Mapping the energetic and allosteric landscapes of protein binding domains

Andre J. Faure[1]*, Júlia Domingo[1,4]*, Jörn M. Schmiedel[1,5]*, Cristina Hidalgo-Carcedo[1], Guillaume Diss[1,6], Ben Lehner[1,2,3†]

*These authors contributed equally to this work

[1] Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Doctor Aiguader 88, 08003 Barcelona, Spain

[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain

[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

[4] New York Genome Center (NYGC), 101 6th Ave, 10013 New York, USA

[5] Genomic Data Science Consulting, Luisenstrasse 14, 96047 Bamberg, Germany

[6] Friedrich Miescher Institute for Biomedical Research (FMI), Maulbeerstrasse 66, 4058 Basel, Switzerland

† email: ben.lehner@crg.eu

Summary paragraph

Allosteric communication between distant sites in proteins is central to biological regulation but still poorly characterised, limiting understanding, engineering and drug development[1–6]. An important reason for this is the lack of methods to comprehensively quantify allostery in diverse proteins. Here we address this shortcoming and present a method that uses deep mutational scanning to globally map allostery. The approach uses an efficient experimental design to infer *en masse* the causal biophysical effects of mutations by quantifying multiple molecular phenotypes—here binding and protein abundance—in multiple genetic backgrounds and fitting thermodynamic models using neural networks. We apply the approach to two of the most common human protein interaction domains, an SH3 domain and a PDZ domain, to produce comprehensive atlases of allosteric communication. Allosteric mutations are abundant with a large mutational target space of network-altering 'edgetic' variants. Mutations are more likely to be allosteric closer to binding interfaces, at Glycines and in specific residues connecting to an opposite surface in the PDZ domain. This general approach of quantifying mutational effects for multiple molecular phenotypes and in multiple genetic backgrounds should allow the energetic and allosteric landscapes of many proteins to be rapidly and comprehensively mapped.

Main text

Proteins with important functions are usually ‘switchable’, with their activities modulated by the binding of other molecules, covalent modifications or mutations outside of their active sites. This transmission of information spatially from one site to another in a protein is termed allostery, which Monod famously referred to as ‘the second secret of life’[7,8]. Allosteric regulation is central to nearly all of biology, including signal transduction, transcriptional regulation, and metabolic control. Many disease-causing mutations, including numerous cancer driver mutations, are pathological because of their allosteric effects[1]. Conversely, many of the most effective therapeutic agents do not directly inhibit the active sites of proteins but modify their activities by binding to allosteric sites. Amongst other benefits, allosteric drugs often have higher specificity than orthosteric drugs that bind active sites conserved in protein families[2,3].

Allosteric sites are difficult to predict, even for highly studied proteins with known active and inactive states[4]. Individual proteins may contain a limited number of allosteric sites, which would be consistent with their physiological regulation by a limited number of ligands and modifications. Alternatively, as has been suggested by theoretical work, allostery might be quite widely distributed throughout protein domains[3,4,9]. This distinction between ‘sparse’ and ‘abundant’ allosteric sites has important implications: abundant allosteric sites would both facilitate the evolution of allosteric control[5] and increase the likelihood of identifying therapeutic molecules that can bind a target protein and regulate its activity[6]. Most known allosteric sites are involved in physiological regulation but ‘orphan’ or ‘serendipitous’ sites without any understood physiological role have been identified for some proteins. Moreover, domain-insertion and mutagenesis also suggest quite extensive long-range communication in protein interaction domains[10], enzymes[11–14], transcription factors[15,16] and receptors[17].

Physical interactions between proteins are critical to most biological processes and represent a potentially vast therapeutic target space[2]. However, allosteric sites are not known for most protein-protein interactions (PPIs), a comprehensive map of allosteric sites has not been produced for any protein interaction domain, and generic methods to identify allosteric sites regulating PPIs do not exist.

Global maps of allosteric communication could be generated for protein binding domains if the effects of all mutations on binding affinity could be quantified: any mutation altering binding affinity but not directly contacting a ligand must be having an allosteric effect. However, changes in affinity cannot be inferred simply by quantifying changes in binding to an interaction partner; even in the simplest genotype-to-phenotype (energy) landscapes, ‘biophysical ambiguities’[18] exist, meaning that changes in a molecular phenotype (e.g. binding to an interaction partner) can be caused by many different changes in the underlying biophysical properties (e.g. folding or binding affinity)[18,19]. To quantify the effects of mutations on binding affinity and so globally map allosteric communication, these ambiguities must be resolved.

Here we present an approach to achieve this for PPIs, allowing us to globally map the energetic and allosteric landscapes of protein interaction domains. The approach takes advantage of the massively parallel nature of deep mutational scanning (DMS) to quantify the phenotypic effects of thousands of perturbations[20]. We use an experimentally efficient strategy that we refer to as ‘multidimensional mutagenesis’ whereby the effects of mutations are quantified for multiple molecular phenotypes and in multiple genetic backgrounds. This method resolves ambiguities where a number of causal biophysical changes could account for an observed mutational effect[18,19] and allows the inference of the *in vivo* biophysical effects of mutations. We harness the flexibility of neural networks to fit thermodynamic models to these experimental measurements, thereby accurately inferring the underlying causal changes in free energy. Applied to two protein domains, the method provides near complete views of their free energy landscapes and the first global maps of allosteric mutations.

ddPCA quantifies abundance and binding

The binding of a protein to an interaction partner depends on both its affinity and the concentration of the active folded state. Existing methods that quantify how a perturbation changes the amount of protein bound to an interaction partner[21] are inadequate for the identification of allosteric sites, because they do not distinguish between mutational effects on binding affinity versus protein abundance[22]. In this situation, they would lead to false positives where changes in binding are caused by changes in concentration and false negatives where changes in affinity are masked by changes in abundance.

We therefore developed a strategy that uses two separate selection assays based on protein fragment complementation (PCA) to quantify the effects of mutations on both the abundance of a protein and its binding to an interaction partner (Fig. 1a). As perturbations to probe the potential for allosteric regulation we use mutations which are a convenient method to introduce diverse changes in chemistry at all sites in a protein[20,23]. In the first assay, *bindingPCA*, the binding between two proteins is quantified by fusing them to different fragments of a reporter enzyme, dihydrofolate reductase (DHFR). Interaction between the proteins brings the DHFR fragments in close proximity allowing them to form a functional enzyme whose activity as measured by cellular growth in selective conditions is proportional to the intracellular concentration of the protein complex[24]. In the second assay, *abundancePCA*, only one protein is expressed and fused to a DHFR fragment with the other DHFR fragment highly expressed. Functional DHFR is now reconstituted by random encounters and growth is proportional to the intracellular concentration of the first protein over >3 orders of magnitude, as validated by applying the assay to >2000 yeast proteins[25]. We refer to the combination of these two assays as *DoubleDeepPCA* (*ddPCA*), a high-throughput method that quantifies the effects of mutations on both the abundance of a protein and its binding to one or more interaction partners. *ddPCA* builds on and extends prior work using PCA to probe the effects of mutations on protein binding and stability[26,27].

We applied *ddPCA* to examples of two of the most common protein interaction domains encoded in the human genome: the C-terminal SH3 domain of the human growth factor receptor-bound protein 2 (GRB2), which binds a proline-rich linear peptide of GRB2 associated-binding protein 2 (GAB2), and the third PDZ domain from the adaptor protein PSD95/DLG4, which binds to the C-terminus of the protein CRIPT (Fig. 1d, Methods).

There are two key principles of the *ddPCA* approach, which we refer to as ‘multidimensional mutagenesis’. First, the effects of mutations on two molecular phenotypes—binding and abundance—are quantified, and second, mutational effects are quantified starting from multiple genetic backgrounds. Both of these strategies are important for correctly inferring (disentangling) the underlying causal free energy changes from the measured mutational effects: many different free energy changes can generate the same change in phenotype[18] and quantifying how mutations interact in double mutants[18,19,24], as well as their effects on two different molecular traits, serves to resolve these biophysical ambiguities (Fig. 2c). Moreover, the relationships between the free energies and folding and binding phenotypes/measurements are nonlinear and plateau at high and low energies[28] (Fig. 2f); quantifying the effects of mutations from different starting genotypes therefore serves to expand the effective dynamic range of individual measured mutational effects.

We generated mutagenesis libraries of the GRB2-SH3 and PSD95-PDZ3 domains containing both single and double amino acid (AA) substitutions (Extended Data Fig. 1a) and quantified their effects on binding to GAB2 and CRIPT, respectively, using *bindingPCA,* and on the intracellular concentration of the free domains using *abundancePCA*. All experiments were performed in biological triplicate, with deep sequencing used to quantify relative changes in binding and abundance in pooled selection assays (Fig. 1b). We calculated abundance and binding fitness scores and associated errors using DiMSum (Methods). Binding and abundance fitness scores were highly reproducible between replicates (Fig. 1b, Pearson's r = 0.87-0.92). Mutational effects also agreed very well with individual growth measurements (Pearson's r = 0.94, n = 14; P = 5e-7, Fig. 1c).

The distributions of mutational effects corresponding to both binding and abundance are bimodal for both domains, with, for example, 28% of single AA substitutions strongly affecting binding of the PDZ domain and 46% having nearly neutral or mild effects (*bindingPCA* fitness within the lower peak < -0.75 and upper peak > -0.25 respectively, Fig. 1e). The mutational effect matrices for binding reveal that mutations with large effects on binding are distributed throughout both domains (Fig. 1f-g). Similarly, the mutational effect matrices for abundance show that mutations throughout both domains also have large effects on protein concentration (Fig. 1f-g). Indeed, plotting the changes in binding against the changes in abundance reveals that most mutations altering binding also alter the concentration of the isolated domains (Fig. 1h), consistent with the expectation that changes in protein stability are a major cause of mutational effects on binding[29].

Inference of free energy changes

We used a neural network formulation that relies on the Boltzmann distribution to fit thermodynamic models to the experimental data obtained using *ddPCA*, thereby inferring the underlying causal free energy changes from the effects of each single AA substitution on these two molecular phenotypes (Fig. 2a,b). Protein binding can be most simply modelled as a three-state equilibrium with unfolded, folded and bound energetic states (Fig. 2a). In this genotype-phenotype model, mutations alter the free energy of folding (ΔG_f) and/or binding (ΔG_b), and changes in free energy combine additively in double AA substitutions. The relationship between the fraction of folded or folded+bound protein and the respective measured phenotypes (*abundancePCA* and *bindingPCA* fitness) is assumed to be linear[19,24] (Fig. 2b; Methods).

The three-state model provides an excellent quantitative fit to the data for both domains (Fig. 2d, R^2 = 0.84-0.91, see Extended Data Fig. 1b-g for similar comparisons shown separately for single and double AA substitutions, as well as for validation data held out during model fitting), strongly supporting the assumption that most changes in the free energy of both folding and binding are additive in double AA substitutions[23,30]. Training models using data corresponding to only one molecular phenotype (binding, Extended Data Fig. 2a), or two molecular phenotypes but only fitness effects from single AA substitutions (Extended Data Fig. 2b), results in worse fits to the data (R^2 = 0.05-0.92 and R^2 = 0.77-0.83 respectively). The number of double AA substitutions for each single AA substitution varies across the four datasets, with the relatively few double AA substitutions in the PSD95-PDZ3 libraries (median = 5 and 4 in the binding and abundance libraries respectively, Extended Data Fig. 1a) still sufficient to infer the underlying free energy changes. Downsampling double mutant data in both datasets illustrates how increasing the number of double mutants improves the model fit (Extended Data Fig. 3a).

To evaluate the quality of the inferred free energy changes upon mutation, we compared them to *in vitro* measurements for the PDZ domain. We find excellent agreement between inferred free energies of folding relative to the wild-type ($\Delta\Delta G_f$) and those corresponding to single AA substitutions determined *in vitro* for PSD95-PDZ3 (F337W background)[31] (Fig. 2e, Pearson's r = 0.79, n = 30; P = 2e-7). Consistent with previous assessments[32,33], computational predictions of mutation effects on protein stability or function only partially explain inferred folding free energy changes (Pearson's r = 0.2-0.7, Extended Data Fig. 4). Binding free energy changes similarly agree with those measured by stopped-flow experiments for the PSD95-PDZ3:CRIPT interaction[34] (Fig. 2e, Pearson's r = 0.91, n = 26; P = 8e-11; see Extended Data Fig. 3b for additional comparisons to smaller-scale *in vitro* validation datasets). Using only the binding or only single AA substitutions data results in worse agreement with the *in vitro* binding and folding free energy changes (Extended Data Fig. 2c,d), as does reducing the number of double mutants used to fit the model (Extended Data Fig. 3a). As a further validation of our method to infer free energy changes from molecular phenotypes, we fit the same three-state model to previously published *in vitro* mutagenesis data for the binding of nearly all single and double AA substitutions of protein G domain B1 to IgG-Fc[35] (Extended Data Fig. 1b,e). Even in the absence of multiple measured phenotypes (binding only), with this depth of double mutant data we find excellent agreement between inferred free energy changes of folding and *in vitro* measurements[33,35] (Fig. 2e, Pearson's r = 0.8-0.91, n = 685 and 80; P < 2.2e-16), similar to a previous analysis[19]. These comparisons further demonstrate the validity of our inferences and the general flexibility of the approach.

Free energy landscapes of SH3 and PDZ

Free energy landscapes of mutational effects (Fig. 3a,b) have important advantages over maps of phenotypic effects (Fig. 1f,g), converting mutational effects to the underlying additive biophysical traits and allowing accurate genetic prediction when mutations are combined in pairs and larger combinations[18]. Furthermore, the free energy landscapes are more complete, as single mutant free energies can be inferred by their effects in different backgrounds (e.g. double AA substitutions) or their effects on a related phenotype (e.g. folding energies from binding phenotype) despite missing single mutant phenotypes.

In general, mutations act asymmetrically on the two inferred biophysical traits, with mutations tending to have stronger effects on the free energy of folding than binding (Fig. 3a,b,d, Extended Data Fig. 5a,b,d). This is less evident in comparisons at the phenotypic level where the fraction of bound protein complex is a nonlinear function of both underlying biophysical traits (Fig. 1f,g). Mutations that affect folding are also more numerous and more widely distributed throughout the domains, with many positions sensitive to perturbation. Effects on binding affinity on the other hand are comparatively less frequent and enriched in residues proximal to the ligand i.e. in the binding interface (Fig. 3a,b, Extended Data Fig. 5a,b, black boxes). Thus, changes in protein binding (change in *bindingPCA* fitness) are predominantly driven by changes in protein stability, especially for mild and intermediate fitness effects (Extended Data Fig. 5f).

Directly comparing free energies stratified by the position of the mutated residues reveals that mutations in core residues (relative solvent accessible surface area, RSASA < 0.25) have the largest effects on folding, whereas mutations in the binding interface (ligand distance < 5 Å) have the strongest effects on binding affinity (Fig. 3c, Extended Data Fig. 5c). Surface residues (RSASA ≥ 0.25) are more tolerant to mutations (Fig. 3c, Extended Data Fig. 5c).

The bimodality of phenotypic effects (Fig. 1e) is much reduced in distributions of relative free energy changes, particularly in the case of binding free energies, whose mode is centred on $\Delta\Delta G = 0$ i.e. no difference from wild-type (Fig. 3d, Extended Data Fig. 5d). The distribution of folding free energies has a positive mode ($\Delta\Delta G > 0$) and a heavy right tail indicating that most single AA substitutions have destabilising effects and many substitutions are strongly destabilising (Fig. 3d, Extended Data Fig. 5d).

Extensive biophysical pleiotropy

Comprehensively quantifying the effects of mutations on both folding and binding provides the first opportunity to assess the extent to which mutations affect multiple biophysical properties i.e. biophysical pleiotropy[18]. Overall, more than two thirds of all mutations altering binding are biophysically pleiotropic (86% and 67% increasing binding and 80% and 77% of mutations decreasing it are biophysically pleiotropic in GRB2-SH3 and PSD95-PDZ, respectively). In both domains, mutations that disrupt binding most often show synergistic pleiotropy, reducing both binding and folding stability (Fig. 3e, Extended Data Fig. 5e). In contrast, mutations that increase binding tend to display antagonistic pleiotropy with the effects on binding and folding free energies in opposing directions (Fig. 3e, Extended Data Fig. 5e). The proportion of different kinds of pleiotropic mutations differs depending on the region of the domain (Extended Data Fig. 5g). For instance in GRB2-SH3 and PSD95-PDZ, compared to core or surface residues, binding interface positions harbour a higher proportion of mutations that disrupt binding despite antagonistic effects on the free energy of folding (see below), consistent with the hypothesis that residues of proteins involved in substrate binding or catalysis are not optimized for stability[36].

This extensive biophysical pleiotropy further emphasises the importance of determining the biophysical effects of mutations. For both genetic prediction and protein engineering, the outcome when combining mutations is often only predictable if the causal biophysical effects can be measured or inferred: combining mutations with the same phenotypic outcomes but different biophysical causes often results in different phenotypic consequences[18].

Surfaces suboptimal for stability

Overlaying the mean folding free energy changes per residue on the domain structures further illustrates that solvent exposed (surface or binding interface) residues tend to be less sensitive to mutations than those that constitute the buried core of the structure (Fig. 4a,b, Extended Data Fig. 6a,b), with per-residue mean folding energies anti-correlated with residue burial (RSASA; Fig. 4c, Extended Data Fig. 6c).

Although mutations overwhelmingly destabilise folding, a small subset of residues are enriched for mutations that increase stability ($\Delta\Delta G < 0$). We find nine residues in GRB2-SH3 and three residues in PSD95-PDZ3 where at least five distinct single AA substitutions are observed to decrease the free energy of folding compared to the wild-type, none of which are classified as core residues (Fig. 4d). That AAs unfavourable for stability have been retained by evolution suggests selective constraints other than stability acting on these sites[36]. Consistent with this, 4/9 destabilising residues in GRB2-SH3 and 1/3 destabilising residues in PSD95-PDZ3 are in the binding interface (ligand distance < 5 Å).

The remaining 7 surface destabilising residues are either highly evolutionarily conserved or uncharacteristically hydrophobic compared to other surface residues (Extended Data Fig. 6e,f), suggesting that they are involved in additional molecular functions or interactions. Indeed, three of these residues in GRB2-SH3 occur within the homodimer interface of the full-length protein and two in PSD95-PDZ3 are involved in extra-domain interactions (Extended Data Fig. 6g).

This illustrates how *abundancePCA* data can help identify functionally important surface sites, even when interaction partners and functions are unknown. Moreover, in general, identifying surface sites that are suboptimal for stability (but otherwise functionally neutral) has therapeutic implications, predicting where the binding of small molecules or biologics may help stabilise proteins carrying disease-causing destabilising mutations[37].

Binding interface identification

Overlaying the mean absolute binding free energy changes per residue on the domain structures shows a strong enrichment for the largest mutational effects in the binding interface (ligand distance < 5 Å, Fig. 5a, Extended Data Fig. 7a). Indeed, the inferred binding free energy changes alone accurately predict binding interface residues (area under the ROC curve, AUC = 0.83-0.94, Extended Data Fig. 7d). Furthermore, the effects of individual mutations at key ligand-contacting residues are consistent with structural information at those positions (Extended Data Fig. 7e). This illustrates how quantifying the effects of mutations on binding and abundance (but neither of these phenotypes alone, Extended Data Fig. 7f) can identify binding interfaces, which may be useful for identifying the interfaces of the very large number of protein interactions without any structural information[38].

Comprehensive maps of allostery

We next asked whether any residues outside the binding interface are also enriched for mutations that modulate binding affinity. We identified two sites in GRB2-SH3 (G15 and G45) and 8 sites in PSD95-PDZ3 (R312, R318, G329, G330, I336, D357, E373 and A375) with mean absolute change in binding free energy greater than that of mutations in the binding interface (Fig. 5b,c, Extended Data Fig. 8a,b). We refer to these ligand-distal residues at which many mutations have strong effects on binding affinity as major allosteric sites (see Extended Data Fig. 7b,c for the GB1 domain).

A previous study identified 9 residues within PDS95-PDZ3 at which AA substitutions have the capacity to switch PSD95-PDZ3 ligand binding class specificity—an observation that can only be explained by an underlying causal change in binding affinity[39]. 6/9 of these class-switching residues (two within and four outside the binding interface) are identified as major allosteric sites by our definition. Moreover, the remaining three class-switching residues (G322, V362 and L379) are enriched for mutations with strong (albeit low confidence) binding free energy changes compared to other residues not classified as major allosteric sites (AUC = 0.76, n = 1,305; *P* = 1e-10, two-sided Mann-Whitney U test). This identification of previously described specificity-determining residues as major allosteric sites further validates our approach.

The two major allosteric sites in GRB2-SH3 and 6/8 in PSD95-PDZ3 (R318, G329, G330, I336, E373, A375) are either in direct physical contact with binding interface residues (i.e. second shell sites) or immediately adjacent to them in the linear AA sequence (i.e. backbone-backbone contacts; see asterisks in Fig. 3a,b). However, two sites in PSD95-PDZ3 (R312 and D357) are on the opposite surface of the domain, with distances of 11-14Å to the closest ligand residue. These sites form a near-contiguous “chain” of residues linking the N-terminal beta strand to the binding interface via a salt bridge formed by residues R312 and D357, where the latter is in close proximity to the second-shell class-switching residue I336 located in the adjacent beta sheet strand (minimum backbone atom distance = 4.1 Å). Thus, whilst the sites most enriched for mutations affecting binding affinity are mostly proximal to the binding interface, in the PDZ domain they also extend throughout the domain to the opposite surface.

Allosteric mutations are abundant

Although these 10 residues are the positions most enriched for allosteric effects, mutations affecting binding affinity actually occur throughout both protein domains (Fig. 3a,b). Defining allosteric mutations as those with effects at least as large as the mean absolute binding free energy change of mutations in the binding interface, we find a total of 55 allosteric mutations in 24 distinct residues in GRB2-SH3 (33 core, 22 surface) and 152 allosteric mutations in 49 residues in PSD95-PDZ3 (83 core, 69 surface). 40% (12/30) and 55% (24/44) of all surface residues have at least one allosteric mutation in GRB2-SH3 and PSD95-PDZ3 respectively (Fig. 6a,b, Extended Data Fig. 8c,d, Extended Data Fig. 9a,b). These results suggest that allosteric mutations are abundant in the core of proteins and also in solvent accessible regions. Moreover, similar to their sub-optimality for folding, surface sites are also often suboptimal for binding at a distal interface.

In addition to the average absolute change in the free energy of binding being higher within residues comprising the ‘sector’ defined by 20 coevolving residues in PSD95-PDZ3[9,39] (AUC = 0.78, n = 84; *P* = 2e-4, two-sided Mann-Whitney U test), allosteric mutations themselves are highly enriched at these sites (odds ratio = 8.3, n = 1,033; *P* < 2.2e-16, two-sided Fisher’s Exact Test, Extended Data Fig. 9e), indicating that this network of physically proximal sites partially identifies the patterns of allostery described here. We also find that the probability that a mutation outside the binding interface will be allosteric significantly depends on its distance to the ligand in GRB2-SH3 and PSD-PDZ3 (Spearman's ρ = -0.3 and -0.58 respectively, Fig. 6c, Extended Data Fig. 9c), further suggesting the propagation of local perturbations to neighbouring residues as an important cause of allosteric effects observed in these domains. Also consistent with this, allosteric coupling scores estimated by a network-based perturbation propagation algorithm using only structural contacts[40] correlate with the proportion of allosteric mutations per residue in both GRB2-SH3 and PSD95-PDZ3 (Spearman's ρ = 0.51 and 0.42 respectively, Extended Data Fig. 10a).

If the disruption of local energetic couplings is indeed an important initiator of allosteric effects, we reasoned that the identities of the original and substituted residues in allosteric mutations should be enriched in specific AA types. Indeed, allosteric mutations are significantly enriched at Glycine residues in all three protein domains considered (odds ratio = 2.4-5.7, n = 470, 1,033 and 740 for GRB2-SH3, PSD95-PDZ3 and GB1 respectively; *P* < 9e-5, two-sided Fisher's Exact Test, Fig. 6c and Extended Data Fig. 9f), whose replacement would increase the local mass, volume and also conformational rigidity. Glycine residues also comprise two major allosteric sites in each domain. In fact, across all three protein domains, four out of five Glycines that occur in secondary structure elements (and outside the binding interface) are major allosteric sites (odds ratio = 20, n = 20; *P* = 0.01, two-sided Fisher's Exact Test, Fig. 6c and Extended Data Fig. 9g). Likewise, changes to Proline are consistently the most enriched mutant AA in allosteric mutations (odds ratio = 2.7-5.8, n = 470, 1,033 and 740 for GRB2-SH3, PSD95-PDZ3 and GB1 respectively; *P* < 0.02, two-sided Fisher's Exact Test, Fig. 6c and Extended Data Fig. 9f) with this residue's exceptional rigidity likely to introduce both local structural distortion and altered dynamics, both of which may be important for allosteric communication[3].

Mutations for network re-wiring

Disease-causing and evolutionarily-selected mutations have previously been conceptualised as perturbing cellular processes by altering either the 'nodes' or 'edges' of PPI networks[41]. However, systematic data quantifying the effects of mutations on network edges is limited to a small number of mutations for any individual protein[42]. We therefore used our data to further investigate the properties of 'edgetic'[41,42] network-altering mutations.

We first considered changes in binding affinity as the trait of interest. Mutations in ligand-contacting sites are very biased towards disruption (Fig. 3a-c, Extended Data Fig. 10b) with many fewer mutations increasing rather than decreasing binding affinity. This is consistent with the binding interface residues being near optimal for this function. Mutations at ligand-distal surface positions tend to have milder effects on the free energy of binding than those at ligand-proximal sites, but their total number is greater and their direction less biased (Extended Data Fig. 10b). Indeed, the number of mutations at surface residues increasing binding affinity is greater than the number of disrupting mutations in the binding interface (110 vs. 91 and 143 vs. 126 for GRB2-SH3 and PSD95-PDZ3 respectively).

We next considered changes in the fraction of bound protein complex as the phenotype of interest i.e. regardless of whether the underlying biophysical mechanism involves a change in binding or folding energy or both (Fig. 6d, Extended Data Fig. 9d). Mutations in the protein core frequently and strongly reduce the fraction of bound protein complex, with those in PSD95-PDZ3 far outnumbering mutations in the binding interface and surface of similar effect size. Finally, we considered mutations that alter binding affinity without a significant change in abundance (Fig. 6e). These mutations are comparatively rare in the protein core and indeed appear to be totally absent in GRB2-SH3. Unsurprisingly, mutations in the binding interface have strongly disruptive effects on binding. However, mutations in solvent exposed surface positions are particular in that they are both numerous and, especially in PSD95-PDZ3, can fine tune binding affinity in both directions without disrupting stability.

In summary, the high density of allosteric mutations throughout these domains suggests a larger mutational target space for ‘edgetic’ network-altering genetic variants than has been previously appreciated: many mutations outside of interaction interfaces should be expected to alter not just protein stability but also the affinities of proteins for their interaction partners.

Discussion

We have presented here a general approach—multidimensional mutagenesis—to infer the *in vivo* biophysical effects of mutations and used a specific implementation of it—*ddPCA*—to produce the first global maps of allosteric mutations for any proteins.

The approach of fitting additive thermodynamic models to mutation scanning data is conceptually similar to previous pioneering work inferring the energetic effects of mutations in regulatory elements from combinatorial mutants[43–45] and antibody affinities using ligand titrations[46]. Combined with a diversity of selection methods[23,47], multidimensional mutagenesis strategies including *ddPCA* should facilitate the rapid and comprehensive mapping of the *in vivo* biophysical effects of mutations and the generation of free energy landscapes for diverse macromolecules, interactions and pathways.

As implemented here, *ddPCA* can likely be applied to many intracellular proteins, but alternative assays will be needed for secreted proteins and those that do not express in yeast. In addition, active degradation signals and aggregation may necessitate different models and additional measurements to be made.

The comprehensive energetic and allosteric landscapes presented here provide a number of important insights into protein function and evolution. First, allosteric mutations are common and their frequency increases closer to binding interfaces, suggesting local propagation of perturbations as an important molecular mechanism. The abundance of mutations that alter binding affinity represents a rich genotypic space for both evolutionary innovation of allosteric control mechanisms and potential therapeutic exploitation. Second, allosteric mutations are strongly enriched for certain AA changes, with mutations at Glycine residues in secondary structure elements particularly likely to be allosteric and mutations to Proline also frequently having allosteric effects. Third, mutations are frequently pleiotropic, affecting both stability (PPI network node) and affinity (PPI network edge). Fourth, mutations in both protein cores and surfaces can tune stability and affinity, suggesting high evolvability for new regulatory mechanisms and diverse opportunities for the modulation of protein abundance and interactions via drug binding.

The application of *ddPCA* and related methods should help accelerate allosteric drug discovery by producing global allosteric maps for therapeutic target proteins, including those currently considered ‘undruggable’[48]. Moreover, systematic maps of spatial information transfer in proteins—and selection experiments using different protein family members, ligands or modifiers of this transfer[11,15,16]—should provide fundamental insights into allosteric mechanisms, evolution, specificity and the bidirectionality of allosteric regulation.

Understanding, predicting and engineering the encoding of biophysical properties by amino acid sequences is one of the most fundamental problems in molecular biology. That such a central problem remains unresolved after decades of research is, we would argue, primarily due to a lack of systematic and unbiased data quantifying how changes in sequence alter the biophysical properties of proteins. For fundamental problems where very large quantitative datasets already exist, dramatic recent progress has been made using deep learning, allowing, for example, the accurate prediction of protein structures from sequence[49]. However, for many other core problems of molecular biology, suitably diverse and quantitative training datasets do not yet exist: we still need to generate them. A key advantage of a general method such as *ddPCA* is that it can be potentially used to quantify the effects of millions of mutations on the biophysical properties of thousands of proteins, allowing three of the fundamental encoding problems of biology – protein folding (sequence-to-stability), binding (sequence-to-affinity) and allostery to be addressed using massive-scale perturbation experiments. We envisage that such large, quantitative datasets will allow machine learning approaches to be effectively brought to bear on the generative functions of molecular biology, including predicting macromolecular stability, affinity, specificity and allostery from sequence. If successful, this combination of brute force experimentation and machine learning will usher in a new era of *predictive* molecular biology, where the biophysical properties of proteins can be accurately determined and engineered. Such predictive ability would open up unprecedented possibilities in industrial, agricultural and environmental biotechnology, and would revolutionise clinical genetics and the development of therapeutics.

References

1. Guarnera, E. & Berezovsky, I. N. Allosteric drugs and mutations: chances, challenges, and necessity. *Curr. Opin. Struct. Biol.* **62**, 149–157 (2020).
2. Arkin, M. R., Tang, Y. & Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. *Chem. Biol.* **21**, 1102–1114 (2014).
3. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331–339 (2014).
4. Xie, J. & Lai, L. Protein topology and allostery. *Curr. Opin. Struct. Biol.* **62**, 158–165 (2020).
5. Kuriyan, J. & Eisenberg, D. The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983–990 (2007).
6. Nussinov, R. & Tsai, C.-J. Allostery in disease and in drug discovery. *Cell* **153**, 293–305 (2013).
7. Monod, J., Changeux, J. P. & Jacob, F. Allosteric proteins and cellular control systems. *J. Mol. Biol.* **6**, 306–329 (1963).
8. Ullmann, A. In Memoriam: Jacques Monod (1910–1976). *Genome Biol. Evol.* **3**, 1025–1033 (2011).
9. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **138**, 774–786 (2009).
10. Dionne, U. et al. Protein context shapes the specificity of SH3 domain-mediated interactions in vivo. *Nat. Commun.* **12**, 1–15 (2021).
11. McCormick, J. W., Russo, M. A., Thompson, S., Blevins, A. & Reynolds, K. A. Structurally distributed surface sites tune allosteric regulation. *eLife* **10**, e68346 (2021).
12. Bandaru, P. et al. Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife* **6**, e27810 (2017).
13. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147**, 1564–1575 (2011).
14. Oakes, B. L. et al. Profiling of engineering hotspots identifies an allosteric CRISPR-

Cas9 switch. *Nat. Biotechnol*. **34**, 646–651 (2016).

15. Leander, M., Yuan, Y., Meger, A., Cui, Q. & Raman, S. Functional plasticity and evolutionary adaptation of allosteric regulation. *Proc. Natl. Acad. Sci. U. S. A*. **117**, 25445–25454 (2020).
16. Tack, D. S. et al. The genotype-phenotype landscape of an allosteric protein. *Mol. Syst. Biol*. **17**, e10179 (2021).
17. Coyote-Maestas, W., He, Y., Myers, C. L. & Schmidt, D. Domain insertion permissibility-guided engineering of allostery in ion channels. *Nat. Commun*. **10**, 1–14 (2019).
18. Li, X. & Lehner, B. Biophysical ambiguities prevent accurate genetic prediction. *Nat. Commun*. **11**, 4923 (2020).
19. Otwinowski, J. Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. *Mol. Biol. Evol*. **35**, 2345–2354 (2018).
20. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
21. Woodsmith, J. et al. Protein interaction perturbation profiling at amino-acid resolution. *Nat. Methods* **14**, 1213–1221 (2017).
22. Cagiada, M. et al. Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. *Mol. Biol. Evol*. **38**, 3235–3246 (2021).
23. Domingo, J., Baeza-Centurion, P. & Lehner, B. The Causes and Consequences of Genetic Interactions (Epistasis). *Annu. Rev. Genom. Hum. Genet.* **20**, 433–460 (2019).
24. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, e32472 (2018).
25. Levy, E. D., Kowarzyk, J. & Michnick, S. W. High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. *Cell Rep*. **7**, 1333–1340 (2014).
26. Pelletier, J. N., Arndt, K. M., Plückthun, A. & Michnick, S. W. An in vivo library-versus-library selection of optimized protein-protein interactions. *Nat. Biotechnol*. **17**, 683–690

(1999).

27. Campbell-Valois, F.-X., Tarassov, K. & Michnick, S. W. Massive sequence perturbation of a small protein. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14988–14993 (2005).
28. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
29. Wei, X. et al. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet.* **10**, e1004819 (2014).
30. Horovitz, A., Fleisher, R. C. & Mondal, T. Double-mutant cycles: new directions and applications. *Curr. Opin. Struct. Biol.* **58**, 10–17 (2019).
31. Calosci, N. et al. Comparison of successive transition states for folding reveals alternative early folding pathways of two homologous proteins. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 19241–19246 (2008).
32. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830–838 (2011).
33. Nisthal, A., Wang, C. Y., Ary, M. L. & Mayo, S. L. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 16367–16377 (2019).
34. Laursen, L., Kliche, J., Gianni, S. & Jemth, P. Supertertiary protein structure affects an allosteric network. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 24294–24304 (2020).
35. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).
36. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 452–456 (1995).
37. Redler, R. L., Das, J., Diaz, J. R. & Dokholyan, N. V. Protein Destabilization as a Common Factor in Diverse Inherited Disorders. *J. Mol. Evol.* **82**, 11–16 (2016).
38. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53 (2013).

39. McLaughlin Jr, R. N. et al. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).

40. Wang, J. et al. Mapping allosteric communications within individual proteins. *Nat. Commun*. **11**, 3862 (2020).

41. Zhong, Q. et al. Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol*. **5**, 321 (2009).

42. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).

43. Kinney, J. B., Murugan, A., Callan Jr, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U. S. A*. **107**, 9158–9163 (2010).

44. Forcier, T. L. et al. Measuring cis-regulatory energetics in living cells using allelic manifolds. *eLife* **7**, e40618 (2018).

45. Tareen, A. et al. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. Preprint at https://www.biorxiv.org/content/10.1101/2020.07.14.201475 (2020).

46. Adams, R. M., Mora, T., Walczak, A. M. & Kinney, J. B. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife* **5**, e23156 (2016).

47. Kinney, J. B. & McCandlish, D. M. Massively parallel assays and quantitative sequence–function relationships. *Annu. Rev. Genomics Hum. Genet*. **20**, 99-127 (2019).

48. Skoulidis, F. et al. Sotorasib for Lung Cancers with KRAS p.G12C Mutation. *N. Engl. J. Med*. **384**, 2371–2381 (2021).

49. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

Figure legends

Fig. 1. *ddPCA* quantifies the effects of mutations on protein abundance and binding.
a, Overview of *ddPCA*. **b,** Reproducibility of fitness estimates from *ddPCA*. **c,** Comparison of individually measured growth rates to those inferred from deep sequencing for selected GRB2-SH3 variants covering a wide range of effects. The red line corresponds to the linear regression model. **d,** 3D structures of GRB2-SH3 bound to GAB2 (PDB entry 2VWF) and PSD95-PDZ3 bound to CRIPT (PDB entry 1BE9). **e,** Fitness density distributions. Total number of singles and doubles are indicated. Vertical continuous and dashed lines indicate the median fitness of the synonymous WT variants and of STOP codon mutations in the central 50% of the coding sequence, respectively. **f-g,** Heatmaps of fitness effects of single AA substitutions for GRB2-SH3 (**f**) and PSD95-PDZ3 (**g**) from *bindingPCA* (upper panel) and *abundancePCA* (lower panel) assays. Fitness values more extreme than ±1.5 were set to this limit. PDB residue numbering differs from UniProt for GRB2-SH3. **h.** Scatterplots comparing abundance and binding fitness of single AA substitutions. *r* = Pearson correlation coefficient.

Fig. 2. From molecular phenotypes to free energy changes.
a, Three-state equilibrium and corresponding thermodynamic model. **b,** Neural network architecture used to fit the thermodynamic model to the *ddPCA* data (target/output data; bottom), thereby inferring the causal changes in free energy of folding and binding associated with single AA substitutions (input values; top). **c,** Combinations of ΔG of binding and folding and the resulting fraction of bound protein complex (colour scale) illustrate how biophysical ambiguities (left) can be resolved by measuring more than one phenotype (middle) or by quantifying the effects of mutations in multiple starting genetic backgrounds (right). **d,** Performance of models fit to *ddPCA* data. R^2 = proportion variance explained. **e,** Comparisons of the confident model-inferred free energy changes to previously reported *in vitro* measurements. *r* = Pearson correlation coefficient. Free energies are from a single model; error bars indicate 95%CI from a Monte Carlo simulation approach (n = 10 experiments). **f,** Non-linear relationships (global epistasis) between observed *abundancePCA* fitness and changes in free energy of folding (top panels) or *bindingPCA* fitness and both free energies of binding and folding (bottom panels). Thermodynamic model fit shown in red. Free energy changes outside the interval [-2,7] are not shown.

Fig. 3. Binding and folding free energy landscapes of the SH3 and PDZ domains.
a-b, Heatmaps showing inferred changes in free energies of binding and folding for GRB2-SH3 (**a**) and PSD95-PDZ3 (**b**), Free energy changes of ligand-proximal residues are boxed and asterisks indicate major allosteric positions. Lower confidence estimates are indicated with dots (95%CI ≥ 1 kcal/mol). Free energy changes more extreme than ± 2.5 kcal/mol were set to this limit. **c,** Scatterplots comparing confident binding and folding free energy changes. Contours indicate estimates of 2D densities using 6 contour bins. Axis limits were adjusted to include the largest contour bin (more extreme data points are not shown). **d,** Distributions of confident binding and folding free energy changes. X-axis limits were adjusted to match those in panel c. **e,** Percentage of mutations that significantly decrease (top) or increase (bottom) fitness in the binding assay (FDR = 0.05) categorised by their biophysical mechanism. Pleiotropic mutations have significant changes in free energies of both folding and binding (FDR = 0.05) and are classified as either synergistic or antagonistic depending on whether their effects are in the same or different direction respectively. See Extended Data Fig. 5a-e for the GB1 domain.

Fig. 4. Mutational effects on protein stability.
a, 3D structures of the GRB2-SH3 and PSD95-PDZ3 domains where residue atoms are colored by the position-wise average change in the free energy of folding. **b,** Violin plots indicating distributions of confident changes in free energy of folding (n = 1,025 and n = 1,148 for GRB2-SH3 and PSD95-PDZ3 respectively; ****P* < 2.2e-16, two-sided Mann-Whitney U test comparing mutations in the core versus the remainder for both protein domains). **c,** Anti-correlation between the position-wise average change in free energy of folding and the solvent exposure of the corresponding residue (RSASA). *r* = Pearson

correlation coefficient. Error bars indicate 95%CI (n = 11-19 for GRB2-SH3; n = 18-19 for PSD95-PDZ3). **d,** Percentage of core, surface or binding interface residues shown separately for de-stabilising residues (positions with ≥ 5 stabilising mutations, folding $\Delta\Delta G < 0$, FDR = 0.05) and the remainder. Inset numbers are total counts. See Extended Data Fig. 6a-d for the GB1 domain.

Fig. 5. Major allosteric sites in protein binding domains.

a, Domain 3D structures where residue atoms are colored by the position-wise average absolute change in the free energy of binding. **b,** Domain structures with orthosteric and major allosteric site residues highlighted. **c,** Relationship between the position-wise average absolute change in free energy of binding and the minimal side chain heavy atom distance to the ligand. Major allosteric sites are defined as non-binding interface residues with weighted average absolute change in free energy of binding higher than the average of binding interface residue mutations. Class-switching residues in PSD95-PDZ3 are those that favour a change in specificity for a T-2F ligand defined in McLaughlin *et al*. 2012[39]. Error bars indicate 95%CI (n = 10-17 for GRB2-SH3; n = 17-19 for PSD95-PDZ3). See Extended Data Fig. 7a-c for the GB1 domain.

Fig. 6. Protein surfaces are frequent sites of binding affinity modulation.

a, Domain structures with highlighted surface allosteric sites and surface residues with allosteric mutations. **b,** Scatterplot showing the binding free energy changes of all mutations coloured according to residue position. **c,** Percentage of allosteric mutations per residue versus ligand proximity, excluding sites within the binding interface. ρ = Spearman rank correlation coefficient. Inset: enrichment (log2 odds ratio) of allosteric mutations at WT (or introducing mutant, Mut.) Glycines and Prolines in positions outside the binding interface or further restricted to those in secondary structure elements. The associated P value from a two-sided Fisher's Exact Test is indicated (*$P < 0.05$, ***$P < 0.001$). Also see Extended Data Fig. 9f,g. **d,** Total numbers of mutations decreasing or increasing *bindingPCA* fitness (i.e. the fraction of bound protein complex) beyond the indicated minimum or maximum thresholds (x-axis; two-sided Z-test $P < 0.05$) respectively. **e,** Similar to (**d**) except only mutations without significant effects on *abundancePCA* fitness are shown. See Extended Data Fig. 9a-d for the GB1 domain.

Acknowledgements

This work was funded by European Research Council (ERC) Advanced (883742) and Consolidator (616434) grants, the Spanish Ministry of Science and Innovation (PID2020-118723GB-I00, BFU2017-89488-P, EMBL Partnership, Severo Ochoa Centre of Excellence), the Bettencourt Schueller Foundation, the AXA Research Fund, Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322), and the CERCA Program/Generalitat de Catalunya. JMS was supported by an EMBO Long-Term Fellowship (ALTF 857-2016) and a Marie Skłodowska-Curie Fellowship (752809, EU Commission Horizon 2020). We thank Mafalda Dias, Jonathan Frazer and Deborah Marks for providing EVE and EVmutation predictions and all members of the Lehner Lab for helpful discussions and suggestions, especially Pablo Baeza-Centurion, Xianghua Li and Aaron M. New.

Author contributions

J.D., J.M.S., G.D. and B.L. conceived the project and designed the experiments; J.D., J.M.S. and C.H. constructed the mutant libraries; J.D. performed the yeast competition experiments with help from C.H.; J.D. constructed the sequencing libraries for NGS; A.J.F. led the data analysis with help from J.D. and J.M.S.; A.J.F., J.M.S and B.L. formulated the thermodynamic model; A.J.F. wrote the code to implement and fit the model; B.L., A.J.F. and J.D. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to B.L. Reprints and permissions information is available at www.nature.com/reprints.

Data availability

All DNA sequencing data have been deposited in the Gene Expression Omnibus with accession number GSE184042. Protein structures were obtained from the PDB and the AlphaFold Protein Structure Database with the following accessions:
GRB2-SH3: 2VWF (https://www.rcsb.org/structure/2VWF)
PSD95-PDZ3: 1BE9 (https://www.rcsb.org/structure/1BE9)
GB1: 1FCC (https://www.rcsb.org/structure/1FCC)
GRB2 homodimer: 1GRI (https://www.rcsb.org/structure/1GRI)
PSD95 AlphaFold prediction: P78352 (https://alphafold.ebi.ac.uk/entry/P78352)

Code availability

Source code used to fit thermodynamic models, to perform all downstream analyses and to reproduce all figures in this work is available at https://github.com/lehner-lab/doubledeepms.

Extended Data

Extended Data Fig. 1. Performance of thermodynamic models.

a, Distribution of the number of double AA substitutions comprising the same single AA substitution in the *abundancePCA* (blue) or *bindingPCA* (red) assays for the GRB2-SH3 (left) and PSD95-PDZ3 (right) protein domains. Median indicated with a dashed line and text label. **b-d**, 2d density plots comparing the *ddPCA* observed fitness and the model predicted fitness of single (left panels) and double AA substitutions (right panels) for the binding (top panels) and when existing, folding assays (bottom panels) of the GB1 (**b**), GRB2-SH3 (**c**) and PSD95-PDZ3 (**d**) domains. R^2 = proportion variance explained. **e-g**, Same as (**b-d**) but using validation data comprising 10% of double mutants held out during model fitting.

Extended Data Fig. 2. Performance of thermodynamic models after restricting data to a single phenotype or a single genetic background.

a, 2d density plots comparing the observed and predicted fitness of the binding (top panels) and abundance (bottom panels) assays when only the *bindingPCA* data is used for training the model for the GRB2-SH3 (left panels) and PSD95-PDZ3 (right panels). **b**, Same as in (**a**), but only using single mutant data from both binding and abundance assays to fit the models. R^2 = proportion variance explained. **c-d**, Comparisons of inferred free energy changes to previously reported PSD95-PDZ3 mutant *in vitro* measurements where only *bindingPCA* data (**c**) or single mutants (**d**) were used to fit thermodynamic models. Free energies are from a single model; error bars indicate 95%CI from a Monte Carlo simulation approach (n = 10 experiments) and the regression error bands indicate 95%CI for predictions from a linear model (panel c top: n = 22, bottom: n = 25, panel d top: n = 32, bottom: n = 29). *r* = Pearson correlation coefficient.

Extended Data Fig. 3. Performance of thermodynamic models after downsampling and comparisons of inferred free energy changes to smaller-scale datasets of *in vitro* measurements.

a, Dashed lines indicate the relationship between the percentage of fitness variance explained by model predictions with respect to held out validation data (10% of doubles) and the percentage of randomly retained double AA mutants used to train the model in the abundance (blue) or binding (red) assay. Results are shown separately for all protein domains. Solid lines indicate the relationship between the percentage variance explained by inferred free energies with respect to previously reported in vitro measurements for GB1 (Nisthal *et al.* 2019[33]) and PSD95-PDZ3 (Laursen *et al*. 2020[34] for ΔΔG binding, red; Calosci *et al*. 2008[31] for ΔΔG folding, blue), where models were trained using varying fractions of randomly downsampled double mutants (x-axis). The top scale indicates the median number of double AA mutants per single AA mutant in the full dataset. **b**, Comparisons of the model-inferred free energy changes to previously reported *in vitro* measurements for GRB2-SH3 (Malagrino *et al.* 2019[56] for ΔΔG binding and Troilo *et al.* 2018[57] for ΔΔG folding) and PSD95-PDZ3 (Chi *et al*. 2008[58]). Note the modest effect sizes of variants assayed in Malagrino *et al.* 2019. Free energies are from a single model; error bars indicate 95%CI from a Monte Carlo simulation approach (n = 10 experiments,*in vitro* error measurement not provided) and the regression error bands indicate 95%CI for predictions from a linear model (top left: n = 11, bottom left: n = 15, top right: n = 11, bottom right: n = 12). *r* = Pearson correlation coefficient.

Extended Data Fig. 4. Correlation of folding free energy changes with computational predictions of mutational effects.

a, High confidence inferred folding free energy changes versus corresponding FoldX[59] predictions upon mutation ("PositionScan" command), excluding substitutions involving potentially large increases in mass/volume (at wild-type Glycine, Alanine, Valine) or the replacement of Histidine (whose charge depends on the pH and local chemical environment). **b**, High confidence inferred folding free energy changes versus corresponding PolyPhen2[60] predictions for amino acid substitutions reachable by single nucleotide substitutions (SNPs). **c**, High confidence inferred folding free energy changes versus corresponding EVE pathogenicity scores[61]. **d**, Same as in (**c**), but scores are based on evolutionary couplings[62]. *r* = Pearson correlation coefficient.

Extended Data Fig. 5. Binding and folding free energy landscapes of the GB1 domain and biophysical mechanism of mutations that affect binding.

a-b, Heatmaps showing inferred changes in free energies of binding (**a**) and folding (**b**) for the GB1 domain. The final row in each heatmap indicates the minimal distance to the ligand (considering the side chain heavy atoms or the alpha carbon atoms in the case of glycine). Free energy changes of ligand-proximal residues (ligand distance < 5 Å) are boxed. Low confidence estimates are indicated with dots (95%CI ≥ 1 kcal/mol). Free energy changes more extreme than ±2.5 were set to this limit. **c**, Scatterplot comparing binding and folding free energy changes of mutations in the core, surface and binding interface. Contours indicate estimates of 2D densities with 6 contour bins. **d**, Distribution of binding (red) and folding (blue) free energy changes. **e**, Percentage of mutations that significantly decrease (top) or increase (bottom) fitness in the binding assay (FDR < 0.05) categorised by their biophysical mechanism. Pleiotropic mutations have significant changes in free energies of both folding and binding (FDR < 0.05) and are classified as either synergistic or antagonistic depending on whether their effects are in the same or different direction respectively. **f**, Changes in free energy of binding (blue) or folding (red) of single AA substitutions with different fitness effects in the binding assay for the three protein domains. **g**, Percentage of core, surface or ligand binding mutations that significantly decrease (top) or increase (bottom) fitness in the binding assay (FDR < 0.05) categorised by their biophysical mechanism. Pleiotropic mutations have significant changes in free energies of both folding and binding (FDR < 0.05) and are classified as either synergistic or antagonistic depending on whether their effects are in the same or different direction respectively.

Extended Data Fig. 6. GB1 mutational effects on protein stability and characterisation of surface de-stabilising residues.

a, 3D structure of GB1 (PDB entry 1FCC) where residue atoms are coloured by the position-wise average change in the free energy of folding. The FC domain of the human Immunoglobulin G is shown as black sticks. **b**, Violin plots indicating distributions of confident changes in free energy of folding (n = 898; ****P* < 2.2e-16, two-sided Mann-Whitney U test comparing mutations in the core versus the remainder). **c**, Anti-correlation between the position-wise average change in free energy of folding and the solvent exposure of the corresponding residue (RSASA) in GB1. Error bars indicate 95%CI (n = 19). *r* = Pearson correlation coefficient. **d**, Percentage of core, surface or binding interface residues in GB1 shown separately for de-stabilising residues (positions with ≥ 5 stabilising mutations, folding $\Delta\Delta G < 0$, FDR < 0.05) and the remainder. Inset numbers are total counts. **e**, Violin plots indicating evolutionary conservation scores (from a multiple sequence alignment of 185, 8,852, 276,481 homologous sequences of the GB1, GRB2-SH3 and PSD95-PDZ3 domains, respectively) shown separately for surface de-stabilising residues and remaining surface or core residues. **f**, Violin plots indicating hydrophobicity score distributions shown separately for surface de-stabilising residues and remaining surface or core residues. **g**, 3D structures of the GRB2-SH3 and PSD95-PDZ3 domains (grey cartoons) with the side-chains of surface de-stabilising residues highlighted in green sticks. Ligands are shown as black sticks. In the insets, in yellow is shown the SH2 domains of the second monomer of GRB2 when found in dimeric form (left, PDB entry 1GRI)[63], and relevant proximal portions of PSD95 C-terminal to the PDZ3 domain (middle and right, PDB entry 1BE9 and AlphaFold Protein Structure Database entry P78352).

Extended Data Fig. 7. Major allosteric sites in the GB1 domain and changes in free energy of binding in ligand binding interfaces.

a, 3D structures of the protein G B1 domain where residue atoms are coloured by the position-wise average absolute change in the free energy of binding. The FC domain of the human Immunoglobulin G is shown as black sticks. **b**, GB1 domain structure with binding interface residues (ligand distance < 5 Å) highlighted in red spheres and major allosteric site residues highlighted in orange spheres **c**, Relationship between the position-wise average absolute change in free energy of binding and the distance to the ligand (minimal side chain heavy atom distance) in the GB1 domain. Major allosteric sites (yellow) are defined as non-binding interface residues with weighted average absolute change in free energy of binding higher than the average of binding interface residue mutations (red). **d**, ROC curves for predicting ligand contacting residues (ligand distance < 5 Å) using (weighted) mean absolute binding $\Delta\Delta G$ considering all variants or those with confident inferred free energies (conf.). AUC = Area Under the Curve. **e**, Inferring changes in free energy of binding provides insights into the interactions that mediate binding between GRB2-SH3 and GAB2 peptide, and how mutations disrupt binding. F7 and Y51 of the GRB2-SH3 domain contact P3 and P4 of the GAB2 peptide through aromatic-proline interactions (left heatmap). In these two positions, only mutations to Y, F, Q and H, which can interact with proline through aromatic-proline or amino-aromatic interactions, are tolerated, while all other amino acid substitutions result in decreased binding affinity (positive binding $\Delta\Delta G$). Residue M46 can tolerate all amino acid substitutions except to positively charged residues (right heatmap). The closest residue of GAB2 is a lysine, and so a repulsive electrostatic interaction likely occurs when a positively charged amino acid occupies position 46 of the SH3 domain (binding $\Delta\Delta G$ of 2.1 and 1.99 for M46K and M46R respectively). **f**, ROC curves for predicting ligand contacting residues using (weighted) mean *bindingPCA* or *abundancePCA* fitness.

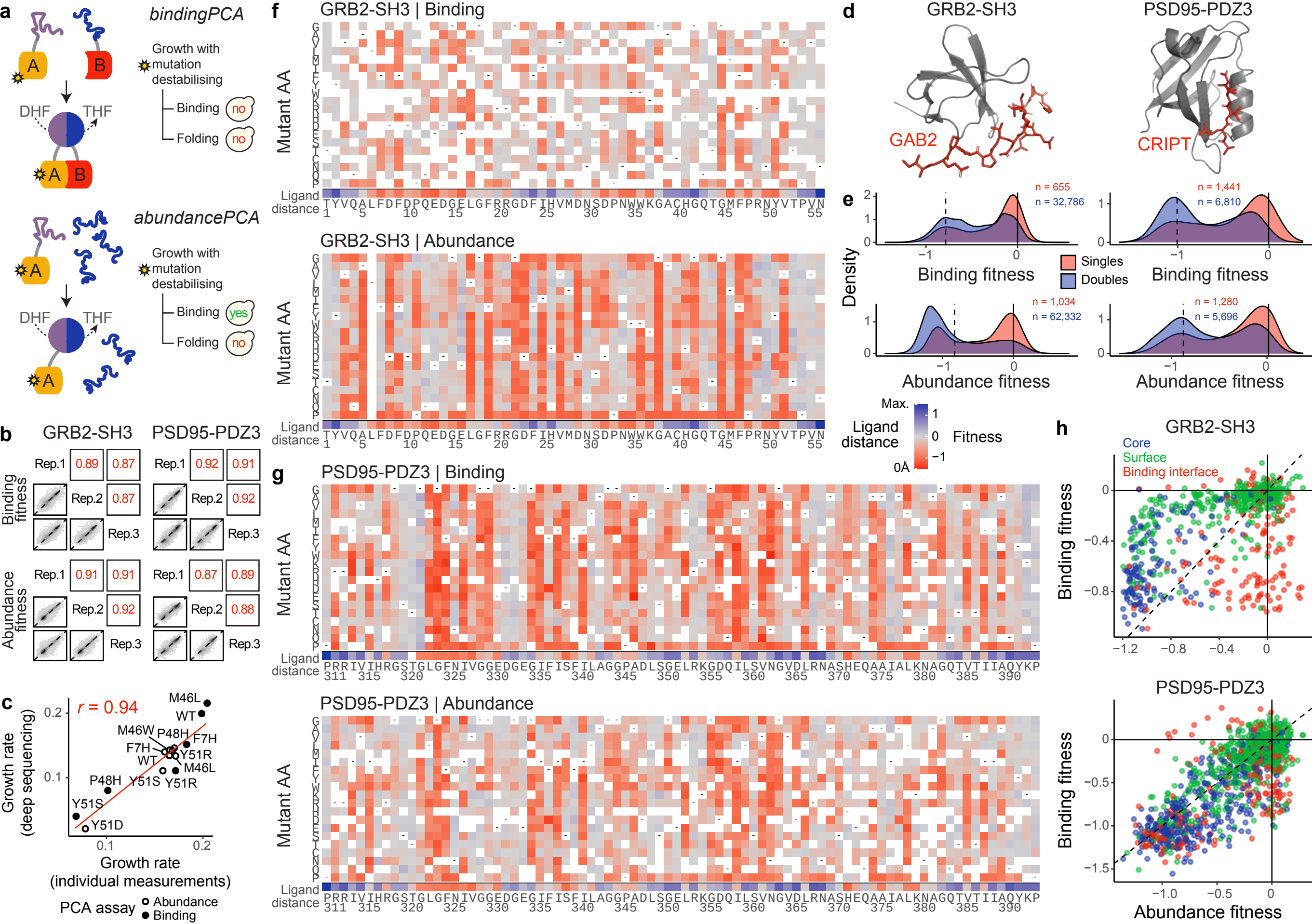
Extended Data Fig. 8. Changes in fitness and free energy of binding and folding of major allosteric sites and allosteric mutations.

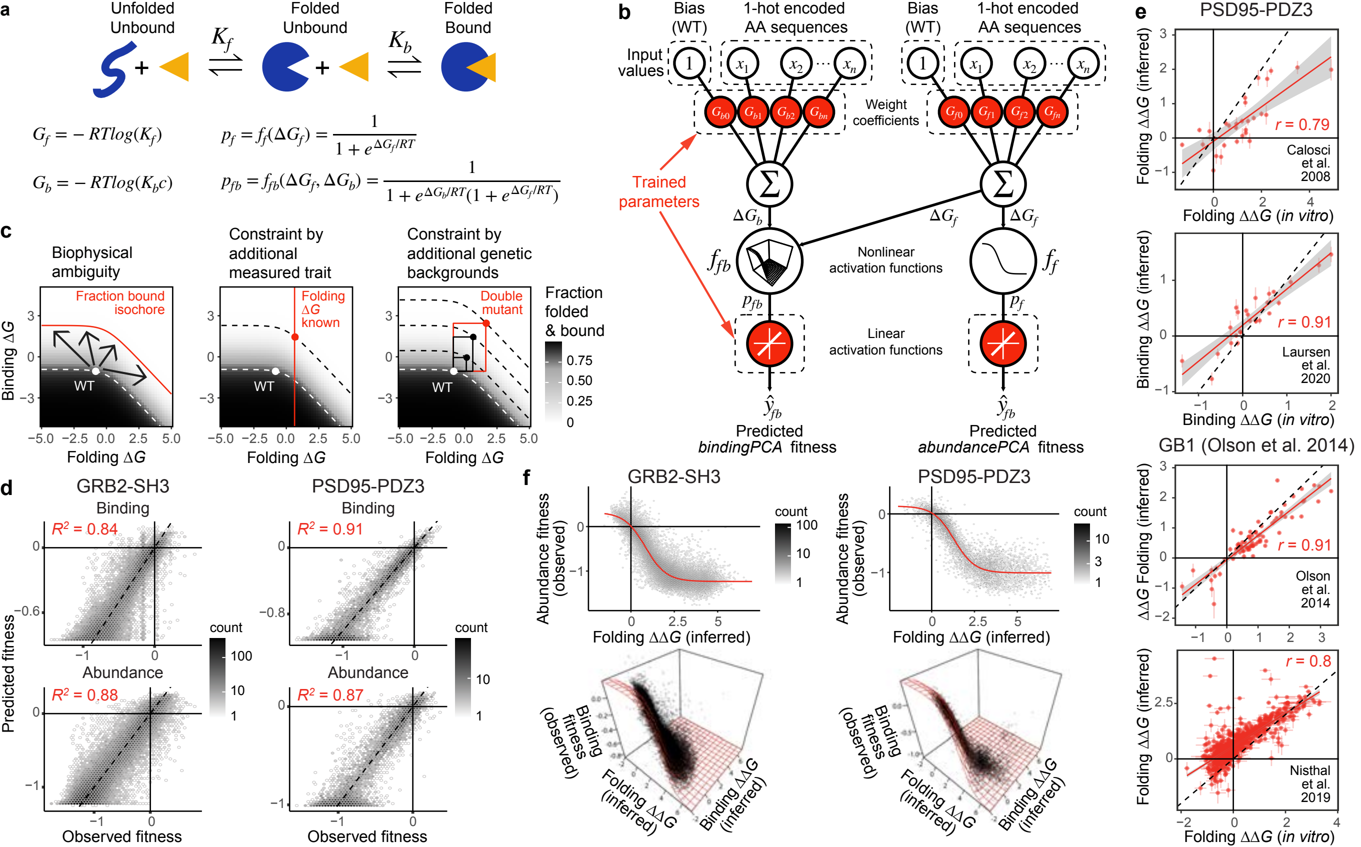
a, Scatterplots of single AA substitutions' changes in free energy of binding and folding for the GB1 (left panel), GRB2-SH3 (middle panel) and PSD95-PDZ3 (right panel) protein domains. Variants are coloured by AA position if found in a major allosteric site. Free energies are from a single model; error bars indicate 95%CI from a Monte Carlo simulation approach (n = 10 experiments). **b**, Scatterplots comparing abundance and binding fitness of single AA substitutions in the GRB2-SH3 (left panel) and PSD95-PDZ3 (right panel). Variants are coloured by AA position if found in a major allosteric site. Data are presented as mean values and error bars indicate 95%CI (n = 3 biological replicates). The red line indicates the model-derived relationship between abundance and binding fitness in the absence of a change in the free energy of binding. **c**, Scatterplots of single AA substitutions' changes in free energy of binding and folding for the GB1 (left panel), GRB2-SH3 (middle panel) and PSD95-PDZ3 (right panel) protein domains. Variants are coloured by AA position if found in a major allosteric site (yellow) or in a position that has allosteric mutations (green). Free energies are from a single model; error bars indicate 95%CI from a Monte Carlo simulation approach (n = 10 experiments). **d**, Scatterplots comparing abundance and binding fitness of single AA substitutions in the GRB2-SH3 (left panel) and PSD95-PDZ3 (right panel). Variants are coloured by AA position if found in a major allosteric site (yellow) or in a position that has allosteric mutations (green). Data are presented as mean values and error bars indicate 95%CI (n = 3 biological replicates). The red line indicates the model-derived relationship between abundance and binding fitness in the absence of a change in the free energy of binding.

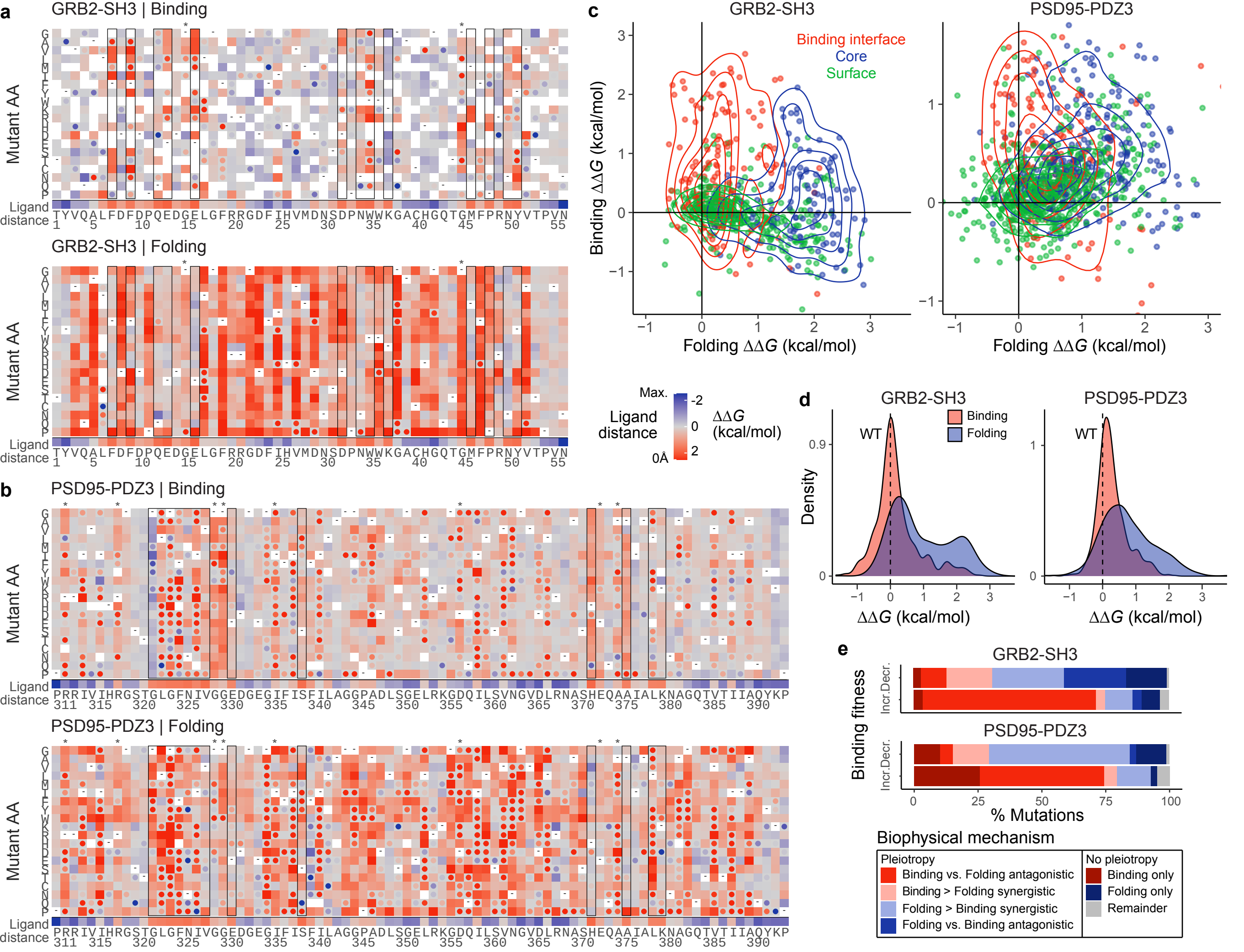
Extended Data Fig. 9. Allosteric mutations in GB1 and enrichment of allosteric mutations in literature allosteric networks and specific residue types and classes.
a, Domain structure of GB1 with surface allosteric sites and surface residues with allosteric mutations highlighted in orange and green respectively. The FC domain of the human Immunoglobulin G is shown as black sticks. **b**, Scatterplot showing the binding free energy changes of all mutations and coloured according to residue position: allosteric site (orange), orthosteric site/mutation (red), core allosteric mutation (blue), surface allosteric mutation (green). **c**, Percentage of allosteric mutations per residue versus ligand proximity, excluding sites within the binding interface. Points are coloured according to residue position and major allosteric sites are indicated (see legend). ρ = Spearman rank correlation coefficient. **d.** Total numbers of mutations decreasing or increasing binding fitness (i.e. the fraction of bound protein complex) beyond the indicated minimum or maximum thresholds (x-axis; two-sided Z-test $P < 0.05$) respectively. **e**, Enrichment of allosteric mutations in sets of residues defined by previously reported allosteric networks in PSD95-PDZ3: Mclaughlin *et al*. 2012[39], Salinas *et al*. 2018[64], Gerek *et al*. 2011[65], Kumawat *et al*. 2017[66], Gianni *et al*. 2011[67], Kalescky *et al*. 2015[68], Du *et al*. 2010[69], Kaya *et al*. 2013[70]. The log2 odds ratio corresponding to a 2x2 contingency table is shown on the x-axis and the associated *P* value from a two-sided Fisher's Exact Test is indicated. Residues within the binding interface (ligand distance < 5 Å) were ignored. Original literature allosteric network sizes are shown in parentheses. **f-g**, Same as (**e**) except sets of residues are defined by the identity of the WT or mutant amino acid (see legend) or their physicochemical properties (hydrophobic i.e. A, V, I, L, M, F, Y, W or charged i.e. R, H, K, D, E). Results are shown for all residues outside the binding interface (**f**) and further restricted to those residues in beta strands or helices i.e. not within loops/turns (**g**). Sets are ranked by their mean effect across the three protein domains.

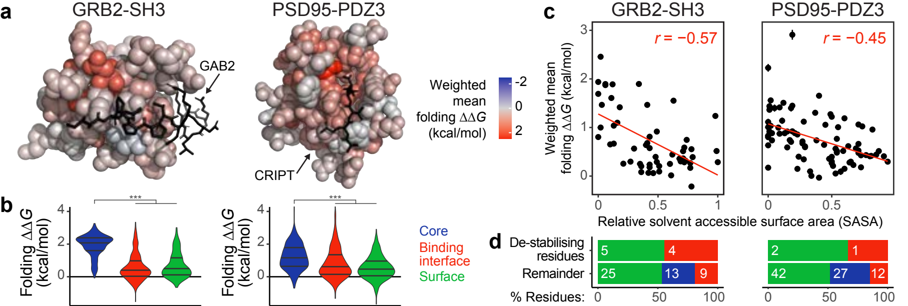
Extended Data Fig. 10. Comparisons to computationally predicted allosteric coupling scores and mutational biases towards increased or decreased binding given the position in the domain structure.
a, Percentage of allosteric mutations per residue versus allosteric coupling scores estimated by a network-based perturbation propagation algorithm[40], where residues in the binding interface (ligand distance < 5 Å) are omitted as they represent the query set. Residues immediately adjacent to binding interface residues in the linear AA sequence (i.e. backbone-backbone contacts which are disregarded by the Ohm algorithm) were given the maximum allosteric coupling score (1.0). Major allosteric sites (in yellow) and Spearman rank correlation coefficients (ρ) are indicated. **b**, Total numbers of mutations decreasing or increasing the free

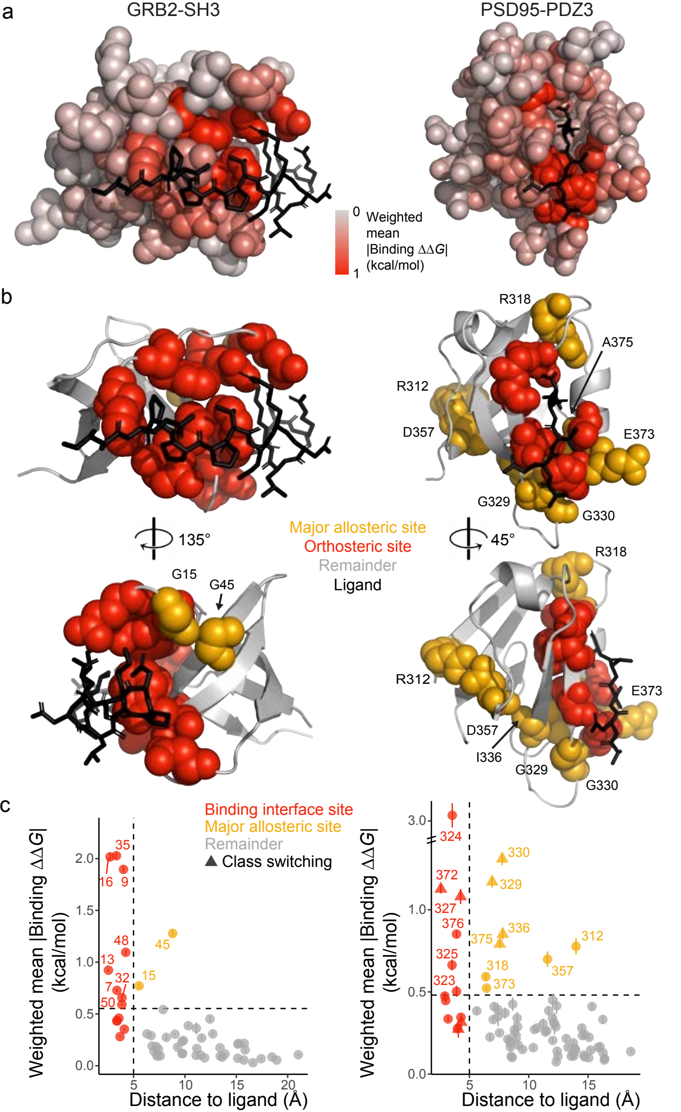
energy of binding beyond the indicated minimum or maximum thresholds (x-axis; two-sided Z-test *P* < 0.05) respectively, stratified by position in the structure considering all variants (regardless of the confidence of inferred free energies).

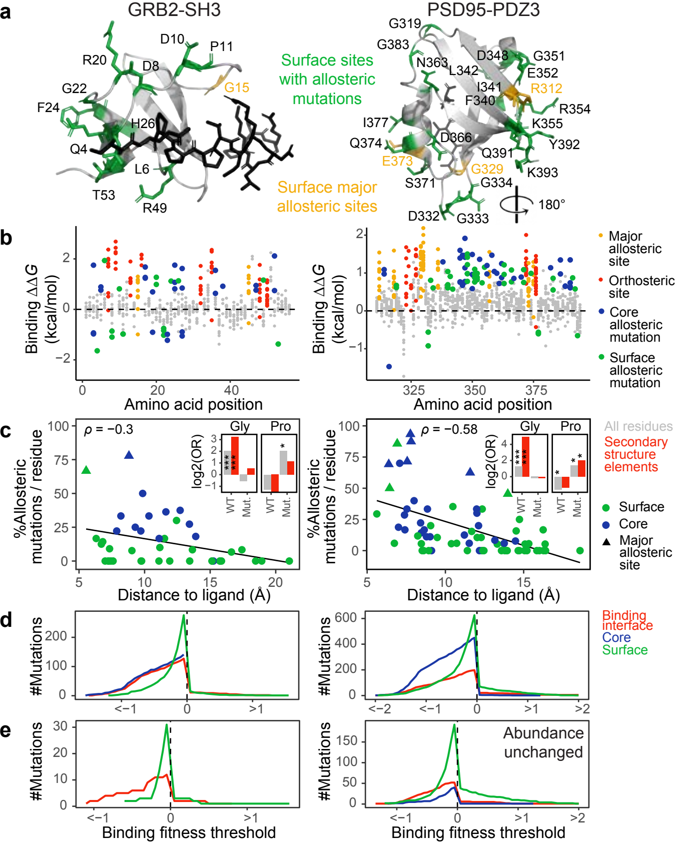

a
bindingPCA
A
B
Growth with mutation destabilising
Binding no
Folding no
DHF
THF
abundancePCA
Binding yes
Folding no
b
GRB2-SH3
PSD95-PDZ3
Binding fitness
Abundance fitness
Rep.1 0.89 0.87
Rep.2 0.87
Rep.3
Rep.1 0.92 0.91
Rep.2 0.92
Rep.3
Rep.1 0.91 0.91
Rep.2 0.92
Rep.3
Rep.1 0.87 0.89
Rep.2 0.88
Rep.3
c
r = 0.94
M46L
WT
M46W
P48H
F7H
F7H
WT
Y51R
M46L
P48H
Y51S
Y51R
Y51S
Y51D
Growth rate (deep sequencing)
Growth rate (individual measurements)
0.1
0.2
PCA assay ○ Abundance ● Binding
f
GRB2-SH3 | Binding
GRB2-SH3 | Abundance
Mutant AA
G A V L M I F Y W K R H D E S T C N Q P
Ligand distance
TYVQALFDFDPQEDGELGFRRGDFIHVMDNSDPNWWKGACHGQTGMFPRNYVTPVN
1 5 10 15 20 25 30 35 40 45 50 55
g
PSD95-PDZ3 | Binding
PSD95-PDZ3 | Abundance
PRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQTVTIIAQYKP
311 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390
d
GRB2-SH3
PSD95-PDZ3
GAB2
CRIPT
e
Density
n = 655
n = 32,786
n = 1,441
n = 6,810
n = 1,034
n = 62,332
n = 1,280
n = 5,696
Binding fitness
Abundance fitness
Singles
Doubles
−1
0
Max.
Ligand distance
0Å
1
0
−1
Fitness
h
GRB2-SH3
Core
Surface
Binding interface
Binding fitness
0
−0.4
−0.8
−1.2
PSD95-PDZ3
−0.5
−1.0
−1.5
Abundance fitness


a
Unfolded Unbound
Folded Unbound
Folded Bound
K_f
K_b
$G_f = -RTlog(K_f)$
$G_b = -RTlog(K_b c)$
$p_f = f_f(\Delta G_f) = \frac{1}{1 + e^{\Delta G_f/RT}}$
$p_{fb} = f_{fb}(\Delta G_f, \Delta G_b) = \frac{1}{1 + e^{\Delta G_b/RT}(1 + e^{\Delta G_f/RT})}$
b
Bias (WT)
1-hot encoded AA sequences
Input values
1
x_1
x_2
x_n
Weight coefficients
G_{b0}
G_{b1}
G_{b2}
G_{bn}
G_{f0}
G_{f1}
G_{f2}
G_{fn}
Trained parameters
Σ
ΔG_b
ΔG_f
f_{fb}
f_f
Nonlinear activation functions
p_{fb}
p_f
Linear activation functions
$\hat{y}_{fb}$
Predicted *bindingPCA* fitness
Predicted *abundancePCA* fitness
c
Biophysical ambiguity
Constraint by additional measured trait
Constraint by additional genetic backgrounds
Fraction bound isochore
Folding ΔG known
Double mutant
WT
Binding ΔG
Folding ΔG
3
0
−3
−5.0
−2.5
0.0
2.5
5.0
Fraction folded & bound
0.75
0.50
0.25
0
d
GRB2-SH3
PSD95-PDZ3
Binding
Abundance
$R^2 = 0.84$
$R^2 = 0.88$
$R^2 = 0.91$
$R^2 = 0.87$
Predicted fitness
Observed fitness
−0.6
−0.8
−1
0
count
100
10
1
e
PSD95-PDZ3
Folding ΔΔG (inferred)
Folding ΔΔG (*in vitro*)
r = 0.79
Calosci et al. 2008
Binding ΔΔG (inferred)
Binding ΔΔG (*in vitro*)
r = 0.91
Laursen et al. 2020
GB1 (Olson et al. 2014)
ΔΔG Folding (inferred)
r = 0.91
Olson et al. 2014
r = 0.8
Nisthal et al. 2019
2.5
f
GRB2-SH3
PSD95-PDZ3
Abundance fitness (observed)
Folding ΔΔG (inferred)
count
100
10
1
3
Binding fitness (observed)
Binding ΔΔG (inferred)


a
GRB2-SH3 | Binding
Mutant AA
G A V L M I F Y W K R H D E S T C N Q P
Ligand distance
TYVQALFDFDPQEDGELGFRRGDFIHVMDNSDPNWWKGACHGQTGMFPRNYVTPVN
1 5 10 15 20 25 30 35 40 45 50 55
GRB2-SH3 | Folding
Mutant AA
Ligand distance
TYVQALFDFDPQEDGELGFRRGDFIHVMDNSDPNWWKGACHGQTGMFPRNYVTPVN
1 5 10 15 20 25 30 35 40 45 50 55
b
PSD95-PDZ3 | Binding
Mutant AA
Ligand distance
PRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQTVTIIAQYKP
311 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390
PSD95-PDZ3 | Folding
Mutant AA
Ligand distance
PRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQTVTIIAQYKP
311 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390
Max.
Ligand distance
0Å
-2
0
2
ΔΔG (kcal/mol)
c
GRB2-SH3
PSD95-PDZ3
Binding interface
Core
Surface
Binding ΔΔG (kcal/mol)
Folding ΔΔG (kcal/mol)
3 2 1 0 −1
−1 0 1 2 3
d
GRB2-SH3
PSD95-PDZ3
WT
Binding
Folding
Density
0.9
1
0
ΔΔG (kcal/mol)
−1 0 1 2 3
e
GRB2-SH3
PSD95-PDZ3
Binding fitness
Incr.
Decr.
0 25 50 75 100
% Mutations
Biophysical mechanism
Pleiotropy
Binding vs. Folding antagonistic
Binding > Folding synergistic
Folding > Binding synergistic
Folding vs. Binding antagonistic
No pleiotropy
Binding only
Folding only
Remainder


a
GRB2-SH3
PSD95-PDZ3
GAB2
CRIPT
Weighted mean folding ΔΔG (kcal/mol)
-2
0
2
b
Folding ΔΔG (kcal/mol)
4
2
0
***
Core
Binding interface
Surface
c
GRB2-SH3
PSD95-PDZ3
r = −0.57
r = −0.45
Weighted mean folding ΔΔG (kcal/mol)
3
2
1
0
0.5
Relative solvent accessible surface area (SASA)
d
De-stabilising residues
Remainder
% Residues:
5
4
25
13
9
2
1
42
27
12
50
100


a
GRB2-SH3
PSD95-PDZ3
0
1
Weighted mean |Binding ΔΔG| (kcal/mol)
b
135°
G15
G45
Major allosteric site
Orthosteric site
Remainder
Ligand
45°
R318
A375
R312
D357
E373
G329
G330
I336
c
Binding interface site
Major allosteric site
Remainder
▲ Class switching
Weighted mean |Binding ΔΔG| (kcal/mol)
Distance to ligand (Å)
0.0
0.5
1.0
1.5
2.0
5
10
15
20
35
16
9
48
13
32
7
50
45
15
3.0
324
330
372
329
327
376
336
375
312
325
318
357
323
373


a
GRB2-SH3
PSD95-PDZ3
D10
P11
R20
D8
G15
G22
F24
H26
Q4
L6
T53
R49
Surface sites
with allosteric
mutations
Surface major
allosteric sites
G319
G383
N363
L342
D348
G351
E352
R312
I341
F340
R354
I377
K355
Y392
D366
Q374
Q391
E373
G329
K393
S371
G334
D332
G333
180°
b
Binding ΔΔG
(kcal/mol)
2
0
−2
0
20
40
Amino acid position
2
1
0
−1
325
350
375
Amino acid position
● Major allosteric site
● Orthosteric site
● Core allosteric mutation
● Surface allosteric mutation
c
%Allosteric
mutations / residue
ρ = −0.3
100
75
50
25
0
5
10
15
20
Distance to ligand (Å)
Gly
Pro
log2(OR)
3
2
1
0
−1
***
***
*
WT
Mut.
WT
Mut.
ρ = −0.58
4
2
0
***
***
*
*
*
5
10
15
Distance to ligand (Å)
All residues
Secondary structure elements
● Surface
● Core
▲ Major allosteric site
d
#Mutations
200
100
0
<−1
0
>1
600
400
200
0
<−2
<−1
0
>1
>2
Binding interface
Core
Surface
e
#Mutations
30
20
10
0
<−1
0
>1
Binding fitness threshold
150
100
50
0
<−1
0
>1
>2
Binding fitness threshold
Abundance unchanged