFold (slate-gray), and inferred (light gray) features for domain-domain interfaces and the outer circle with hatched pattern represents ligand interface contacts with phosphotyrosine (pTyr)-containing ligands.

to build comprehensive information about a proteome of interest, including the addition of PTMs and mutations, also in the CoDIAC package. The CoDIAC components dedicated to integrating features can be used outside of contact mapping to merge different resources. For example, we used it to translate manually extracted features from prior research, which detailed the function effects of phage display mutagenesis on ligand binding (23) and the surface contacts of the SH2 domain that interact with phospholipids (24). Finally, although our motivation and demonstrated use of CoDIAC here is domain-centric analysis, the extraction of contact maps by the middle layer of the CoDIAC pipeline, is agnostic as to what the structure file is and one can analyze all or parts of regions of contact maps. Hence, the modular and flexible framework of different aspects of CoDIAC provides an open source Python-based toolkit for a variety of protein and domain-based analyses.

CoDIAC contact mapping. A major component of CoDIAC is converting a structure file into a fingerprint of non-covalent interactions within the structure. In the full pipeline, this means performing contact mapping across all PDB and AlphaFold structures that were identified as having a domain of interest. For any one structure, CoDIAC uses the python package Arpeggio (25) to convert structure files to a flat text file of all possible interatomic interactions in the structure. This "adjacency file" contains information about the distance and contact type extracted for each pair of interacting residues across entities (different molecular entities) and chains (replicates of the same entity) in the structure file. Next, we create binary adjacency files indicating if a pairwise residue-level interaction has sufficient (binary value 1) or insufficient (binary value 0) evidence. Evidence is based on user-defined parameters of the maximum distance of an interaction or the type of contact present between residues. Additionally, for experiment-based structures, where it is common to have multiple representations of the macromolecules in different assemblies, it is also necessary to aggregate across these to define relevant contact maps. In this case, CoDIAC uses evidence from across replicates of an assembly to de-

termine whether there is sufficient evidence of residue-level contacts (also a user-defined parameter). For the analysis of SH2 domains in this work, we selected to retain interactions of distances less than 5Å (typically the range for including all types of non-covalent interactions) and we kept interactions if they existed across at least 25% of the chains. For structures containing multiple entities (i.e. domain and ligand entities), we selected to retain contacts when shared by at least 50% of the domain-ligand pairs in the structure. The outcome of this aggregation approach is a final contact map, represented as a binary adjacency file that details interactions, with each line indicating the PDB ID, Residue numbers, Residue IDs, Entity IDs and whether the binary value of whether an interaction was found based on evidence. Contact mapping by CoDIAC can use mmCIF formatted files from any resource, including PDB and AlphaFold, and produces both the adjacency files and binary adjacency files (i.e. contact maps). A benefit of CoDIAC is that it allows for high flexibility, giving control over the parameters for contact mapping (e.g. inter-atomic distance cut-off, intra/inter-protein thresholds, type of inter-atomic interactions), in a common open source programmatic environment and it produces human-readable and machine-readable file formats.

In order to perform analysis for a domain of interest, CoDIAC uses the contact maps generated for all available structures and the annotated structure file, which allows for the exact identification of domain regions in a structure, to analyze the interactions between the domain and other regions of interest. In this work, the regions of interest are between the SH2 domain and other domains within a protein or with phosphotyrosine-containing ligands. CoDIAC contains a module that creates Jalview-style feature files of contact maps – a convenient format for both visualization and analysis (CoDIAC contains a host of tools for reporting and integrating on feature files). Since there is frequently more than one structure for proteins that have been studied experimentally, we additionally aggregate across independent experiments to produce a single set of contact features for every available interaction between the SH2 domain and other domains or ligands. Since we expect that the more structures, the more total features we will find, we hypothesized that requiring some degree of agreement between independent structures would allow us to avoid introducing spurious contacts that results from study bias for some SH2 domains. Measuring the number of features extracted versus the threshold for inclusion across all available structures, we found that total extracted feature loss changes rates (i.e. creates an elbow Fig S1) as we increase the stringency of around a 30% cutoff. We therefore selected a requirement that the final contact features extracted for SH2 domain interfaces are shared in at least 30% of available structures, when more than one structure is available. This final aggregation threshold is also flexible. The outcome of this pipeline is a final set of feature files that indicate which SH2 domain residues interact with either a ligand or another domain in the same protein. These reference-sequence oriented features, along with PTMs, mutations, and other hand annotations then enables an in-depth exploration of how the central domain of interest is possibly regulated by PTMs or mutations and what this means for protein regulation and ligand binding.

SH2 domain analysis of interfaces. We wished to understand how the systematic extraction of domain interfaces from structures could help us understand binding specificity, overlap and conservation of binding interfaces, and infer the function of mutations and post-translational modifications on the SH2 domain. We used the CoDIAC pipeline to identify all available experimentally resolved structures from the PDB and all predicted structures from AlphaFold, that contain a human SH2 domain. Fig. 1B depicts the human SH2 domain family (119 domains spanning 109 protein) as it relates to the protein architectures they are part of and the coverage of interfaces we were able to extract from the PDB, AlphaFold, or by inference through conservation (described later). From the PDB, we recovered 467 total experimental structures, covering 61 unique human SH2 proteins, where 135 PDB structures included a complex with a pTyr-containing ligand. The majority of the structures (101) containing ligands occur as a complex in trans (ligand and domain are different molecular entities) and the remaining 34 are bound in cis, where the ligand is part of the same protein chain (Fig S2). For the purposes of SH2 contact mapping, we did not consider structures that included variants within the SH2 domain, or during domain-domain contact mapping, variants in either of the domains being mapped. Domain-ligand experimental structures provided 78 unique SH2-pTyr ligand pairs and 207 PDB structures across 32 unique proteins included coverage of a domain-domain interface, where 144 of these represent full protein architectures (Fig S3). Although AlphaFold cannot, at this time, give insight into ligand binding requiring modified residues, we recovered 109 predicted structures for full protein architectures for domain-domain interface analysis. We used CoDIAC to identify all of these SH2-domain containing structures and for converting the structures to contact maps for analysis of domain-ligand and domain-domain interfaces. We used PROMALS3D (26) to create a structure-based alignment of the SH2 domain family and all data including the human reference, the PDB and AlphaFold annotated references, alignments, contact maps, features and analysis produced are available on Figshare project (https://figshare.com/account/home#/projects/211453).

Comprehensive ligand contact mapping. We used CoDIAC to extract the ligand interacting residues of the SH2 domain across all available structures. Figure 2A depicts the residue-level contacts extracted for the 28 unique SH2 domains and their ligands from 111 PDB structures. Canonical SH2 domain binding results from a central engagement of the pTyr residue with an invariant arginine in the binding pocket (alignment position 62) along with specificity provided by residues surrounding the pTyr in the ligand interacting with a relatively shallow pocket (27). It has been shown that half the binding energy of the interaction occurs from the pTyr interaction, with the remaining half of the binding energy

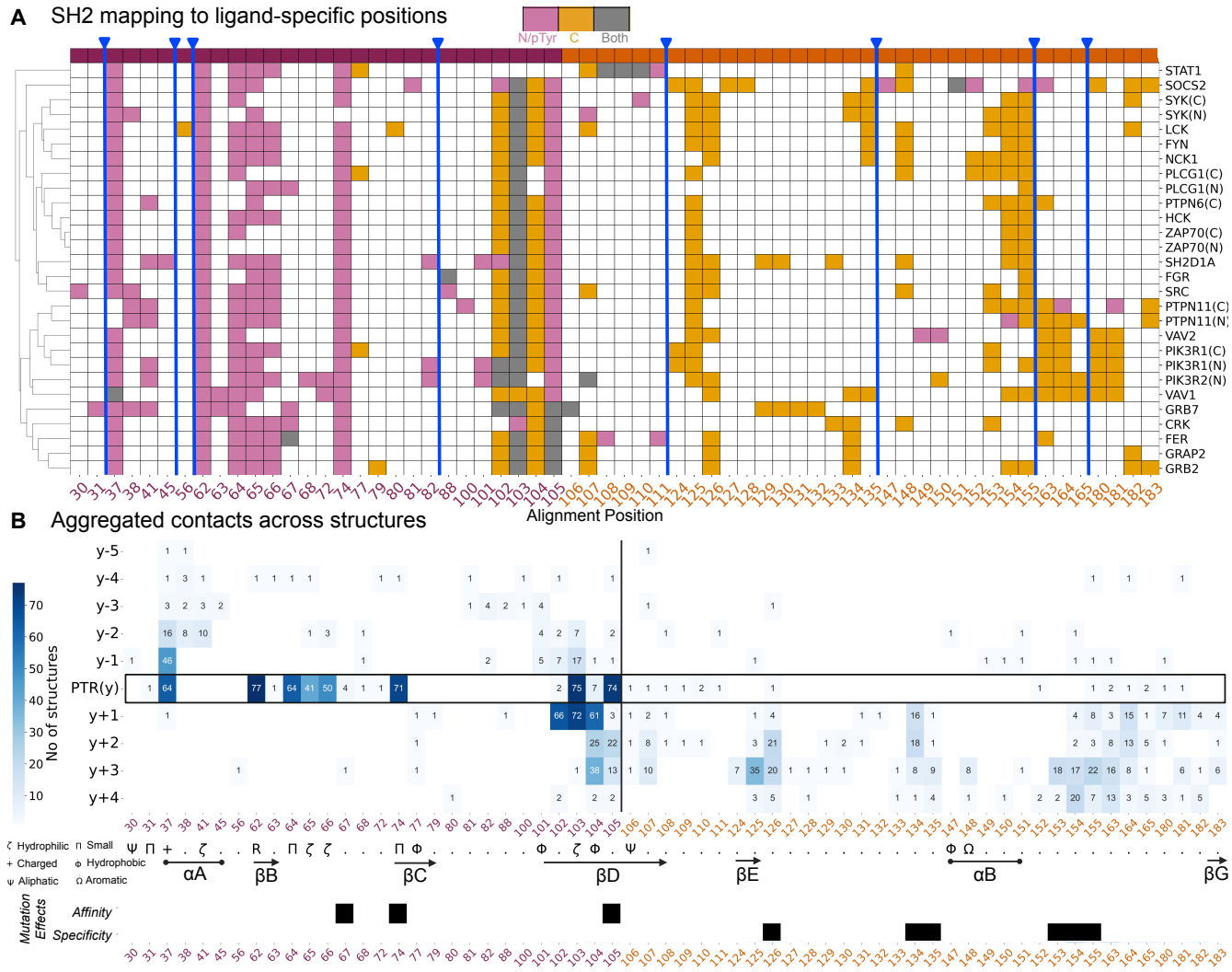

Fig. 2. **A)** A matrix of ligand-based contacts of the SH2 domain relative to the reference alignment of the SH2 domain family (selected examples). Contacts are colored, based on if the residue makes contact to the N-terminal side or the pTyr on the ligand (pink), to the C-terminal side (orange), or to both sides (gray). Position 105 is the most C-terminal position we see consistently interact with the N-terminal portion of the ligand. **B)** Matrix representation of the combined feature set for canonical binding interface of SH2-pTyr interactions, relative to the reference alignment of SH2 domains and oriented by the pTyr for ligands. Features viewed here are aggregated from across 78 unique SH2-pTyr ligand pairs and each feature is projected as an interaction between the residues on domain (x-axis) and ligand (y-axis). Secondary structures and amino acid acid consensus annotations are obtained from the PROMALS3D alignment. Bottom panel: Representative mutations and their effects from phage display studies separated according to whether they increase affinity (i.e. "superbinder") or alter specificity.

contributed by interactions with the nearby ligand residues (28). Systematic analysis of ligand pocket mapping recovers the interaction of the pTyr with the invariant arginine in almost all cases. There were nine exceptions to this interaction, which upon evaluation, we found are for interactions occurring under high local concentrations (Fig. S4). Some of these exceptions were for tandem SH2 domains complexed with bivalent pTyr ligands (SYK, ZAP70, and RASA1), where we saw cases that one of the ligands did not engage the invariant pTyr-arginine interaction (RASA1) or we saw one pTyr making contact with both SH2 domains (SYK and ZAP70). Other exceptions to canonical binding includes a class of interactions occurring in cis, where single SH2 domains (CRK, CRKL) are tethered to their pTyr ligands in the same entity. Although these data are interesting and highlight the possibility that binding of ligands can occur in the absence of the invariant arginine engagement, we excluded these structures for the purposes of contact mapping of the canonical binding pocket.

We explored how important the sequence identity of the ligand was to the mapping of the SH2 domain ligand binding pocket. We encoded the SH2 domain residues that make contact with the ligands according to whether they make contact with the N-terminal or pTyr of the ligand, the C-terminal side of the ligand, or both N-term/pTyr and C-term (Fig. S5 and Fig. 2A). This encoding recapitulates individual structure studies that noted that the first N-terminal portion of the SH2 domain engages the N-terminal and pTyr positions of the ligand and C-terminal SH2 binds the C-terminal ligand positions. We noted that the primary bifurcation point of the transition from the pTyr pocket to the specificity binding region across the majority of domains occurs at position 105 in our reference alignment (Fig. 2A). If we apply a similar encoding to the ligands, indicating on the ligand residue whether it

interacts with residues before or after position 105 (N-term, C-term, or both) of the SH2 domain and cluster these data, we see that ligand interactions are driven by SH2 domain identity and not by ligand sequences (Fig. S6), consistent with a flexible ligand fitting into a rigid globular domain.

Although the specific ligand sequence does not appear to affect the SH2 domain residues mapped, we next asked if ligand length had an effect. We explored the residues mapped across the SH2 domain by the various ligands available – including the residues mapped by just the pTyr Fig. S7 and additional flanking residues (Fig. S5). First, we found that amino acids to the N-terminal side of the ligand pTyr were unnecessary for recovering unique contacts. For example, ligands consistently mapped the same pocket residues with no N-terminal amino acids as those ligands that had N-terminal residues (e.g. cases in STAT1 and PIK3R1(C) domain binding). Second, we often found that even very small ligands could map the same SH2 domain residues as longer ligands. For example, ligands with only the pTyr and two additional C-terminal residues, i.e. +1 and +2 ligand positions mapped the same interface of GRB2 as a ligand that was twice as long on the C-terminal side, suggesting that these additional ligand amino acids did not map novel SH2 domain pocket residues. However, we did occasionally see differences in mapping of ligands that were too small, recovering only the pTyr pocket and some specificity region, but not the entirety of the specificity interaction region (e.g. a five amino acid ligand recovered novel contacts from SRC and LCK that three and four amino acid peptides did not). Given that residues mapped in the SH2 domain are highly consistent across independent experiments and with different ligand sequences and lengths generally capturing the entirety of the pocket, there is strong evidence that the ligands available in the PDB give a robust mapping of the SH2 domain ligand binding pocket, especially as we aggregate the information across available structures for those with length limitations.

Next, we wished to explore more deeply the interactions between SH2 domain residues and their contacts with specific ligand positions. Figure 2B is a 2-dimensional view of contact mapping with SH2 domain positions on the x-axis alignment positions, ligand positions on the y-axis, and a value that indicates the total number of structures that map a contact between the position-specific pair. The largest number of pairwise residue interactions across available structures occurs for alignment position 62, the βB-strand, invariant arginine interacting with the pTyr ligand residue. Similarly, the highest agreement across structures occur for interactions involving conserved structural residues of the SH2 domain interacting with the ligand pTyr residue. For example, positions 103 (polar/His dominant) and 105 (hydrophobic/Lys dominant), both on the βD-strand, engage with pTyr in >95% of the structures (Fig. 2B). Position 37 on the αA helix, occupied by positively charged residues, (arginine being the most dominant among them), engage with the pTyr in 83% of structures. These highly enriched positions that engage with pTyr from our combined feature set agree with the experimentally known details of SH2-pTyr binding mechanism (29–31). We additionally find that the residues at positions 64 (last residue of βB)and 74 (occupied mostly by small residues on the βC) also interact strongly with the pTyr moiety in 85% of the structures (Fig. 2B). As the binding transitions into the C-terminal portion, the diversity of binding modes, which contributes to diversity in binding partners, becomes highly apparent in this matrix view, and again highlights the bifurcation point of alignment position 105 as being the most C-terminal residue of the SH2 domain involved in engaging the pTyr moiety. Additionally, this visualization highlights that the ligand -1 position engagement predominantly occurs in conjunction with adjacent pTyr engagement (alignment position 37) and that the ligand +1 position is concurrently engaged with the same residue at position 103. This suggests that both ligand residues in the -1 and +1 would have to satisfy properties of binding that would also enable the SH2 domain residues at position 37 and 103 to bind the negatively charged pTyr residue. However, given that other SH2 domain residues can independently interact with the +1 residue (but not the -1 residue) such as positions 102 and 104, suggests there may be slight differences in the dual constraints needed for -1 versus the +1 residues. Comprehensive contact mapping across all available structures acts to bridge the knowledge gained from individual structures and highlights the places of domain-specific variability helping drive different ligand specificities.

We next asked if comprehensive structure analysis can describe the results from phage display mutagenesis experiments. These experiments were undertaken by researchers seeking to create SH2 "superbinders" for the purpose of pTyr-pull down reagents (7, 23). When we translate features from phage-based mutation studies on the SH2 domain (Fig. S8), the mutations that lead to increased binding affinity of the SH2 domain (i.e. the "superbinder" mutations) map directly to the N-terminal part of the SH2 domain that coordinates pTyr binding, including positions 74 and 105 – two of the triplet mutations (resulting in 100-fold higher affinity), currently being used to produce improved reagents for pTyr pull down (7) (Fig. 2B). The third mutation in the triple mutant, position 67 in our alignment, is directly adjacent to a conserved pTyr interaction residue, suggesting that immediately adjacent positions are likely also important to shaping the ligand binding pocket. Positions that were found not to be evolvable (i.e. they lose function) include the invariant arginine at position 62, but also include highly conserved positions not captured here as being involved or near binding contacts – likely because they effect the fold of the SH2 domain instead of the binding interface. Finally, positions found to change the specificity of SH2 domains occur at and beyond positions 124, solidly in the variable structural contact region of the SH2 domain (Fig. 2B and Fig. S8). Hence, systematic extraction across all available structures by CoDIAC helps to identify the deeply conserved structural positions of the pTyr pocket and identify the more variable, SH2-specific positions that contribute to the ligand selection of binding across the SH2 domain family. This information could be helpful in de-

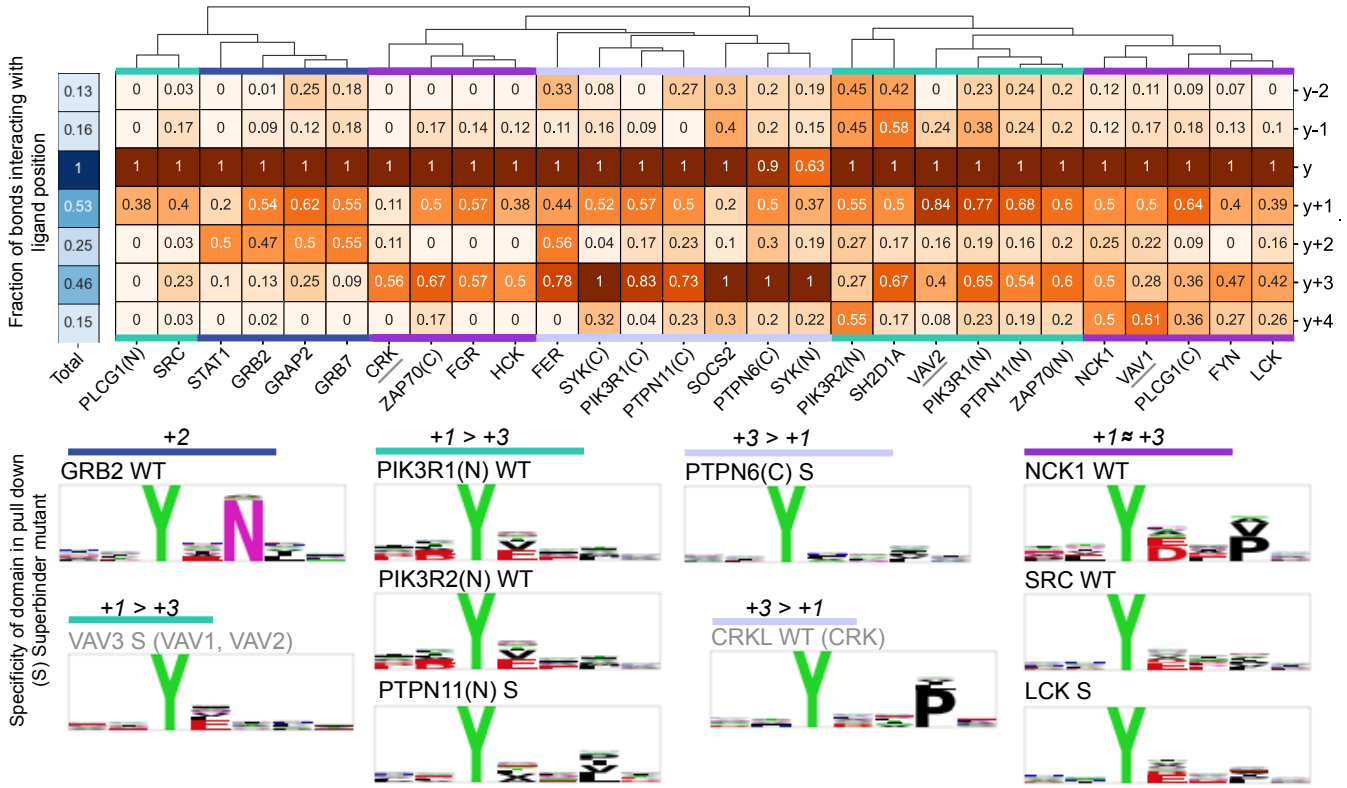

Fig. 3. Testing the hypothesis that the fraction of bonds dedicated to coordinating specific ligand positions is related to ligand specificity. For each domain complexed with ligand in available structures, we normalized the number of residue-level interactions made to ligand positions by the maximum number of residue-level bonds made (almost always a value of 1 for the pTyr interaction). 'Total' indicates the average across all available structures. The matrix was clustered and we used this grouping and relative emphasis on the fraction of bonds to define groupings, such as +2 binding for the GRB2/GRAP2-containing group. Bottom half of the figure are motif logos generated for peptides pulled down from Martyn et al.'s superbinder experiments of stimulated and pervanadate treated Jurkat cell lysates (7). Here, we included any domain with more than 50 peptides and randomly downsampled to 100 total for controlling entropy and WT indicates the pull down is from wild type SH2 domains and S indicates data is from a "superbinder" mutant domain. The emphasis above the bars are based on the fraction of bonds. VAV3 and CRKL are not directly represented by available structure, but have close homologs (VAV1/VAV2 and CRK, respectively).

riving new SH2 domain products, providing residue locations for evolving different affinity or new specificities of SH2 domains.

The relative proportion of residue-level interactions determines importance of ligand binding. In the 2-dimensional structure encoding of the SH2 ligand interface (Fig. 2B), we observed that the total number of residues in the SH2 domain engaged with ligand positions appeared to be highly correlated with known binding engagement – the most interactions occur with the pTyr residue and predominantly with the C-terminal part of the ligand. To explore this further, we extracted the number of residue-level bonds made to each position of the ligand and normalized to the maximum number of bonds observed in any position, predominantly the pTyr position, Fig. 3. Averaging this information across all SH2 domains reinforces the observation that the majority of bonds occurs foremost in coordinating the pTyr residue, followed by the +1 and +3 position, then the +2 position, with relatively negligible average bonds made to the N-terminal side of the pTyr or beyond the +3 side of the ligand. This information is consistent with SH2 domain binding profiling, such as from degenerate peptide libraries, where the C-terminal positions of the ligand, in particular the +3 position, is an important selection criteria for protein-protein interactions between specific domains and their ligands (32, 33).

We hypothesized that the fraction of bonds used to coordinate ligand binding for each domain predicts interaction selectivity. To test this hypothesis, we used data from a recent study that performed pTyr pull down from pervanadate-treated Jurkat cell lysates (7) with SH2 domains, which included testing of superbinder mutations across a suite of SH2 domains. In order to test for overall position-level preferences of SH2 domains, we generated motif logos from pull down experiments for any pull down that had at least 50 peptides. If a pull down resulted in significantly more peptides for any one SH2 domains, we subsampled to 100 peptides to control for entropy differences (we did not see differences on the general motif logos across random subsamples). When available, we used the wild type SH2 domain structure. However, for those domains where the wild type structure recovered too few peptides for useful motif generation, we used data from the triplet superbinder mutations. These positions were shown by the authors not to have a significant effect on binding specificity, which is also soundly in agreement with the identification that these positions are exclusively coordinating the pTyr residue, according to our structure extraction

(Fig 2B). Fig. 3 includes motif logos for all domains that met the information criteria and for which we had a ligand-containing structure. Strikingly, we saw agreement between the fraction of bonds and the specificity induced in binding, based on having higher information content in those ligand positions. Additionally, this allowed a test of the hypothesis that the +1 position amino acid might need to be complementary to the residue 103 position that also binds the pTyr. Most of the motif logos (8/10) show a preference for acidic amino acids in the +1 position, consistent with a physiochemical constraint on this position acting through the shared SH2 domain residue interaction.

Although there is a suggestion that global engagement of SH2 domains rely most on the 0 to the +3, with special emphasis on the +1 and +3 positions, which is consistent with binding behavior of SH2 domains generally, there is diversity in specific engagement of family members – highlighted by hierarchical clustering of this view of the fraction of bonds to ligand positions. For example, although the overall average fraction of bonds made in the +2 position for the available structures is relatively low, for some family members (e.g. GRB2, GRAP2) the fraction of bonds made to the +2 position is significantly higher than most other family members. The GRB2 binding motif is the only motif with a significant +2 determinant, consistent with the emphasis of this being the group member of the +2 binding mode. Additionally, the total fraction of bonds in the +3 position for GRB2 and related members have fractionally low +3 position interactions, compared to other SH2 domains, which is also consistent with prior work that showed that a bulky tryptophan residue in the EF-loop occludes access to the +3 binding pocket area (34). Hence, the individual residue-level bonds appear to be useful for interpreting an individual SH2 domain binding positional preference, along with making relative interpretations in comparison to other family members.

Additional groupings and trends suggest that +1 and +3 are important to most SH2 domains (not including GRB2-group members). However, the emphasis between the +1 and +3 position varies across the SH2 domain family, as does the degree of bonds that reside outside of the 0 to +3 range. For example, we see strong +3 emphasis in the group consisting of PTPN6(C) along with CRK (CRK likely clusters as an outgroup due to very low bonds outside the 0 and +3 position). The motif logos for PTPN6(C) and homolog of CRK, CRKL, suggest that specificity of binding is driven to a larger extent by the +3 than the +1, directly correlated with the fraction of bonds formed with the SH2 domain. Additionally, CRK which distributes relatively few bonds elsewhere, presents has having a stronger emphasis on the +3, compared to the PTPN6(C) domain. This suggests that one can learn about the importance of individual ligand positions, in addition to the conditional context of other ligand positions from the fraction of bonds committed to the interaction at specific residue locations.

Two additional patterns of binding clearly emerge from the clustered heatmap, which includes a group with equivalent emphasis on the +1 and +3 position and a group where +1 is stronger than +3. Motif logos tend to match the bond fraction observations, for example PIK3R1(N) and PIK3R2(N), which have stronger +1 bond emphasis, show higher selectivity for the +1 position then they do for the +3 in their motif logos. Additionally, these domains have the most bond fractions accumulated in the -1 and -2 position (in addition to SH2D1A, which has no pull down information) and these logos are the strongest examples of information content being carried in these positions N-terminal to the pTyr residue. Although we see some occurrence of high fractional residue-level bonds at the +4 position, we saw no motif logos that suggested that +4 position information was selected for specifically. Pascal et al. found that PLCG1(C), along with other Group II binders (SH2 domains that have small amino acids in the βD5, our alignment position 104), allow for an extended interaction with the ligand (up to and past the +4 ligand position) (35). By bond fraction, PLCG1(C) and a few other structures show higher bonds to the +4 position than the remaining family members. However, no motif logos, for these or other members, suggest that the +4 position is discriminatory for binding. This might be explained by the view of contact maps in Fig. 2B, which demonstrates that the +4 position binding occurs more independently than the -1 to +3 positions, i.e. the same residues in the SH2 domain are concurrently coordinating more than one ligand residue between -1 and +3, possibly creating higher discrimination by creating interdependence between multiple positions in the SH2 domain and the ligand. The bond fraction view of PLCG1(C) suggests that the +1 position is very important to binding and this is consistent with the Pascal study, which demonstrated through mutagenesis that only the +1 position provides significant binding energy outside of the pTyr interaction (35), consistent with the +1 PLCG1(C) having high bond fractions to the SH2 domain. Taken together, these results demonstrate that comprehensive analysis across available structures recovers key information about how the SH2 domain coordinates its ligand and how individual domains differ in their ligand binding, but it also highlights that a deeper understanding of how combinatorial residue-level interactions result in ligand partner specificity will likely be needed.

Intraprotein, domain-domain, interaction interfaces. For the purposes of understanding the role of PTMs, mutations, and non-ligand interfaces to the regulation of full protein architectures, we next evaluated the interfaces between SH2 domains and other modular domains within the protein. This analysis also allowed the testing of domain-domain contact similarity occurring between experiments and AlphaFold predicted structures as well as how conserved a domain-domain interface remains as it is reused in different overall protein architectures (e.g. the comparison of SH3-SH2 domain interfaces as SH3-SH2 architectures are reused in adaptor proteins and enzymes). The PDB structures provided contact maps for 9 domain-domain interfaces, spanning 21 unique SH2 domains (Fig S9). Predicted domain interfaces from AlphaFold structures yielded 5 additional domain-domain interfaces, along with interfaces for comparison with the 9 identified by PDB (Fig S10). All together the

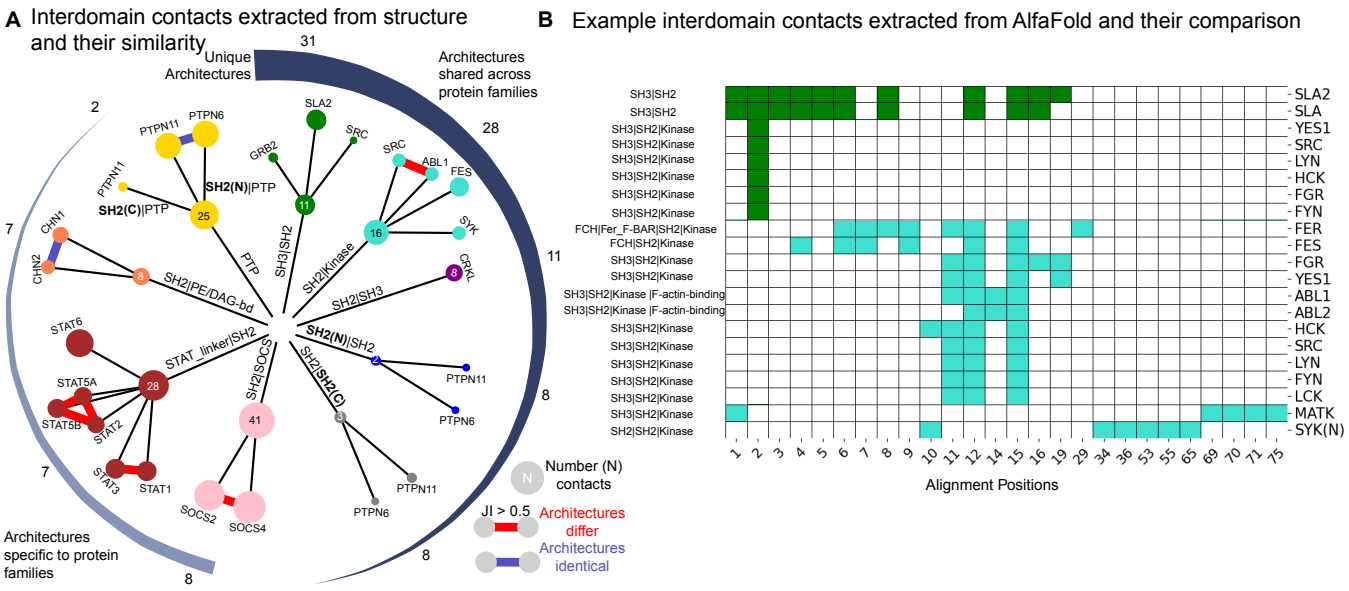

Fig. 4. A) All domain-domain interactions extracted from available structures. The size of the nodes are proportional to the total number of contacts made with that domain interface and the SH2 domain. If the Jaccard index (JI) of the overlap between features was greater than 0.5 for multiple proteins, the line connects those proteins (red if the full protein architectures are identical, blue if the full protein architectures are different). The interfaces are sorted, in a circular format, according to the size of the total number in the proteome (e.g. 31 SH2 domain-containing proteins have an SH3-SH2 pairing). **B)** AlphaFold predictions for the interfaces of SH3-SH2 (dark green) and SH2-Kinase interfaces (cyan) across diverse full protein architectures. Protein architecture is given on the left and protein name on right. Positions on x-axis are relative to the reference SH2 domain alignment.

interfaces we extracted from AlphaFold included 43 unique SH2 domains and their adjoining domains.

To understand the usability of AlphaFold predicted structures, we compared domain-domain contact maps extracted from PDB structures and AlphaFold structures. First, we measured the similarity between the experimentally- and AlphaFold-derived interfaces using a Jaccard Index (JI), the number of residue features shared between the methods normalized by the total number of residues extracted from both methods. Experimental and predicted contacts share a high similarity (JI >0.5) for 16 out of 19 domain-domain interfaces (Fig S11A). The outcome of this comparison prompted us to test whether the low contact similarity occurred due to poor predicted structural fold of the domain-domain interfaces. For this, we used the Predicted Aligned Error (PAE), provided by AlphaFold. A PAE value is available for every residue pair of the protein structure and it represents the distance error between those residues – the larger the PAE value the larger the estimated error. We employ PAE values as an indicator for determining predicted high-confidence regions by measuring the proportion of the domain-domain interface's spatial assembly that has been predicted with a low error (PAE <= 10Å). Three domain-domain interfaces across two proteins had low contact similarity. The SH2-PE/DAG-bd in CHN2 shared only half of the contacts between experiment and prediction, but the average PAE value was high (Table S5). The SH2-SH2 domain interface of PTPN6 also differed – predicted structures suggest that the interface consists of 5 residue-level interactions, but multiple experimental structures only extract one residue-level contact. PTPN6 predictions have poor PAE values on average across the domains of interest (Table S5), indicating that PAE is a useful indicator for the usability of a predicted structure to extract contacts between domains (Fig S11). Therefore, we use interfaces contacts from predicted structures only if our PAE estimate results in capturing at least 50% of the domain-domain interface within the tolerable error range. Since experimental and predicted structures generally have high agreement, and prediction error as a stringency for inclusion can likely reduce error, AlphaFold predictions are likely useful representations of domain-domain interfaces. Hence, we generated contact maps for an additional 13 unique interfaces between SH2 domains and other modular domains from the protein family using AlphaFold predictions.

Having identified a method to use both experimental and predicted structures for domain-domain contact mapping, we next looked at the composition of interfaces across the SH2 family and the similarity of interfaces across proteins with common domain-domain pairings. We see that the degree of SH2 domain engaged in domain-domain interfaces is highly variable. For example, the SH2-SOCS_box interface is the largest, with 41 contacts with the SOCS_box in SOCS protein families, whereas some interfaces consist of just a few contacts, such as the tandem SH2 domain interface in the PTPN family (Fig. 4A). The extent of the interface may be indicative of the importance of overall protein regulation. For example, the extensive contacts made between the SH2(N) domain with the PTP catalytic domain in PTPN6/11 is consistent with the importance of the SH2 domain in sustaining an inactive conformation of the phosphatase (36). Next, we measured the similarity of domain-domain contacts across the family and found consistently high conservation of contact interfaces for proteins that share entire protein domain architectures (Fig. 4A). For exam-

ple, we observe high similarity in the contact interfaces across protein homologs with identical overall protein architectures, including: PTPN11/PTPN6, CHN1/2, STAT5A/B. We also see high conservation of domain-domain interfaces in some families where the overall architectures are high, but for which some families have small alterations (e.g. SOCS2/4, JAKs, GRB7/10/14, and STAT1/2/3 Fig. S9 and Fig. S10). However, where domain-domain pairings (i.e. sub-architectures) occur across highly variable full protein compositions, the domain-domain interfaces are very different. Sub-architecture interfaces such as SH2-SH3 interface within the SLA/SLA2 and SRC family kinases are conserved only among their respective protein families, and not across (Fig. 4B). The SH2-Kinase interface is predicted to have a core of approximately 4-6 amino acids near the N-terminal part of the SH2 domain, and this contact core is highly similar between ABL and SRC family kinases. The high conservation between ABL and SRC kinases also suggests that the addition of the F-actin-binding domain specific to ABL family, relative to SRC family kinases, has a relatively low impact on reordering the overall SH2 structure interactions with the kinase domain. ABL and SRC family kinases do not share the same level of conservation of the SH2-Kinase interface with other protein families such as FER/FES and SYK (Fig. 4B). Hence, we find that domain-domain interface mapping yields information about the extent of domain-domain interactions and that these tend to remain the same under highly homologous protein duplications, with some ability to retain those interfaces under small expansions of protein architectures, but rarely under larger alterations to overall protein architectures.

Given the ability of some domain additions to alter the structural arrangement of shared sub-architectures, we wished to explore how dependent contacts are on the experimental design (i.e. the length of the protein expressed and studied). Consistent with domain additions that cause rearrangements, we find deeply divergent contacts emerge when considering partial protein structures, compared to full protein structures. For example, PTPN11 is a well studied protein structurally, consisting of two SH2 domains followed by a tyrosine phosphatase catalytic domain (SH2-SH2-PTP_cat). When the full protein is expressed and resolved, the contacts between the two SH2 domains and between each SH2 domain and the PTP domain are highly consistent across all available structures. In two structures, where just the tandem SH2 domains were expressed, they are also consistent with each other, but deeply divergent from contacts suggested by full length protein structures (Fig S12). However, in some partial protein structures for STAT and SOCS proteins, we find contact interfaces are minimally affected in the absence of other domains that together would result in a full protein domain architecture (confirmed by high confidence AlphaFold predicted structures). Since structures of partial proteins can result in an incorrect contact map of domain-domain interfaces, we increased the stringency of our contact maps to use the full domain architectures and only use partial protein structures if the domain-domain pairs align well structurally (have low RMSD) within a full protein structure (Table S5B). If there is no experimentally determined full protein structure, we use AlphaFold, but only when the PAE estimates are within the tolerance range. The outcome of this work is a confident set of contact maps between the SH2 domain and other protein domains across the entire SH2 domain family, which can be used to understand more about how the SH2 domain helps to regulate overall protein function and fold (Fig. 1B).

Extending contact maps by evolution and structural similarity. Although there are a relatively large number of SH2 domain structures, the overall coverage of the SH2 domain family is low– only ∼24% and ∼18% of the SH2 domain have resolved experimental structures with ligands or more than one domain, respectively. Therefore, we next sought to develop a method that can confidently group SH2 domains for the purposes of projecting contact maps from available structure to those missing structures. This is especially a challenge for the ligand binding pockets, given the variability of the specificity region of the ligand binding pocket. We used evolutionary distances to hierarchically group SH2 domains, calculated by a Neighbor-Joining method (Fig S13), which identified 26 clusters. Within these clusters, we found evidence that further sub-clustering was necessary. For example, a cluster consisting of PIK3R(C) family and VAV family SH2 domains (Fig S14), there was strong sequence homology within a group, but not across. Since the cases where this occurs could not be systematically determined by a different cut of the hierarchical tree, we performed clustering within the 26 clusters to identify and possible improvements by sub-clustering. This resulted in partitioning the SH2 domains into 58 clusters (Table S1). We also tested clustering using sequence identity alone or structure-based clustering (minimizing RMSD scores within clusters) (Fig S14). The sequence based clustering reproduced identical clustering and structure-based clustering reproduced 79% of the clusters (37/47 sub-clusters are identical that comprises of more than two proteins). We used identical sub-clusters found by evolution-, sequence-, and structure-based approaches, grouping the remaining domains into a sub-cluster (Table S2). For the purpose of inference, we focus on first projecting features within these sub-clusters since these represent highly related domains and later extend our search to related clusters only when we had no scope for projection within the sub-cluster.

To test the ability of our sequence and evolutionary-based groupings for use in projecting ligand contacts, we measured the similarity of extracted contacts within the groups, for groups with more than one structure. Generally, we found good agreement of ligand contacts within groups. For example, (GRB2, GRAP2), (VAV1, VAV2), and (HCK, LCK, FGR) groups shared high residue-based contacts (JI values greater than 0.5, Fig S15). However, the structural contacts of the tandem SH2 domains do not agree with their sequence/evolution-based groupings. In these families, the clustering approach indicates that the between protein domains are more similar (e.g. PTPN6(N) domain is more related to the PTPN11(N) domain than it is to the PTPN6(C)

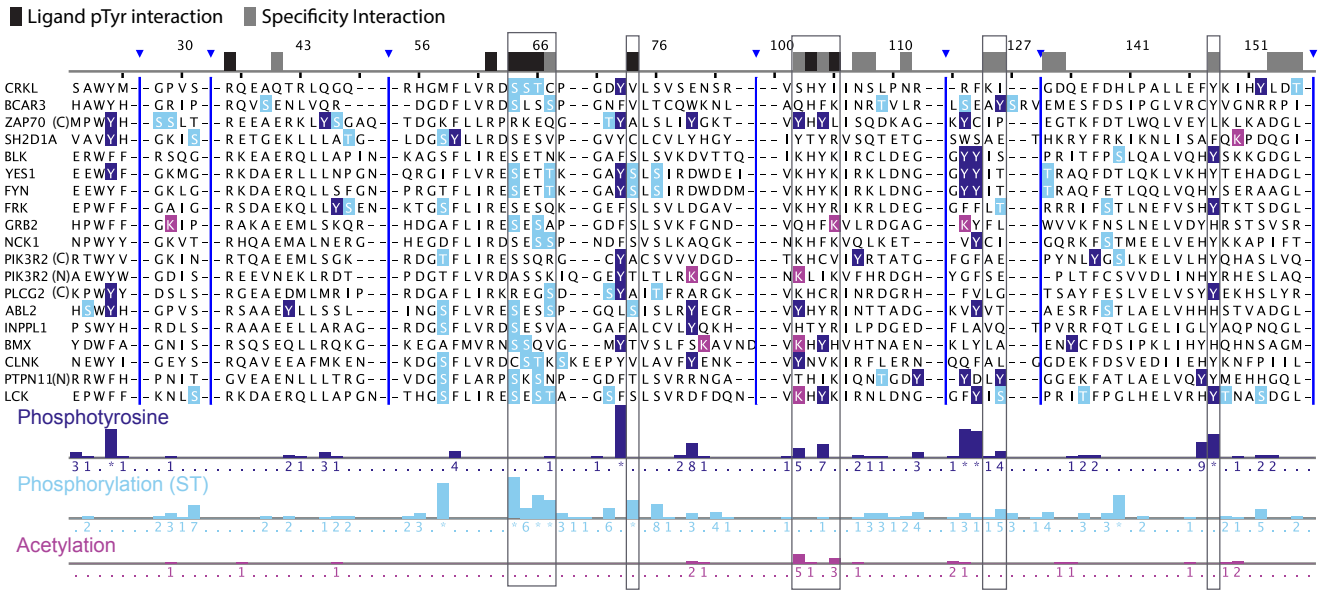

Fig. 5. Post-translational modifications (PTMs) in a Jalview-based visualization on the reference SH2 domain alignment for a subset of SH2 domains that cover the diversity of PTMs. Ligand features were collapsed across structures and indicated if they were shared by 10 or more unique SH2-ligand pairs. Ligand interactions were labeled as interacting with pTyr (black) or specificity (gray) based on whether the interaction was predominantly with the pTyr residue or with other ligand positions (-1 to +3 predominantly). Tracks at the bottom indicate number of PTMs found in that alignment position across the family. Numbers on the tracks indicate the number of PTMs identified in that position, * indicates more than 10 (pTyr: position 73 has 28, position 123, 124, and 148 have 16, 14, 13, respectively; pSer/pThr position 58 has 19, and the positions between 64 and 67 have 22, 6, 13, and 10 phosphorylation sites, 74 has 10 pSer/pThr sites. Dark blue indicates phosphotyrosine; Light blue indicates a phospho-serine/threonine; Maroon indicates N6-acetyllysine. Boxes around alignment and tracks indicate locations of high conservation of PTMs and overlap with ligand interacting residues.

domain). However, the structural contacts actually suggest that the PTPN6(C) domain binds its ligand the same as its PTPN6(N) domain. This pattern held true for all tandem domains for which we had structural data, including: SYK, ZAP70, PTPN11, and PIK3R1 (Fig. S15). Therefore, as exceptions, we clustered the tandem SH2 domains in their own groups (Table S3) and the remaining family is in a final set of 42 clusters for the purposes of ligand contact projection (Table S2). Given the findings of the domain-domain contact interfaces as being most conserved when full protein architectures are shared, we used the protein architecture as the grouping approach for domain-domain contact projection. By using the projections from available structures within a cluster for ligand contacts, and within a cluster with identical full protein architectures for domain-domain contacts from structures or AlphaFold projections, we were able to extend contact maps to 31 (ligand contacts) and 10 (domain-domain contacts) additional SH2 domains, respectively (Fig. 1B). It is possible to lower the constraints for family-identified projections, increasing total coverage, but for the purposes of this work we selected to require high structural and evolutionary similarity for extending contact maps.

Inferring the effect of post-translational modifications by conserved structural analysis. SH2 domains are ubiquitously modified by post-translational modifications. Across the 119 human SH2 domains, 191 pTyr sites, 164 ubiquitinated lysines, 168 phosphorylated serines, 65 phosphorylated threonines, 34 acetylated lysines, as well as a small number of methylations and sumoylation sites have been identified. Despite the abundance of annotated modifications within the SH2 domain, there are relatively few reports regarding the effect of these modifications. One of the examples of prior literature in this area are several independent reports that demonstrate that in at least three of the SRC family kinases, SH2 domain phosphorylation sites alters binding. Phosphorylation of the LCK SH2 domain at Y192 was shown to decrease binding affinity to pY ligands, with the interaction between the SH2 domain and the amino acid in the +3 position of the ligand being the key interaction affected by the presence of the phosphate group (12) and leads to lower LCK activity in TCR signaling (37). Similarly, LYN SH2 domain binding to some peptides was decreased when the SH2 domain was phosphorylated at Y194, regardless of whether the peptide was known to bind with high or low affinity (13). Finally, SRC Y213 phosphorylation was shown to reduce binding of the SRC C-terminal tail phosphotyrosine (Y527), but had no affect binding to another ligand of SRC's SH2 domain (11). These three phosphotyrosines are in the same structural position (alignment position 124) – a position that is immediately adjacent to a conserved +3 binding site at position 125 (Fig. 2A). Therefore, the structural position of these residues, squarely in the specificity determining part of the SH2 domain and proximal to the +3 binding residues of the SH2 domain family, is consistent with the available studies that demonstrate that the SRC family kinase SH2 domains retain their capacity to bind upon phosphorylation of alignment position 125 tyrosine, but alter their specificity, specifically through +3 determinant changes. Given the importance of the +3 generally, this can also likely appear to create a large loss in binding, as seen in the Jin et al. study for LYN Y194 (13). Hence, we hypothesize that integrating PTMs with compre-

hensive structural characterization of the SH2 domain might allow us to more quickly hypothesize function for the myriad of PTMs on the SH2 domain.

We used CoDIAC modules for ProteomeScout (3), PhosphoSitePlus (19), and Jalview to map PTMs onto the reference sequences of SH2 domains and evaluate them within the common PROMALS3D alignment. From this view, it becomes clear that there are patterns of conserved PTMs across the SH2 domain structure, i.e. alignment positions that contain coordinated information of multiple PTMs annotated in a structural position across many SH2 domain family members. We developed reports that assess each PTM and its respect to how many PTMs of that type are in the same structural position (by alignment), the distance of that PTM to a ligand binding interface and number of interface residues within 2 amino acids to the N- and C-terminal, and whether the PTM overlaps directly with a domain-domain or a phospholipid binding interface residue (Table S6). Using this information, we are able to generate hypotheses about the impact of these sites on protein binding and function. For example, due to the size and charge differences introduced by PTMs, we hypothesize that sites nearest to and overlapping with the region of the binding interface that directly interacts with the ligand pTyr may block domain-ligand binding all together, while sites further away from the ligand pTyr are likely involved in specificity/affinity tuning. There are 53 modifications that occur directly on residues that engage the pTyr and hence may result in loss or alteration of the SH2 domain canonical ligand binding capacity. Of these 53 PTMs, 18 are pY, 29 are pS/pT, and 7 are N6-acetyllysine modifications. Interestingly, despite small numbers of overall acetylation currently identified, several occur directly on the interaction interface, in particular on a highly conserved lysine at position 105 which is involved in the coordination of the pTyr (and 5 additional acetylation modifications occur on ligand binding residues in the specificity region). Hence, the removal of the positive lysine from the pTyr coordination by acetylation and the emergence of some degree of conservation may suggest mechanisms of acetylation-based signaling to directly perturb pTyr-mediated signaling. We observe interesting patterns around phosphorylation as well. Conserved phospho-serine/threonine occurs predominantly in the N-terminal half of the SH2 domain and frequently occurs on serine/threonine residues that directly interact with the ligand pTyr, suggesting the possibility that pSer/pThr-signaling might act as a direct method to dampen SH2-pTyr signaling interactions. In contrast to pSer/pThr, where conserved phosphorylation predominantly occurs on residues directly interacting with the ligand pTyr, SH2 domain tyrosine phosphorylation has a tendency to be conserved and close to, or on residues, that bind the ligand in the +1 to +3 region, suggesting that pTyr might help globally regulate or "tune" the specificity of ligand binding, although the most conserved site of tyrosine phosphorylation (alignment position 73 Fig. 5) is immediately adjacent to a conserved structural position that binds the pTyr on the ligand on the βC-strand, suggesting some tyrosine phosphorylation sites in SH2 domains may also act to change binding in a more switch-like manner.

Given the conservation and high degree of phosphorylation at or adjacent to ligand binding residues, we next looked for studies regarding these modifications and found evidence in literature that some of these pSer/pThr sites (38) and pTyr sites (39) have been shown to have an effect on binding and signaling in a regulated manner. Lee et al. studied two phosphoserine sites in the tandem SH2-containing protein PIK3R1 (p85a), in the same structural position of both domains (alignment position 65, a site of conserved pSer/pThr). Upon phosphorylation of either one of the sites, SH2 domain ligand binding was dramatically decreased and PI3K signaling in cells was abrogated (38). When both sites were phosphorylated, PI3K dissociated from its upstream activators, leading to decreased AKT activation. In total, this region of the SH2 domain contains 51 annotated phosphoserine and phosphothreonine sites that directly interact with ligand binding, suggesting that all might have similar effects across many SH2 domains. Similarly, for tyrosine phosphorylation, we recover the previously studied SRC family kinase regulatory sites by using conservation of modifications in structural positions – alignment position 124 contains 14 modified sites and extends well beyond the SFK SH2 domains (e.g. including ZAP70(C), NCK1, and ABL2, Table S6). This position is adjacent to conserved specificity residue contacts at positions 125 and 126, consistent with phosphorylation at 124 effecting +2 and +3 interactions (Fig. 2B). Additionally, we note that many SH2 domains contain a pair of tyrosines, both known to be phosphorylated (positions 123 and 124), including some of the SRC family kinases. Weir et al. evaluated three tyrosines in FYN (Y185, Y213, and Y214, corresponding to alignment positions 73, 123, and 124 (all sites of conserved tyrosine phosphorylation), concluding that these SH2 domain phosphorylation sites lead to reduced binding capacity of the SH2 domain (39). Taken together, these studies highlight the power of comprehensive ligand binding contact mapping as one considers the effects of PTMs. Using this approach, we hypothesize that 54% and 35% of the SH2 domains can be regulated in their binding to ligands by tyrosine and serine/threonine phosphorylation, respectively, and can include multiple points of regulation on any one domain, including both tyrosine and serine/threonine phosphorylation.

Beyond the ligand binding interface, we tested whether we could hypothesize effects of PTMs by their appearance at domain-domain interaction interfaces, possibly suggesting larger protein regulation and control. In this case, given the protein architecture-specific nature of these interfaces, we required PTMs occur directly on a domain-domain interaction interface, which was somewhat rare, compared to ligand interfaces (40 total PTMs occur on SH2 domains at a domain-domain interface, Table S6). However, using this approach we recover key insights into regulation of larger protein structures. For example, Burmeister et al. showed that phosphorylation of two sites in PTPN11, PTPN11(N) T73 and PTPN11(C) S189 (alignment position 140) by PKA, inhibits catalytic activity of PTPN11 (40). The authors proposed that phosphorylation inhibit phosphatase activity by stabilizing

the closed conformation of PTPN11 and by altering binding to ligands (this was done by phosphomimic mutagenesis to aspartic acids). CoDIAC contact map-based analysis finds that PTPN11(N) T73 falls at the SH2-PTP catalytic domain interface and PTPN11(C) occurs at an SH2-SH2 domain interface, highlighting their possible importance in overall protein fold and consistent with prior studies. All together, we identify 28 post-translational modifications occurring on SH2 domain interaction interfaces with other domains (Table S6), highlighting the possibility of producing directed hypotheses for functional tests.

Beyond the ligand and domain-domain interaction interfaces, some SH2 domains have cationic patches that have been shown to interact with phospholipids (24). We hand annotated these phospholipid binding features from Park et al. (24) and used CoDIAC to translate them onto our integrated alignment of PTMs and contact maps. Unfortunately, the small cationic patches are not conserved (24), so we were only able to characterize about 10 SH2 domains with phospholipid binding residues. With the data that is currently available, we hypothesize that five pTyr sites across four SH2 domains may play a role in altering lipid binding specificity due to their immediate proximity to positively charged amino acids found at the phospholipid binding interface (Table S6). Interestingly, several of phospholipid binding residues (6 total lysines) have been directly annotated as undergoing ubiquitination (Table S6). Ubiquitination might play a role in modulating phospholipid binding, but may also be explained by its surface accessibility.

Integrated analysis of clinically relevant mutations. CoDIAC can also be used to connect mutations to regulation of SH2 domains, similar to the PTM-centric analysis, assessing the relationship between mutations and interaction interfaces, as well as their relationship to PTMs in SH2 domains (Table S7). We identified clinically significant mutations from different databases (OMIM, gnomAD, and PDB) that occurred on human SH2 domains (111 total mutations). All together, we found 29 mutant sites to occur on a ligand contact, 40 within 2 amino acids of ligand binding, 16 at a domain-domain contact, and 9 on a residue known to be modified. When we consider the severity of the mutation, we see that 83%, 85%, and 44% of the mutant sites resulted in a mutation that drastically changes the physiochemical properties of the residue across ligand, domain-domain contacts, and PTMs, respectively (Table S7). By integrated analysis, we see a range of possible effects. For example, SRC K206L (alignment position 105) results in loss of a positive charge at a ligand-pTyr coordinating residue, STAT1 K637E (alignment position 134) results in a charge switch at a ligand binding residue (specificity region) and loss in a lysine known to be ubiquitinated, and PTPN11 Y62D (alignment position 112) is adjacent to the ligand binding interface, at the PTP catalytic interface, and removes the ability to regulate a known tyrosine phosphorylation site. PTPN11 Y62D and mutations at Y63 (also a residue at the PTP catalytic interface and known site of tyrosine phosphorylation) have been associated with several pathologies such as Noonan syndrome, Leopard syndrome, and RASopathy (ClinVar Reference Accessions : VCV000013329.48, VCV000013333.86). Mutations known to disrupt the SH2-PTP interface in PTPN11 increase catalytic activity by disrupting the closed conformation of PTPN11 (41). CoDIAC analysis reveals that these PTPN11 mutations not only lie at the SH2-PTP interface, but that they are also adjacent to ligand binding residues and they may act to disrupt the role of tyrosine phosphorylation. In particular, Y62D may effectively be a form of PTPN11 that results from the phosphorylation, mimicking the negative charge of a phosphorylation at Y62. These possible effects are significantly easier to synthesize from the comprehensive aggregation of information across the SH2 domain family, including the appearance of conservation of mutations in shared positions at or near ligand binding interfaces, suggesting that clinically relevant mutations frequently occur on or immediately adjacent to key ligand binding residues (Table S7).

Another example of the power of integrated analysis, especially with the ability to pull information from AlphaFold and project contacts by conservation, reveals interesting insights into SOCS1 mutations. In SOCS1, P123R and Y154H mutations have been shown to activate JAK-STAT pathway in tumor cells and have also been associated to B cell lymphomas (42). In our contact extraction, we found that P123 (alignment position 101) is directly adjacent to conserved ligand residue +1 binder 102 and is on the same side of the βD-strand as pTyr/+1 interacting residue 103 (Fig. 2B) and that Y154 (alignment position 148) is both a ligand binding contact and interacts with the SOCS_box domain. Additionally, although Y154 in SOCS1 has not been identified as being phosphorylated at this time, it is aligned with a conserved position of tyrosine phosphorylation (alignment position 148), which is predicted to alter binding specificity (this tyrosine directly interacts with ligand +3 residues). That this site is also a rare event of the ligand pocket residues also making intraprotein interactions, it suggests that phosphorylation could play a role in opening the SH2 domain for ligands. This tyrosine is part of the doublet 'YY' that undergoes double phosphorylation (by conservation) and mutation, which in addition to altering a modifiable residue, could also impact the regulation of the first tyrosine. Hence, the Y154H mutation could operate on three independent functions: 1) disruption of the SH2-SOCS_box interface and alteration of catalytic activity, 2) changing ligand specificity, and 3) altering the possible regulation of the ligand and domain-domain interface by tyrosine phosphorylation. Though previous work hypothesized loss of ligand contacts (42), this work can exactly map the specific interactions that are affected and highlights new possibilities for the role of that mutation based on CoDIAC generated contact maps and projections and integrated feature analysis.

Discussion

CoDIAC is a flexible framework for extracting contact maps from structures (experiments and predictions), for annotating domains covered within structures, and for harnessing all available structures for a protein domain family. Here,

we focused on using it to explore the myriad of interaction interfaces of modular domains, creating a comprehensive overview of the SH2 domain and the intersection of post-translational modifications and mutations with SH2 binding interfaces. In addition to recovering information that has emerged from individual structure-based studies, the comprehensive evaluation across the entire family identified emergent properties of the SH2 domain. However, a key limitation of this, and all structure-based studies, is that the interaction interfaces (particularly domain-domain contact mapping) is based on a single low energy configuration of the protein and does not capture the dynamic nature of protein interaction interfaces. However, this low energy state is still likely relevant for considering how mutations and PTMs may affect a protein configuration. Although we did not consider the role of linker regions in regulating overall protein function in this work, a well known phenomenon in most of the SH2 domain containing tyrosine kinases (43), CoDIAC can easily be used for these purposes – interdomain regions are also effectively annotated by the CoDIAC pipeline. Instead, we chose here to focus on the modular interfaces that were shared by more of the SH2 protein family. Additionally, there are some protein structures that note that domain-domain interfaces are altered in active conformational states, such as an SH2-kinase interface that occurs in active ABL, which is thought to help stabilize the open conformation (43). This interface is clearly also important, including the possibility that PTMs at these locations could increase or decrease the stability of this conformational state. By choosing to select contacts shared by the majority of structures, seeking to reduce study bias in the final maps, we also biased the contact maps towards those structures that are most represented (i.e. mostly the inactive conformations of proteins). However, that comprehensive extraction allows for the direct identification and comparison between structures and conservation between highly conserved family members, means that CoDIAC can easily be used to compare and contrast interfaces for such purposes. In addition to the ability of using CoDIAC to map interfaces in a protein that include non-domain regions and to assess the variety of interfaces available, CoDIAC can also be extended to map interfaces with other kinds of partners, such as lipids, small molecules, RNA, and DNA.

Although structures are powerful for many applications, a few additional caveats are important to consider, which were highlighted by this study. For example, structures of partial protein coverage should be used carefully. This became evident from both differences in contact maps from proteins with representations of partial and full protein coverage and differences in contact interfaces between subarchitectures that are shared between diverse families (i.e. that the addition of new domains can cause structural interface rearrangements). Additionally, a caveat of ligand contact mapping from structures arose from our observation of interaction interfaces that changed when ligands were presented in cis, on the same chain, or via a multivalent interaction (such as in the tandem SH2 domains). That SH2 ligand engagement can occur in the absence of the canonical pTyr-invariant arginine interaction suggests that contact mapping from structures of isolated domain-ligand pairs may be somewhat limited to understanding physiologically-relevant binding interactions in signaling networks. Beyond the tandem SH2 domain family, many domain-ligand interacting modules work in concert with other modules (e.g. the SH3-SH2 pairing found across a large part of the family), suggesting that this type of non-canonical binding could be seen across a much larger part of the protein interaction space then what might be suggested by isolated domain studies.

Finally, conservation analysis of PTMs within SH2 domains highlighted patterns that may provide insight into SH2 domain regulation and suggests that there are extensive mechanisms by which a number of signaling systems directly regulate interaction interfaces of the SH2 domain. Some of these insights are potentially surprising, contradicting a prevailing thought that modifications primarily occur in intrinsically disordered segments – many of the conserved sites fall directly on β-strands within the SH2 domain. The breadth of these PTMs, both across independent experiments and across homologs, along with data that suggests they are regulated directly by drugs (44), growth factors (45, 46), or other stimuli (47), builds strong evidence that these structurally conserved modifications are likely transiently regulated, important to signaling, and not simply a function of false positives from mass spectrometry-based experiments. Additionally, literature that has described the effect of modifications in conserved positions, on sites not yet annotated in databases, also hints at the likelihood that other modifiable residues in these conserved positions may be found to be modified in the future. Although we highlighted individual PTM effects, as they relate to regulating ligand or domain-domain interactions, there is some evidence that they co-occur and offer multifactorial control of the SH2 domain. One particularly intriguing pattern we noticed were the many instances of phosphoserine sites falling directly adjacent to phosphotyrosine sites near key ligand binding regions, including alignment positions 72 and 73, which directly interact with the ligand pTyr. Recent work profiling the entire complement of human serine/threonine kinases revealed that priming phosphorylation may be a relatively common motif for driving subsequent phosphorylation on serines and threonines (48). Using ProteomeScout, we looked for evidence of whether both modifications have been observed and found a doubly phosphorylated tryptic fragment of those positions in the SH2D1A SH2 domain in breast cancer samples (49), suggesting that one may be involved in priming the kinase motif for the other. Although there is still much to be tested about PTM-regulation of SH2 domain interactions, the comprehensive integration of structure and PTMs enables the more direct hypothesis of the effect of a PTM and the prioritization of these based on the nature of the conservation and impact on protein function as well as identifying a myriad of effects that should be considered with respect to considering disease-relevant mutations.

Materials and Methods

UniProt, InterPro, and structure reference generation of the SH2 domain family. We developed and used the CoDIAC 'InterPro' module to gather all SH2 domain containing proteins for 'Homo Sapiens' with the InterPro ID 'IPR000980', resulting in 109 UniProt identifiers. We then used the 'UniProt' module of CoDIAC to build reference files of the proteins and a reference FASTA file of just the SH2 domain regions for the 119 unique SH2 domains. Given that there can sometimes be discrepancies around specific domain boundaries, we included the ability to alter the boundaries (systematically for all domains) when producing a FASTA sequence file of domains. Based on alignment quality, we found that truncating the boundary of the SH2 domains defined in the reference by one amino acid on the C-terminal resulted in a significantly better alignment and we selected an 'n-terminal offset' of 0 and a 'c-terminal offset' of -1. Additionally on domain quality checks, given its high specificity for only the PIK3 regulatory family, we manually removed the InterPro region in the PIK3R1/2/3 proteins 'PI3K_P85_iSH2:IPR032498', which is an alpha-helical region of between the tandem SH2 domains. We also manually removed an InterPro defined domain in SUPT6H that indicates (Spt6_SH2_C:IPR035018) as it overlapped with the parent SH2 family. Finally, we found that using the SMART (50) defined boundaries, which were an average between UniProt and InterPro domains for the atypical SH2 domains (JAK family) resulted in a better overall alignment. At the time of this analysis (June 2024), UniProt returned 1,477 experimental structures associated with the UniProt set of proteins.

We created the structural reference datasets containing structures for SH2 domain proteins obtained through experiments (Integrated PDB reference file) and predictions (AlphaFold reference file). For the generation of this dataset, we used CoDIAC's 'PDB' followed by 'IntegrateStructure_Reference' module that captured an exhaustive list of experimental details for each of the PDB identifiers associated to the UniProt IDs of human proteins containing an SH2 domain. Using CoDIAC's 'IntegrateStructure' module, which aligns structure data with the reference data and annotates domains found within structures, we found 467 total structures for analysis that covered an SH2 domain in its entirety. Gaps, and variants of experimentally-derived sequences, relative to reference, were noted in the annotated file for consideration of exclusion criteria during contact mapping. We performed an identical process of capturing all predicted structures for the UniProt IDs identified in the family, annotating their regions relative to the reference and generated the AlphaFold reference file. The 109 predicted structures were downloaded as mmCIF files using AlphaFold database version 2 (v2). We used CoDIAC's 'PTM' module to capture PTMs found within the domain boundary regions of SH2 domains for all PTMs for which there were five or more PTMs in the family, producing a unique set of PTMs from both ProteomeScout (3), using the ProteomeScoutAPI (51), and PhosphoSitePlus (19) using an API that we created and incorporated in CoDIAC that operates similarly to the ProteomeScoutAPI, then combined them and kept the unique set of all PTMs for final analysis. We used CoDIAC's 'mutations' module to gather mutations within the SH2 domain regions defined in the reference from gnomAD and OMIM. We used PROMALS3D (26) to align the reference SH2 FASTA file and then used CoDIAC's 'Jalview' module to translate and integrate feature files produced from PTMs along with contact map feature files, along with creating annotation tracks of features based on summing features along the columns of the alignment. Integrated features and annotation tracks were used to create analysis reports concerning PTMs and mutations (compiled in Tables S6 and S7, respectively). All data derived from this pipeline are available at https://doi.org/10.6084/m9.figshare.26321968 as well as on our GitHub repository, where we also have provided all Python code we developed for SH2 domain analysis at https://github.com/NaegleLab/SH2_contact_analysis.

Adjacency File and Contact Map generation. We generated binary text files using the 'AdjacencyFiles' module that are a simplified representation of interatomic interactions. To make these files, we first make these calculations using a Python package Arpeggio (25), which outputs a JSON-formatted file comprising all types of interatomic interactions that exists within an input protein structure. For the purposes of domain focused contact extraction, we kept all non-covalent interactions (aromatic, carbonyl, hydrophobic, ionic, polar, van der Waals, halogen, and hydrogen bond) occurring at a distance < 5Å between atoms that reside on SH2 domains and their interacting domains/ligands. We do not include any interactions that may occur between protein entities and small molecules in this work. All components (chains, residues, residue positions, atom pairs, distance, and contact type) of the filtered non-covalent interactions are saved as adjacency text files and using our set threshold (retain interactions if found in >=25% of the chains for domain-domain analysis and in >=50% of bound chain replicates for domain-ligand complexes) we aggregate contacts across chains of specific entities. We represented whether the contact exists (binary value '1') or not (binary value '0') between or across protein entities in the final binarized adjacency text files. We were unable to generate Arpeggio json files for 12 PDB structures (PDB IDs: 7UNC, 7UND, 8H36, 8H37, 8OEU, 8OEV, 8OF0, 6GMH, 6TED, 7OOP, 7OPC, 7OPD) and the adjacency files we successfully generated for 455 PDB and 109 AlphaFold structures generated in this work can be found here (https://doi.org/10.6084/m9.figshare.26309674). Across the 475 PDB structures, we identified 17 unique post translational modifications (PTM). We generated a PTM dictionary that is used by the 'contactMap' module to replace these modified residue three letter codes by their native residue's single-letter abbreviation in the structure sequence and allowed for precise comparisons between structure and reference sequence. This PTM dictionary can be modified based on the ones extracted within the structure datasets.

We used CoDIAC's 'contactMap' module to generate contact maps for domain-domain and ligand binding interfaces of SH2 domains. This module creates a Python object that stores all the structural information gathered from both the structure datasets (PDB/AlphaFold reference files) and the binary adjacency files. By incorporating annotated and interaction data for each structure, contact maps were constructed in the form of nested python dictionaries whose outer and inner dictionary keys indicate the residues that are non-covalently linked and the values for the inner dictionary represents the binary value of the interaction. We printed contacts from these dictionaries to Jalview supported feature files that bind SH2 domains and its other interacting domains/ligands. Using the 'analysis' module we aggregated contacts across various PDB structures that represent an identical interface. We evaluated after what threshold (percent) the loss of features upon aggregation becomes minimal and found 30% to be suitable threshold for merging contacts across SH2 domain containing structures (Fig S1).

We found 207 PDB structures spanning 32 unique proteins that allowed us to examine domain-domain interfaces for contacts. We removed PDB structures for LCK, PLCG1, ZAP70 from our analysis since we found structural discrepancies while evaluating RMSDs and were unsure whether the structures represented a native conformation. We do not extract contacts across SH2 domains and other modular domains within PIK3R1, CRK and SUPT6H structures. Given these removals and the evaluation of wild-type structures for domain-domain binding interfaces resulted in maps for 19 unique SH2 domain proteins (Fig. 1B). We failed to generate contact maps due to a mismatch between the structure and reference sequence alignments for two of the HCK structures (PDB IDs: 2HCK and 1AD5), due to a 10 amino acid unmodeled region within their kinase domains.

PAE estimation for predicted structural fold of domain-domain interfaces. AlphaFold prediction system provides Predicted Aligned Error (PAE) values for every residue pair in a protein structure that reflects the distance error (Å) between the residues oriented in three dimensional space. Given a domain interface of interest, we first retrieve the PAE values between the residue pairs r_i - r_j, where r_i and r_j reside on either of the domains. Since AlphaFold presents the PAE values in the form of an asymmetrical matrix, we calculate the average of the PAE estimates for the two relative structural configurations of the domain-domain interface (i.e. from both directions of the interface).

$$PAE_{avg} = \frac{PAE_{12} + PAE_{21}}{2} \quad \textbf{(1)}$$

Where PAE_{12} is the fraction of total number of residue pairs (r_i - r_j) that have a low PAE (<= 10Å), given r_i is on the SH2 domain and r_j on the partner domain and PAE_{21} is the estimate of the same interface when r_i is on the partner domain and r_j is the SH2 domain.

We report the final PAE estimate as this average value calculated from both directions of the domain-domain's structural orientation. We consider assessing interfaces only if our PAE estimate results in capturing at least 50% of the domain-domain interface within our defined tolerable error range.

Methods for measuring the evolutionary and structural properties of SH2-domains. First, we used evolutionary distances as the basis to classify SH2-domain containing proteins into families. We implemented a distance-based Neighbor-Joining algorithm to calculate the evolutionary distances among all the known 119 SH2 domains. We then applied Ward's method on the distance matrix to hierarchically cluster the proteins. Using the elbow method identified an optimal linkage distance for clustering – elbow point occurring at 0.5 indicates a decrease in variance within the clusters beyond this linkage distance (Fig S16). This threshold produced 40 clusters (Table S4), but we were unable to sufficiently project contacts, so we reassessed clustering for this purpose at the next elbow point that occurs at linkage distance 0.8 that produced 26 clusters (Table S1). The clusters generated from a threshold of 0.5 and 0.8 were highly similar. Each of these 26 clusters were subdivided into 58 sub-clusters and we used silhouette score to identify the number of sub-clusters that can be formed within each of these clusters. We benefited by working with a higher linkage threshold of 0.8, since it did not affect the protein grouping and allowed us to find more closely related neighbors for enhancing the chances of contact projection. We gained features for 9 additional proteins at a threshold 0.8, compared to what we would achieve with 0.5. Upon performing these steps, we were able to split the SH2 domain family into 26 clusters and 58 sub-clusters (Table S1). We next validated clusters by closely examining the sequence and structure of the domains to evaluate whether the evolutionary-based agglomerative clustering would be beneficial in identifying and separating intra-clustered homologues efficiently. For primary sequence comparison we calculated amino acid sequence identities between proteins within each of these 26 clusters. We hierarchically grouped each of the 26 clusters based on the sequence identity scores. This resulted in exactly the same sub-grouping we had attained using evolutionary distances.

For structural comparisons, we estimate RMSD values using a PyMOL utility 'super' that involves sequence-independent programming alignment. We make these calculations between several structures by using Python controlled PyMOL scripts. The experimental structures do not cover the entire SH2 domain family, so we covered for the unresolved protein domains by incorporating the predicted structures of the domains from AlphaFold, since we observed very low RMSD scores (<2Å) between the SH2 domain structures from experiments and AlphaFold (Fig S17). We next repeated similar hierarchical clustering method for the 26 clusters to gain the intra-clustered structural homologues using the RMSD scores.

pTyr ligand peptide clustering. We represented every pTyr peptide in our dataset as DPPS vectors (52) and then measured the average Euclidean distance between these DPPS vectors of the peptides. These average Euclidean distances indicate how closely associated the peptides are by

their physiochemical properties. We hierarchically clustered the peptides based on these average distances. We adapted this method in our work from (8) to categorize pTyr peptides of length 7 amino acids (-2 to +4 of pTyr).

ACKNOWLEDGEMENTS

We would like to acknowledge: Margaret Ryan for early work on PDB interfaces for identifying relevant SH2 domain containing proteins; Dr. Eli Draizen for help on initial scripts for parsing and contact extractions using mmCIF files; and Dr. Philip Bourne and Dr. Mohammad Fallahi-Sichani for helpful discussions. Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM138127 and the National Institute of Allergy and Infectious Disease under Award Number R01AI153617. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Bibliography

1. R. Dustin Schaeffer, Jing Zhang, Lisa N. Kinch, Jimin Pei, Qian Cong, and Nick V. Grishin. Classification of domains in predicted structures of the human proteome. *Proceedings of the National Academy of Sciences of the United States of America*, 120(12):e2214069120, March 2023. ISSN 1091-6490. doi: 10.1073/pnas.2214069120.
2. David Jakubec, Miroslav Kratochvíl, Jiří Vymětal, and Jiří Vondrášek. Widespread evolutionary crosstalk among protein domains in the context of multi-domain proteins. *PLOS ONE*, 13(8):e0203085, August 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0203085.
3. Matthew K. Matlock, Alex S. Holehouse, and Kristen M. Naegle. ProteomeScout: A repository and analysis resource for post-translational modifications and proteins. *Nucleic Acids Research*, 43(D1):D521–D530, 2015. ISSN 13624962. doi: 10.1093/nar/gku1154.
4. Roman Sloutsky and Kristen M Naegle. Proteome-Level Analysis Indicates Global Mechanisms for Post-Translational Regulation of RRM Domains. *Journal of Molecular Biology*, 430(1):41–44, 2018. ISSN 0022-2836. doi: 10.1016/j.jmb.2017.11.001.
5. Melis Kayikci, A. J. Venkatakrishnan, James Scott-Brown, Charles N.J. Ravarani, Tilman Flock, and M. Madan Babu. Visualization and analysis of non-covalent contacts using the Protein Contacts Atlas. *Nature Structural and Molecular Biology*, 25(2):185–194, 2018. ISSN 15459985. doi: 10.1038/s41594-017-0019-z.
6. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.
7. 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.
8. Tom Ronan, Roman Garnett, and Kristen M Naegle. New analysis pipeline for high-throughput domain-peptide affinity experiments improves SH2 interaction data. *J. Biol. Chem.*, 295(32):11346–11363, August 2020.
9. An-dong Liu, Hui Xu, Ya-nan Gao, Dan-ni Luo, Zhao-feng Li, Courtney Voss, Shawn S. C. Li, and Xuan Cao. (Arg)9-SH2 superbinder: A novel promising anticancer therapy to melanoma by blocking phosphotyrosine signaling. *Journal of Experimental & Clinical Cancer Research*, 37(1):138, July 2018. ISSN 1756-9966. doi: 10.1186/s13046-018-0812-5.
10. Awa Diop, Daniele Santorelli, Francesca Malagrinò, Caterina Nardella, Valeria Pennacchietti, Livia Pagano, Lucia Marcocci, Paola Pietrangeli, Stefano Gianni, and Angelo Toto. SH2 Domains: Folding, Binding and Therapeutical Approaches. *International Journal of Molecular Sciences*, 23(24):15944, December 2022. ISSN 1422-0067. doi: 10.3390/ijms232415944.
11. David R. Stover, Pascal Furet, and Nicholas B. Lydon. Modulation of the SH2 Binding Specificity and Kinase Activity of Src by Tyrosine Phosphorylation within Its SH2 Domain. *Journal of Biological Chemistry*, 271(21):12481–12487, May 1996. ISSN 00219258. doi: 10.1074/jbc.271.21.12481.
12. Clément Couture, Zhou Songyang, Thomas Jascur, Scott Williams, Pankaj Tailor, Lewis C. Cantley, and Tomas Mustelin. Regulation of the Lck SH2 Domain by Tyrosine Phosphorylation. *Journal of Biological Chemistry*, 271(40):24880–24884, October 1996. ISSN 00219258. doi: 10.1074/jbc.271.40.24880.
13. Lily L. Jin, Leanne E. Wybenga-Groot, Jiefei Tong, Paul Taylor, Mark D. Minden, Suzanne Trudel, C. Jane McGlade, and Michael F. Moran. Tyrosine Phosphorylation of the Lyn Src Homology 2 (SH2) Domain Modulates Its Binding Affinity and Specificity*. *Molecular & Cellular Proteomics*, 14(3):695–706, March 2015. ISSN 15359476. doi: 10.1074/mcp.M114.044404.
14. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235.
15. Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunić, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, and Alex Bateman. InterPro in 2022. *Nucleic acids research*, 51(D1):D418—D427, January 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac993.
16. Andrew M. Waterhouse, James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, May 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp033.
17. Alex Bateman. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 2019. ISSN 13624962. doi: 10.1093/nar/gky1049.
18. Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–3, June 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp163.
19. Peter V Hornbeck, Jon M Kornhauser, Sasha Tkachev, Bin Zhang, Elzbieta Skrzypek, Beth Murray, Vaughan Latham, and Michael Sullivan. PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, 40(Database issue): D261–70, January 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1122.
20. A. Hamosh. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517, December 2004. ISSN 1362-4962. doi: 10.1093/nar/gki033.
21. Siwei Chen, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A. Watts, Christopher Vittal, Laura D. Gauthier, Timothy Poterba, Michael W. Wilson, Yekaterina Tarasova, William Phu, Riley Grant, Mary T. Yohannes, Zan Koenig, Yossi Farjoun, Eric Banks, Stacey Donnelly, Stacey Gabriel, Namrata Gupta, Steven Ferriera, Charlotte Tolonen, Sam Novod, Louis Bergelson, David Roazen, Valentin Ruano-Rubio, Miguel Covarrubias, Christopher Llanwarne, Nikelle Petrillo, Gordon Wade, Thibault Jeandet, Ruchi Munshi, Kathleen Tibbetts, Genome Aggregation Database Consortium, Anne O'Donnell-Luria, Matthew Solomonson, Cotton Seed, Alicia R. Martin, Michael E. Talkowski, Heidi L. Rehm, Mark J. Daly, Grace Tiao, Benjamin M. Neale, Daniel G. MacArthur, and Konrad J. Karczewski. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):92–100, January 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06045-0.
22. Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(Database issue): D980–985, January 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1113.
23. 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.
24. Mi-Jeong Park, Ren Sheng, Antonina Silkov, Da-Jung Jung, Zhi-Gang Wang, Yao Xin, Hyunjin Kim, Pallavi Thiagarajan-Rosenkranz, Seohyeon Song, Youngdae Yoon, Wonhee Nam, Ilshin Kim, Eui Kim, Dong-Gyu Lee, Yong Chen, Indira Singaram, Li Wang, Myoung Ho Jang, Cheol-Sang Hwang, Barry Honig, Sungho Ryu, Justin Lorieau, You-Me Kim, and Wonhwa Cho. SH2 Domains Serve as Lipid-Binding Modules for pTyr-Signaling Proteins. *Molecular Cell*, 62(1):7–20, April 2016. ISSN 10972765. doi: 10.1016/j.molcel.2016.01.027.
25. Harry C Jubb, Alicia P Higueruelo, Bernardo Ochoa-Montaño, Will R Pitt, David B Ascher, and Tom L Blundell. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology*, 429(3):365–371, February 2017. ISSN 0022-2836. doi: 10.1016/j.jmb.2016.12.004.
26. Jimin Pei, Bong-Hyun Kim, and Nick V. Grishin. PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Research*, 36(7):2295–2300, April 2008. ISSN 0305-1048. doi: 10.1093/nar/gkn072.
27. Rachel Jaber Chehayeb and Titus J. Boggon. SH2 Domain Binding: Diverse FLVRs of Partnership. *Frontiers in Endocrinology*, 11:575220, September 2020. ISSN 1664-2392. doi: 10.3389/fendo.2020.575220.
28. Wenxun Gan and Benoît Roux. Binding specificity of SH2 domains: Insight from free energy simulations. *Proteins: Structure, Function, and Bioinformatics*, 74(4):996–1007, March 2009. ISSN 0887-3585, 1097-0134. doi: 10.1002/prot.22209.
29. R. A. Grucza, J. M. Bradshaw, K. Fütterer, and G. Waksman. SH2 domains: From structure to energetics, a dual approach to the study of structure–function relationships. *Medicinal Research Reviews*, 19(4):273–293, 1999. ISSN 1098-1128. doi: 10.1002/(SICI)1098-1128(199907)19:4<273::AID-MED2>3.0.CO;2-G.
30. Luc E. M. Marengere and Tony Pawson. Structure and function of SH2 domains. *Journal of Cell Science*, 1994(Supplement_18):97–104, January 1994. ISSN 0021-9533. doi: 10.1242/jcs.1994.Supplement_18.14.
31. Michelangelo Marasco and Teresa Carlomagno. Specificity and regulation of phosphotyrosine signaling through SH2 domains. *Journal of Structural Biology: X*, 4:100026, January 2020. ISSN 2590-1524. doi: 10.1016/j.yjsbx.2020.100026.
32. Zhou Songyang, Steven Blechnert, Nicole Hoagland, Merl F Hoekstra, Helen Piwnica-worms, and Lewis C Cantley. Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Current Biology*, 4(11), 1994.
33. J. C. Obenauer, Lewis C Cantley, and Michael B Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, 31(13):3635–3641, July 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg584.
34. Joseph Rahuel, Brigitte Gay, Dirk Erdmann, André Strauss, Carlos García-Echeverría, Pascal Furet, Giorgio Caravatti, Heinz Fretz, Joseph Schoepfer, and Markus G. Grütter. Structural basis for specificity of GRB2-SH2 revealed by a novel ligand binding mode. *Nature Structural Biology*, 3(7):586–589, July 1996. ISSN 1072-8368. doi: 10.1038/nsb0796-586.
35. S. M. Pascal, A. U. Singer, G. Gish, T. Yamazaki, S. E. Shoelson, T. Pawson, L. E. Kay, and J. D. Forman-Kay. Nuclear magnetic resonance structure of an SH2 domain of phospholipase C-gamma 1 complexed with a high affinity binding peptide. *Cell*, 77(3):461–472, May 1994. ISSN 0092-8674. doi: 10.1016/0092-8674(94)90160-0.
36. Maria I Kontaridis, Kenneth D Swanson, Frank S David, David Barford, and Benjamin G Neel. PTPN11 (Shp2) Mutations in LEOPARD Syndrome Have Dominant Negative , Not

Activating , Effects * □. *Journal of Biological Chemistry*, 281(10):6785–6792, 2006. doi: 10.1074/jbc.M513068200.

37. Adam H. Courtney, Jeanine F. Amacher, Theresa A. Kadlecek, Marianne N. Mollenauer, Byron B. Au-Yeung, John Kuriyan, and Arthur Weiss. A Phosphosite within the SH2 Domain of Lck Regulates Its Activation by CD45. *Molecular Cell*, 67(3):498–511.e6, August 2017. ISSN 1097-2765. doi: 10.1016/j.molcel.2017.06.024.
38. Jennifer Y. Lee, Yu-Hsin Chiu, John Asara, and Lewis C. Cantley. Inhibition of PI3K binding to activators by serine phosphorylation of PI3K regulatory subunit p85α Src homology-2 domains. *Proceedings of the National Academy of Sciences*, 108(34):14157–14162, August 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1107747108.
39. Marion E. Weir, Jacqueline E. Mann, Thomas Corwin, Zachary W. Fulton, Jennifer M. Hao, Jeanine F. Maniscalco, Marie C. Kenney, Kristal M. Roman Roque, Elizabeth F. Chapdelaine, Ulrich Stelzl, Paula B. Deming, Bryan A. Ballif, and Karen L. Hinkle. Novel autophosphorylation sites of Src family kinases regulate kinase activity and SH 2 domain-binding capacity. *FEBS Letters*, 590(8):1042–1052, April 2016. ISSN 0014-5793, 1873-3468. doi: 10.1002/1873-3468.12144.
40. Brian T. Burmeister, Li Wang, Matthew G. Gold, Randal A. Skidgel, John P. O'Bryan, and Graeme K. Carnegie. Protein Kinase A (PKA) Phosphorylation of Shp2 Protein Inhibits Its Phosphatase Activity and Modulates Ligand Specificity. *Journal of Biological Chemistry*, 290(19):12058–12067, May 2015. ISSN 00219258. doi: 10.1074/jbc.M115.642983.
41. P Hof, S Pluskey, S Dhe-Paganon, M J Eck, and S E Shoelson. Crystal structure of the tyrosine phosphatase SHP-2. *Cell*, 92(4):441–50, February 1998. ISSN 0092-8674.
42. Jérôme Hadjadj, Carla Noemi Castro, Maud Tusseau, Marie-Claude Stolzenberg, Fabienne Mazerolles, Nathalie Aladjidi, Martin Armstrong, Houman Ashrafian, Ioana Cutcutache, Georg Ebetsberger-Dachs, Katherine S Elliott, Isabelle Durieu, Nicole Fabien, Mathieu Fusaro, Maximilian Heeg, Yohan Schmitt, Marc Bras, Julian C Knight, Jean-Christophe Lega, Gaetan Lesca, Anne-Laure Mathieu, Marion Moreews, Baptiste Moreira, Audrey Nosbaum, Matthew Page, Cécile Picard, T Ronan Leahy, Isabelle Rouvet, Ethel Ryan, Damien Sanlaville, Klaus Schwarz, Andrew Skelton, Jean-Francois Viallard, Sebastien Viel, Marine Villard, Isabelle Callebaut, Capucine Picard, Thierry Walzer, Stephan Ehl, Alain Fischer, Bénédicte Neven, Alexandre Belot, and Frédéric Rieux-Laucat. Early-onset autoimmunity associated with SOCS1 haploinsufficiency. *Nat. Commun.*, 11(1):5341, October 2020.
43. Panagis Filippakopoulos, Susanne Müller, and Stefan Knapp. SH2 domains: Modulators of nonreceptor tyrosine kinase activity. *Current Opinion in Structural Biology*, 19(6):643–649, 2009. ISSN 0959440X. doi: 10.1016/j.sbi.2009.10.001.
44. Jennifer Asmussen, Elisabeth A. Lasater, Cheryl Tajon, Juan Oses-Prieto, Young-wook Jun, Barry S. Taylor, Alma Burlingame, Charles S. Craik, and Neil P. Shah. MEK-Dependent Negative Feedback Underlies BCR–ABL-Mediated Oncogene Addiction. *Cancer Discovery*, 4(2):200–215, February 2014. ISSN 2159-8274, 2159-8290. doi: 10.1158/2159-8290.CD-13-0235.
45. Alejandro Wolf-Yadlin, Neil Kumar, Yi Zhang, Sampsa Hautaniemi, Muhammad Zaman, Hyung-Do Kim, Viara Grantcharova, Douglas A Lauffenburger, and Forest M White. Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Molecular Systems Biology*, 2(1):54, January 2006. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb4100094.
46. Lily A. Chylek, Vyacheslav Akimov, Jörn Dengjel, Kristoffer T. G. Rigbolt, Bin Hu, William S. Hlavacek, and Blagoy Blagoev. Phosphorylation Site Dynamics of Early T-cell Receptor Signaling. *PLoS ONE*, 9(8):e104240, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0104240.
47. Joseph A. Caruso, Paul M. Stemmer, Alan Dombkowski, Nicholas J. Caruthers, Randall Gill, and Allen J. Rosenspire. A systems toxicology approach identifies Lyn as a key signaling phosphoprotein modulated by mercury in a B lymphocyte cell model. *Toxicology and Applied Pharmacology*, 276(1):47–54, April 2014. ISSN 1096-0333. doi: 10.1016/j.taap.2014.01.002.
48. Jared L. Johnson, Tomer M. Yaron, Emily M. Huntsman, Alexander Kerelsky, Junho Song, Amit Regev, Ting-Yu Lin, Katerina Liberatore, Daniel M. Cizin, Benjamin M. Cohen, Neil Vasan, Yilun Ma, Konstantin Krismer, Jaylissa Torres Robels, Bert van de Kooij, Anne E van Vlimmeren, Nicole Andree-Busch, Norbert Kaufer, Maxim V Dorovkov, Alexey G. Ryazanov, Yuichiro Takagi, Edward R. Kastenhuber, Marcus D. Gonclvaes, Olivier Elemento, Dylan J. Taatjes, Alexandre Maucuer, Akio Yamashita, Alexie Degterev, Rune Linding, John Blenis, Peter V. Hornbeck, Bejnjamin E. Turk, Michael B. Yaffe, and Lewis C. Cantley. A global atlas of substrate specificities for the human serine/threonine kinome. *BioRxiv*, pages 1–32, 2022.
49. Philipp Mertins, Feng Yang, Tao Liu, D.R. Mani, Vladislav A. Petyuk, Michael A. Gillette, Karl R. Clauser, Jana W. Qiao, Marina A. Gritsenko, Ronald J. Moore, Douglas A. Levine, Reid Townsend, Petra Erdmann-Gilmore, Jacqueline E. Snider, Sherri R. Davies, Kelly V. Ruggles, David Fenyo, R. Thomas Kitchens, Shunqiang Li, Narciso Olvera, Fanny Dao, Henry Rodriguez, Daniel W. Chan, Daniel Liebler, Forest White, Karin D. Rodland, Gordon B. Mills, Richard D. Smith, Amanda G. Paulovich, Matthew Ellis, and Steven A. Carr. Ischemia in Tumors Induces Early and Sustained Phosphorylation Changes in Stress Kinase Pathways but Does Not Affect Global Protein Levels. *Molecular & Cellular Proteomics*, 13(7):1690–1704, July 2014. ISSN 15359476. doi: 10.1074/mcp.M113.036392.
50. Ivica Letunic, Supriya Khedkar, and Peer Bork. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Research*, 49(D1):D458–D460, 10 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa937.
51. Alex S Holehouse and Kristen M. Naegle. Reproducible Analysis of Post-Translational Modifications in Proteomes - Application to Human Mutations. *PLoS ONE*, 10(12):1–19, 2015.
52. F Tian, L Yang, F Lv, Q Yang, and P Zhou. In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach. *Amino Acids*, 36(3):535–554, March 2009.