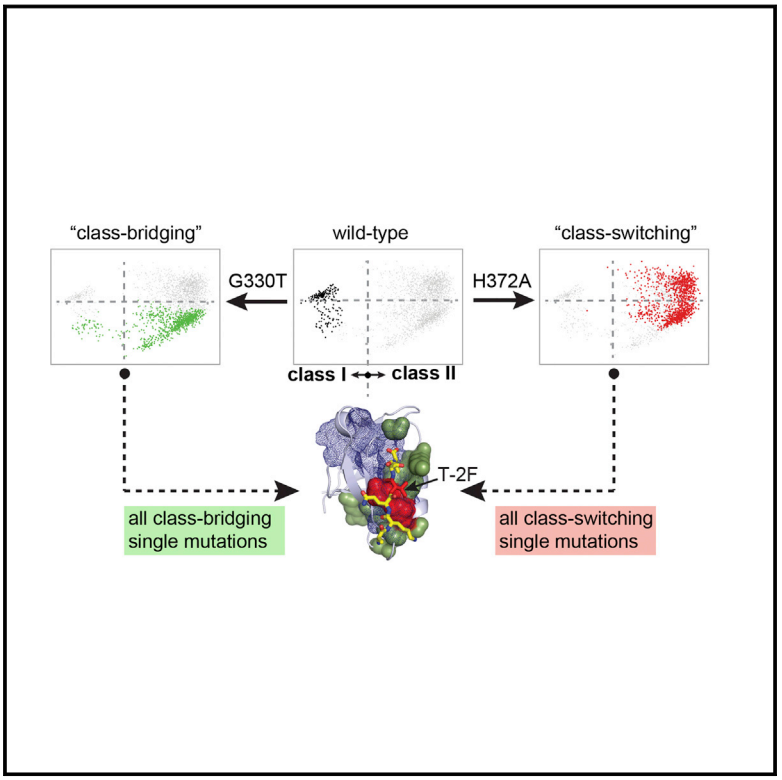
Article

Cell

Origins of Allostery and Evolvability in Proteins: A Case Study

Graphical Abstract



Authors

Arjun S. Raman, K. Ian White, Rama Ranganathan

Correspondence

rama.ranganathan@utsouthwestern.edu

In Brief

Mutations that are neutral but that potentiate functional protein adaptation by a subsequent mutation result in enhanced conformational plasticity at the distal "business end" of the protein. The data suggest a model in which the origin of allostery lies in the capacity for adaptation.

Highlights

- Adaptation in a PDZ domain involves intermediate mutations that bridge ligand classes
- A simple model shows that class-bridging mutations are evolutionarily preferred
- Class-bridging mutations act allosterically
- The origin of allostery in proteins could be in the capacity to adapt

Accession Numbers

5HEB
5HED
5HET
5HEY
5HF1
5HFB
5HFC
5HFF

Raman et al., 2016, Cell *166*, 468–480
July 14, 2016
http://dx.doi.org/10.1016/j.cell.2016.05.047

CellPress

Cell

Article

Origins of Allostery and Evolvability in Proteins: A Case Study

Arjun S. Raman,[1] K. Ian White,[1] and Rama Ranganathan[1,2,*]
[1]Green Center for Systems Biology
[2]Departments of Biophysics and Pharmacology
University of Texas Southwestern Medical Center, 6001 Forest Park Road, Dallas, TX 75390, USA
*Correspondence: rama.ranganathan@utsouthwestern.edu
http://dx.doi.org/10.1016/j.cell.2016.05.047

SUMMARY

Proteins display the capacity for adaptation to new functions, a property critical for evolvability. But what structural principles underlie the capacity for adaptation? Here, we show that adaptation to a physiologically distinct class of ligand specificity in a PSD95, DLG1, ZO-1 (PDZ) domain preferentially occurs through class-bridging intermediate mutations located distant from the ligand-binding site. These mutations provide a functional link between ligand classes and demonstrate the principle of "conditional neutrality" in mediating evolutionary adaptation. Structures show that class-bridging mutations work allosterically to open up conformational plasticity at the active site, permitting novel functions while retaining existing function. More generally, the class-bridging phenotype arises from mutations in an evolutionarily conserved network of coevolving amino acids in the PDZ family (the sector) that connects the active site to distant surface sites. These findings introduce the concept that allostery in proteins could have its origins not in protein function but in the capacity to adapt.

INTRODUCTION

Proteins display the capacity to fold, often into well packed three-dimensional structures, and to carry out biologically essential activities such as catalysis, signal transmission, and allosteric regulation. The amino acid sequence reflects the constraints arising from these properties, and considerable prior work has focused on understanding how the sequence encodes folding and biochemical function (Anfinsen, 1973; Bowie et al., 1990; Halabi et al., 2009; Morcos et al., 2011; Süel et al., 2003). However, it has been appreciated for decades that there may be non-trivial pressures on proteins that come not just from the physics of folding and function but also from the process of evolution itself (Anfinsen, 1973; Bershtein et al., 2006; Smith, 1970; Tokuriki and Tawfik, 2009; Wagner, 2005). For instance, proteins must arise through a random, iterative, stepwise process of mutation and selection, and they must be capable of adaptive variation as conditions of fitness vary in the environment. These considerations may place unique and yet unknown "design" constraints on evolved proteins, a missing aspect of our current understanding. An example of such a constraint might be functional connectivity in adaptive paths—the requirement that adaptation between fitness peaks proceeds through intermediates that maintain function above the threshold of selection (Smith, 1970).

Recent work has begun to elucidate the general properties of adaptation that are likely critical for the evolution of biological systems (Aakre et al., 2015; Bershtein et al., 2006; Bloom et al., 2005, 2006; Draghi et al., 2010; Draghi and Plotkin, 2011; Hayden et al., 2011; Tokuriki and Tawfik, 2009; Wagner, 2005). One such property is conditional neutrality, a special case in which mutations have no significant effect in the existing genetic or environmental background but have a significant effect upon subsequent changes in either the genome or environment (Draghi et al., 2010; Draghi and Plotkin, 2011; Hayden et al., 2011; Wagner, 2005). Such variations are "cryptic" in the sense that they hide their effects on fitness until exposed in the right setting and can therefore accumulate and pre-exist in populations as standing genetic variations (Luria and Delbrück, 1943). Because conditionally neutral mutations arise without selection and only express their fitness advantages upon subsequent events, they are said to be pre-adaptive (or "exaptive"; Gould and Vrba 1982) and represent a pool of variants that can facilitate the emergence of novel adaptive phenotypes. Indeed, conditional neutrality has been convincingly demonstrated to facilitate adaptation both theoretically (Draghi et al., 2010) and experimentally (Hayden et al., 2011) and, conceptually, represents the key link between the two major driving forces for genetic variation in populations: neutral drift and selection (Draghi and Plotkin, 2011). Understanding the prevalence and structural principles of conditional neutrality in protein molecules represents a key step in linking biophysical variation at the molecular level to evolutionary viability.

In this work, we carry out a series of experiments and simulations in the PSD95, DLG1, ZO-1 (PDZ) family of protein interaction modules to explore the prevalence, structural origin, importance, and mechanism of conditional neutrality. We show that conditional neutrality emerges from a distributed network of physically contiguous amino acids that permits a specific set of distantly positioned residues to unlock new structural and functional states at the ligand binding site. These adaptive positions are contained within the PDZ protein sector—a conserved

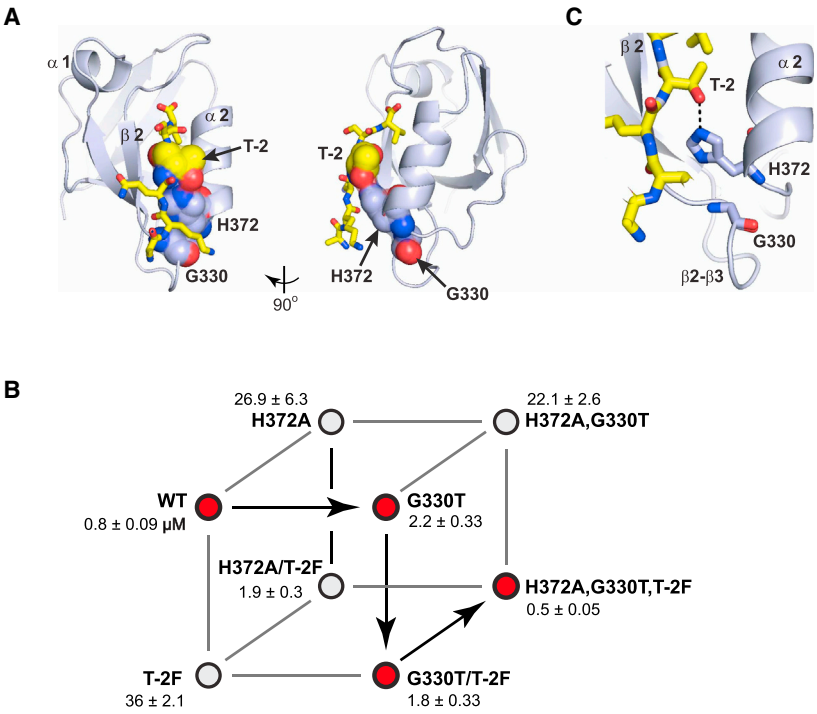


Figure 1. A Two-Mutation Path to New Functional Specificity in a PDZ Domain

(A) The structure of the PDZ domain (PSD95^{pdz3}, PDB: 1BE9) bound to the CRIPT C-terminal peptide (yellow stick bonds). Positions G330 and H372 in the protein and T-2 in the ligand peptide are shown as spheres with an overlaid van der Waals surface.

(B) A thermodynamic cube showing the effects of the G330T and H372A mutations in the context of the wild-type CRIPT ligand (top face) and the T-2F ligand (bottom face). Wild-type PSD95^{pdz3} shows a 45-fold preference for the CRIPT ligand, whereas the G330T,H372A double mutant shows a 45-fold preference for the T-2F ligand.

(C) Stereochemical details around ligand position −2. H372 makes a hydrogen bond with the class I-defining threonine side chain of ligand position −2, and G330 occurs on a surface loop (β 2-β 3) that is packed against the region of position 372.

and coevolving network of amino acid positions in the entire protein family (Halabi et al., 2009; McLaughlin et al., 2012). These findings allow us to propose basic structural principles of adaptation, and they open the hypothesis that the capacity to adapt may be the origin of allosteric mechanisms in proteins.

RESULTS

A Two-Mutation Path to New Specificity

A case study of protein adaptation is evident in PSD95^{pdz3}, a member of the PDZ protein family (Harris and Lim, 2001; Sheng and Sala, 2001). PDZ domains are roughly 100-amino acid mixed α/β folds that typically recognize the C-terminal region of target proteins at a groove formed between the β2 strand and α2 helix (Figure 1A); for simplicity, we refer to this groove as the "active site." Ligand specificity is largely defined by the amino acid sequence in the last four positions (numbered in reverse order from the carboxyl terminus [position 0]), with the major determinant of specificity being the identity of the residue at the antepenultimate (−2) position (Songyang et al., 1997; Stiffler et al., 2007; Tonikian et al., 2008). For example, class I domains bind ligands with a consensus of -X-**S/T**-X-ϕ-COOH (X is any amino acid, ϕ is hydrophobic) and class II domains with a consensus of -X-$\boldsymbol{\phi}$-X-ϕ-COOH. PSD95^{pdz3} is an archetypical class I domain, binding its cognate ligand (-TKNYKQ**T**SV-COOH, derived from the cysteine-rich interactor of PDZ (CRIPT) (Niethammer et al., 1998) with an affinity (indicated by the equilibrium dissociation constant, K_d = 0.8μM) that is 45-fold higher than to a class-switching mutant peptide (T-2F, -TKNYKQ**F**SV-COOH, specifying a Thr-to-Phe substitution at the −2 position) (Figure 1B). Because PDZ domains operate with dissociation constants in the 1–15μM range (Stiffler et al., 2007), this dynamic range represents a physiologically significant degree of specificity. Previous work shows that a double mutant of PSD95^{pdz3} (G330T,H372A) displays a near-complete reversal of specificity for these two peptides (McLaughlin et al., 2012). It shows roughly a 45-fold preference for the T-2F ligand and an affinity (K_d = 0.5μM) that is similar to that of the wild-type protein for the CRIPT ligand (Figure 1B).

These data frame a minimal instance of the problem of adaptive mutational paths. H372A and G330T represent two distinct strategies for achieving the class-switching specificity change. Position 372 is directly located at the ligand binding site and interacts with the residue at the −2 position of the target peptide. For example, in PSD95^{pdz3}, the histidine residue at this position makes a hydrogen bond with the hydroxyl group on the −2 threonine side chain of the ligand (Figure 1C), an interaction that is thought to represent the basis for the Ser/Thr specificity of class I PDZ domains (Doyle et al., 1996). Not surprisingly, the H372A mutation shows a substantial (34-fold) loss of affinity for the CRIPT ligand (Figure 1B). It also shows gain of function for the T-2F ligand, and, thus, this single active site mutation on its own provides a partially class-switching phenotype (13-fold T-2F preference) that is then enhanced by the G330T mutation to quantitatively complete the specificity switch.

In contrast, the G330T-first path is strikingly different. Position 330 is positioned on a surface loop (β 2-β 3) behind the active site and makes no direct interactions with ligand (Figures 1A and 1C), but mutation at this site creates a dual-function protein capable of binding both the CRIPT ligand and the T-2F ligand with equally high affinity (Figure 1B). In the language of evolutionary biology, G330T is a potential example of a "conditionally neutral" or cryptic mutation (Draghi et al., 2010; Hayden et al., 2011; Wagner, 2005)—essentially neutral with regard to existing function (binding of the CRIPT ligand) but significantly favorable with regard to a new function (binding of the class-switching T-2F ligand). Subsequent acquisition of the H372A mutation establishes class II specificity by reducing the affinity for the class I CRIPT ligand (Figure 1B).

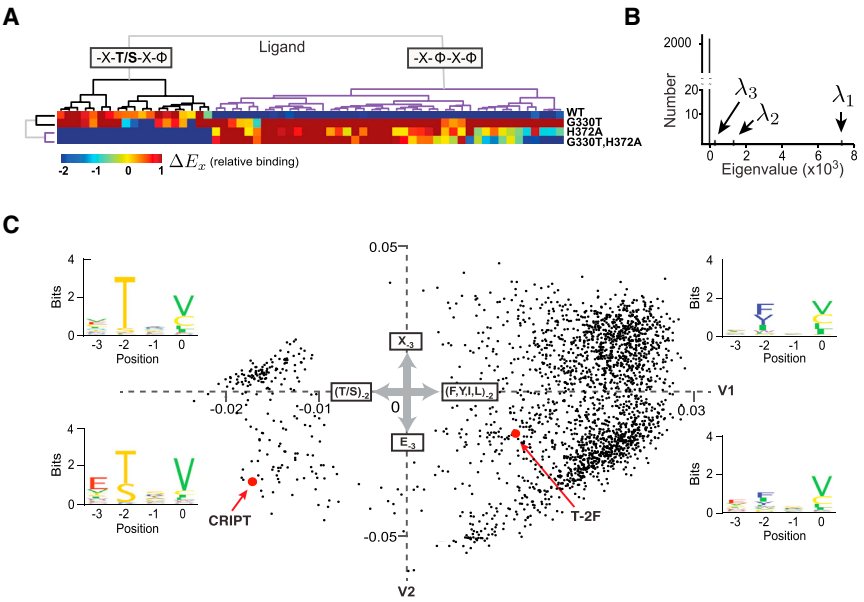


Figure 2. A Global Mapping of Primary Ligand Specificity in the PDZ Domain

Shown is the outcome of a quantitative bacterial two-hybrid assay in which we measured the binding of wild-type, G330T, H372A, and the double mutant variants of PSD95^{pdz3} to a library of C-terminal peptides randomized in the terminal four residues (154,521 of 160,000 ligands measured). Of these, 2,359 show better than 15μM binding to at least one PSD95^{pdz3} variant and are analyzed here.

(A) A clustered heatmap showing a small sampling of data. Each pixel shows the binding of one ligand x (ΔE_x, log scale) normalized so that zero represents wild-type binding ($\sim$1μM). Ligands (columns) cluster by known class specificities, and proteins (rows) show profiles consistent with the study of the CRIPT and T-2F ligands (Figure 1B): wild-type PSD95^{pdz3} shows class I specificity, H372A and G330T,H372A show class II specificity, and G330T shows dual specificity.

(B) The eigenvalues of the correlation matrix of ligand profiles, showing that the top two eigenmodes account for essentially all relationships between ligands.

(C) The ligand space defined by the top two eigenmodes. Dots correspond to the 2,359 physiologically relevant ligands, and proximity of two dots indicates similarity in the binding profile over the four PDZ proteins assayed. The insets show amino acid motifs for the peptides in each quadrant of the map. The first eigenvector ($V1$) separates ligands by identity at the class-defining position -2, with class I ligands in the left half and class II ligands in the right half. The second eigenvector ($V2$) separates ligands by a motif involving both positions -2 and -3.

A Comprehensive View of Ligand Specificity in PDZ Domains

To test the notion that G330T acts as a functional bridge between class specificities, we measured the binding of wild-type, G330T, H372A, and the double-mutant protein to a library of all possible peptide ligands defined by randomizing the C-terminal four amino acid positions (20^4 = 160,000 total ligands). This analysis is made possible by a quantitative bacterial two-hybrid (BTH) assay for PDZ function in which transcription of a reporter gene is tuned to be linearly proportional to the binding free energy of PSD95^{pdz3} to target ligands (Figure S1; McLaughlin et al., 2012). The reporter gene is chloramphenicol acetyl transferase (CAT), and thus the binding profile of any PDZ variant over the full space of all ligands can be assessed simply by selecting bacterial cells carrying the BTH on chloramphenicol and deep sequencing of the library before and after selection. The binding between PDZ and each ligand is given by ΔE_x, the normalized log ratio of observing ligand x in the selected and unselected libraries (Experimental Procedures; Figure S1). We obtained excellent counting statistics for 154,521 of the 160,000 possible ligands for all four PSD95^{pdz3} variants (Table S1)—a near-complete global profile of ligand specificity over the adaptive path defined by the G330T,H372A double mutant.

Hierarchical clustering of a few top binding sequences provides an intuitive preview of the full dataset (Figure 2A). Consistent with the data in Figure 1B, PSD95^{pdz3} binds ligands with T/S preference at the -2 position (class I), H372A and the double mutant bind ligands with hydrophobic amino acids at -2 (class II), and G330T binds ligands in both classes. The full dataset is very high-dimensional and impossible to directly visualize. Accordingly, we used principal components analysis (PCA) to project all peptides bound with better than 15 μM affinity by any of the four PSD95^{pdz3} variants (2,359 total) onto a two-dimensional space based on their profile of binding (Figure 2C). This low-dimensional projection is well justified. The first two principal components ($V1$ and $V2$) capture nearly 97% of the total variance representing relationships between the 2,359 ligands with physiologically relevant affinity (Figure 2B). The principal components also have clear biochemical meaning. $V1$ separates ligands by amino acid preference at the class-defining -2 position (S/T [class I] to left of the origin, bulky hydrophobic [class II] to the right), and $V2$ separates ligands by a combination of preference at both -2 and -3 (Figure 2C). Thus, the PCA provides a statistically accurate and intuitive representation of ligand specificity over the adaptive path.

Figure 3 shows the binding profiles of wild-type, G330T,H372A, and the double mutant variants of PSD95^{pdz3} projected onto the two-dimensional space. With the 15 μM cutoff for physiological relevance, the wild-type exclusively binds peptides in the class I space (Figure 3A), and both H372A and the double mutant bind ligands exclusively in the class II space (Figures 3C and 3D). This confirms that both the H372A single mutant and the double mutant are able to switch class specificity. In contrast, G330T shows a distinct binding profile that provides a link between the space of class I and class II ligands (Figure 3B). Thus, G330T is indeed a functional bridge—a single mutation that makes PSD95^{pdz3} capable of recognizing both classes of ligands. Amino acid profiles of ligands bound by the four PSD95^{pdz3} variants over a range of threshold binding affinities reinforce these findings (Figure S2). These data also provide a data-rich illustration of the concept of conditional neutrality—a mutation that preserves

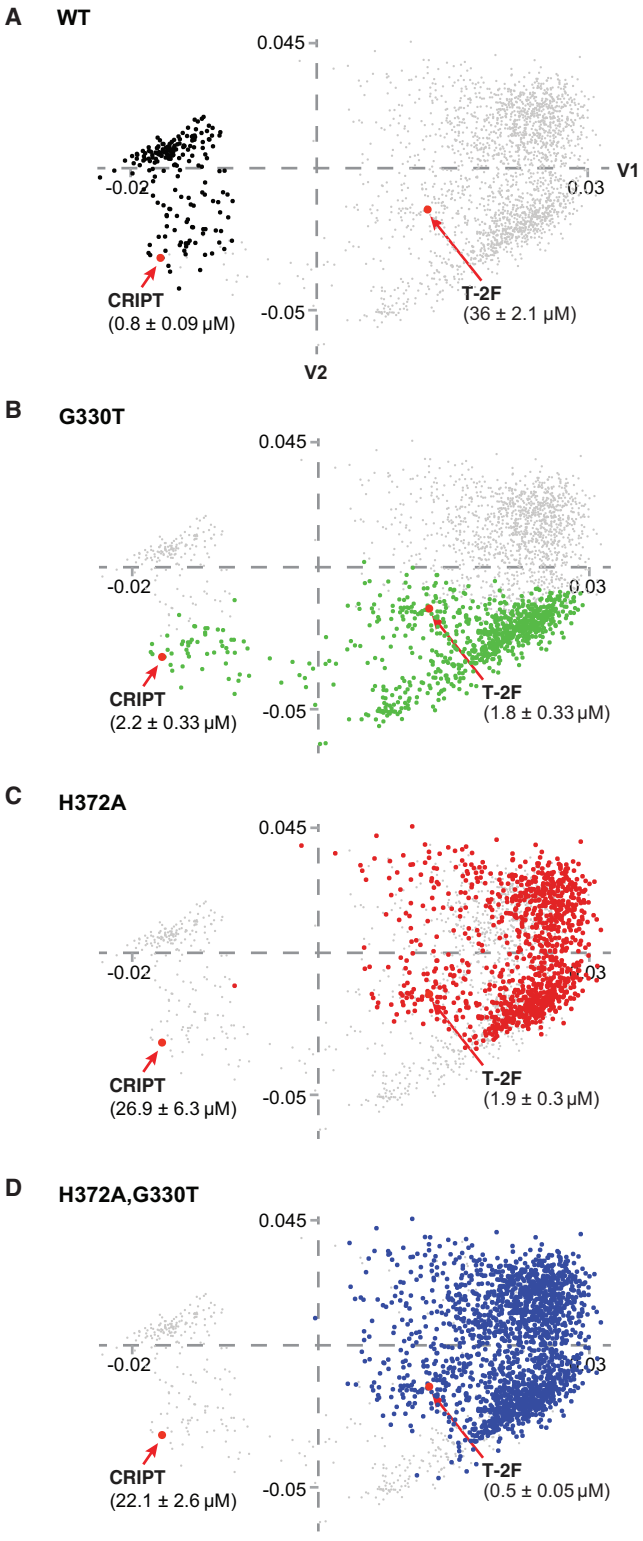


Figure 3. Binding Phenotypes of PDZ Variants along the Adaptive Path

(A–D) The ligand space defined in Figure 2, with bolded dots indicating the peptides recognized by wild-type (A, black), G330T (B, green), H372A (C, red), and the double mutant (D, blue) variants of PSD95^{pdz3}, respectively. The boundary is defined by affinity $\leq 15\mu$M. The data show that the wild-type binds exclusively to class I ligands, that H372A and the double mutant bind exclusively to class II ligands, and that G330T is a bridge between class specificities, binding a subset of ligands in both class I and class II regions.

existing function while opening up new functions; in this case, binding to physiologically distinct ligand variants.

Note that G330T is not a "generalist" or "promiscuous" protein in any typical sense. It maintains high-affinity binding and only recognizes 846 ligands of 154,521 total, a number that is in the same range as the other, more class-specific PSD95^{pdz3} variants (Table S1). It also recognizes ligands only in the lower half of the ligand space spanned by the top principal components (Figure 3B), showing that the effect is mainly in the shape, rather than the breadth, of the binding profile. Thus, G330T is a targeted class-bridging mutation capable of providing a near-neutral path from the native class I ligand (CRIPT) to a specific region of class II specificity. Subsequent acquisition of H372A to make the double mutant then localizes and reconfigures the binding specificity completely in the class II space (Figure 3D).

Preferential Adaptive Path

Which path—G330T first or H372A first—is evolutionarily more likely to achieve the specificity switch defined by the double mutant in PSD95^{pdz3}? Given that H372A provides a partial class switch in just one mutation (Figure 1B), it is important to understand what advantage (if any) there is in the class-bridging phenotype of G330T as an intermediate in the adaptive path. To study this, we carried out computational simulations of evolutionary dynamics between the wild-type and G330T,H372A double mutant states over a range of mutation rates and ligand-switching rates. Each trial of simulation is initiated with a population of wild-type PSD95^{pdz3} genotypes (N = 1000), which is large given the small number of genotypes considered (four; Figure 4A). At each generation, single mutations between genotypes are allowed with a probability μ and double mutations with probability μ^2, the target ligand switches between the class I CRIPT peptide and the class II T-2F peptide every τ generations, and fitness at every generation is defined as a fraction-bound of ligand determined from the experimentally defined equilibrium dissociation constants (Figure 1B). The total ligand concentration is set to 10μM, a value in the middle of the specificity range of wild-type PSD95^{pdz3}—the relevant regime for non-trivial dynamics. In essence, this simulation gives an opportunity to study how the flux between the two paths to the double mutant state depends on both internal parameters (mutation rate and population size) and external parameters (environmental switching between the two class-distinct ligands).

A representative simulation trajectory at one particular mutation and ligand-switching rate illustrates properties of the adaptive process (Figure 4B). In this case, τ = 500 and $N\mu$ = 1, meaning that the ligand switches every 500 generations, and one single mutation is made, on average, at every generation. As expected, the wild-type genotype (black trace) is the most fit in the presence of the class I CRIPT ligand, with a small fraction of other genotypes stochastically occurring in the population according to the mutation rate and on their fitness relative to the wild-type. Switching to the T-2F ligand causes the population to ultimately switch to the double mutant state (blue), the genotype that is most fit for the

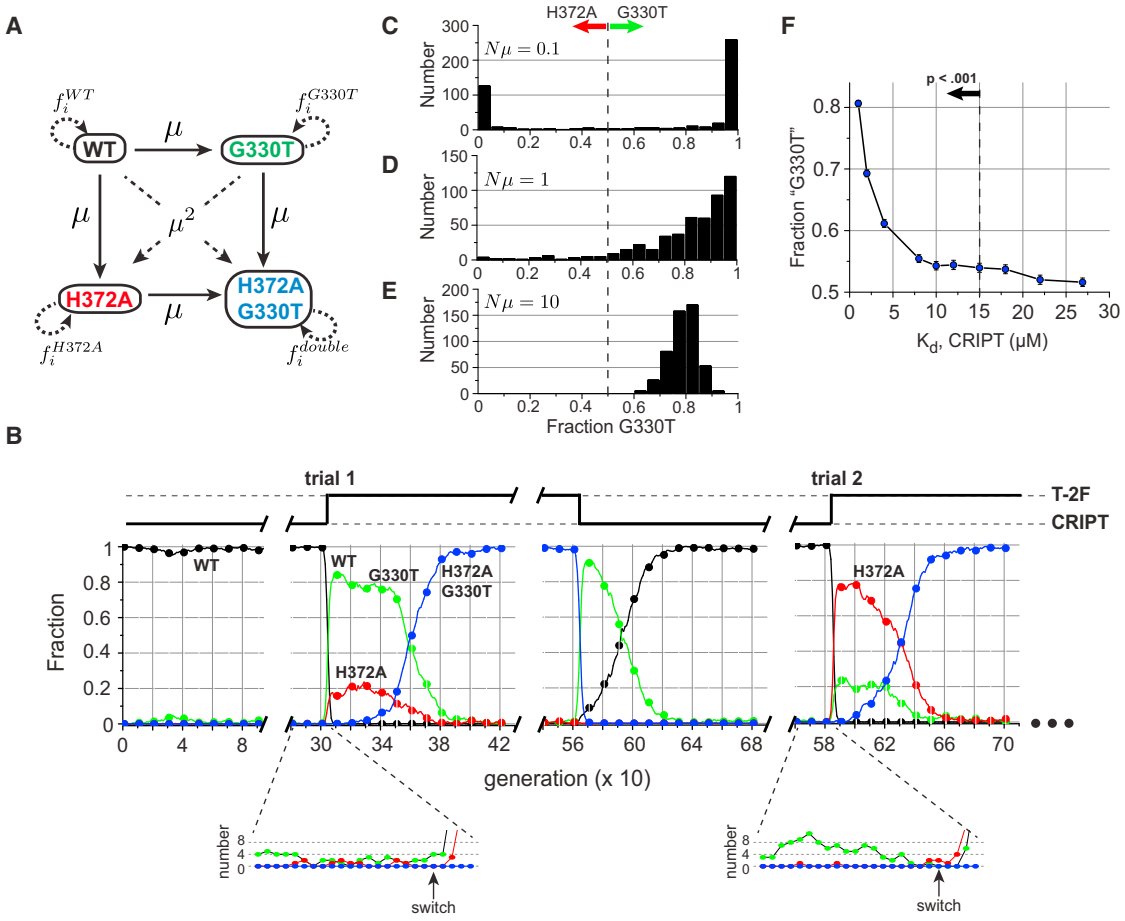


Figure 4. The Preferred Path of Adaptation

(A) A population dynamics model for the path of adaptation between the wild-type and G330T,H372A double mutant genotypes. Simulations are initiated with a population of N wild-type individuals, and, at each generation, single mutations are allowed with rate μ and double mutations with rate μ^2, ligands switch between CRIPT and T-2F every τ generations, and the fitness of each genotype is defined as the fraction bound of ligand. The simulation permits a quantitative analysis of the relative flux through G330T or H372A along the path to the double-mutant state.

(B) One simulation trajectory at a particular mutation and ligand-switching rate ($N\mu = 1$ and $\tau = 500$), showing two trials of adaptation in response to switching from the CRIPT ligand (class I) to the T-2F ligand (class II). These examples show that different proportions of the two single mutants can act as intermediates, depending on the pre-existing population of G330T and H372A variants at the moment of ligand switching (insets). Circles mark every 10 generations in main panels and every generation in insets.

(C–E) Histograms of the fraction of G330T ($n^{G330T}/(n^{G330T} + n^{H372A})$, where n represents integrated counts over the period of switching to the double mutant state for ~500 trials of switching from the CRIPT to T-2F ligands. The analysis is shown for three regimes of mutation rate: (1) $N\mu = 0.1$, where mutations are rare (C), (2) $N\mu = 1$ (D), and (3) $N\mu = 10$, where mutations are abundant (E). G330T is always the preferred path of adaptation to the double-mutant state upon ligand switching.

(F) The preference of a hypothetical "G330T" variant (thus in quotation marks) over H372A for adaptation to the T-2F ligand as a function of computationally varying the affinity for the CRIPT ligand from 1μM to 26.9μM, the same affinity as H372A. Thus, given model parameters, affinities as low as 15μM can still provide a statistical advantage over H372A in facilitating adaptation (Wilcoxon rank-sum test).

T-2F ligand under these simulation conditions. However, the path of switching can show considerable trial-by-trial variability with regard to intermediates. For example, in this trajectory, G330T (green) is more prevalent in trial 1 and H372A (red) more prevalent in trial 2. Averaged over many trials of switching (~500 events) from CRIPT to T-2F, we find that G330T is by far the preferred path of adaptation to the double-mutant state given the selected mutation and ligand-switching rates (Figure 4D).

How can we understand this result mechanistically? Because both G330T and H372A can bind the T-2F ligand about equally well (Figure 1B), the path simply depends on the relative availability of these genotypes in the population at the moment of switching (Figure 4B, insets). This property, in turn, depends on the fitness of G330T and H372A while in the CRIPT environment, a factor that heavily favors G330T over H372A (Figure 1B; $K_d^{G330T} = 2.2 \pm 0.33\mu$M and $K_d^{H372A} = 26.9 \pm 6.3\mu$M). As a consequence, G330T typically comprises the majority of the cryptic genetic variation in the CRIPT environment, more likely to be present and able to support transition to the double-mutant state when the environment switches to T-2F.

How does this result depend on mutation rate and the switching rate of target ligand? Simulations show that conversion to the

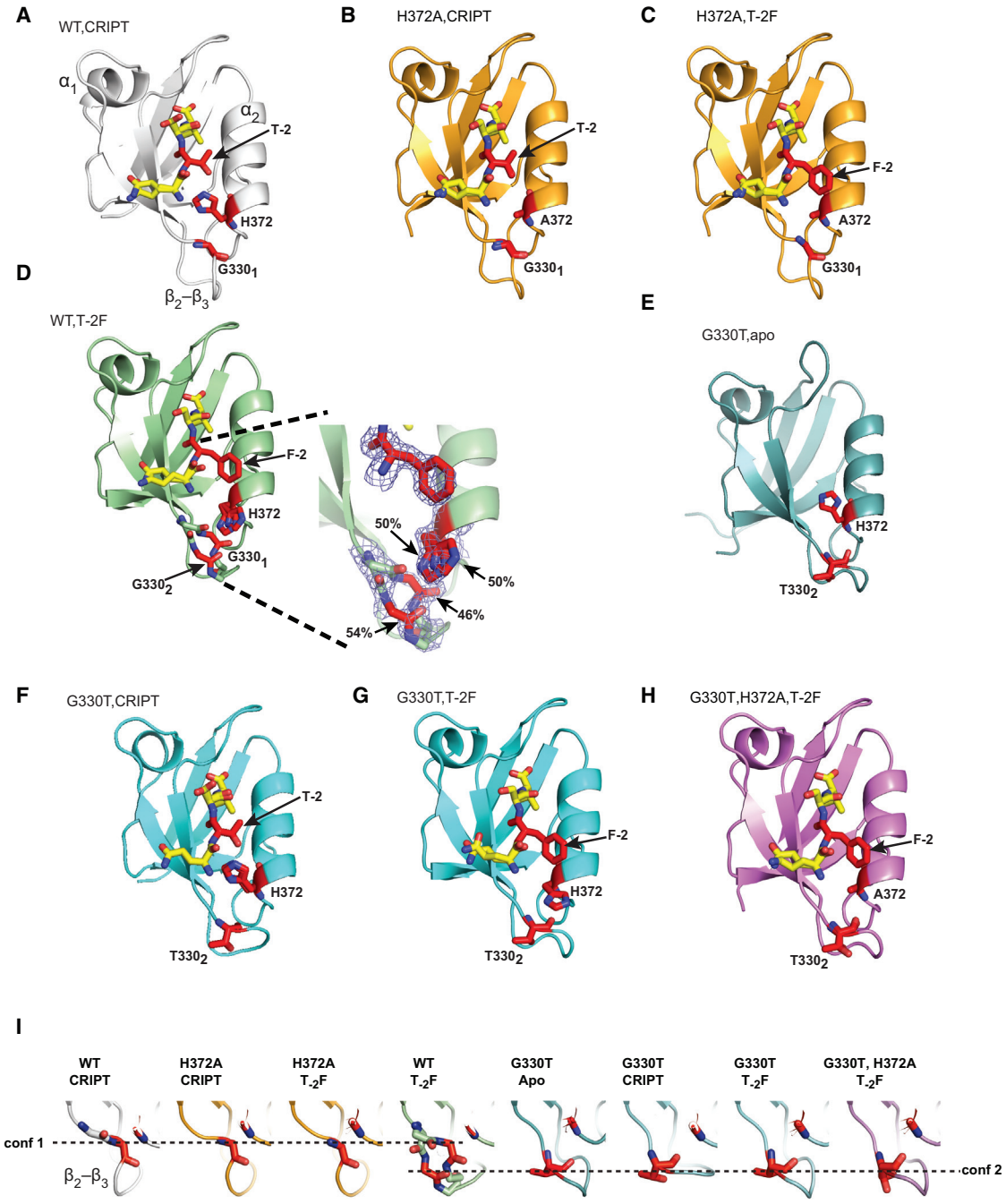


Figure 5. The Structural Basis for Ligand Specificity Switching

Shown are high-resolution crystal structures of wild-type (WT), H372A, G330T, and double mutant variants of PSD95^{pdz3}, either unliganded (apo) or bound to CRIPT or T-2F ligands as labeled.

(A) The WT-CRIPT structure, recapitulating features of class I ligand recognition. The threonine hydroxyl of −2 is hydrogen-bonded to histidine 372, and G330 is located on a well ordered $\beta2$–$\beta3$ loop packed against the region of 372 ($G330_1$, the subscript indicates conformation 1).

(B and C) The structures of H372A bound to CRIPT (B) or T-2F (C) show truncation of the 372 side chain and little other conformational change, a local perturbation permitting accommodation of the phenylalanine side chain at −2 without steric clash. The loss of both bulk and hydrogen bonding potential at position 372 is consistent with the partially class-switching phenotype of H372A.

(D) Binding of the T-2F ligand to wild-type PSD95^{pdz3} causes rotation of H372 to a new, non-native rotamer state (to prevent steric clash), and induction of two partially occupied conformational states of the $\beta2$–$\beta3$ ($G330_1$ and $G330_2$).

(E–G) The G330T mutation (apo state) stabilizes the $\beta2$–$\beta3$ loop in the alternate conformation 2 (E), a state that can permit either rotamer state of H372 without steric clash. Thus, in G330T, the H372 side chain occupies the native rotamer in binding CRIPT (F) and the alternative rotamer in binding T-2F (G).

(legend continued on next page)

double-mutant state is only achieved over a certain regime of ligand-switching rate. This makes sense: if ligand switches too rapidly to permit fixation of the double mutant, the population converges to the only genotype that is fit for the average of both ligand environments—G330T (Figure S3). However, in any regime of ligand-switching rate in which the double mutant fixes in the population, G330T is always more preferred than H372A in mediating adaptation (Figures 4C–4E). This is true when mutations are rare ($N\mu << 1$; Figure 4C) and when mutations are abundant ($N\mu >> 1$; Figure 4E). Thus, it is the neutral, class-bridging genotype rather than the direct class-switching genotype that represents the likely intermediate in adaptation to new ligand class specificity.

But how "neutral" does a mutation have to be to be statistically preferred over a class-switching mutant such as H372A in mediating adaptation? Indeed, even G330T is not strictly neutral in the CRIPT environment ($K_d^{G330T} = 2.2 \pm 0.33\mu$M and $K_d^{WT} = 0.8 \pm 0.09\mu$M); this is the reason why it is considerably less competitive than the wild-type (Figure 4B). To study this, we carried out a series of simulations in which we examined the effect of varying the affinity of G330T for the CRIPT ligand from near wild-type (1μM) to that of the class-switching mutant (26.9μM; Figure 1B). The data show that, given the conditions of the simulation, affinities up to the limit of physiological PDZ binding (<15μM) will be statistically preferred to H372A (Figure 4F). This result relaxes the notion of conditional neutrality, defining a limit of protein function at which a mutant can still contribute to the cryptic genetic variation and be distinguished in adaptive capacity from direct class-switching mutations.

Structural Basis for Class-Bridging Ligand Binding

The dominance of the class-bridging G330T mutation in adaptation to new ligand specificity is interesting because it is not structurally obvious. Position 330 occurs on a surface loop (β2–β3) that lies behind the substrate binding pocket and makes no direct contact with ligand (Figures 1A and 1C). How does mutation at this distant site create a dual-function PSD95^{pdz3} binding pocket capable of recognizing both class I and class II ligands? To address this issue, we solved the high-resolution crystal structures of the four PSD95^{pdz3} variants in either the apo (unliganded) state or bound to either CRIPT or T-2F ligands—a total of eight structures (Figure 5; Tables S2 and S3). All structures were solved under near-isomorphous conditions—in the same crystal form ($P4_132$) with unit cell constants within 0.5% of each other—and models were refined to a similarly high resolution ($\leq$2.0 Å) with excellent statistics and geometry (Tables S2 and S3). Thus, we are in a position to make statements about the mechanism of action of the mutations from a comparative study of atomic structures.

The structure of PSD95^{pdz3} bound to the class I CRIPT ligand—the wild-type complex—shows H372 in a rotameric state in which it can hydrogen-bond with T-2 and a well ordered β2–β3 loop that makes backbone hydrogen bonds with the H372 region (Figure 5A; Doyle et al., 1996). Not surprisingly, mutation of position 372 to alanine abrogates class I recognition by eliminating the hydrogen-bonding partner for the Thr/Ser residue at the −2 position (Figure 5B) but also creates space for accommodating a bulky hydrophobic side chain at the −2 position without steric clash (Figure 5C). No other conformational changes are evident, indicating that the direct class-switching phenotype of H372A is due to effects that are spatially localized to the site of adaptive challenge.

The origin of the class-bridging phenotype of G330T is qualitatively different. To explain, consider the effect of the T-2F ligand in binding to wild-type PSD95^{pdz3}, a low-affinity complex. Binding of the T-2F ligand involves a propagated structural perturbation in which the side chain of H372 (presumably due to steric clash) is forced to adopt a new rotamer state with two split conformations with roughly equal occupancy, and the β2–β3 loop (containing position 330) is, in turn, induced to partially adopt an alternate conformational state (Figures 5D and 5I). The conformational heterogeneity at both H372 and the β2–β3 loop are consistent with the poor affinity of the wild-type protein for the T-2F ligand (Figure 1B). How does G330T provide for high-affinity binding of both CRIPT and T-2F ligands? The G330T mutation stabilizes the β 2-β 3 loop in the non-native alternate conformation (Figures 5E and 5I), a structural change that permits the H372 side chain to adopt either rotamer state without steric penalty (Figures 5F and 5G). Thus, the G330T variant can recognize both class I and class II ligands with high affinity, switching the rotameric state of H372 in a ligand-dependent manner (compare Figures 5F and 5G). To complete the path of adaptation, it is straightforward to see that addition of H372A in the background of G330T would (just as in the wild-type background) abrogate class I ligand recognition, resulting in the class II specificity observed in the double mutant (Figure 5H).

In summary, H372A works directly and locally at the binding pocket to simultaneously eliminate class I ligand binding and to promote class II ligand binding—the phenotype of direct switching. In contrast, G330T works allosterically to open up conformational plasticity at the binding pocket, which enables both class I and class II recognition—the phenotype of class-bridging binding. It is important to note that this plasticity does not come in the form of general flexibility of the binding pocket reminiscent of low-affinity, broad-specificity interfaces, such as in the immature, germline isoforms of antibodies (Wedemayer et al., 1997). Instead, it opens up just one additional macroscopic conformational state (Figures 5F and 5G), which results in high-affinity, dual-class ligand specificity (Figure 3B).

Spatial Distribution of Conditional Neutrality

The detailed study of G330T and H372A motivates a comprehensive analysis of all adaptive mutations to define the general structural rules. Such a study is made possible by a

(H) The structure of the G330T,H372A double mutant bound to T-2F is similar to that of H372A alone (C), except that the β2–β3 loop is in conformation 2, consistent with G330T.

(I) Conformational states of the β2–β3 loop in all eight PSD95^{pdz3} variants presented in this work. The loop is in one conformation (conf 1) in the wild-type PSD95^{pdz3}-CRIPT, H372A-CRIPT, and H372A-T-2F complexes; in two partially occupied conformations (conf 1 and conf 2) in the WT-T-2F complex; and in conf 2 in all variants in a G330T background. The dashed lines are drawn through the C_α atom of position 330.

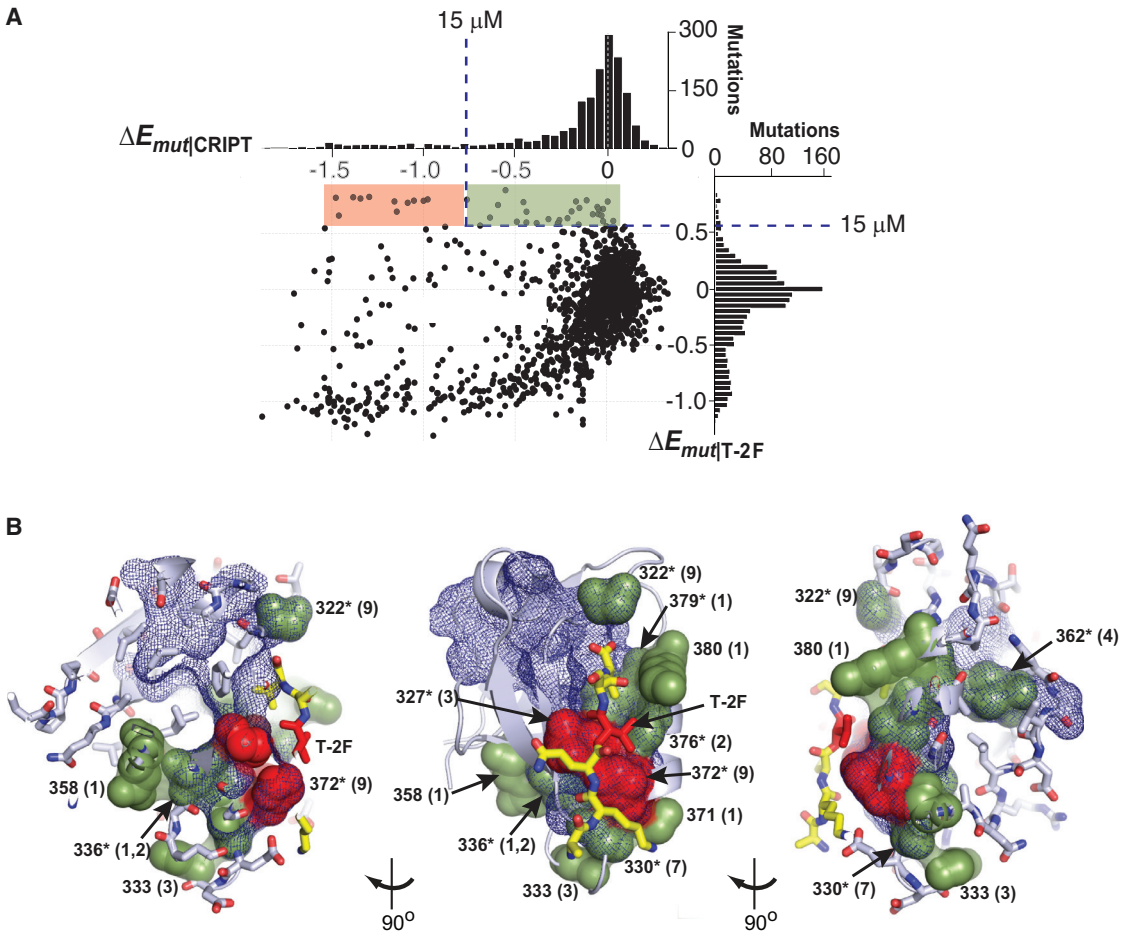


Figure 6. Spatial Architecture of Adaptive Mutations in Response to T-2F

(A) The effect of all possible single mutations in PSD95^{pdz3} on binding to either the class I CRIPT ligand (ordinate) or the class II T-2F variant (abscissa). The shaded regions describe physiologically significant ($\leq 15\mu$M) binding to the T-2F ligand, with either associated loss of function (red) or physiological neutrality (green) for the CRIPT ligand.

(B) Three rotations of PSD95^{pdz3}, with all positions containing adaptive mutations for T-2F in sphere representation with colors as shown in (A). Thus, red spheres correspond to positions with direct switching mutational phenotypes, and green spheres indicate positions with class-bridging phenotypes. The number of mutations at each position with that phenotype are shown in parentheses. The blue mesh indicates the protein sector—the network of coevolving positions in the PDZ family. The data show that nearly all adaptive mutations are contained in the protein sector (marked with asterisks), that direct-switching phenotypes localize to the site of adaptive challenge (ligand −2), and that class-bridging phenotypes arise from a distributed, contiguous network of residues leading away from the binding pocket through the sector.

dataset comprising a total saturation mutagenesis of PSD95^{pdz3} (McLaughlin et al., 2012), reporting the effect of every possible amino acid substitution at every position in the PDZ domain (1,598 total) on the binding of either the CRIPT ligand or the T-2F variant (Figure 6A). This dataset permits enumeration of every mutation in PSD95^{pdz3} that shows direct class-switching (like H372A) or class-bridging ligand recognition (like G330T). Using the 15 μM cutoff for physiological binding, this analysis shows that, although the vast majority of mutations are either neutral or destabilizing for both ligands, a subset of 44 mutations shows gain of function for the T-2F ligand (shaded regions, Figure 6A). Of these adaptive mutations, 12—like H372A—show loss of function for the CRIPT ligand (class-switching phenotype, red shade, Figure 6A), and 32—like G330T—show near-neutrality for the CRIPT ligand (class-bridging phenotype, green shade, Figure 6A). Mapping of the positions corresponding to these 44 mutations on the tertiary structure of PSD95^{pdz3} shows the global spatial distribution of adaptation for the T-2F ligand (Figure 6B) and leads to a simple conclusion. All class-switching mutations directly contact the site of adaptive challenge (T-2F), and all class-bridging mutations are invariably outside of the contact environment of T-2F (Figure 6B). Almost no positions (336 excepted) contain mutations with both phenotypes, arguing that the distinction between class-switching and class-bridging phenotypes is a characteristic of the position rather than of the specific substitution at that position.

Interestingly, adaptive mutations are organized in the tertiary structure into a physically contiguous, wire-like network of

residues linking the class-switching active site residues to class-bridging regions distributed throughout the protein structure (Figures 6B and 7). The network is not isotropically organized in space around the T-2F site in a manner consistent with a simple model of spatial proximity to the site of adaptation. Instead, it is an anisotropic network that fractures through the protein structure to include some distantly positioned residues at the expense of some more proximal ones (Figures 6B and 7). For example, position 373 is in the immediate vicinity of T-2F but shows no mutations capable of adaptation. In contrast, position 362 is nearly 15 Å from ligand position −2 but has four mutations that create a binding pocket capable of recognizing both class I and class II ligands. Thus, the data argue that all class-bridging mutations are fundamentally allosteric in nature, forming specific networks of amino acids within the protein structure to influence active site function from a distance.

The Protein Sector as the Origin of Adaptive Mutations

What principle of protein structure underlies the complex spatial organization of conditionally neutral mutations? The special relevance of this class of mutations for evolutionary dynamics and the finding that they originate from wire-like amino acid networks within the protein structure (Figures 6B and 7) provide an important clue. Previous studies have introduced the concept of "protein sectors," groups of amino acids that are conserved and coevolve over the long-term evolutionary history of a protein family (Halabi et al., 2009; Smock et al., 2010). Sectors comprise sparse, contiguous networks of amino acids that typically link the protein active site to distantly positioned surface sites (Lockless and Ranganathan, 1999; Süel et al., 2003) and have been linked to the constraints on protein folding (Reynolds et al., 2013; Socolich et al., 2005) and functional properties such as catalysis (Halabi et al., 2009; Reynolds et al., 2011), binding (McLaughlin et al., 2012; Russ et al., 2005), signal transmission (Ferguson et al., 2007; Lee et al., 2008, 2009; Süel et al., 2003), and allosteric regulation (Reynolds et al., 2011; Süel et al., 2003). In short, sectors represent a model for the relevant cooperative action of amino acid positions in proteins.

We compared the pattern of adaptive mutations—both direct class-switching and class-bridging—with the pattern of coevolution in the PDZ domain family (McLaughlin et al., 2012). The sector in the PDZ domain family (blue mesh, Figures 6B and 7) comprises a group of 20 amino acid positions (~20% of total residues, default parameters, SCA5.0) that form a network linking the ligand binding pocket to three regions: the β2–β3 loop, the α1 helix, and the end of the β4 strand (positions 362–363) (Figures S4 and S5B). Essentially all of the adaptive mutations, both directly class-switching and class-bridging, are contained within the sector (38 of 44 mutations at 8 of 12 positions; Figures 6B and 7) (p = 0.001, Fisher's exact test; Figure S5), and the six remaining mutations occur at four surface sites (358, 333, 371, and 380; marked in yellow in Figure 7) that contact the peripheral edges of the sector. Importantly, keeping the number of top-scoring positions the same as for the sector (i.e., 20), neither spatial proximity to the site of adaptive challenge (p = 0.063, Fisher's exact test) or position-specific conservation (p = 0.47, Fisher's exact test) show such significant correlation with adaptive positions (Figures S5D–S5F). Thus, at least for the primary specificity site, the capacity to adapt in the PDZ domain emerges from an evolutionarily ancient coevolving network of residues. The coevolution of amino acids within the sector is consistent with the cooperativity and allosteric effect of class-bridging mutations and generalizes the role of protein sectors as not only functional units of proteins but as adaptive units of proteins.

DISCUSSION

A Structural Model for Protein Adaptation

The motivation for this work is the concept that evolutionary dynamics place non-trivial constraints on the design of natural proteins. An example of such a constraint is the existence of the class of epistatic mutations termed conditionally neutral—mutations that do not influence the existing functional activity but that open up new activities that can be selected under the right environmental conditions (Hayden et al., 2011; Luria and Delbrück, 1943; Wagner, 2005). Such mutations can contribute to the standing genetic variations in populations and can facilitate the acquisition of new phenotypes as selection conditions fluctuate (Draghi et al., 2010; Hayden et al., 2011). Thus, the elucidation of structural principles of conditional neutrality in proteins is a key next step in understanding their mechanisms and origin.

In this work, we demonstrate the existence, evolutionary relevance, and mechanism of conditionally neutral mutations in a member of the PDZ family of protein interaction modules. The main result is that conditional neutrality is generally allosteric in nature, working from a distance through a network of amino acid interactions to open up new conformational states at the ligand binding pocket. In contrast, adaptive mutations located at the active site have the property of direct switching of ligand class-specificity—new ligand binding is gained at the expense of binding to the existing ligand. Simulations of evolutionary dynamics confirm the notion that it is the class-bridging mutations, not the direct switching ones, that are likely to serve as intermediates in adaptation. Thus, we conclude that, in addition to its contributions to functional properties such as signal transmission and regulation, intramolecular allostery plays a key role in facilitating the evolutionary process.

Recent high-throughput methods for mutagenesis (Fowler and Fields, 2014; McLaughlin et al., 2012; Stiffler et al., 2015) will facilitate testing of the generality of this conclusion. However, the findings here are consistent with data in at least one other protein system—TEM-1 β-lactamase, an enzyme that confers resistance to specific antibiotics in bacteria (Salverda et al., 2010). Deep mutational scanning reveals a class of mutations underlying adaptation that shows conditional neutrality; that is, neutral with regard to the existing substrate but gain of function toward a new substrate (Stiffler et al., 2015). As in PSD95^{pdz3}, these mutations occur at sites that are distant from the active site, connecting through physically contiguous networks within the protein structure. Interestingly, in TEM-1, distance from the active site correlates with the robustness of neutrality to increasing selection pressure, a finding that might help explain why conditionally neutral mutations form wire-like networks that extend far from the active site.

Taken together, these findings suggest an "outside-in" structural principle for protein adaptation. The idea is that the path of

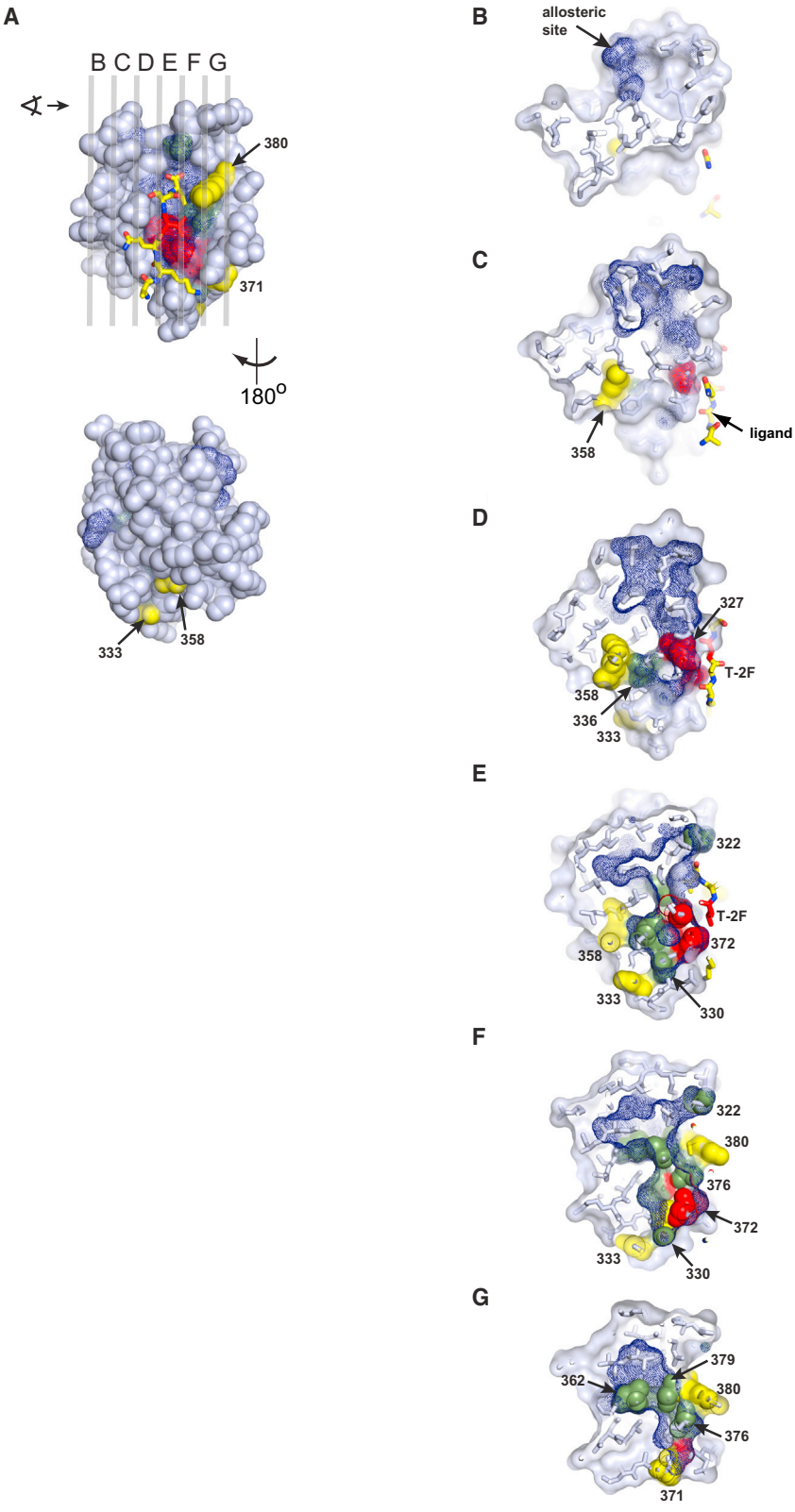


Figure 7. Relationship of the Protein Sector to Adaptive Positions

(A) A space-filling representation of PSD95^{pdz3}, with the protein sector shown as blue mesh and positions capable of adaptation to the T-2F ligand colored red (direct-switching, in sector), green (class-bridging, in sector), or yellow (class-bridging, non-sector). The data show that the four non-sector adaptive mutations (labeled) occur at surface-exposed sites distant from ligand position −2 (shown as red stick bonds). (B–G) Serial slices through PSD95^{pdz3} at the planes indicated in (A). The views are from the left as indicated. The data show that adaptive positions are nearly all contained within the protein sector (overlap of blue mesh with red and green spheres), and the four surface-exposed positions with class-neutral mutations are connected to the peripheral regions of the protein sector. Overall, adaptive positions comprise a wire-like system of physically connected residues that connects the site of adaptive challenge (ligand −2) through the protein structure.

adaptation likely starts from the acquisition of mutations at positions distant from the active site but that are wired up through a pre-existing network of cooperative amino acid interactions. Through action at a distance, these mutations have the capacity for opening up active-site conformational states that can introduce new functional phenotypes without abrogating existing function—the property of conditional neutrality. From a point of view of evolutionary dynamics, the key benefit of such mutations is that they can temporally unlink the appearance of adaptive mutations from environmental fluctuations that alter selection pressures. The neutrality protects against purifying selection and enables adaptive mutations to pre-exist in populations as cryptic variations. In contrast, mutations that cause direct phenotypic switching (even when structurally more obvious) can only support adaptation with temporal coupling of mutation and selection pressures. More generally, we propose that the degree of neutrality of adaptive mutations toward existing function will set the timescale (relative to the mutation rate and environmental switching rate) over which it can support adaptation. Thus, conditionally neutral mutations would seem to represent a pool of natural variation that is the engine for the evolution of new phenotypes.

Implications for Protein Engineering

The outside-in concept for protein adaptation is interesting because it is essentially opposite to the current practice of structure-guided protein engineering. For example, classic work on switching the primary (P1 site) substrate specificity of the serine protease trypsin to that of chymotrypsin (Hedstrom, 2002; Hedstrom et al., 1994) began with mutations at sites directly contacting the P1 side chain (the S1 pocket). The result was initial loss of enzyme function, a phenotype explained by collapse of the S1 pocket upon mutation (Perona et al., 1995). Successful transfer of specificity required the subsequent addition of mutations at positions distributed through the protein structure, which has the effect of restoring stability (and new functional specificity) to the S1 pocket. Similarly, attempts to switch the activities of type II restriction enzymes (Lukacs et al., 2000; Morgan and Luyten, 2009), transcription factors (Poelwijk et al., 2011), and β-lactamases (Stiffler et al., 2015) show that active site mutations tend to display loss of function, whereas combinations with structurally non-obvious peripheral mutations facilitate acquisition of new function. In light of the work presented here, a useful avenue for protein engineering might be to target mutations not by the principle of spatial proximity to the active site but by the spatial pattern of adaptive mutations.

In this regard, it is interesting that the adaptive mutations, both direct-switching and class-bridging, occur within the network of coevolving positions (the sector) in the PDZ and serine protease families (Halabi et al., 2009). This finding strongly argues that the pattern of adaptive mutations is not merely an idiosyncratic feature of each model system but is, instead, a deeply conserved aspect of the entire protein family that can be predicted through sequence analysis alone. It will be interesting to combine sector predictions and the principle of outside-in mutagenesis to explore new general strategies for the evolution and engineering of new protein functions.

Origins of Allostery

The finding that the protein sector contains adaptive mutations offers an interesting hypothesis about the origin of this cooperative internal architecture within protein tertiary structures. Sectors are coevolving units of protein structures and have been associated with various functional properties of proteins—catalysis, binding, and allosteric signaling (Halabi et al., 2009; Reynolds et al., 2011; Smock et al., 2010; Süel et al., 2003). A natural inference might be that the wire-like architecture of sectors, connecting active sites to distant surface sites through the protein core (Figure 7), emerged in evolution as a consequence of selection for the corresponding functional property. However, it is not obvious how such a network of cooperative amino acid interactions could be built through a process of stepwise variation and selection given that intermediate genotypes are not guaranteed to be functional.

The data presented here suggest another model. Sectors are primarily a consequence of a history of adaptation to fluctuating conditions of fitness, with the wire-like distributed architecture evolving simply because conditional neutrality is enabled by non-local allosteric mechanisms. That is, we propose that the origins of allostery lie in evolvability, not in function. According to this model, functional properties that make use of allostery (e.g., long-range regulation and signal transmission) are derivatives that emerge easily at multiple surface sites through engagement of the pre-existing allosteric network. Indeed, experiments suggest that it is possible to naively engineer new allosteric control into proteins through engagement of sector-connected surface sites (Lee et al., 2008; Reynolds et al., 2011). The recent development of techniques for fast continuous evolution of proteins (Esvelt et al., 2011) may help with designing experiments to test these ideas.

EXPERIMENTAL PROCEDURES

Global Analysis of PDZ Ligand Specificity

The comprehensive study of PDZ binding specificity is made possible by a modified version of a BTH system (McLaughlin et al., 2012) in which transcription of the chloramphenicol acetyltransferase (pZE1RM-CAT plasmid) reporter gene is made quantitatively dependent on the binding between each PDZ domain variant (fused to the bacteriophage λ-c1 DNA binding domain, pZS22 plasmid) and a library of ligands (fused to the N-terminal domain of the *Escherichia coli* RNA polymerase α subunit [pZA31 plasmid, total theoretical library complexity $20^4 = 160{,}000$]) (Figure S1; Table S1). The details of construction of the ligand library and the BTH assay are given in the Supplemental Experimental Procedures. After selection, cultures were subject to plasmid DNA isolation, PCR amplification of the ligand region of pZA31, and standard preparation for Illumina Hi-Seq 2500 sequencing (University of Texas Southwestern [UTSW] genomics core). Sequencing data were analyzed using home-written codes and MATLAB (MathWorks, available upon request) and used to compute $\Delta E_x = \log(f_x^s/f_x^u) - \log(f_o^s/f_o^u)$, the enrichment of each ligand x in the selected (s) and unselected (u) libraries relative to a reference sequence o with similar affinity for each PDZ variant. The reference sequence was CRIPT for wild-type and G330T variants and T-2F for H372A and the double mutant variants.

Expression and Purification of PSD95^{pdz3} Proteins

Wild-type or mutant PSD95^{pdz3} (amino acid range 297–415) were expressed as glutathione S-transferase (GST)-fusions in *Escherichia coli* BL21(DE3) cells and purified to near homogeneity through sequential affinity chromatography, cleavage of the GST tag, source 15Q anion exchange chromatography, and

size exclusion chromatography. Complete details are given in the Supplemental Experimental Procedures. Purified proteins were concentrated to 35 mg/ml and subsequently either flash-frozen in liquid N_2 for storage at −80°C or used immediately for crystallization. Substrate peptides for co-crystallization (CRIPT [acetyl-TKNYKQTSV-COOH], T-2F [acetyl-TKNYKQFSV-COOH]) were synthesized using standard fluorenylmethyloxycarbonyl chloride (FMOC) chemistry (UTSW Proteomics Core Facility), high-performance liquid chromatography (HPLC)-purified, and lyophilized.

Crystallization and Structure Determination of PSD95^{pdz3} Variants

Crystallization of PSD95^{pdz3} variants was performed by the vapor diffusion hanging drop method at 16°C. Details are given in the Supplemental Experimental Procedures, and specific crystallization conditions for each mutant are shown in Table S4. Diamond-shaped crystals appeared either spontaneously or with microseeding after 1–5 days and grew to 100–200 μm in length over several weeks. Single crystals were cryoprotected by serial equilibration into crystallization buffer with increasing amounts of glycerol (up to 25%) and flash-frozen in liquid N_2.

Diffraction data were collected at 100 K at either at the UTSW structural biology laboratory or at the Advanced Photon Source (Argonne National Laboratory, 19-ID) and indexed and scaled in HKL-2000 (Otwinowski and Minor, 1997) (HKL Research). Resolution cutoffs were chosen based on I/σ and *CC* 1/2 (Tables S2 and S3). Phasing and refinement were carried out using PHENIX (Python-based hierarchical environment for integrated xtallography) (Adams et al., 2010) with manual modeling in COOT (crystallographic object-oriented toolkit) (Emsley et al., 2010). The data collection and refinement statistics are summarized in Tables S2 and S3. An initial model was obtained from rigid body and temperature factor refinement using published structures of PSD95^{pdz3} (PDB: 1BFE and 1BE9, with ligand removed) and subjected to 0.5-Å coordinate randomization followed by Cartesian simulated annealing to minimize phase bias. Further computational refinement steps involved iterative rounds of positional and temperature factor minimization, manual model building, solvent placement, and translation/libration/screw (TLS) refinement, guided by decrease in crystallographic R-factors. Figures were prepared with PyMol (DeLano, 2002). The atomic coordinates and structure factors were deposited in the PDB: 5HEB (PSD95^{pdz3}(WT)-CRIPT), 5HED (PSD95^{pdz3}(WT)-T-2F), 5HET (PSD95^{pdz3}(G330T)-apo), 5HEY (PSD95^{pdz3}(G330T)-CRIPT), 5HF1 (PSD95^{pdz3}(G330T)-T-2F), 5HFB (PSD95^{pdz3}(H372A)-CRIPT), 5HFC (PSD95^{pdz3}(H372A)-T-2F), 5HFF (PSD95^{pdz3}(G330T, H372A)-T-2F).

Computational Simulations

The population dynamics model shown in Figure 4A was simulated using custom-written code developed in MATLAB and executed on a LINUX high-performance computing cluster (BioHPC, UTSW Medical Center). The complete annotated codes and details of the simulation are given in the Supplemental Experimental Procedures. The model simulates the dynamics of a constant-sized population (here, $N = 1000$, large relative to the number of genotypes) comprising the four PDZ variants (wild-type, G330T, H372A, and the double mutant), with mutation and selection under a fluctuating environment of either CRIPT or T-2F ligands. For each trial of switching from CRIPT to T-2F in which the double mutant ultimately goes to fixation in the population, we computed the fraction of G330T and H372A in the interval from the switch to fixation of the double mutant. Limits for integration were automatically determined by empirical fitting of the probability density of the double mutant in each trial (see codes). The data in Figures 4C–4F were obtained from ~ 500 trials of CRIPT to T-2F ligand switching each.

Statistical Tests

Fisher's exact tests of the association between adaptive positions and classification of PSD95^{pdz3} positions by sector, positional conservation, or spatial proximity were carried out using MATLAB. Positional conservation and sector were defined as reported previously (McLaughlin et al., 2012), and spatial proximity was computed using coordinates from a high-resolution crystal structure of PSD95^{pdz3} (PDB: 1BE9 (Doyle et al., 1996)) as the closest distance between any pair of atoms of ligand T-2 and every other amino acid; see Figure S5 for tables and the contingency matrices.

ACCESSION NUMBERS

The accession numbers for the atomic coordinates and structure factors reported in this paper are PDB: 5HEB (PSD95^{pdz3}(WT)-CRIPT), 5HED (PSD95^{pdz3}(WT)-T-2F), 5HET (PSD95^{pdz3}(G330T)-apo), 5HEY (PSD95^{pdz3}(G330T)-CRIPT), 5HF1 (PSD95^{pdz3}(G330T)-T-2F), 5HFB (PSD95^{pdz3}(H372A)-CRIPT), 5HFC (PSD95^{pdz3}(H372A)-T-2F), and 5HFF (PSD95^{pdz3}(G330T, H372A)-T-2F).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.05.047.

AUTHOR CONTRIBUTIONS

A.S.R. and R.R. developed the research plan and experimental strategy. A.S.R. and K.I.W. performed all experiments, and A.S.R. and R.R. carried out the computational simulations. A.S.R., K.I.W., and R.R. interpreted the data and wrote the paper.

ACKNOWLEDGMENTS

We thank members of the R.R. lab for critical review of the manuscript; M.A. Socolich, F.J. Poelwijk, and S. Subramanian for technical help; and K.A. Reynolds for discussions. We also thank the UT Southwestern Genomics Core for assistance with sequencing and the High Performance Computing Group (BioHPC) at UT Southwestern for facilitating the computational simulations. This work was supported by NIH Grant RO1EY018720-05 (to R.R.), Robert A. Welch Foundation Grant I-1366 (to R.R.), and the Green Center for Systems Biology at UT Southwestern Medical Center. A.S.R. and K.I.W. were supported in part through pre-doctoral fellowships (NIGMS T32 GM008203).

Received: January 29, 2016
Revised: March 23, 2016
Accepted: May 13, 2016
Published: June 16, 2016

REFERENCES

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. Cell *163*, 594–606.

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. *66*, 213–221.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223–230.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D.S. (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature *444*, 929–932.

Bloom, J.D., Silberg, J.J., Wilke, C.O., Drummond, D.A., Adami, C., and Arnold, F.H. (2005). Thermodynamic prediction of protein neutrality. Proc. Natl. Acad. Sci. USA *102*, 606–611.

Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. (2006). Protein stability promotes evolvability. Proc. Natl. Acad. Sci. USA *103*, 5869–5874.

Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A., and Sauer, R.T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science *247*, 1306–1310.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

Doyle, D.A., Lee, A., Lewis, J., Kim, E., Sheng, M., and MacKinnon, R. (1996). Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell *85*, 1067–1076.

Draghi, J.A., and Plotkin, J.B. (2011). Molecular evolution: Hidden diversity sparks adaptation. Nature *474*, 45–46.

Draghi, J.A., Parsons, T.L., Wagner, G.P., and Plotkin, J.B. (2010). Mutational robustness can facilitate adaptation. Nature *463*, 353–355.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501.

Esvelt, K.M., Carlson, J.C., and Liu, D.R. (2011). A system for the continuous directed evolution of biomolecules. Nature *472*, 499–503.

Ferguson, A.D., Amezcua, C.A., Halabi, N.M., Chelliah, Y., Rosen, M.K., Ranganathan, R., and Deisenhofer, J. (2007). Signal transduction pathway of TonB-dependent transporters. Proc. Natl. Acad. Sci. USA *104*, 513–518.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

Gould, S.J., and Vrba, E.S. (1982). Exaptation - a Missing Term in the Science of Form. Paleobiology *8*, 4–15.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774–786.

Harris, B.Z., and Lim, W.A. (2001). Mechanism and role of PDZ domains in signaling complex assembly. J. Cell Sci. *114*, 3219–3231.

Hayden, E.J., Ferrada, E., and Wagner, A. (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. Nature *474*, 92–95.

Hedstrom, L. (2002). Serine protease mechanism and specificity. Chem. Rev. *102*, 4501–4524.

Hedstrom, L., Perona, J.J., and Rutter, W.J. (1994). Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. Biochemistry *33*, 8757–8763.

Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J., and Ranganathan, R. (2008). Surface sites for engineering allosteric control in proteins. Science *322*, 438–442.

Lee, S.Y., Banerjee, A., and MacKinnon, R. (2009). Two separate interfaces between the voltage sensor and pore are required for the function of voltage-dependent K(+) channels. PLoS Biol. *7*, e47.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295–299.

Lukacs, C.M., Kucera, R., Schildkraut, I., and Aggarwal, A.K. (2000). Understanding the immutability of restriction enzymes: crystal structure of BglII and its DNA substrate at 1.5 A resolution. Nat. Struct. Biol. *7*, 134–140.

Luria, S.E., and Delbrück, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics *28*, 491–511.

McLaughlin, R.N., Jr., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. USA *108*, E1293–E1301.

Morgan, R.D., and Luyten, Y.A. (2009). Rational engineering of type II restriction endonuclease DNA binding and cleavage specificity. Nucleic Acids Res. *37*, 5222–5233.

Niethammer, M., Valtschanoff, J.G., Kapoor, T.M., Allison, D.W., Weinberg, R.J., Craig, A.M., and Sheng, M. (1998). CRIPT, a novel postsynaptic protein that binds to the third PDZ domain of PSD-95/SAP90. Neuron *20*, 693–707.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. *276*, 307–326.

Perona, J.J., Hedstrom, L., Rutter, W.J., and Fletterick, R.J. (1995). Structural origins of substrate discrimination in trypsin and chymotrypsin. Biochemistry *34*, 1489–1499.

Poelwijk, F.J., de Vos, M.G., and Tans, S.J. (2011). Tradeoffs and optimality in the evolution of gene regulation. Cell *146*, 462–470.

Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. Cell *147*, 1564–1575.

Reynolds, K.A., Russ, W.P., Socolich, M., and Ranganathan, R. (2013). Evolution-based design of proteins. Methods Enzymol. *523*, 213–235.

Russ, W.P., Lowery, D.M., Mishra, P., Yaffe, M.B., and Ranganathan, R. (2005). Natural-like function in artificial WW domains. Nature *437*, 579–583.

Salverda, M.L., De Visser, J.A., and Barlow, M. (2010). Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. FEMS Microbiol. Rev. *34*, 1015–1036.

Sheng, M., and Sala, C. (2001). PDZ domains and the organization of supramolecular complexes. Annu. Rev. Neurosci. *24*, 1–29.

Smith, J.M. (1970). Natural selection and the concept of a protein space. Nature *225*, 563–564.

Smock, R.G., Rivoire, O., Russ, W.P., Swain, J.F., Leibler, S., Ranganathan, R., and Gierasch, L.M. (2010). An interdomain sector mediating allostery in Hsp70 molecular chaperones. Mol. Syst. Biol. *6*, 414.

Socolich, M., Lockless, S.W., Russ, W.P., Lee, H., Gardner, K.H., and Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. Nature *437*, 512–518.

Songyang, Z., Fanning, A.S., Fu, C., Xu, J., Marfatia, S.M., Chishti, A.H., Crompton, A., Chan, A.C., Anderson, J.M., and Cantley, L.C. (1997). Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science *275*, 73–77.

Stiffler, M.A., Chen, J.R., Grantcharova, V.P., Lei, Y., Fuchs, D., Allen, J.E., Zaslavskaia, L.A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. Science *317*, 364–369.

Stiffler, M.A., Hekstra, D.R., and Ranganathan, R. (2015). Evolvability as a function of purifying selection in TEM-1 β-lactamase. Cell *160*, 882–892.

Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat. Struct. Biol. *10*, 59–69.

Tokuriki, N., and Tawfik, D.S. (2009). Protein dynamism and evolvability. Science *324*, 203–207.

Tonikian, R., Zhang, Y., Sazinsky, S.L., Currell, B., Yeh, J.H., Reva, B., Held, H.A., Appleton, B.A., Evangelista, M., Wu, Y., et al. (2008). A specificity map for the PDZ domain family. PLoS Biol. *6*, e239.

Wagner, A. (2005). Robustness, evolvability, and neutrality. FEBS Lett. *579*, 1772–1778.

Wedemayer, G.J., Patten, P.A., Wang, L.H., Schultz, P.G., and Stevens, R.C. (1997). Structural insights into the evolution of an antibody combining site. Science *276*, 1665–1669.

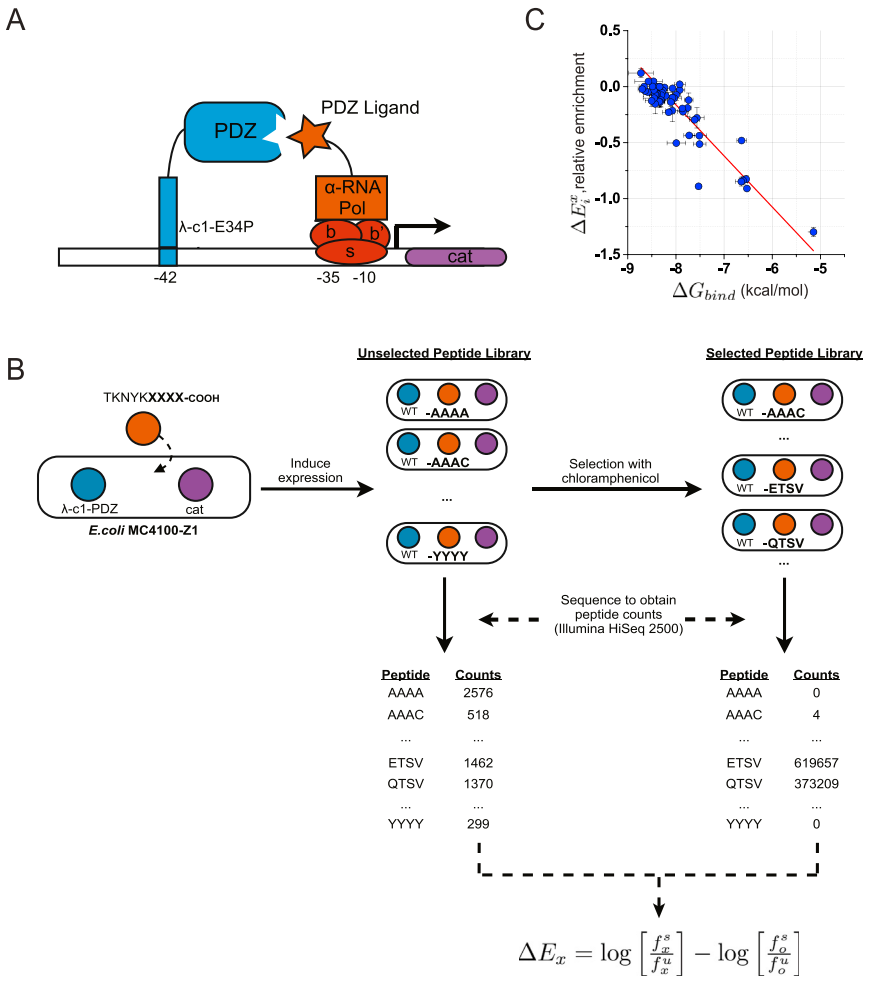


Figure S1. The BTH Assay and the PDZ Ligand Screen, Related to Figures 2 and 3

(A) The BTH assay is modified from previous work (Dove et al., 1997; McLaughlin et al., 2012) to give a quantitative readout of the affinity between a PDZ domain and its peptide ligand in the expression of a target gene. The PDZ domain is expressed as a C-terminal fusion with the DNA binding domain of the bacteriophage λ-c1 protein (with the E34P mutation to optimize dynamic range (McLaughlin et al., 2012)). The peptide ligand is expressed as a C-terminal fusion to the N-terminal domain of the *E.coli* RNA polymerase α-subunit, and the target gene is chloramphenicol acetyl transferase (cat).

(B) A library of potential PDZ ligands (randomized in terminal four residues, 20^4 = 160,000 total) is transformed into *E.coli* MC4100-Z1 cells expressing the other components of the BTH assay, and are selected on chloramphenicol for PDZ function (see methods for details). The ligand library is subject to Illumina HiSeq sequencing to count the frequency of each C-terminal peptide in the library in both the unselected and selected cell populations. As shown for a few examples, legitimate ligands are highly enriched in the selected population while non-binding ligands are absent. Quantitatively, we compute the enrichment of each ligand *x* relative to reference ligand *o* ("relative enrichment," ΔE_x) according to the equation shown. The reference ligand is chosen for same absolute affinity over PDZ variants to put datasets on the same scale for quantitative comparison; thus, CRIPT for wild-type and G330T and T-2F for H372A and G330T,H372A.

(C) a standard curve showing the relationship of relative enrichment to the equilibrium free energy of binding for a set of 83 single point mutations in the $PSD95^{pdz3}$ domain to the CRIPT ligand. The experiment shows that the sequencing based assay provides a quantitative measure of ligand binding.

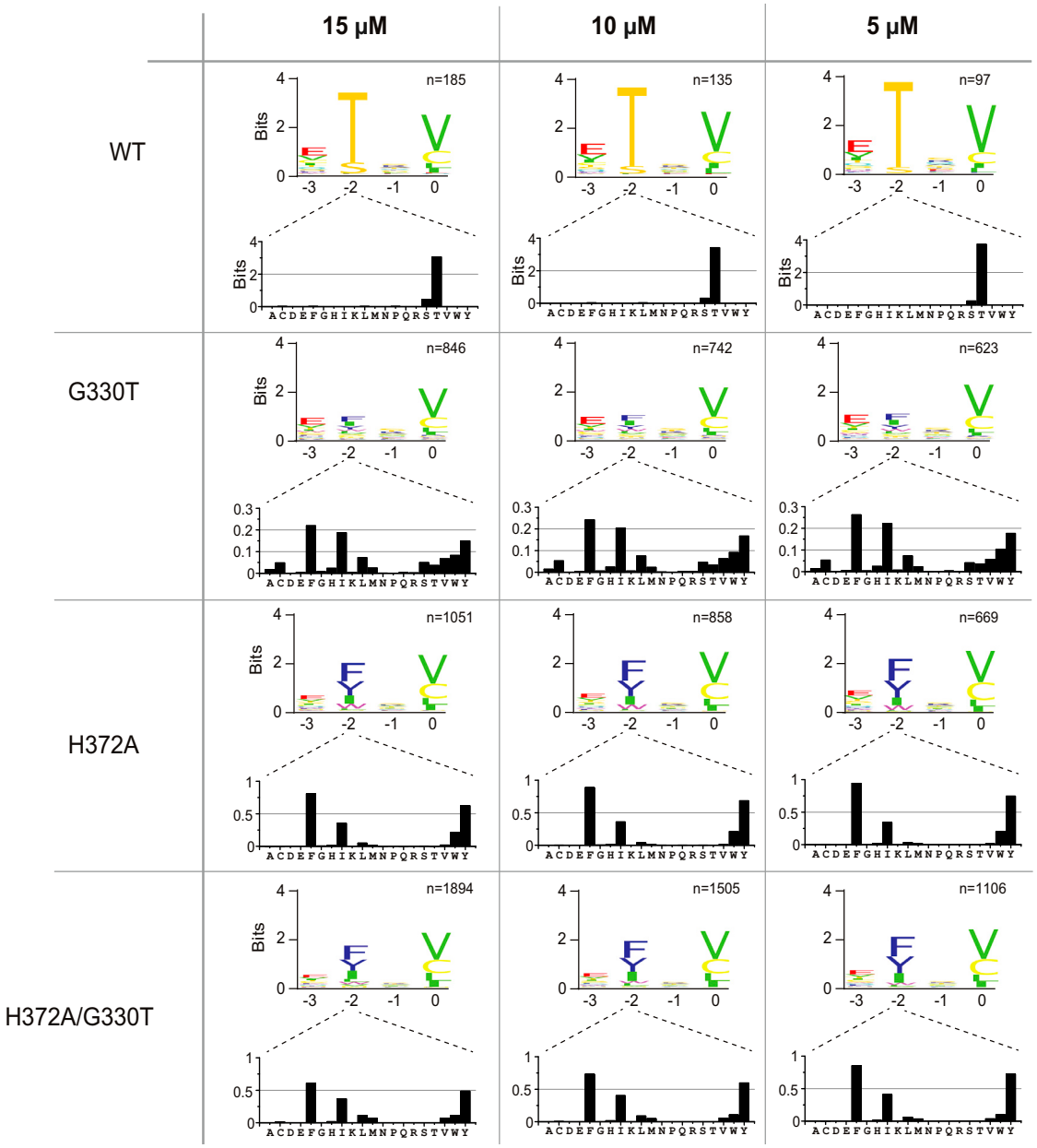


Figure S2. Sequence Profiles of Ligand Peptides Bound by PDZ Variants along the Adaptive Path, Related to Figures 2 and 3

Each panel shows the pattern of amino acid preferences at ligand positions (labeled −3 to 0) as a "sequence logo" plot for the PDZ variant and affinity cutoff indicated. In each panel, the bar graph shows a more detailed, quantitative view of the amino acid profile at position −2, the main specificity determinant. The units of amino acid preference are "bits," where a frequency of unity for a particular residue would give a bit score of log_2 (20). The data permit two conclusions. First, that the amino acid profiles are robust to affinity cutoff. Second, that wild-type PSD95^{pdz3} (WT) displays class I (S/T) preference at position −2, H372A and the G330T,H372A double mutant display class II (bulky hydrophobic) preference, and G330T displays a class-bridging phenotype, capable of binding both classes of ligands.

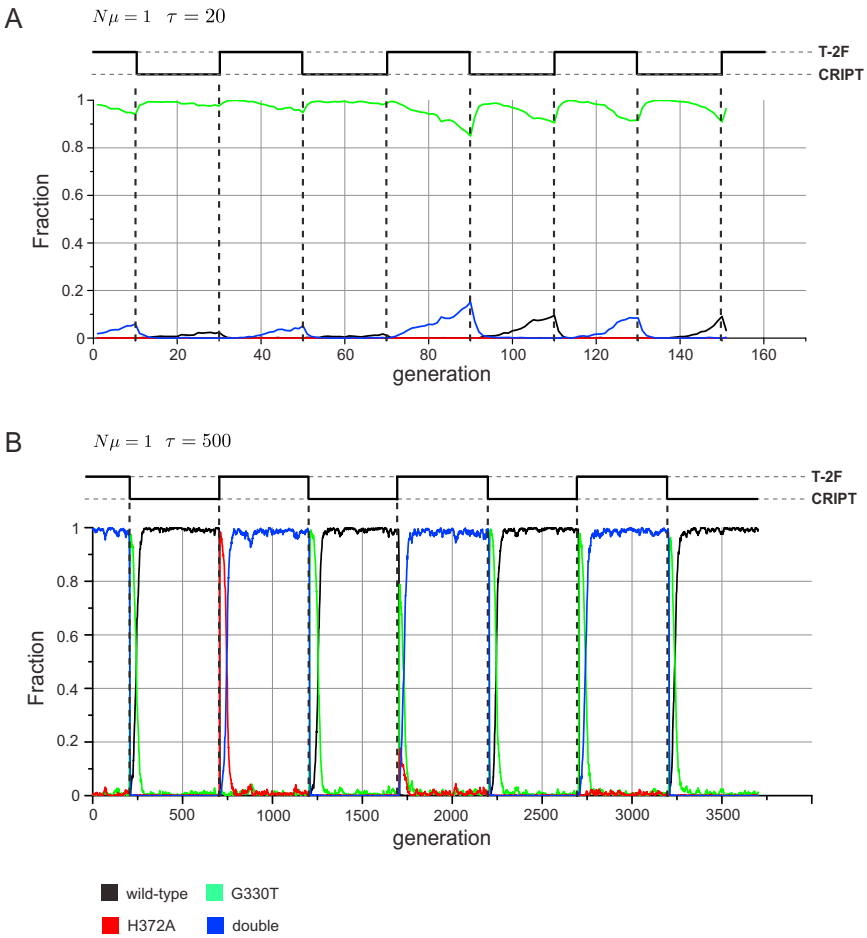


Figure S3. The Population Dynamics Model as a Function of τ, the Ligand Switching Rate, Related to Figure 4

(A) A portion of one simulation trajectory, mutation rate $N\mu$ = 1 (meaning production of about one single mutation per generation) and the wait-time for ligand switching τ = 20 (meaning the ligand switches between CRIPT and T-2F every 20 generations). At this fast switching rate, the only genotype that dominates in the population is the one capable of binding both ligands well, G330T.

(B) For comparison, a slower ligand switching rate (here, τ = 500) permits adaptation to the genotypes that are most fit for the CRIPT and T-2F environments (wild-type and double mutant, respectively), with varying proportion of G330T and H372A as intermediates.

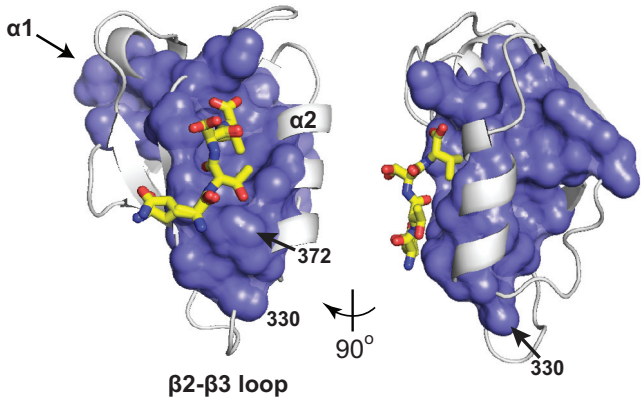


Figure S4. The Protein Sector in the PDZ Domain Family, Related to Figures 6 and 7

Statistical coupling analysis (SCA (Halabi et al., 2009; Lockless and Ranganathan, 1999)) reveals a network of coevolving amino acids in the PDZ domain family - a "sector" - that links the ligand binding pocket to a few distant surface sites. Residues G330 and H372 are indicated.

A

adaptive position	n_mut
322	9
372	9
330	7
362	4
327	3
333	3
336	3
376	2
358	1
371	1
379	1
380	1
total	44

B

sector	positional cons	spatial proximity
372	357	372
329	324	327
336	363	326
351	356	376
347	338	325
386	362	380
362	359	328
325	347	373
322	351	339
363	372	377
376	353	379
375	341	337
323	325	340
330	375	329
364	388	375
353	379	338
388	364	324
379	327	336
359	323	374
327	345	331

C

classification	fraction adaptive
sector	**38/44**
positional cons	**17/44**
spatial proximity	**19/44**

D

sector

adaptive \ sector	Y	N	
Y	8	4	12
N	12	57	69
	20	61	81

p = 0.001

E

positional cons

adaptive \ positional cons	Y	N	
Y	4	8	12
N	16	53	69
	20	61	81

p = 0.47

F

spatial proximity

adaptive \ spatial proximity	Y	N	
Y	6	6	12
N	14	55	69
	20	61	81

p = 0.063

Figure S5. The Relationship of Adaptive Positions to the Protein Sector, Positional Conservation, and Spatial Proximity to the Site of Ligand Variation, Related to Figures 6 and 7

(A) A list of positions in PSD95^{pdz3} that contain mutations capable of adaptation to the T-2F ligand (shaded regions, Figure 7). The number of adaptive mutations at each position is indicated (total, 44).

(B) As defined in (McLaughlin et al., 2012), the group of statistically co-evolving positions (the "sector") in the PDZ protein family comprises 20 positions, indicated at left (shown in Figure S4). For comparison, the middle and right columns show a ranked list of the top 20 conserved positions and the top 20 spatially closest positions to ligand position −2, respectively. Positional conservation is computed by the Kullback-Liebler relative entropy (Cover and Thomas, 2006), and spatial proximity is measured as a closest distance between atoms comprising the T-2 ligand residue and each residue in PSD95^{pdz3}. Yellow highlight indicates a match with an adaptive position.

(C) The fraction of total adaptive positions included in each indicated classification of PSD95^{pdz3} positions. These data show that the majority of adaptive mutations occur in the protein sector, and that fewer are accounted for by the other classifications.

(D–F) 2 × 2 contingency matrices describing the relationship between classifications of PSD95^{pdz3} positions by adaptive capacity (rows) and by sector, positional conservation, or spatial proximity to the T-2F ligand position (columns). These data permit a Fisher Exact Text, with the null hypothesis that the two classifications are statistically independent. p values show that the null hypothesis is rejected at a $p < 0.05$ level for the PDZ sector, and is not rejected otherwise. Note that the near-significant p value for spatial proximity makes sense; since physical interactions between atoms arise largely from very short-range forces, we expect that adaptive mutations for T-2F will be enriched in spatially local sites. But the data show that the concept of spatial proximity is incomplete, discounting the adaptive effect of a number of positions (e.g., 330, 362, 322, etc.) that work through allosteric mechanisms. The data suggest that the sector provides a better model for the adaptive effect of mutations by taking such mechanisms into account.

Cell, Volume *166*

Supplemental Information

Origins of Allostery and Evolvability in Proteins:

A Case Study

Arjun S. Raman, K. Ian White, and Rama Ranganathan

Origins of allostery and evolvability in proteins: Supplementary Information

Arjun S. Raman[1], Kristopher I. White[1] & Rama Ranganathan[1,2*]

[1]The Green Center for Systems Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390

[2]Departments of Pharmacology and Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390

Contents

I. SUPPLEMENTARY EXPERIMENTAL PROCEDURES

A. Global analysis of PDZ ligand specificity.

Comprehensive study of PDZ binding specificity is made possible by a modified version of a bacterial two-hybrid system (McLaughlin et al., 2012) in which transcription of the chloramphenicol acetyl transferase (CAT) reporter gene (pZE1RM plasmid, pRM+ promoter, ampicillin resistant) is made quantitatively dependent on the binding between a PDZ domain (fused to the pRM+ promoter-binding bacteriophage λ-c1 DNA binding domain, pZS22 plasmid, IPTG inducible, trimethoprim (trm) resistant) and its ligand (fused to the N-terminal domain of *E.coli* RNA polymerase α subunit, pZA31 plasmid, anhydrotetracycline (aTC) inducible, kanamycin resistant) (Fig. S1). Electrocompentent MC4100-Z1 cells containing pZE1RM-CAT and pZS22-PDZ3 variant plasmids were transformed with 1 µl of 20 ng/µl pZA31-RNAα-ligand library (see below), recovered for one hour in LB media, grown in 20 µg/ml trm, 50 µg/ml kan, 100 µg/ml amp to OD_{550} of 0.04, and induced using 50 ng/ml doxycyline plus antibiotics for 3 hours to an OD_{550} of 0.1. 10ml of the induced culture was used to innoculate 100mL LB + antibiotics as above for selection; the remainder was reserved as the pre-selection population for deep sequencing. Selection was carried out with 150 µg/ml chloramphenicol for 6 hours (taking care that $OD_{550} \leq 0.1$), washed in LB medium, and grown overnight at 37°C. Both pre- and post-selection cultures were subject to plasmid DNA isolation, PCR amplification of the ligand region of pZA31, and standard preparation for Illumina Hi-Seq 2500 sequencing (UT Southwestern genomics core). Sequencing data were analyzed using home-written codes and MATLAB (Mathworks Inc., provided upon request) and used to compute $\Delta E_x = \log(f_x^s/f_x^u) - \log(f_o^s/f_o^u)$, the enrichment of each ligand x in the selected (s) and unselected (u) libraries relative to a reference sequence o with similar affinity for each PDZ variant. The reference sequence was CRIPT for wild-type and G330T variants, and T-2F for H372A and the double mutant variants.

B. Construction of the ligand library

The library of PDZ ligands (randomized in the C-terminal four amino acid positions, total theoretical library complexity $20^4 = 160,000$) was generated as C-terminal fusions with the N-terminal domain of *E.coli* RNA polymerase α subunit. The library was made using NNS oligonucleotide-directed mutagenesis with a pZA31-RNAα template containing a non-binding PDZ ligand (N-TKNYKQGGG-COOH) to eliminate background binding. Two oligonucleotides (one sense, one antisense) were synthesized (IDT) with each sequence complementary to 15 base-pairs (bp) on either side but with one oligo containing four consecutive NNS codons at the target positions; N is a mixture of A, T, C, and G, and S is a mixture of G and C. This results in 32 codons at each position encoding all 20 amino acids. The

oligos incude a type IIs restriction site (BsaI), designed to optimize cloning efficiency by enabling a unimolecular ligation protocol. We carried out a single round of PCR, amplifying the entire plasmid while encoding the full library of ligand sequences. This product was subsequently restricted with BsaI, subject to a unimolecular ligation reaction (1 ml, incubated overnight at 16°C), and purified into a final volume of 10 µl (Zymo purification kit). Ten individual transformations into MaxDH10B *E. coli* (Invitrogen) were made, grown overnight after recovery, and plasmid DNA prepped so as to minimize any possible bottlenecking effect. Transformation of the final library into MC4100-Z1 cells for selection yielded greater than 10^8 transformants, and a near complete representation of the theoretical complexity (Table S1).

C. Expression and purification of PSD95^{pdz3} proteins.

pGEX-4T-1 plasmids containing Glutathione-S-transferase (GST)-fusions of wild-type or mutant PSD95^{pdz3} (amino acid range 297-415) were transformed into *Escherichia coli* BL21(DE3) cells and grown overnight on LB plus 100 μg/mL ampicillin (amp) plates. Streaks of colonies were used to start overnight cultures (LB + amp), used to innoculate 1L cultures (Terrific Broth or ZYM-5052 auto-inducing medium (Studier, 2005) + 100 µg/mL amp). Cultures were grown to an optical density (600 nm) of 0.6-0.8 at 37 ^{o}C, induced overnight at 18 ^{o}C (supplemented with 1 mM isopropyl--D-thiogalactopyranoside if manual induction), and then harvested by centrifugation. Pellets were resuspended in lysis buffer composed of phosphate-buffered saline (PBS; 140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.3) supplemented with 1% glycerol, 1 mg/ml hen egg white lysozyme, 1 mM dithiothreitol (DTT), EDTA-free protease inhibitor cocktail (Roche). The cell suspension was subjected to sonication and centrifugation, and clarified lysate was then incubated with glutathione sepharose 4B resin (GE Healthcare). Bound protein was washed with PBS supplemented with 1% glycerol and 1 mM DTT, and the GST tag was cleaved through bovine thrombin (Calbiochem) proteolysis overnight at room temperature in PBS supplemented with 10% glycerol and 1 mM DTT. Thrombin was removed by benzamidine sepharose (GE Healthcare) and the PDZ domain was purified to near-homogeneity using a Source 15Q anion exchange (GE Healthcare) column employing a linear gradient from low salt (20 mM Tris HCl pH 7.5, 1% glycerol, 1.0 mM DTT) to high salt (20 mM Tris HCl pH 7.5, 1 M NaCl, 1% glycerol, 1 mM DTT). The protein was dialyzed into 10 mM HEPES pH 7.2, 10 mM NaCl, concentrated and subject to size-exclusion chromatography (Superdex 75, GE Healthcare). Peak fractions were pooled, concentrated to 35 mg/mL, and subsequently either flash frozen in liquid N_2 for storage at -80°C or used immediately for crystallization. Substrate peptides for co-crystallization (CRIPT (Acetyl-TKNYKQTSV-COOH), T-2F (Acetyl-TKNYKQFSV-COOH)) were synthesized using standard FMOC chemistry (UTSW Proteomics Core Facility), HPCL purified, and lyophilized.

D. Crystallization and structure determination of PSD95^{pdz3} variants.

Crystallization of PSD95^{pdz3} variants was performed by the vapor diffusion hanging drop method. In all cases, purified protein was diluted to a final concentration of $\sim$ 9 mg/ml in protein buffer (10 mM HEPES pH 7.2, 10 mM NaCl). Where applicable, peptide was included in protein buffer to a final molar ratio of 2:1 relative to protein. Reservoir solutions typically contained 1 M sodium citrate, pH 7.0; specific crystallization conditions for each mutant are shown in Table S4. Equal amounts (1.5 µl) of protein and reservoir solution were mixed and equilibrated against 500 µl of crystallization buffer at 16°C. Diamond-shaped crystals appeared either spontaneously or with microseeding after 1–5 days and grew to 100–200 µm in length over several weeks. To prepare microseeding solutions, wild-type crystals of the appropriate state were crushed and resuspended in crystallization buffer. Single crystals were cryoprotected by serial equilibration into crystallization buffer with increasing amounts of glycerol (up to 25%) and flash frozen in liquid N_2.

Diffraction data were collected at 100 K at either at the UT Southwestern structural biology laboratory or at the Advanced Photon Source (Argonne National Laboratory, 19-ID) and indexed and scaled in HKL-2000 (Otwinowski and Minor, 1997) (HKL Research). Resolution cutoffs were chosen based on I/σ and CC 1/2 (Tables S2-S3). Phasing and automated refinement was carried out using PHENIX (Adams et al., 2010) with manual modeling in COOT (Emsley et al., 2010); the data collection and refinement statistics are summarized in Tables S2-S3. An initial model was obtained from rigid body and temperature factor refinement using published structures of PSD95^{pdz3} (PDB 1BFE and 1BE9, with ligand removed), and subject to 0.5 Å coordinate randomization followed by Cartesian simulated annealing to help reduce phase bias. Further computational refinements steps involved iterative rounds of positional and temperature factor minimization, manual model building, solvent placement, and TLS refinement, guided by decrease in crystallographic R-factors. Figures were prepared with PyMol (DeLano, 2002). The atomic coordinates and structure factors have been deposited in the Protein Data Bank with the following acces-

sion numbers: 5HEB (PSD95^{pdz3}(WT)-CRIPT), 5HED (PSD95^{pdz3}(WT)-T-2F), 5HET (PSD95^{pdz3}(G330T)-apo), 5HEY (PSD95^{pdz3}(G330T)-CRIPT), 5HF1 (PSD95^{pdz3}(G330T)-T-2F), 5HFB (PSD95^{pdz3}(H372A)-CRIPT), 5HFC (PSD95^{pdz3}(H372A)-T-2F), 5HFF (PSD95^{pdz3}(G330T, H372A)-T-2F).

E. Computational simulations and codes

The model shown in Fig. 4 and Fig. S3 simulates the dynamics of a constant sized population (here, $N = 1000$, large relative to the number of genotypes) comprising the four PDZ variants (wild-type, G330T, H372A, and the double mutant) with mutation and selection under a fluctuating condition of fitness. In each generation, single mutations occur with probability μ, double mutations with probability μ^2, and selection re-draws the frequency of each genotype according to its ability to bind ligand relative to all other available genotypes. The fractional binding of each genotype is determined from the experimentally measured equilibrium dissociation constants (Fig. 1B), and the ligand (CRIPT or T-2F) switches every τ generations. For each trial of switching from CRIPT to T-2F in which the double mutant ultimately goes to fixation in the population, we compute the fraction of G330T and H372A in the interval from the switch to fixation of the double mutant; limits for integration were automatically determined by empirical fitting of the probability density of the double mutant in each trial (see codes). The data in Figs. 4C-F were obtained from ∼500 trials of CRIPT to T-2F ligand switching each. The codes were executed using a custom shell script on a high-performance LINUX cluster (BioHPC, UT Southwestern Medical Center).

```
function [out] = evSim_RWR(pinit, mu, tau, nswitches,Kds,switch_mode)
% [sim]=evSim_RWR([1000;0;0;0],mu,tau,nswitches, switch_mode,f);

% The population dynamics model in Raman et al, "Origins of allostery and
% evolvability in proteins: a case study".  This function models the
% dynamics of population shifts in WT, G330T, H372A, and the double mutant
% given an initial population structure, mutation rate, ligand switching
% time, and other parameters, described below.

%  Inputs:

%   (1) "pinit" is the initial population vector, in order wild-type,
%   G330T, H372A, and the double mutant. For example, pinit=[1000;0;0;0] to
%   start with 1000 WT individuals. sum(pinit) gives the total population
%   size, N. In paper pinit = [1000;0;0;0].

%   (2) "mu" is the mutation rate - the probability of a single mutation at
%   each generation.  So, for example, mu=0.001 gives N*mu=1. The paper
%   describes simulations at mu = 0.0001, 0.001, and 0.01.

%   (3) "tau" is the wait time in generations for ligand switching. So,
%   tau=100 means the ligand switches every 100 generations.

%   (4) "nswitches" is total number of ligand switches, and so
%   nswtiches*tau is the total number of generations simulated.
%
%   (5) "Kds" is a 2 X 4 matrix of equilbrium dissociation constants,
%   assumed in units of micromolar. The columns correspond to WT, G330T,
%   H372A, G330T,H372A in order, and rows to CRIPT or T-2F ligand in
%   order. Values are given in Fig. 1B. Kds(:,1)=[0.8 2.2 26.9 22.1] and
%   Kds(:,2)=[36 1.8 1.9 0.5].

%   (5) "switch_mode" is a flag that determines whether ligand switching is
%   regular (every tau generations, switch mode=0), or is Poisson
%   distributed with a mean wait time of tau generartions
%   (switch_mode~=%0).  Default switch_mode=0.

% Outputs:

% "out" is a structure with two fields...out.input has all the inputs and
% out.output has "P", the population vector at each generation,
% "gen_switch", the generation at which each ligand switch happens
% (always starts with CRIPT).

% ****************************************

% ****************************
% Step 1: Set Inputs
% ****************************
if nargin>6
    switch_mode=0;
end
% Determine fractions bound from the input dissociation constants (from
% experiments in Fig. 1, in units of micromolar).  Ligand concentration is
% fixed at 10 micromolar, as explained in the main text.
```

```
56
57 if size(Kds,1)==2; Kds=Kds';end % in case Kds matrix needs transposition
58 f=10./(10+Kds);
59
60 % ******************************
61 % Step 2: Initialization
62 % ******************************
63
64 len=(nswitches+1)*tau; % set number of generations
65
66 % Defines svect, which stores the generations at which ligands switch.
67 % Regular switching every tau generations if switch_mode=0 and Poissonian
68 % switching otherwise.
69 svect=zeros(1,len);
70 if switch_mode==0
71     svect([1:tau:len])=1;svect(1)=0;
72 else
73     draws=round(exprnd(tau,1,nswitches));
74     index=0;
75     for i=1:numel(draws)
76         svect(draws(i)+index)=1;
77         index=index+draws(i);
78     end
79     svect(1)=0;
80 end
81 % set initial ligand environment, 1 is CRIPT, -2 is T-2F
82 env=1;sel=1;
83
84 % initialize population vector (numbers of each species, a 4 X 1 vector)
85 % and individual vector (the actual members of the population, labelled
86 % according to type, a 1000 X 1 vector).
87 P=zeros(4,len);
88 P(:,1)=pinit;
89 psize=sum(pinit); % total population size
90 labels = [0 1 2 3]; % the genotypes (labels) in decimal. 0-WT, 1-G330T, 2-H372A, 3-the double mutant
91 f_vect=pinit./sum(pinit); % frequncies
92 i_vect=(labels*mnrnd(1,f_vect,psize)')'; % make initial individual vector from frequencies (in decimal)
93 m_tmp=dec2bin(i_vect(:,1),2); % binary version of the population vector
94 m_vect=[str2num(m_tmp(:,1)) str2num(m_tmp(:,2))]; %change from char to num. This is the initial popoulation vect
       in binary
95
96 % ********************************
97 % Step 3: The simulation
98 % ********************************
99
100 % The idea is to pick an environment (that defines the selections),
101 % randomly draw single and double mutations using binary representation,
102 % convert to decimal population, apply selection to adjust frequencies,
103 % redraw population, convert to binary representation, and iterate.  Very
104 % likely more efficient ways to do this...
105
106 for i=2:len
107     if svect(i)==1
108         env=-env; % switch ligands if at the generation specified in svect
109         sel=sel-env;
110     end
111     fit_vect=(f(:,sel))./(sum((P(:,i-1)./psize).*f(:,sel))); % convert fraction bound to relative fitness
112     m_vect=abs((poissrnd(mu,2,psize)~=0)'- m_vect); % apply random mutation given mu
113     p_vect=bin2dec(num2str(m_vect)); % switch to decimal
114     P_tmp=[numel(find(p_vect==0));numel(find(p_vect==1));numel(find(p_vect==2));numel(find(p_vect==3))]; % get
          population counts after mutation
115     f_vect=(fit_vect.*P_tmp)./sum(fit_vect.*P_tmp); %get frequencies with applying selection on species in
          population
116     i_vect=(labels*mnrnd(1,f_vect',psize)')'; % generate new population, now after mutation and selection
117     m_tmp=dec2bin(i_vect,2); % back to binary matrix
118     m_vect=[str2num(m_tmp(:,1)) str2num(m_tmp(:,2))]; % switch from char to num
119     P(:,i)=[numel(find(i_vect==0));numel(find(i_vect==1));numel(find(i_vect==2));numel(find(i_vect==3))]; %write
          new population vector after mutation and selection
120 end
121
122 % ********************************
123 % Step 4: Make output structure
124 % ********************************
125
126 % first, we determine the generations at which CRIPT changes to T-2F (every
127 % other one)
128 gen_switch_either=find(svect);
129 gen_switch=gen_switch_either(1:2:numel(gen_switch_either));
130 n_trials=numel(gen_switch);
131
132 % the output structure
133 out.input.pinit = pinit;
134 out.input.mu = mu;
```

```
out.input.tau = tau;
out.input.nswitches = nswitches;
out.input.f = f;
out.input.switch_mode = switch_mode;
out.output.P = P;
out.output.gen_switch = gen_switch;

% ************************************
% Step 5: Analysis
% ************************************
% This is optional, and is best avoided for high-throughput study of many
% conditions.  Comment out below if not desired.

% a plot of the population dynamics over the simulation length
h_sim=figure;clf;hold on;grid on;
plot(P(1,:),'k','LineWidth',1.5);
plot(P(2,:),'g','LineWidth',1.5);
plot(P(3,:),'r','LineWidth',1.5);
plot(P(4,:),'b','LineWidth',1.5);
plot(svect*psize,'--k');
hold off;

% analysis of the simulation. Here, we extract the population dynamics
% following each trial of ligand switching (CRIPT to T-2F), isolate the
% events in which the double mutant goes to near fixation, determine the
% inteval over which the intermediate genotypes should be intergrated, and
% compute the fractional flux through the G330T state.

% pre-allocate variables
events_post=zeros(4,tau,numel(gen_switch)); % population dymamics in each trial
lim=zeros(1,numel(gen_switch)); % limit for integration in each trial
Npost=zeros(2,numel(gen_switch)); % integrated counts of G330T and H372A per trial
frac_G330T=zeros(1,numel(gen_switch)); % fraction of G330T per trial
count_double=0;

for k=1:numel(gen_switch)
    events_post(:,:,k)=P(:,gen_switch(k):gen_switch(k)+(tau-1));
    if max(events_post(4,:,k))>800 % minimal double mutant level to consider trial (arbitrary)
        count_double=count_double+1;
        ydat=smooth(diff(events_post(4,:,k)),20);
        xdat=[1:numel(ydat)];
        try
            pd=fit(xdat',diff(events_post(4,:,k))','gauss1'); % Guassian fit of double mut population
            lim(k)=pd.b1+(2*pd.c1); % limit for integrating G330T, H372A populations
            if lim(k)<tau
                Npost(1,k)=sum(events_post(2,1:floor(lim(k)),k),2); %number G330T
                Npost(2,k)=sum(events_post(3,1:floor(lim(k)),k),2); %number H372A
                frac_G330T(k)=Npost(1,k)./(Npost(1,k)+Npost(2,k)); % the fraction G330T
            else
                frac_G330T(k)=-Inf; % if limit is not less than tau
            end
        catch
            frac_G330T(k)=-Inf; % if fitting throws an exception
        end
    end

end

% clean up events for exceptions
ind=find(~isinf(frac_G330T) & ~isnan(frac_G330T) & frac_G330T~=0 & frac_G330T~=-Inf); % the indices of valid
    trials
frac_G330T_clean=frac_G330T(ind); % clean frac_G330T for valid switching trials
lim_clean=lim(ind); % clean lim for valid switching trials
mean_fracG330T=mean(frac_G330T_clean); % mean fraction G330T over simulation
sem_fracG330T=std(frac_G330T_clean)/sqrt(numel(ind)); % standard error of the mean, since we want confidence in
    the mean value, and not the scatter over trials.

% plotting
h_analysis=figure;clf;
plot(frac_G330T_clean,'ok-','LineWidth',1.5);hold on;grid on
plot([0 numel(gen_switch)],[mean_fracG330T mean_fracG330T],'--or','LineWidth',1);
if ~isempty(frac_G330T_clean)
    text(1,(0.05*max(abs(frac_G330T_clean))),['fraction G330T = ' num2str(mean_fracG330T)],'FontWeight','bold','
        FontSize',12);
    axis([0 numel(gen_switch) 0 1]);
end
P_double=count_double./numel(gen_switch);
[mean_fracG330T sem_fracG330T P_double]

% ************************************
% Step 6: Add analysis to output structure
% ************************************

```

```
215  out.analysis.frac_G330T=frac_G330T_clean;
216  out.analysis.index_valid=ind;
217  out.analysis.integration_limits=lim_clean;
218
219  end
```

II. SUPPLEMENTARY TABLES

Unselected Library Statistics

	Number of reads	Number of Ligands (> 50 counts)	% Library Coverage
Total Input Library	1.07 x 10^8	154,521	96.7

Selected Library Statistics

	Number of reads	Number of Ligands	Number of Ligands (> 50 counts)	Number of Ligands Bound by Protein (> 15 μM)
WT	46,598,840	56,640	55,278	185
G330T	51,195,397	90,735	83,056	846
H372A	29,419,042	59,935	43,488	1051
H372A/ G330T	48,989,769	123,295	86,255	1894

TABLE S1 **Sequencing statistics of the ligand library selection experiments**. Related to Figures 2-3. Sequencing statistics from Illumina HiSeq2500 runs for the unselected and selected populations of peptide libraries for wild-type, G330T, H372A, and the double mutant experiments. The unselected populations were combined over all experiments. The total number of reads, (1.07×10^8) represented approximately 97% coverage of all peptides in the library.

	WT-CRIPT	WT-T$_2$F	G330T-Apo	G330T-CRIPT
Data Collection				
PDB ID	5HEB	5HED	5HET	5HEY
Source	UTSW SBL	UTSW SBL	APS 19-ID	APS 19-ID
Wavelength (Å)	1.54178	1.54178	0.97937	0.97937
Resolution range (Å)	40.15–1.65 (1.709–1.65)	31.66–1.7 (1.761–1.7)	36.52–2.001 (2.073–2.001)	27.21–1.5 (1.554–1.5)
Space group	$P4_132$	$P4_132$	$P4_132$	$P4_132$
Unit cell (Å, º)	89.771 89.771 89.771 90 90 90	89.551 89.551 89.551 90 90 90	89.445 89.445 89.445 90 90 90	90.23 90.23 90.23 90 90 90
Total reflections	209981	128220	87993	454547
Unique reflections	15083 (1156)	13972 (1299)	8697 (805)	20265 (1645)
Multiplicity	13.9 (2.5)	9.2 (3.0)	10.0 (8.4)	21.9 (14.7)
Completeness	0.98	0.99	0.99	0.98
Mean I/σ	61.571 (2.478)	60.017 (2.627)	35.095 (2.077)	30.057 (1.448)
Wilson B-factor	14.71	15.35	25.36	14.99
R-merge	0.034 (0.289)	0.031 (0.302)	0.064 (N/A)	0.110 (N/A)
R-meas	0.034 (0.353)	0.033 (0.361)	0.068 (N/A)	0.112 (N/A)
R-pim	0.008 (0.199)	0.010 (0.193)	0.022 (0.397)	0.035 (0.665)
CC1/2	0.87	0.885	0.713	0.555
Refinement				
Reflections used in refinement	15082 (1156)	13971 (1299)	8696 (805)	20265 (1645)
Reflections used for R-free	1509 (116)	1398 (130)	871 (80)	2024 (165)
R-work	0.1666 (0.2273)	0.1701 (0.2291)	0.1837 (0.2309)	0.1724 (0.2705)
R-free	0.1945 (0.2745)	0.1973 (0.2694)	0.2130 (0.2586)	0.2026 (0.2765)
Number of non-hydrogen atoms	1220	1262	1054	1268
Macromolecules	1086	1110	956	1121
Ligands	12	N/A	N/A	N/A
Protein residues	126	126	118	127
RMS (bonds)	0.007	0.014	0.006	0.011
RMS (angles)	1.22	1.46	0.83	1.84
Ramachandran favored (%)	97	99	98	95
Ramachandran allowed (%)	2.1	0.72	1.6	4.2
Ramachandran outliers (%)	0.71	0	0	0.69
Rotamer outliers (%)	2.6	3.4	2.9	3.3
Clashscore	3.19	7.29	1.06	4.02
Average B-factor	19.78	20.48	32.04	22.33
Macromolecules	18.51	19.21	31.36	21.16
Ligands	62.39	N/A	N/A	N/A
Solvent	26.84	29.79	38.7	31.27
Number of TLS groups	3	3	1	16

TABLE S2 **Crystallographic data collection and refinement statistics - Part 1**. Related to Figure 5. All proteins were crystallized in the same space group ($P4_132$) and showed unit cell constants within 0.5% of each other. Ligand-bound datasets were collected to a resolution higher than 2.0 Å.

	G330T-T$_2$F	H372A-CRIPT	H372A-T$_2$F	G330T-H372A-T$_2$F
Data Collection				
PDB ID	5HF1	5HFB	5HFC	5HFF
Source	APS 19-ID	UTSW SBL	UTSW SBL	APS 19-ID
Wavelength (Å)	0.97918	1.54178	1.54178	0.97918
Resolution range (Å)	40.13–1.747 (1.81–1.747)	40.07-1.617 (1.675-1.617)	36.52–1.851 (1.918–1.851)	36.57–1.749 (1.812–1.749)
Space group	$P4_132$	$P4_132$	$P4_132$	$P4_132$
Unit cell (Å, º)	89.724 89.724 89.724 90 90 90	89.592 89.592 89.592 90 90 90	89.445 89.445 89.445 90 90 90	89.575 89.575 89.575 90 90 90
Total reflections	358023	203441	68030	264373
Unique reflections	13007 (1217)	15108 (556)	10759 (914)	12879 (1179)
Multiplicity	27.3 (26.3)	13.3 (1.6)	6.3 (2.5)	20.3 (20.1)
Completeness	0.99	0.93	0.98	0.99
Mean I/σ	67.804 (1.844)	65.515 (0.936)	35.543 (1.000)	50.222 (1.826)
Wilson B-factor	19.04	15.74	21.65	18.64
R-merge	0.055 (N/A)	0.035 (0.338)	0.040 (0.448)	0.063 (N/A)
R-meas	0.056 (N/A)	0.036 (0.470)	0.044 (0.547)	0.065 (N/A)
R-pim	0.013 (0.581)	0.009 (0.325)	0.017 (0.304)	0.016 (0.478)
CC1/2	0.721	0.695	0.689	0.706
Refinement				
Reflections used in refinement	13006 (1217)	15109 (556)	10759 (914)	12880 (1179)
Reflections used for R-free	1301 (122)	1513 (56)	1078 (92)	1285 (114)
R-work	0.1993 (0.2911)	0.1751 (0.3004)	0.1835 (0.2894)	0.1740 (0.2314)
R-free	0.2179 (0.3097)	0.2033 (0.3281)	0.2051 (0.3146)	0.2233 (0.2761)
Number of non-hydrogen atoms	1182	1204	1161	1191
Macromolecules	1080	1055	1037	1080
Ligands	N/A	N/A	N/A	12
Protein residues	124	123	124	128
RMS (bonds)	0.006	0.006	0.012	0.011
RMS (angles)	0.94	0.95	1.46	1.27
Ramachandran favored (%)	99	99	95	97
Ramachandran allowed (%)	0.73	0.73	4.5	0.71
Ramachandran outliers (%)	0	0	0.76	2.1
Rotamer outliers (%)	4.4	13.3 (1.6)	2.7	6.2
Clashscore	4.63	1.91	10.18	4.16
Average B-factor	28.71	22.31	30.76	25.02
Macromolecules	27.95	20.99	30.12	24.07
Ligands	N/A	N/A	N/A	57.25
Solvent	36.72	31.72	36.13	31.53
Number of TLS groups	5	5	8	7

TABLE S3 **Crystallographic data collection and refinement statistics - Part 2**. Related to Figure 5. All proteins were crystallized in the same space group ($P4_132$) and showed unit cell constants within 0.5% of each other. Ligand-bound datasets were collected to a resolution higher than 2.0 Å.

Protein	[sodium citrate] (M)	crystallization buffer pH	[protein] (mg/ml)
WTapo	1.0	6.9	9
WT,CRIPT	1.0	7.0	9
WT,T-2F	1.125	7.1	9
G330Tapo	1.05	7.0	7
G330T,CRIPT	1.2	7.0	8
G330T,T-2F	1.2	6.8	9
H372Aapo	0.95	7.0	9
H372A,CRIPT	1.05	7.0	7
H372A,T-2F	1.05	7.0	7
G330T,H372Aapo	1.0	7.0	13
G330T,H372A,CRIPT	1.25	7.0	9
G330T,H372A,T-2F	1.2	6.75	7

TABLE S4 **Crystallization conditions for PSD95^{pdz3} variants**. Related to Figure 5. Details of protein expression, purification, and general aspects of crystallization are given in the methods section.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. and Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr *66*, 213–21.

DeLano, W. L. (2002). The PyMol molecular graphics system.

Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr D Biol Crystallogr *66*, 486–501.

McLaughlin, Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–42.

Otwinowski, Z. and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. In Methods in Enzymology, (Carter, Jr, C. W. and Sweet, R. M., eds), vol. 276, pp. 307–326. Academic Press New York macromolecular crystallography, part A edition.

Studier, F. W. (2005). Protein production by auto-induction in high density shaking cultures. Protein Expr Purif *41*, 207–34.