



Figures and figure supplements

Coevolution-based inference of amino acid interactions underlying protein function

Victor H Salinas and Rama Ranganathan

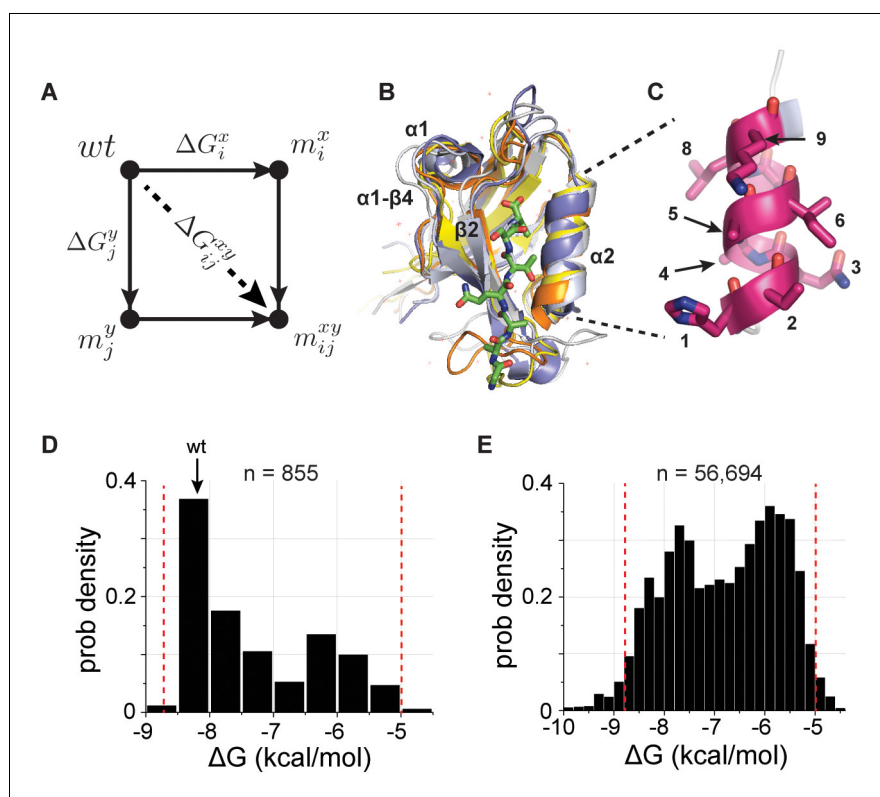


Figure 1. A deep coupling scan (DCS) for the PDZ binding pocket. (A), The thermodynamic double mutant cycle (TDMC), a formalism for studying the energetic coupling of pairs of mutations in a protein. Given two mutations (x at position i and y at position j), the coupling free energy between them is defined as the extent to which the effect of the double mutation (ΔG_{ij}^{xy}) is different from the summed effect of the mutations taken individually ($\Delta G_i^x + \Delta G_j^y$), a measure of the interaction (or epistasis) between the two mutations (see **Equation 1**, main text). (B), Structural overlay of the five PDZ homologs used in this study (PSD95^{pdz3} (1BE9, white), PSD95^{pdz2} (1QLC, orange), ZO1^{pdz} (2RRM, yellow), Shank3^{pdz} (5IZU, gray), and Syntrophin^{pdz} (1Z86, blue)), emphasizing the conserved $\alpha\beta$ -fold architecture of these sequence-diverse proteins (33% average identity, **Table 1**). Structural elements discussed in this work are indicated. (C), The nine-amino acid $\alpha 2$ -helix, which forms one wall of the ligand-binding site. (D–E), The distribution of experimentally determined binding free energies, ΔG_{bind} , for all single mutations (D, 855/855) and nearly all double mutations (E, 56,694/64,980) in the $\alpha 2$ -helix for the 5 PDZ homologs, with the affinity of wild-type PSD95^{pdz3} indicated (wt). The red lines indicate the independently validated range of the assay (**Figure 1—figure supplement 1**); essentially all measurements fall within this range. These data comprise the basis for a deep analysis of conserved thermodynamic coupling in the PDZ family.

DOI: <https://doi.org/10.7554/eLife.34300.002>

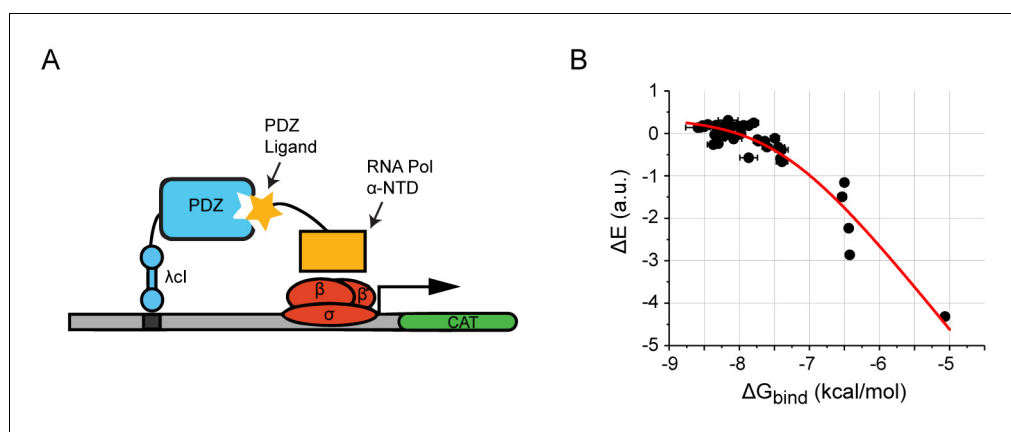


Figure 1—figure supplement 1. The bacterial two-hybrid assay for PDZ ligand binding. (A) The bacterial two-hybrid system consists of three parts, (1) a PDZ domain fused to the C-terminus of a λ -cl DNA binding domain, (2) a ligand peptide as a C-terminal fusion to the α -subunit of RNA polymerase, and (3) a reporter plasmid encoding chloramphenicol acetyltransferase (CAT). Ligand binding induces transcription of the reporter gene, which enables growth in the presence of the antibiotic chloramphenicol. (B) Binding free energies for 45 PSD95^{pdz3} single mutants plotted against the relative enrichment values ΔE derived from deep sequencing before and after selection (see Supplementary Methods). Horizontal bars around each point represent the range of values from replicate binding experiments, and the red curve represents a fit to a model in which the relative enrichment is proportional to the log fraction bound of ligand ($\Delta E \propto \ln f_B$, see Supplementary Methods). Parameters of the assay (inductions conditions, temperature, length of growth, promoter structure) are optimized such relative enrichment values quantitatively report the binding free energy (fit shows an adjusted $r^2 = 0.91$).

DOI: <https://doi.org/10.7554/eLife.34300.003>

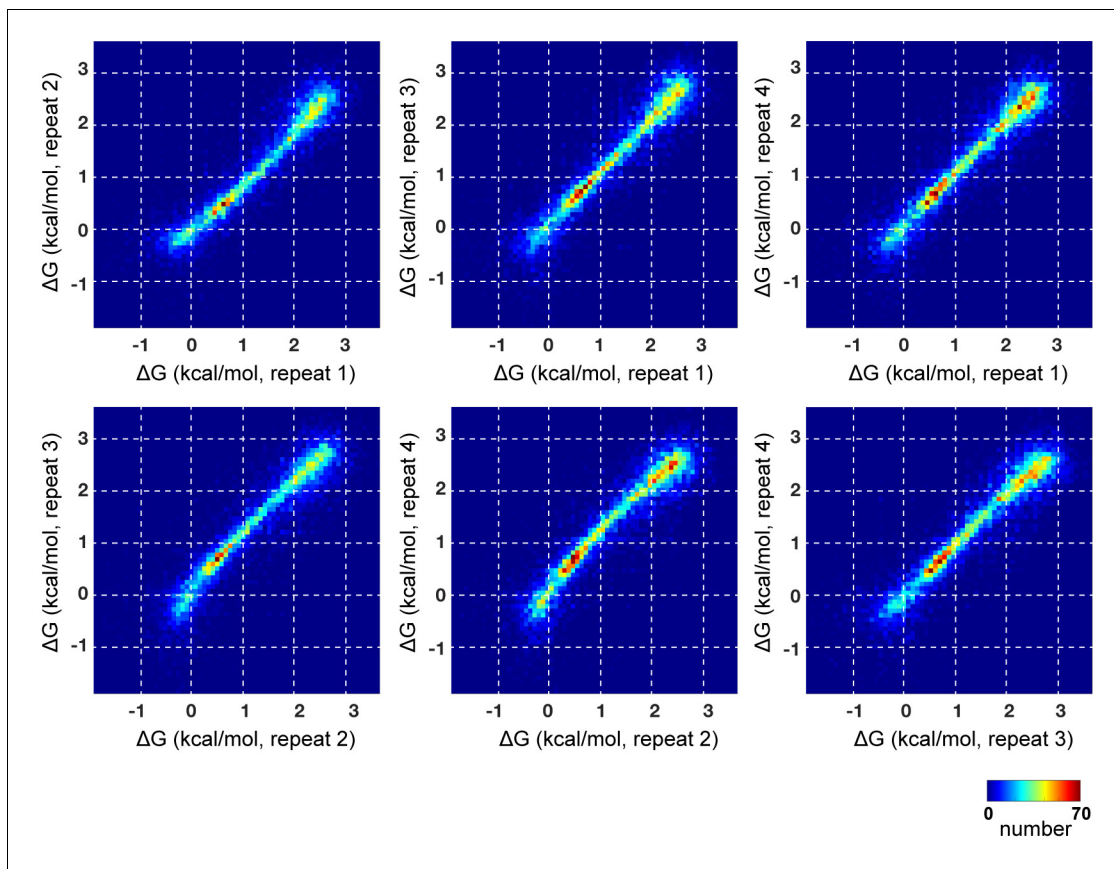


Figure 1—figure supplement 2. Reproducibility and quality of the bacterial two-hybrid assay. The panels show 2D histograms of all single and double mutant binding free energies for four independent experimental trials of the bacterial two-hybrid assay for PSD95^{pdz3}. The data show that the assay is remarkably reproducible, with >95% of values with a variance less than 0.31 kcal/mol. See **Table 1** for counting statistics of all mutants and number of double mutant cycles estimated per PDZ homolog.

DOI: <https://doi.org/10.7554/eLife.34300.004>

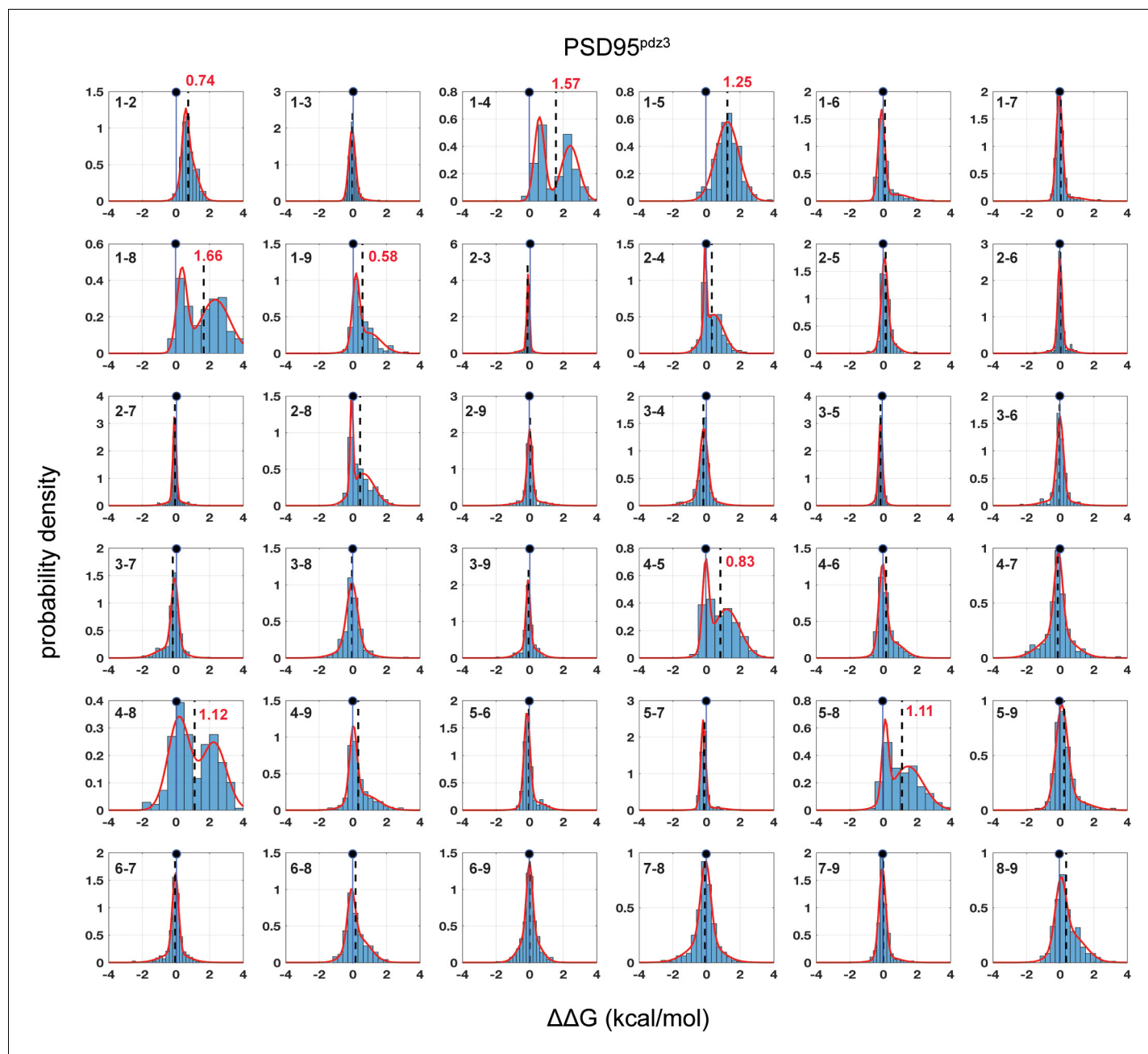


Figure 2. Distributions of pairwise thermodynamic couplings in a single PDZ homolog (PSD95^{pdz3}). Each subplot shows the distribution of coupling free energies ($\Delta\Delta G$, see Equation 1, main text) for all measured mutants at one pair of positions in the $\alpha 2$ -helix (numbering per Figure 1C) in PSD95^{pdz3}. The distributions are fit to single or double Gaussians, using the Bayes Information Criterion to justify choice of model, and the position of zero coupling is indicated by the solid line and circle above. Population-weighted mean values are represented by dashed lines. The data are remarkably well defined by the fitted models. Most position pairs have distributions centered close to zero, with only eight pairs comprising all pairwise couplings between positions 1, 4, 5, and 8, and 1-2, 1-9 showing deviations. For these pairs, distributions of mutational coupling follow either a single mode (1-2, 1-5) or two modes with one centered at zero (1-4, 1-8, 1-9, 4-5, 4-8, 5-8); population-weighted mean values for these pairs are indicated in red.

Figure 2—figure supplement 1–4 show similar data for each of the other homologs taken individually.

DOI: <https://doi.org/10.7554/eLife.34300.006>

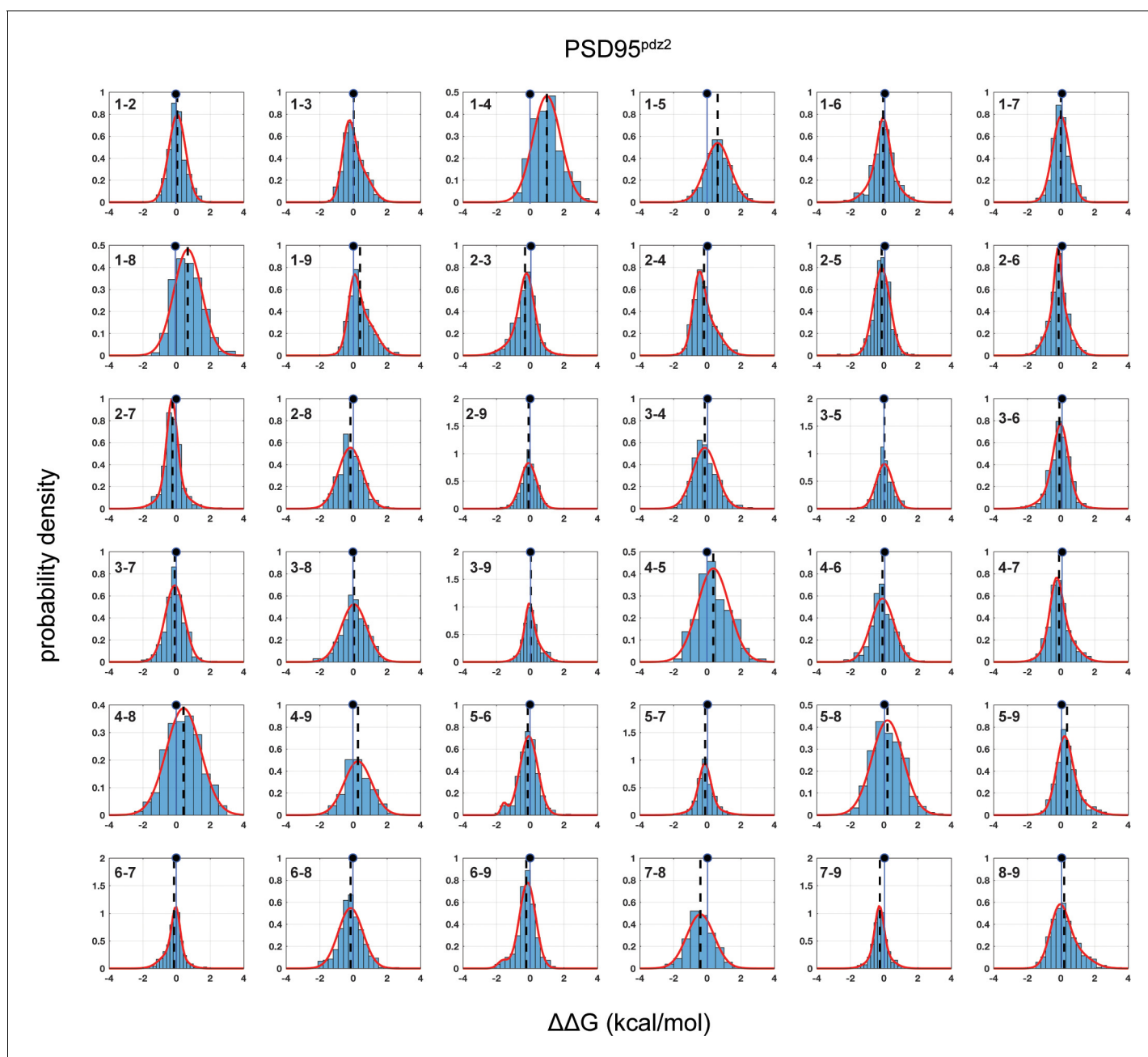


Figure 2—figure supplement 1. Distributions of double mutant cycles for each $\alpha 2$ helix position pair (numbering as in *Figure 1C*) for PSD95^{pdz2}. As in *Figures 2–3*, distributions are fit to single or double Gaussians, using the Bayes Information Criterion to choose the model. The position of zero coupling is indicated by the solid line and circle above, and the population weighted mean is represented in dashed lines.

DOI: <https://doi.org/10.7554/eLife.34300.007>

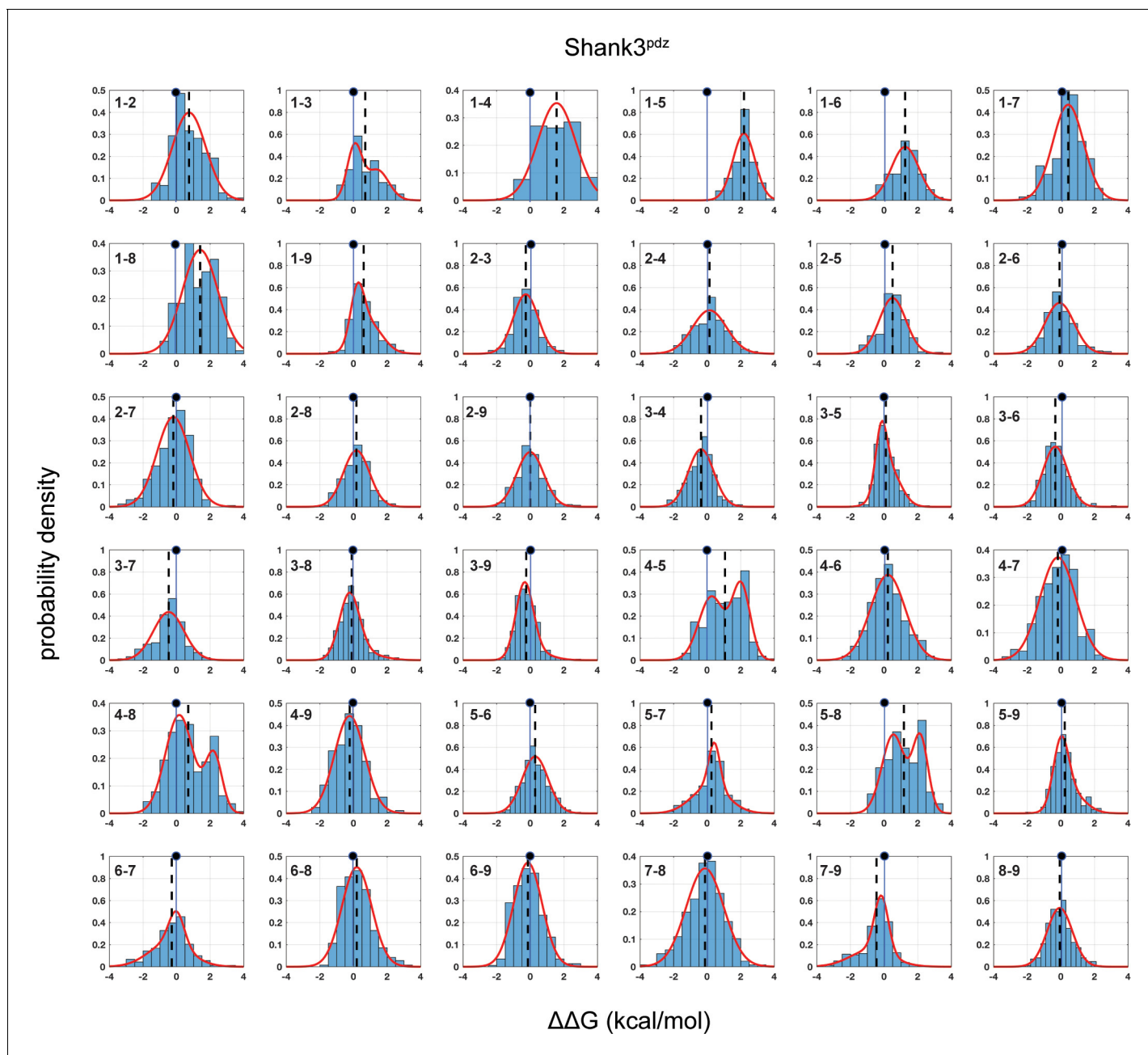


Figure 2—figure supplement 2. Distributions of double mutant cycles for each $\alpha 2$ helix position pair (numbering as in **Figure 1C**) for Shank3^{pdz}. As in **Figure 3**, 2–3, distributions are fit to single or double Gaussians, using the Bayes Information Criterion to choose the model. The position of zero coupling is indicated by the solid line and circle above, and the population weighted mean is represented in dashed lines.

DOI: <https://doi.org/10.7554/eLife.34300.008>

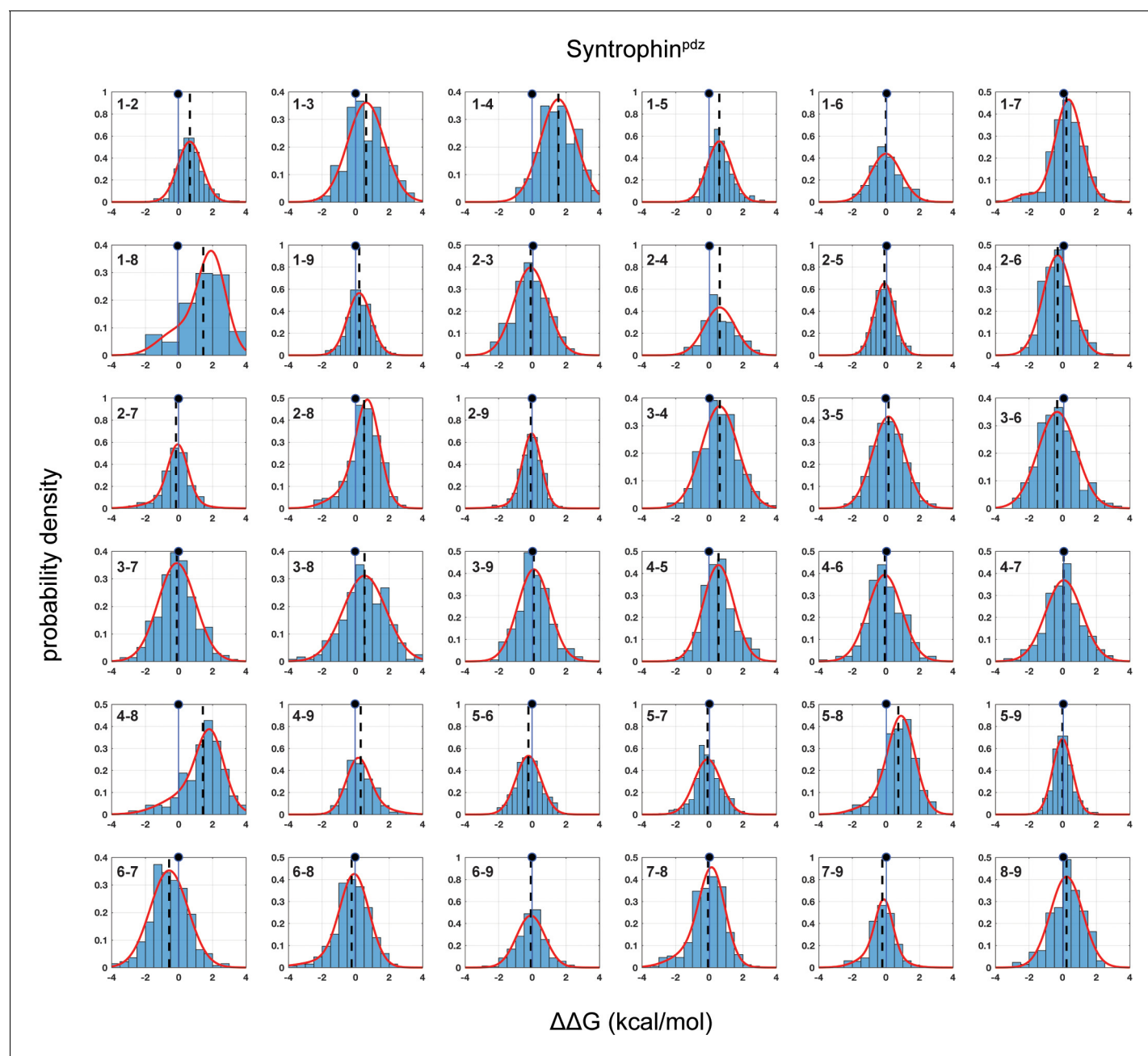


Figure 2—figure supplement 3. Distributions of double mutant cycles for each $\alpha 2$ helix position pair (numbering as in **Figure 1C**) for Syntrophin^{pdz}. As in **Figures 2–3**, distributions are fit to single or double Gaussians, using the Bayes Information Criterion to choose the model. The position of zero coupling is indicated by the solid line and circle above, and the population weighted mean is represented in dashed lines.

DOI: <https://doi.org/10.7554/eLife.34300.009>

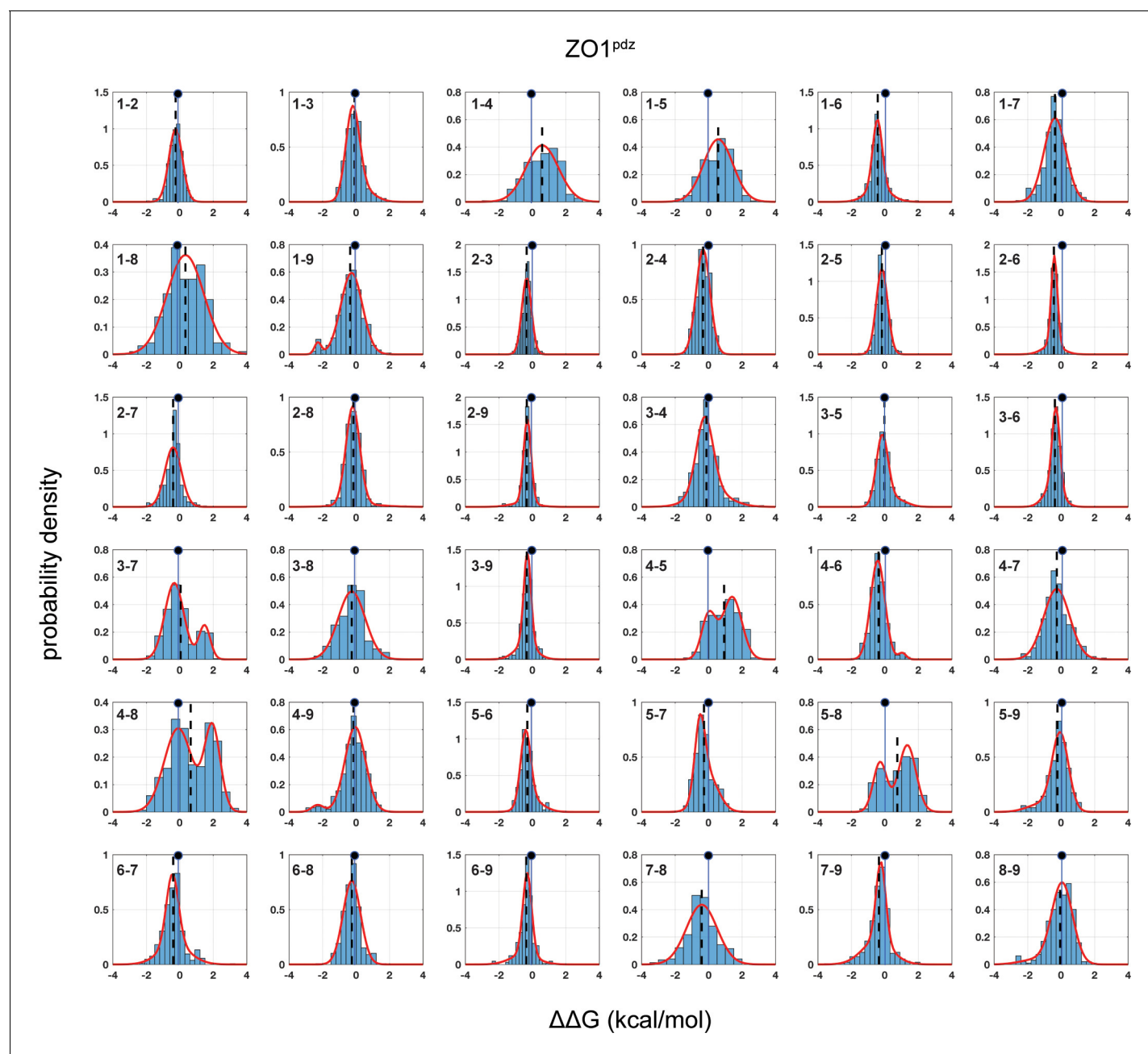


Figure 2—figure supplement 4. Distributions of double mutant cycles for each $\alpha 2$ helix position pair (numbering as in **Figure 1C**) for $ZO1^{pdz}$. As in **Figures 2–3**, distributions are fit to single or double Gaussians, using the Bayes Information Criterion to choose the model. The position of zero coupling is indicated by the solid line and circle above, and the population weighted mean is represented in dashed lines.

DOI: <https://doi.org/10.7554/eLife.34300.010>

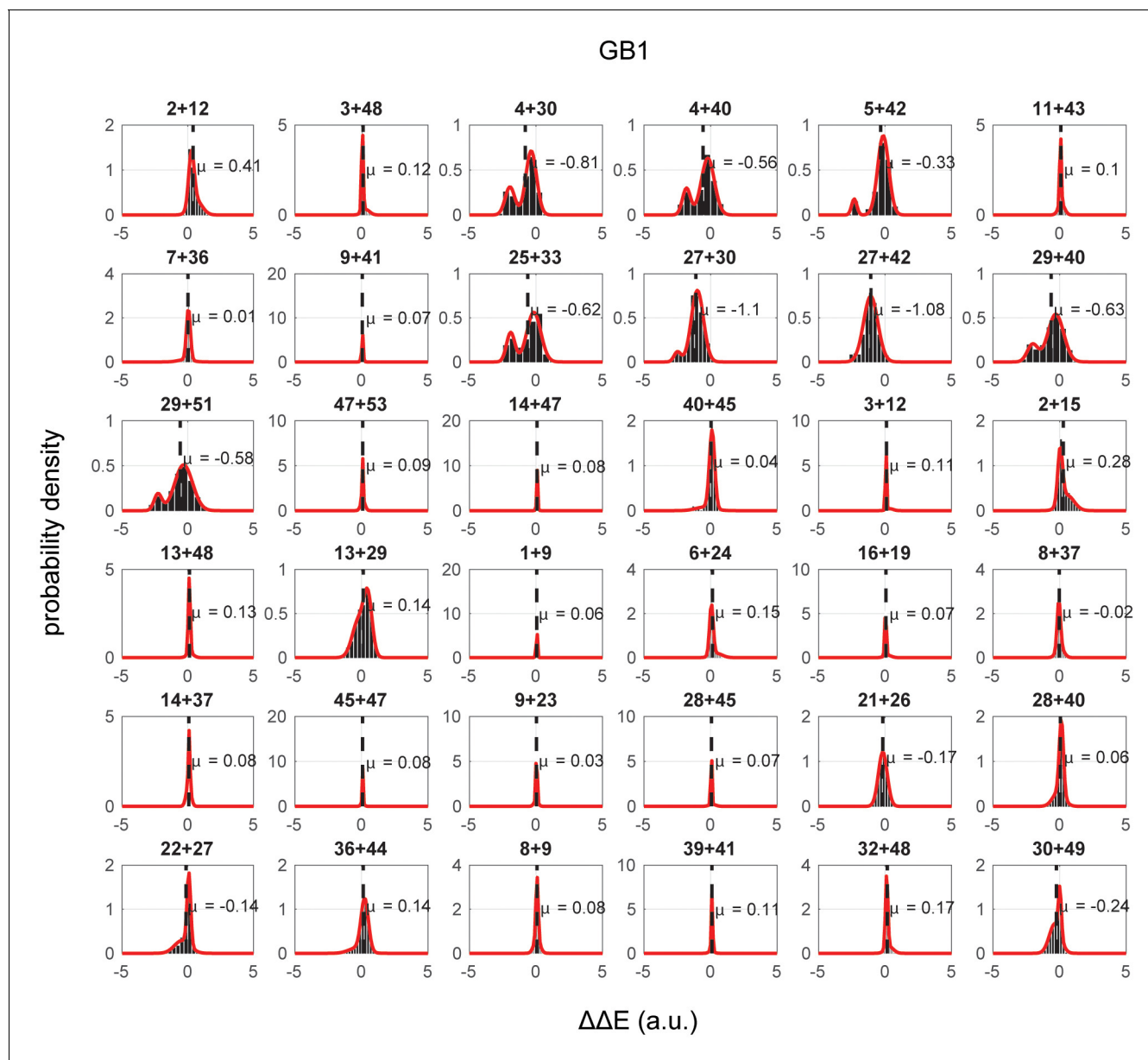


Figure 2—figure supplement 5. Distributions of coupling in relative enrichment ($\Delta\Delta E$) for a sampling of position pairs in the GB1 domain of the immunoglobulin-binding protein G. The data are from (Olson et al., 2014), and position numbering is as given in that work; note that coupling is calculated here directly from enrichment scores. As in Figures 1 and 2, distributions are fit to single or double Gaussians, using the Bayes Information Criterion to choose the model. The population weighted mean is represented in dashed lines, and values for each position pair are shown. Note that coupling energies are reported with the opposite sign due to the inverse relationship between E in Olsen et al. and ΔG in this work.

DOI: <https://doi.org/10.7554/eLife.34300.011>

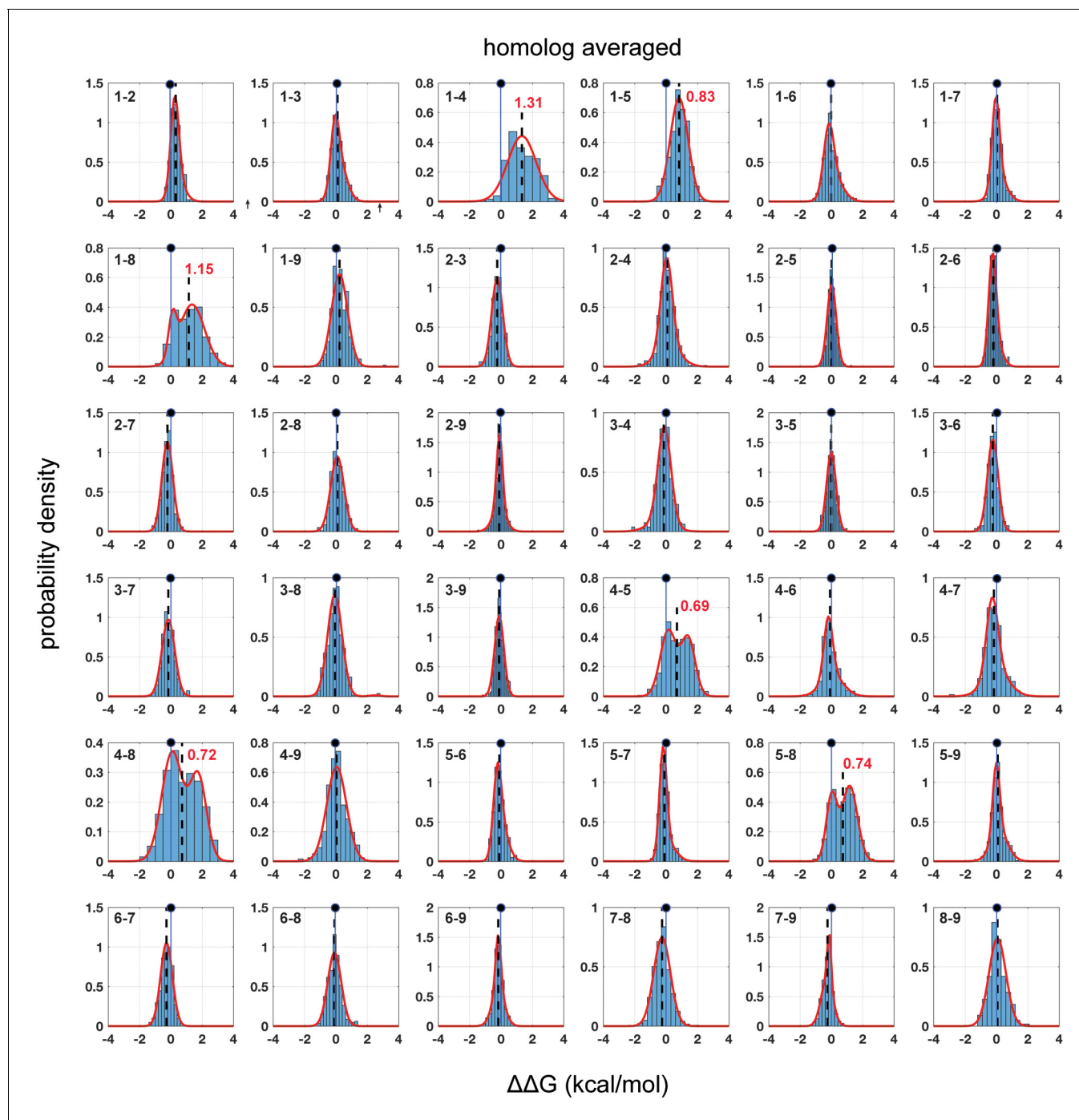


Figure 3. Homolog-averaged pairwise thermodynamic couplings in the PDZ domain. Each subplot shows the distribution of coupling free energies ($\Delta\Delta G$, see [Equation 1](#), main text) for all measured mutants at one pair of positions in the $\alpha 2$ -helix, but here averaged over the five homologs. As in [Figure 2](#), the distributions are fit to single or double Gaussians, using the Bayes Information Criterion to justify choice of model. The position of zero coupling is indicated by the solid line and circle above and population-weighted mean values are represented by dashed lines. Averaging over homologs reveals the conserved pattern of couplings; now, only six pairs comprising all pairwise couplings between positions 1, 4, 5, and 8 show deviations from zero. For these pairs, distributions of mutational coupling follow either a single mode (1-4, 1-5) or two modes with one centered at zero (1-8, 4-5, 4-8, 5-8); population-weighted mean values for these pairs are indicated in red.

DOI: <https://doi.org/10.7554/eLife.34300.012>

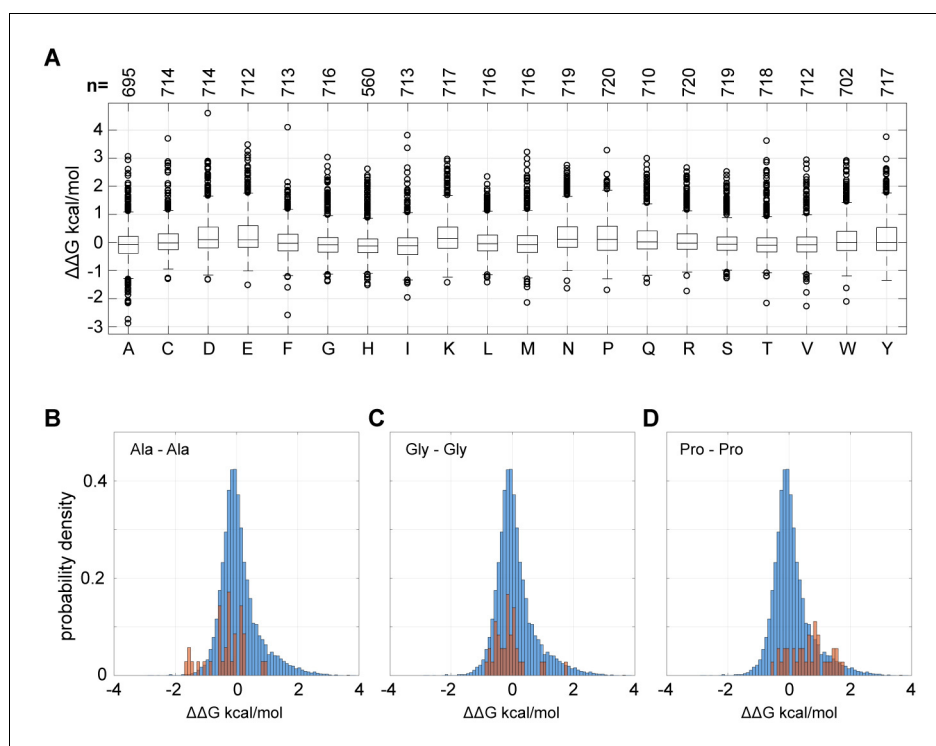


Figure 3—figure supplement 1. Amino acid contributions to distributions of double mutant cycle coupling energies for the PDZ $\alpha 2$ helix. (A), Boxplots showing the couplings for each amino acid substitution with every other amino acid substitution in the homolog-averaged dataset. The data show that every amino acid substitution is involved in essentially the full range of observed couplings, and mean couplings for each mutation is near zero. These findings indicate that the magnitude of couplings is not a simple property of the average chemical properties of different amino acids, and that mutations on average are subtle perturbations to protein function. (B–D), Each graph shows the distribution of pairwise thermodynamic couplings for all pairs of positions in the homolog-averaged dataset (blue), with couplings for Ala – Ala (B), Gly – Gly (C), and Pro – Pro (D) mutations highlighted in red. The data suggest a broad range of couplings for all these substitutions, arguing that the magnitude of coupling energies is not obviously defined by intuitive classifications of mutations expected to be subtle (Ala) and mutations expected to be disruptive for the $\alpha 2$ helix (Gly, Pro).

DOI: <https://doi.org/10.7554/eLife.34300.013>

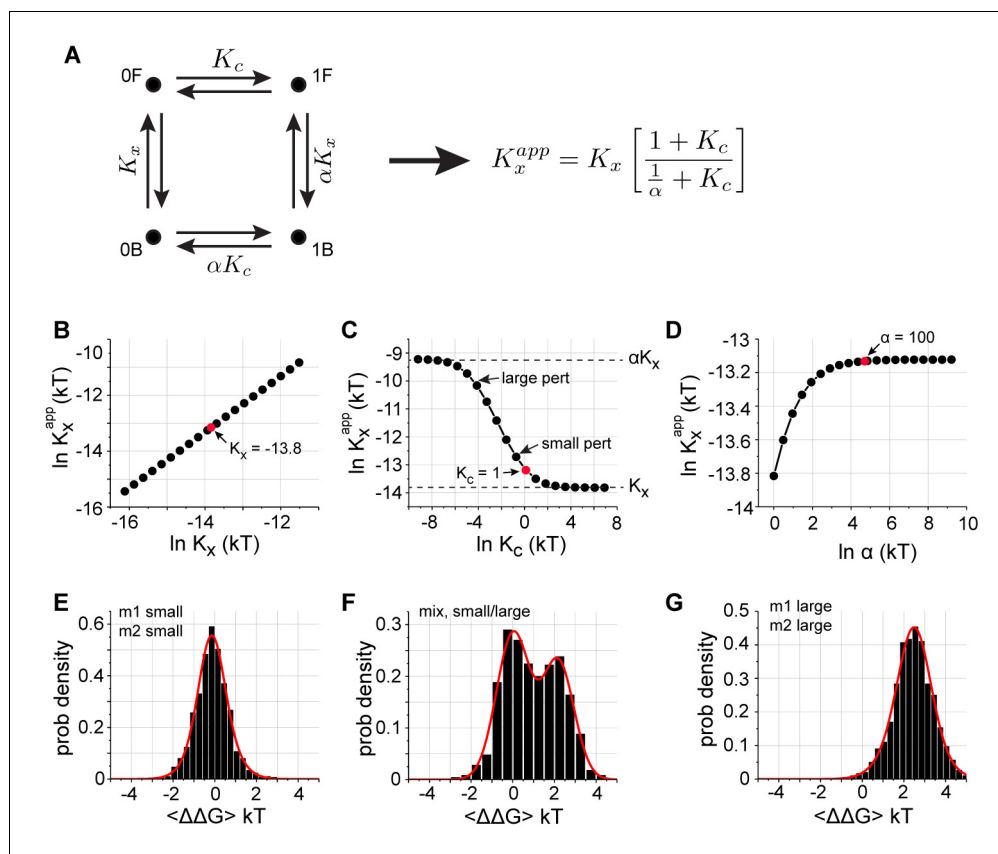


Figure 4. A basic model for observed distributions of double mutant cycle coupling energies. (A), A schematic representation of two coupled equilibria in a protein molecule – a reaction with equilibrium constant K_x corresponding to function (here, binding), an internal two-state conformational equilibrium defined by K_c , and a coupling parameter α linking the two. The equation at right shows the general analytic solution for how the apparent equilibrium constant K_x^{app} depends on these three parameters, and panels (B–D) show graphs of these relationships over a relevant range of values. Note that K_x (and αK_x) are defined as dissociation constants, and $K_c \equiv [0F]/[1F]$ and $\alpha K_c \equiv [0B]/[1B]$. (B–D), K_x^{app} shows a linear dependence on K_x , a saturating relationship with α , and a sigmoidal relationship with K_c . For a range of K_c , K_x^{app} ranges between K_x and αK_x , the two extreme limits set by the reaction diagram in panel (A). (E–G), distributions of coupling energies for simulations in which we choose a set of ‘wild-type’ values of K_x , K_c , and α (red dots, panels B–D) and consider mutations that cause random Gaussian perturbations of K_x and α , but either small or large perturbations of K_c (indicated in panel C). If all mutations cause small effects in K_c , we obtain unimodal distributions centered at zero coupling energy (E), and if all mutations cause large effects in K_c , we obtain unimodal distributions centered at a non-zero coupling energy (G). However, if mutations cause a mix of small and large effects on K_c , we obtain bimodal distributions with one mode centered at zero (F). These three types recapitulate all the observed distributions for all PDZ homologs (main Figure 2 and Figure 1—figure supplement 1–4), for the GB1 protein (Figure 1—figure supplement 5), and for the average over homologs (Figure 3). Note that higher order cooperativity between amino acids specifying K_c (a plausible scenario), would further steepen the relationship shown in panel (C) and would cause the all-or-nothing character of mutations with regard to K_c with even less distinction between large and small perturbations. This model is not intended as a proof of mechanism for the observed distributions, but instead provides a logical scheme that explains the observations in light of known two-state allosteric equilibria in some PDZ domains (Mishra et al., 2007; Raman et al., 2016).

DOI: <https://doi.org/10.7554/eLife.34300.014>

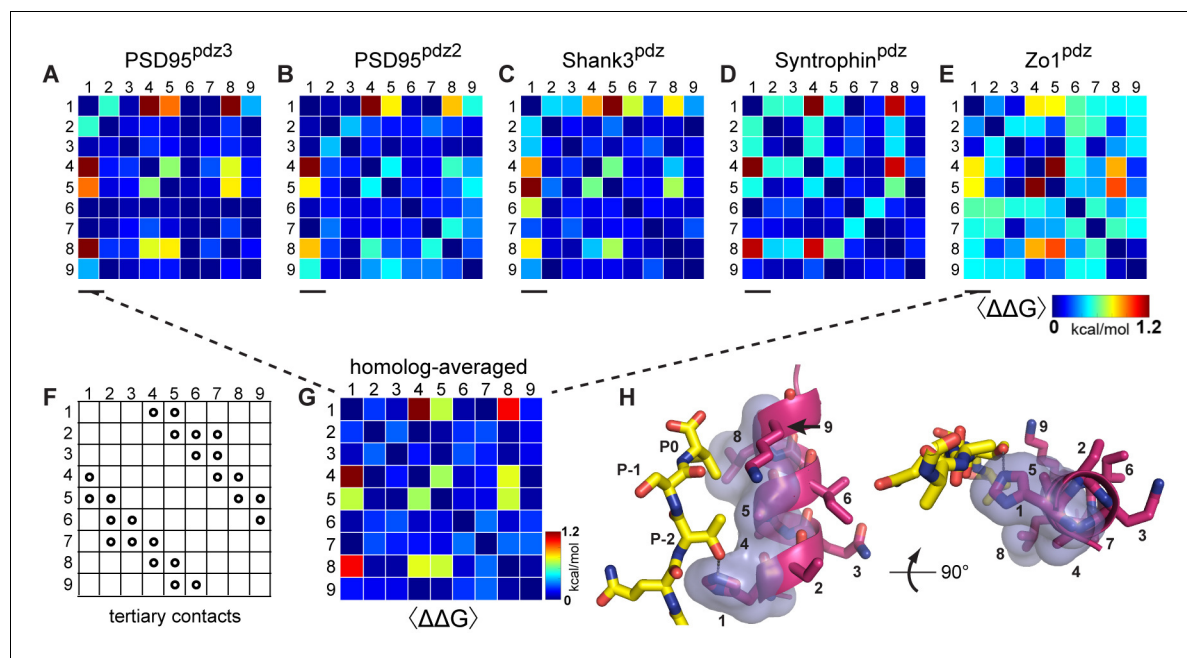


Figure 5. Conservation and idiosyncrasy in the pattern of energetic couplings over PDZ homologs. (A–E), Matrices of mutation averaged pairwise thermodynamic couplings for the α 2-helix in each PDZ homolog. The color scale is chosen to represent the full range of measured energetic couplings. The data show that some couplings are specific to individual homologs or shared by a subset of homologs, but that couplings between positions 1, 4, 5, and 8 are conserved over homologs. (F), the pattern of direct tertiary contacts between amino acid positions in the PDZ α 2 helix. By convention (Morcos et al., 2011), trivial contacts between residues with sequence distance less than three are not shown. (G), The homolog and mutation averaged couplings (corresponding to Figure 3), displaying the conserved interactions between amino acids in the PDZ α 2-helix. (H), Two views of the α 2-helix, with the four interacting positions in the homolog-averaged dataset shown in transparent surface representation, and ligand in yellow stick bonds. These include three positions in direct contact with ligand (1, 5, 8) and one allosteric position buried in the core of the protein (4).

DOI: <https://doi.org/10.7554/eLife.34300.015>

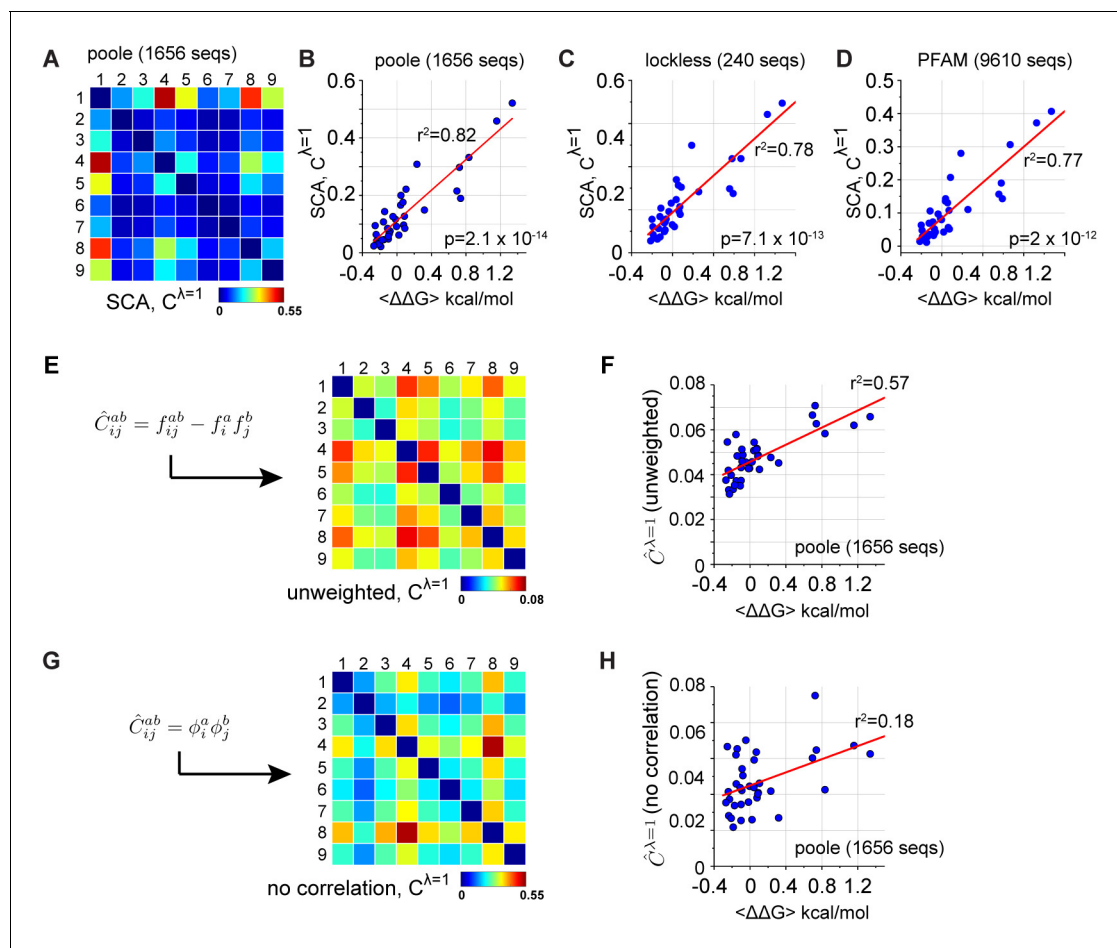


Figure 6. Coevolution-based inference of energetic couplings - SCA. (A), Coevolution of sequence positions corresponding to the top eigenmode of the SCA matrix, derived from an alignment of 1656 eukaryotic PDZ domains (the 'Poole' alignment). The data show that a subset of positions coevolve within the PDZ α 2-helix. (B–D), The relationship between experimental homology-averaged energetic couplings ($\langle\Delta\Delta G\rangle$) and SCA-based coevolution computed for three different alignments that differ in size and method of construction. The p-values give the significance of the coefficient of determination (r^2) by the F-test. (E–H), The basic calculation in SCA is to compute a conservation-weighted correlation matrix $\tilde{C}_{ij}^{ab} = \phi_i^a \phi_j^b [f_{ij}^{ab} - f_i^a f_j^b]$, where f_i^a and f_{ij}^{ab} represent the frequency and joint frequencies of amino acids a and b at positions i and j , respectively, in a multiple sequence alignment. The term $f_{ij}^{ab} - f_i^a f_j^b$ gives the correlation of amino acids at each pair of positions, and ϕ represents a weighting function for each amino acid at each position that is related to its conservation (Halabi et al., 2009; Rivoire et al., 2016). We compared the relationship of the experimental energetic couplings ($\langle\Delta\Delta G\rangle$) with measures of coevolution that leave out the conservation weights (E–F), or that leave out the correlations (G–H). The analysis shows that both terms contribute to predicting native energetic couplings between amino acids.

DOI: <https://doi.org/10.7554/eLife.34300.016>

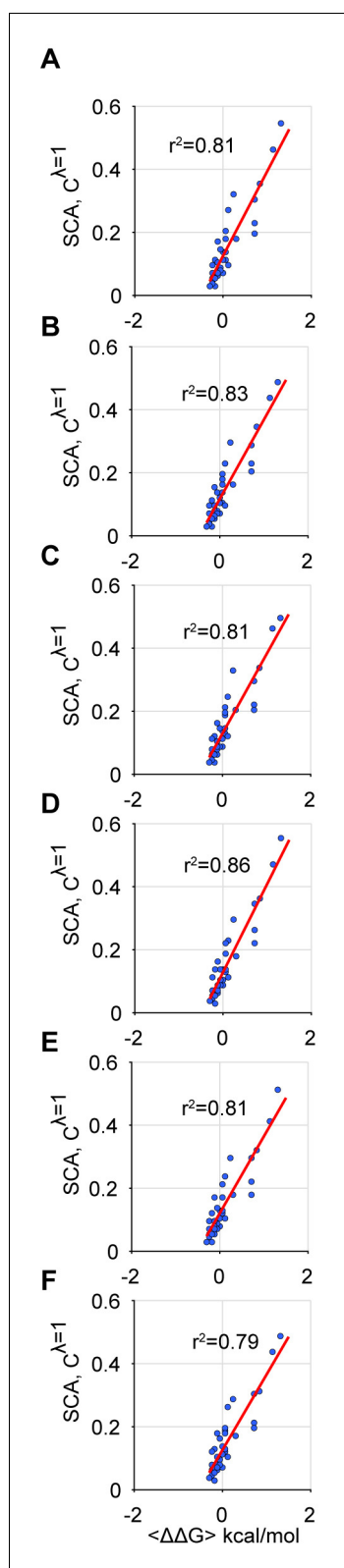


Figure 6—figure supplement 1. Robustness of the SCA to alignment size. (A–F) The graphs show scatterplots of homolog-averaged experimental couplings in the $\alpha 2$ helix against the first eigenmode of Figure 6—figure supplement 1 continued on next page

Figure 6—figure supplement 1 continued

the SCA coevolution matrix for six independent trials of randomly sub-sampling the Poole PDZ alignment (1656 seqs) to retain half the sequences. SCA focuses on conserved correlations and is expectedly robust to alignment size.

DOI: <https://doi.org/10.7554/eLife.34300.017>

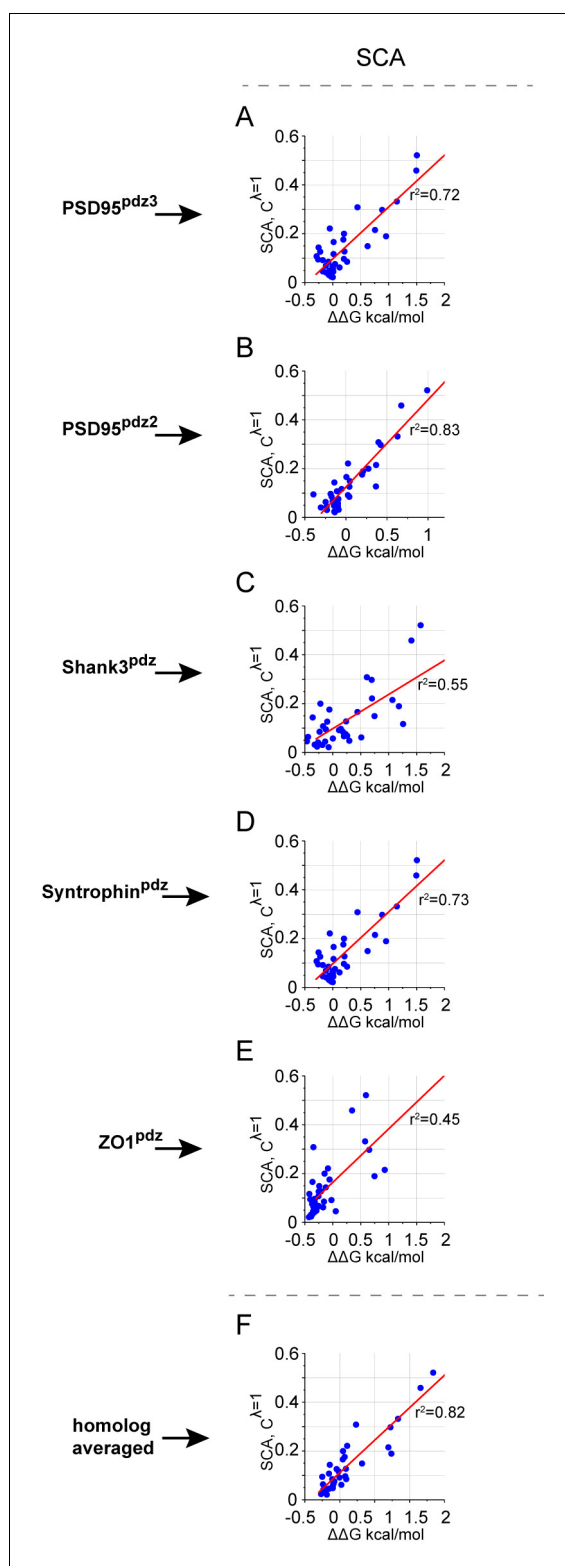


Figure 6—figure supplement 2. The relationship between mutation-averaged couplings and predictions from SCA for individual domains. (A–E), Scatterplots of experimental couplings in the $\alpha 2$ helix of each individual PDZ homolog against the first eigenmode of Figure 6—figure supplement 2 continued on next page

Figure 6—figure supplement 2 continued

the SCA coevolution matrix. Linear fits are shown in red lines with adjusted Pearson correlation coefficients indicated. (F), For comparison, the relationship of SCA with the homolog averaged data (same as **Figure 6B**).

DOI: <https://doi.org/10.7554/eLife.34300.018>

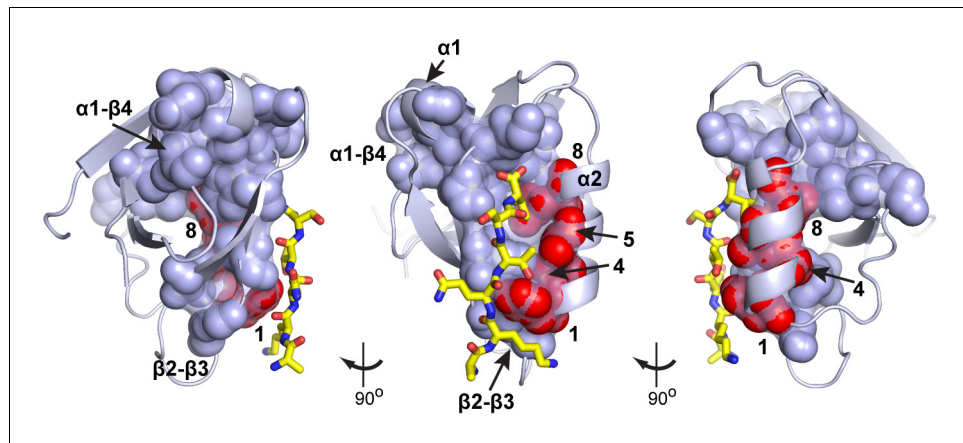


Figure 7. Homolog-averaged thermodynamic couplings and protein sectors. Analysis of the top eigenmodes of the SCA coevolution matrix exposes groups of coevolving amino acids that empirically are found to form physically contiguous networks in the tertiary structure, often connecting the main functional site to remote allosteric sites (Halabi et al., 2009; Lockless and Ranganathan, 1999; Rivoire et al., 2016; Süel et al., 2003). In the PDZ family, the protein sector (shown as CPK spheres and transparent surface on three rotations of a representative structure, PDB 1BE9) connects the ligand binding pocket to two known allosteric sites, one in the $\alpha1$ - $\beta4$ loop (Peterson et al., 2004) and the other in the $\beta2$ - $\beta3$ loop (Raman et al., 2016); the ligand is shown in yellow stick bonds. The homolog-averaged thermodynamic couplings in the $\alpha2$ helix (positions 1, 4, 5, and 8) precisely correspond to the portion of the PDZ sector contributed by the secondary structure element. The selective cooperative action of these residues is consistent with the idea that the sector represents a global collective mode in the PDZ structure associated with function, embedded within a more independent environment.

DOI: <https://doi.org/10.7554/eLife.34300.019>

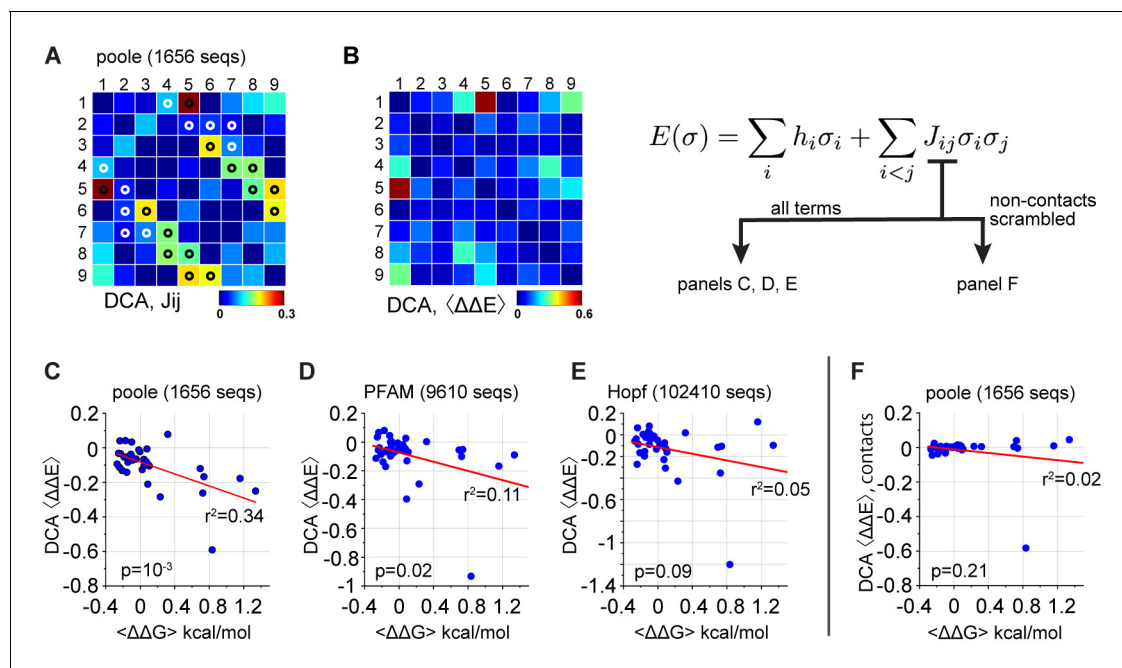


Figure 8. Coevolution-based inference of energetic couplings - DCA. (A), The matrix of direct couplings (J_{ij}) from the DCA method, with tertiary contacts in the PDZ structure (1BE9) indicated by white or black circles. By convention (*Morcos et al., 2011*), trivial contacts between residues with sequence distance less than three are not shown. The data show that all top direct couplings identified by DCA are indeed tertiary structural contacts. (B), The DCA method involves the inference of a statistical energy function $E(\sigma)$ that for each sequence σ , is parameterized by a set of intrinsic constraints on amino acids (h_i) and pairwise interactions between amino acids (J_{ij}). These parameters are optimized to reproduce the observed alignment frequencies and pairwise correlations. Using the model, the matrix shows mutation- and homolog-averaged energetic couplings, computed precisely as for the experimental data; see Materials and methods for details. (C–E), The relationship between experimental ($\langle\Delta\Delta G\rangle$) and DCA-inferred ($\langle\Delta\Delta E\rangle$) couplings in the PDZ α 2-helix, for three PDZ alignments that differ in size and method of construction. The p-values give the significance of the coefficient of determination (r^2) by the F-test. (F), The relationship between experimental and DCA-inferred couplings from J_{ij} in which top couplings defining contacts are preserved and all non-contact couplings are randomly scrambled. The DCA model used for this analysis is from the Poole alignment, as in panel D. The data show that pairwise couplings in the DCA model between non-contacting positions contribute significantly to prediction of protein function.

DOI: <https://doi.org/10.7554/eLife.34300.020>

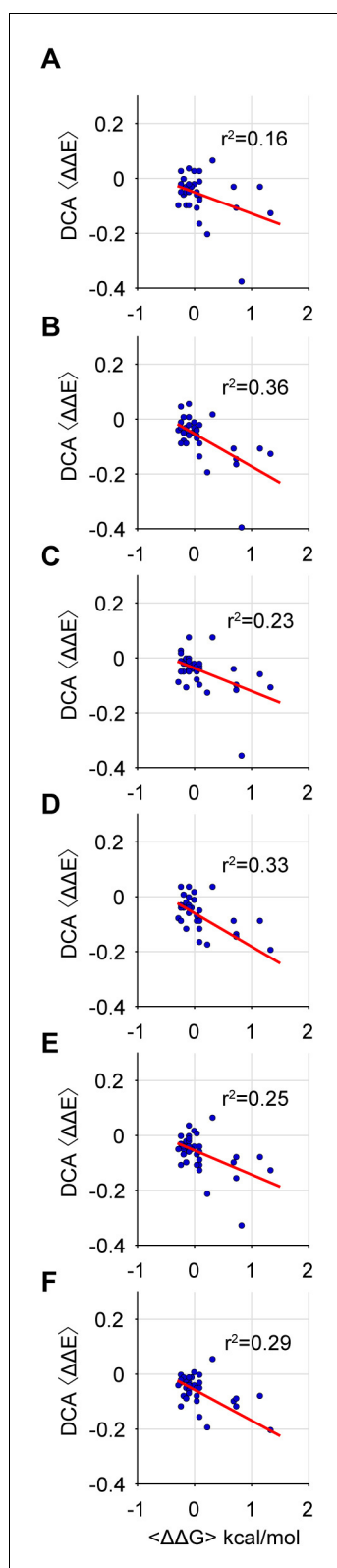


Figure 8—figure supplement 1. Robustness of DCA to alignment size. (A–F), The graphs show scatterplots of homolog-averaged experimental couplings in the $\alpha 2$ helix against computed DCA model coupling energies

Figure 8—figure supplement 1 continued on next page

Figure 8—figure supplement 1 continued

for six independent trials of randomly sub-sampling the Poole PDZ alignment (1656 seqs) to retain half the sequences. DCA expectedly shows more sensitivity to alignment size. The Poole alignment provides the best correspondence between experimental data and DCA and is therefore chosen for this analysis.

DOI: <https://doi.org/10.7554/eLife.34300.021>

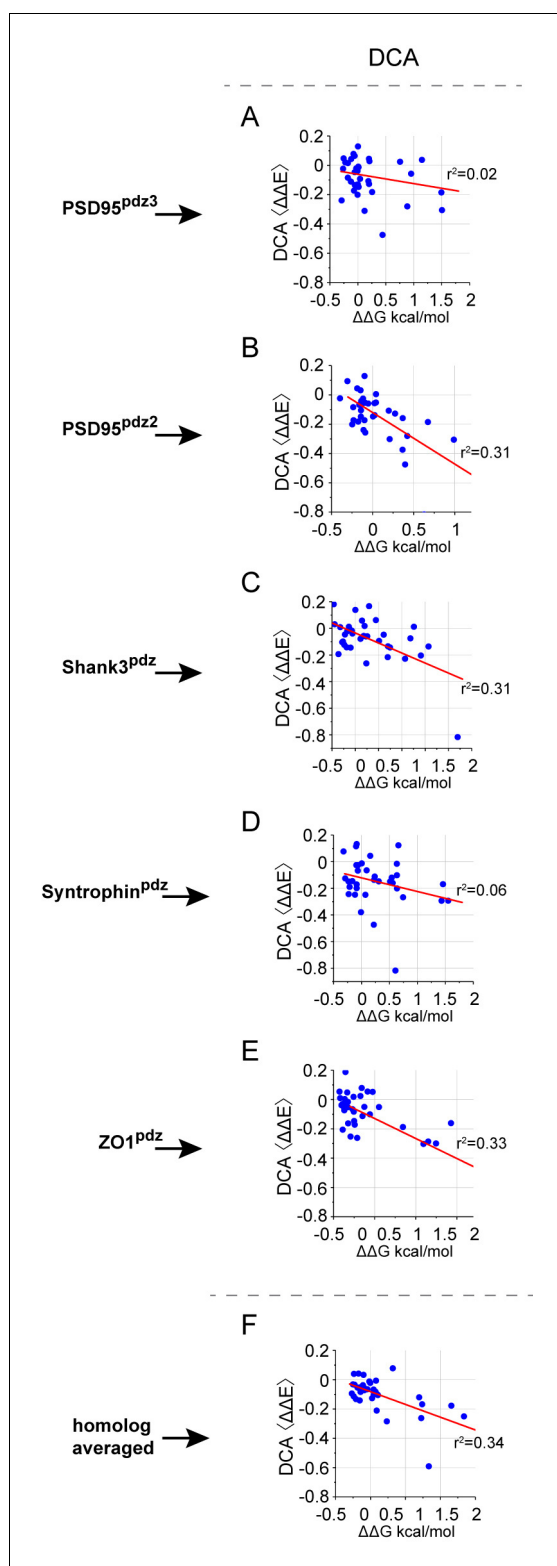


Figure 8—figure supplement 2. The relationship between mutation-averaged couplings and predictions from DCA for individual domains. (A–E), scatterplots of experimental couplings for individual PDZ homologs against the coupling values computed from the DCA

Figure 8—figure supplement 2 continued on next page

Figure 8—figure supplement 2 continued

model. Linear fits are shown in red lines with adjusted Pearson correlation coefficients indicated. (F), For comparison, the relationship of DCA predictions with the homolog averaged data (same as **Figure 8C**).

DOI: <https://doi.org/10.7554/eLife.34300.022>