A complete map of specificity encoding for a partially fuzzy protein interaction

Taraneh Zarin[1] and Ben Lehner[1,2,3,4]

[1]Centre for Genomic Regulation (CRG), Barcelona Institute for Science and Technology (BIST), Barcelona, Spain
[2]Wellcome Sanger Institute, Cambridge, UK
[3]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Abstract

Thousands of human proteins function by binding short linear motifs embedded in intrinsically disordered regions. How affinity and specificity are encoded in these binding domains and the motifs themselves is not well understood. The evolvability of binding specificity - how rapidly and extensively it can change upon mutation - is also largely unexplored, as is the contribution of 'fuzzy' dynamic residues to affinity and specificity in protein-protein interactions. Here we report the first complete map of specificity encoding for a globular protein domain. Quantifying >200,000 energetic interactions between a PDZ domain and its ligand identifies 20 major energetically coupled pairs of sites that control specificity. These are organized into six modules, with most mutations in each module reprogramming specificity for a single position in the ligand. Nine of the major energetic couplings controlling specificity are between structural contacts and 11 have an allosteric mechanism of action. The dynamic tail of the ligand is more robust to mutation than the structured residues but contributes additively to binding affinity and communicates with structured residues to enable changes in specificity. Our results quantify the binding specificities of >1,800 globular proteins to reveal how specificity is encoded and provide a direct comparison of the encoding of affinity and specificity in structured and dynamic molecular recognition.

Introduction

Specific physical interactions between proteins underlie nearly all aspects of biology from transcription and signaling to mechanics and neuronal information processing[1]. Many protein-protein interactions are mediated by intrinsically disordered regions (IDRs) of proteins which do not form stable secondary and/or tertiary structures[2,3]. While some IDRs adopt conditional structures upon binding to their partners[4–6], others remain disordered upon binding, forming so-called 'fuzzy' complexes[7,8]. The ensemble of conformations afforded by fuzzy binding has been shown to facilitate diverse regulatory functions in eukaryotes [7,9–11] but how dynamic binding and conformational heterogeneity contribute to binding affinity and/or specificity is not well understood[8,12–15].

One of the most frequent modes of protein-protein interaction in human cells is the binding of a globular domain to a short linear motif (SLiM) embedded within an IDR[3,16,17]. These peptide recognition domains typically exist in large protein families and have affinity for many potential binding targets[18,19]. How binding specificity is encoded in these domains and how these large protein families evolve without causing interaction cross-talk is a major area of interest[20] and is crucial for understanding human disease and drug development[21,22]. PDZ (postsynaptic density 95, PSD-95; discs large, Dlg; zonula occludens-1, ZO-1) domains, for example, are the largest family of human protein interaction domains with more than 270 PDZ domains in 155 human proteins[19]. These domains typically bind to IDRs at the C-termini of proteins[18,19,21]. Upon binding, the last four amino acids of a PDZ ligand adopt a well-defined structure[23] and have historically been used to classify PDZ domains into distinct groups based on the identity of the P0 (C-terminal) and P-2 positions[24]. The adjacent residues of the IDR however typically remain dynamic in the complex[25], with X-ray structures revealing progressively increasing motion in the bound complex moving away from the C-terminus (Fig. 1a). While these adjacent N-terminal residues have not been nearly as well-studied, there is evidence that they are also important for binding[21,25,26], consistent with an emerging theme that the context around SLiM consensus sites is important for function[27]. The PDZ domain-peptide interaction thus presents an elegant model system in which to understand not only how affinity and specificity are encoded, but also how this encoding is distributed between structured versus dynamic binding modes.

Here we present the first complete map of how binding affinity and specificity are encoded in a globular protein domain interacting with a disordered peptide. Our map reveals that specificity encoding in the domain is highly modular, with distinct residues determining specificity for each position in the ligand by both direct and allosteric mechanisms. The more dynamic, 'fuzzy', tail of the peptide is more robust to mutation but can be used to additively tune affinity. These more dynamic residues make only small contributions to specificity through interactions with the domain, but they do contribute to specificity via interactions with the structured part of the peptide.

Results

Quantifying a combinatorial genetic landscape for peptide recognition

The third PDZ domain (PDZ3) from PSD-95 binds to the C-termini of proteins matching the consensus motif -X-S/T-X-Φ-COOH (where X is any amino acid, aa, and Φ is a hydrophobic residue)[22,28,29]. The bound peptide is structured at the C-terminus but increasingly dynamic before the last four residues (Fig. 1b), a general property of PDZ domains bound to short ligands (Fig. 1a). To better define the binding specificity of PDZ3 for the canonical structured C-terminal part of the ligand, we quantified its binding to >100,000 variants of a 9 aa peptide from the CRIPT protein in which the sequence of each of the last four aa (positions 0 to -3) was fully randomized to any of the 20 aa (Fig. 1c-f). Binding was quantified using a highly-validated protein fragment complementation assay (PCA)[30] (Fig. S1b) and binding scores were very reproducible between replicate experiments (Fig. S2b-c, e, h-i). The position-weight matrix for the top 1% of binders matched the reported consensus[22,28,29] from a previous selection experiment[22] (Fig. 1g). The distribution of binding fitness (Fig. 1f) and proportion of peptides binding to PDZ3 (Fig. 1h) sharply decreased with an increasing number of substitutions. However, because of the exponential increase in the size of sequence space when combining mutations (Fig. 1h), there are actually more peptides with 4 aa changes that bind PDZ3 than peptides with one mutation (179 vs 73; 4.8-fold more, Fig. 1h). Position-weight matrices for peptides containing between one and four mutations show that the preference for T/S at the -2 position and hydrophobic residues (including cysteine) at the 0 position do not change with increasing mutation order (Fig. 1g-h).

Additive energy models accurately predict peptide binding

Precisely quantifying the binding of PDZ3 to >100,000 peptides provides an opportunity to evaluate the extent to which binding to each of the four C-terminal residues is independent of sequence variation at the other three sites. We used MoCHI[31] to fit a two-state thermodynamic model to our binding data. The model accounts for the non-linear relationship between the Gibbs free energy of binding (dG) and the fraction of ligand bound to PDZ3 but otherwise assumes that the energetic effects of mutations (ddG) combine additively with no pairwise or higher order energetic couplings between mutations at different sites (Fig. 2a and Methods). Fitted to a balanced set of >7,000 binding and non-binding genotypes, the model provides very good predictive performance (R^2=0.75 evaluated by 10-fold cross validation, Fig. 2b-iii). Not accounting for the non-linear relationship between the fraction of ligand bound and the binding energy resulted in worse predictive performance (R^2=0.56, Fig. S3c) and systematically biased predictions (Fig. S3c residuals). Moreover, allowing energetic couplings between mutations in different sites did not improve predictive performance (R^2=0.5 by 10-fold cross-validation, Fig. S3c). The additive energy model represents a very large compression of the binding data (>1000-fold, 110,000 genotypes/76 model coefficients) and formally shows that the effects of

mutations in the C-terminus of the ligand have largely energetically independent effects on binding.

Dynamic residues are important for binding affinity

We next focused on the N-terminal residues in the CRIPT peptide (position -4 to -7) which display a progressive increase in dynamicity in the PDZ3-CRIPT complex (Fig. 1b). To test the importance of these more dynamic residues to binding, we quantified the binding of >90,000 variants of CRIPT in which the four aa before the consensus binding motif were randomized to all other aa (positions -4 to -7) (Fig. 1c-f). Binding measurements were well correlated across independent experiments as with the C terminus (Fig. S2a, d, f-g). The distribution of binding scores was, however, very different to that of the C-terminus library (Fig. 1d), with a more gradual decrease in binding with an increasing number of substitutions (Fig. 1f). Thus, although mutations in these four aa typically have smaller effects on PDZ3 binding than mutations in the C-terminal four aa, the region still makes an important contribution to binding affinity, with a clear preference for positively charged and aromatic residues at the -4 and -5 positions (Fig. 1g), and the combined effects of multiple substitutions frequently being very detrimental (Fig. 1f, h).

Fitting an additive thermodynamic model to the data, we again found that mutations have largely independent energetic effects (R^2=0.73 by 10-fold cross-validation, Fig. 2b-i). Ignoring the non-linear relationship between free energy (dG) and binding again resulted in biased predictions (Fig. S3c). Allowing energetic couplings between mutations provided almost the same predictive performance as the additive model (R^2=0.76 Fig. S3c). Therefore, as for mutations in the C-terminus, the effects of mutations in this more dynamic region of the ligand are largely energetically independent.

The mutational energy matrix for PDZ3 binding

The combined energy matrices for positions 0 to -3 and -4 to -7 provide a complete description of the energetic effects of substitutions in all eight positions of the PDZ3 ligand (Fig. 2c, Fig. S3d-e). Substitutions in positions 0 and -2 are most detrimental for binding, consistent with the description of the consensus motif as X-S/T-X-Φ-COOH[22,24,28]. However, not all substitutions in positions 0 and -2 have the same energetic effects and many changes at positions -1 and -3 also cause large changes in binding energy. Moreover, mutations at position -4, which is outside of the canonical motif, have the third largest energetic effects and multiple substitutions in positions -5 and -7 cause detrimental (ddG>0) energy changes (Fig. 2c).

The energy matrix also reveals that CRIPT harbors sites where many mutations are energetically favorable (ddG<0), with a concentration in all of the first five positions (Fig. 2c). For example, E is energetically favored at position -3; W, F, R and H are all favored at position -5; K and R favored at -6; and R is favored at -7 (Fig. 2c). Indeed, in the first four positions but not in the C-terminal four positions, the net charge of the peptide is a strong predictor of binding strength (Spearman's rho=0.57 for N, -0.04 for C, Fig. 1i, Fig. S2j), with positively charged

peptides having higher binding scores. The ability of many substitutions in the N-terminal region of the peptide to increase affinity further emphasizes the importance of this region for binding, with the longer range of electrostatic interactions[32] compared to other non-covalent interactions consistent with a dynamic or 'fuzzy' contribution[8] to affinity.

Quantifying >200,000 energetic interactions between a PDZ domain and its ligand

We next designed an experiment to understand how the binding specificity of PDZ3 is encoded in the globular protein domain (Fig. 3a-b, Fig. S4). We performed a comprehensive set of double mutant cycles, mutating every position in PDZ3 and CRIPT to every other aa alone and in *trans*-double mutant combinations and measured binding (Fig. 3c). The binding measurements were highly correlated with measurements from the combinatorial CRIPT mutagenesis libraries (R^2=0.937 and 0.875 for N and C, n=36 and n=68, respectively, Fig. S2f-i), allowing us to normalize binding scores across independent experiments (Methods). Consistent with previous data[33,34], mutations in PDZ3 detrimental for binding are strongly enriched in the binding interface with CRIPT (Fig. 3d-e).

In total this dataset precisely defines the binding specificity of >1,800 globular proteins: we quantified the binding specificity of nearly every mutation in PDZ3 for nearly every mutation at every position in the ligand (Fig. S4f). Formally, changes in specificity are identified when there is an energetic coupling (or genetic interaction) between a mutation in the PDZ domain and a mutation in the ligand i.e. when binding is not well predicted by an additive energy model (Fig. 3f). The binding scores in this much larger experiment were also extremely well correlated across three replicate selections (Pearson's r>0.91, Fig. S4d), as were the inferred free energy changes (ddGs, Fig. S4g), allowing us to quantify >200,000 energetic couplings between mutations in PDZ3 and its ligand.

Hundreds of mutations in a PDZ domain change its binding specificity

In total, we identified over 600 non-zero energetic couplings between mutations in PDZ3 and mutations in the ligand (Z-test, Benjamini-Hochberg FDR<0.1, Fig. S5a). We define a specificity-changing mutation as one where the observed binding has a positive and significant (Z-test, FDR<0.1) residual to our additive energy model (Fig. 3f) and the observed binding score passes our threshold for non-binding variants (i.e. is binding) (Fig. S5b-c). These non-additive energetic interactions involve 340 distinct mutations in 73 different positions in the PDZ domain, and 114 distinct mutations in all 8 residues of the ligand (Fig. 3g). Mutations in PDZ3 can therefore alter its specificity for all 8 positions within the ligand.

Major specificity encoding residues and the modular encoding of binding specificity

We next identified the residues in PDZ3 most important for encoding specificity for each ligand position. In total, 20 pairs of PDZ3-ligand residues are enriched for specificity-changing mutations (hypergeometric test, FDR<0.1, Fig. 4a). We refer to the PDZ3 residues within these pairs as major specificity encoding residues. There are two major specificity-encoding residues for ligand position 0, six for position -1, five for position -2, three for position -3, three for position -4, and one for position -5 (Fig. 4a, Fig. S5f).

Strikingly, the major specificity encoding residues are largely distinct for each ligand position (Fig. 4a, Fig. S5e). There are 17 unique PDZ positions across the 20 pairs of positions. 14/17 PDZ positions only act as a major specificity determinant for a single ligand position, with 3 acting as major specificity determining sites for two ligand positions (Fig. S5e). The encoding of specificity in the PDZ domain is thus highly modular, with specificity for each ligand residue largely encoded by a distinct subset of residues in the domain.

The 3 exceptions are positions N326 and S339, which are both major specificity-encoding residues for positions -1 and -3, and G330, which defines the specificity for positions -2 and -4. N326 and S339 are contacts of each other and of -3 (N326 contacts -1 as well) (Fig. S5h). Similarly, G330 contacts H372 (Fig. S5h), which is a contact of position -2. G330 is also adjacent to E331, which is a contact of -4 (Fig. S5g). These positions therefore form small networks of specificity-defining residues, with pairs of residues in PDZ3 interacting with pairs of residues in the ligand.

Direct and allosteric reprogramming of specificity

Visualizing the 20 energetically coupled pairs of residues on the structure of the PDZ3-CRIPT complex (Fig. 4b-j), shows that not only are the major specificity encoding residues distinct for each ligand position, but they are also spatially clustered. Indeed 11/17 of the major specificity encoding residues constitute the PDZ3 binding interface (Fig. 4a) and out of the 20 major energetically coupled pairs of residues, 9 are directly contacting each other (<5Å apart and/or predicted to be specifically contacting each other, Fig. 4c-j, Fig. S5g), suggesting local energetic coupling as the mechanism of action.

A further 11 pairs of major specificity encoding residues are not structural contacts. These residues must therefore encode specificity allosterically. Most of these allosteric energetic couplings are, however, local involving PDZ3 residues at the binding interface that contact other residues in the ligand (Fig. 4c-j). Others are the contacts of PDZ3 residues that contact the ligand (Fig. S5g-h).

In summary, specificity is encoded both modularly and locally within the PDZ domain. Mutations in a discrete set of spatially clustered sites reprogram specificity for each ligand residue.

A comprehensive map of specificity-changing mutations

For each pair of positions enriched in specificity-changing mutations, we were interested to see if and how the coupled mutations in the domain and ligand are related to each other (Fig. 5a, Fig. S6), as well as how energetic couplings are related to changes in binding (compared to the wildtype preference) and raw binding scores (Fig. 5a). The majority of mutations that we identify in major specificity encoding residues change specificity such that the mutated ligand residue is preferred over wildtype, similar to 'class-switching' phenotypes[22,33] (Fig. 5a, middle panel). Others bind the mutated and wildtype ligand with similar energy, acting similarly to 'class-bridging' specificity changes[22,33].

Interestingly, the diversity of specificity-changes that could be accommodated at each position in the ligand seems loosely related to the dynamicity of the ligand position, except for position 0 (Fig. 5b). The most diverse specificity changes are carried out through position -2, followed closely by -1 and -3 (Fig. 5b). This is consistent with previous reports that found specificity "modulators" of PDZ domains at non-canonical motif positions in the ligand[35]. It is also consistent with a study in which all positions common to the binding site across PDZ domains (N=10 sites) were mutated and found that only one common binding residue enacts a specificity change through V0[21], suggesting that this is a general effect across PDZ domains.

Overall, we find a large diversity of mutational couplings. While some positions are clearly dominated by charge-charge interactions (Fig. 5a, position -3), we hypothesize others are mediated by a gain/loss of sidechain interactions and/or increased movement in the binding pocket (Fig. 5a). We highlight some of the coupled sites and mutations below for each of the six modules defining specificity for each ligand residue.

Module 0

Mutations in PDZ3 that change specificity for position 0 are quite distributed in the domain with relatively weak effects (Fig. 5a). For example, L379 mutated to V/I changes specificity to V0F, whereas mutating L379 to polar residues G/S/T changes specificity to V0L. More physicochemically diverse changes to position 0 specificity can occur through A343 mutations to M/W, both of which enable a preference for V0D (but interestingly, not V0E). These changes increase the binding fitness of V0D from quite detrimental in the wildtype PDZ3 context to being preferred in the mutant PDZ3 (Fig. 5a). L379 is in the binding interface, close (<5Å) in the structure to V0 and contacts K380, a contact of neighboring ligand residue -1. However, A343 is far in the structure and must have a more indirect allosteric mechanism (Fig. 5a). Within the full map of PDZ3 energetic couplings, there is also a small cluster of mutations that enables a specificity change to V0K: E305F, E310R, G333P, and F325T (Fig. S7).

Module -1

The determinants for specificity to position -1 are comprised of N326 (a contact of S-1), as well as S339 (a contact of N326), F325 (a contact of V0), and farther (>5Å) away in the structure are

G324, G345, and L342. These positions are mainly concentrated in the B2 and B3 beta-sheets and are in the binding interface (Fig. 4a). Many mutations in N326 cause the specificity-change to W and/or F/I/L (Fig. 5a). The broadest effect (W/M/I/L/V) is seen for N326P and N326G, both amino acids with special conformations. Interestingly, negatively charged (D/E) and polar mutations (G/S/T) at N326 change the -1 specificity to positively charged residues (R/K).

Similarly to N326, G324 mutations to D/E change specificity in -1 to R/K (but also W/H), and G324R/M/L change specificity to L. G345 is another case where mutations to aromatic or hydrophobic residues (W/M/L/C) and Q cause specificity changes (to F/W/L at -1). A similar pattern is true for L342, where changes to hydrophobic/aromatic residues favor changes in specificity to hydrophobic residues at position -1, and there is increased binding to positively charged residues (H/R/K). Similarly, F325L and M greatly increase preference to bind H and I, respectively.

Module -2

Position -2 binding specificity is encoded by H372 (a contact of T-2) and A376 (a contact of H372 and K380, contacts of T-2 and V0, respectively) (Fig. 5, Fig. S5g-h). It is also encoded by G329 (a contact of K-4 and H372), G330, and I336, all in the B2 and B3 beta-strands that contain residues directly involved in binding. 4/5 of these sites overlap with a previous study that found these positions as epistatic in PDZ3[33] and 3/4 overlap with those previously identified to strongly change specificity at -2 to F[33].

The PDZ3 residue most strongly energetically coupled to position -2 is H372 (FDR<0.001, hypergeometric test). Indeed, PDZ3 position 372 and ligand position -2 are the most strongly coupled pair of positions in the dataset (table S1). This directly contacting position pair is connected by more than 70 energetic couplings. H372A was previously found to be a “class-switching” mutation[22,33]. changing the preference at the T-2 position to an aromatic residue (F). As noted in a previous study, it is not only H372A that achieves this change in specificity, but many other mutations as well[22,33]. We gain resolution into the -2 preference and find that it can also be broadened to include other aromatic residues (W/Y), as well as hydrophobic and positively charged residues (Fig. 5a). For example, H372V/I/L/G/P/A change the position -2 specificity to aromatic, hydrophobic, and positively charged residues. H372C/T have more specific preference to aromatic or hydrophobic (but not positively charged) residues, and H372D shifts specificity to aromatic residues only. Interestingly, a fourth class of specificity change (T-2D, to an acidic residue) can be achieved with a A376C/G mutation.

In terms of residues that are not direct contacts, hydrophobic and aromatic mutations in G329 change -2 specificity to hydrophobic residues (M/I/L). Similarly, mutations in the adjacent residue, G330, to V or A change specificity to I, and G330W changes specificity to I/C/V/F. Consistent with previous reports that G330T is a class-bridging mutation[22], we find that it clusters (along with G330S and G330W) with class-switching mutations in H372 when considering the complete map of PDZ domains with at least one significant specificity encoding mutation (Fig. S7).

Module -3

The main specificity-encoding positions for -3 are N326 (a contact of Q-3 and S-1), S339 (a contact of Q-3), and F340 (a contact of N326 and S339), meaning that the network of specificity-changing mutations around position -3 is fully connected by hydrogen bonds (Fig. S5g-h). N326 mutations to negatively charged residues (D/E) change -3 specificity to positively charged residues (R/K) and there is a similar pattern with S339, where mutations to D/E (or Q, and more weakly C) change preference at -3 to R/K. The same strong charge-coupling pattern can be seen for F340, where mutations to D/E create a preference for Q-3K/R.

Module -4

Specificity changes for position -4 occur via mutations at E331 and E373, both negatively charged contacts of K-4, located in the B2-B3 loop and the A2 helix, respectively, as well as G330 (a shared specificity-determinant of position -2). E331W/Y changes specificity of -4 to Y/F/W, as does G330W. In contrast, E331N changes specificity of -4 to the negatively charged residues D/E. Similarly, the polar mutation E373Q enables a -4E specificity change. The interactions between position -4 and E331 are interesting, as these residues were found to form transient salt bridges in molecular dynamics simulations[25] as well as through our analysis of contacts from the crystal structure (Fig. S5g).

Module -5

Position -5 has a single major specificity determinant: V328. Once again, we find interesting charge-based couplings: V328E (negatively charged) enables binding to positively charged R/H, and V328R (positively charged) enables binding negatively charged D/E. Though V328 does not interact with Y-5 according to the crystal structure, V328 does contact PDZ positions G329 and I336, both of which encode specificity for T-2, indicating an allosteric mechanism of action.

Specificity encoding through a disordered tail

Our energetic coupling analysis between the domain and peptide yielded a limited role for the dynamic tail of CRIPT in defining specificity via energetic interactions with the globular domain. We were thus motivated to construct another library of variants in which we made all possible double mutants across the entire peptide (Fig. 6a), yielding more than 8,000 energetic coupling measurements across the 8 aa of the peptide (Fig. 6b-c). Once again, we found excellent reproducibility across replicate experiments and libraries (Fig. S8a-e). We used the same modelling approach (Fig. 6d, Fig. S8f-g) and quantification of residuals as above to find pairs of sites in CRIPT that are enriched in energetic couplings (Fig. 6e). This identified a significant coupling (FDR<0.1, hypergeometric test) between the structured and dynamic parts of the peptide, between the -3 and -5 positions (Fig. 6e). This pair of sites contains more than 20 significant (FDR<0.1, Z-test) specificity-changing mutations (Fig. 6f), most of which are linked to a change in position -3 to a G residue. Substitution to a G at -3 enables a specificity change to any residue that is not positively charged or aromatic at the -5 position (Fig. 6f). We hypothesize

that the mechanism for this interaction could be via the gain/loss of (dynamic) sidechain-sidechain interactions, as Q-3 and Y-5 sidechains are oriented toward the same side in the ligand and glycine would remove the sidechain and add flexibility to the chain. Another potential mechanism could be through the Y-5 interaction with V328 as an intermediary, as this position also has Van der Waal's contacts with Q-3 (Fig. S5g).

Interestingly, quantifying second order interactions for residues in the dynamic N-terminus in our combinatorial library also identified high confidence (95% CI < 1 kcal/mol) dddGs involving mutations at Y-5, with both being examples of positive sign epistasis (Fig. S6). Y-5G and K-7W/Y are both weakly detrimental as single mutations, but together they are favored energetically (Fig. S6d). Y-5 E/D (and to a lesser extent G/N/S) are coupled to K-4G (Fig. S6f), again individually being detrimental but energetically favored when combined. These results suggest that not only is Y-5 coupled to Q-3, as seen above, but it also communicates with K-7 and K-4. An energetic coupling network thus links residues as distant as K-7 to the structured region of the ligand at the binding site, and in turn to the domain itself.

Discussion

Here we have constructed the first comprehensive map of how binding specificity is encoded in a globular protein domain. The map required the binding specificities of >1,800 proteins to be experimentally measured and provides a comprehensive and interpretable energy model describing how mutations throughout a protein domain reprogram its binding specificity.

Quantifying >200,000 energetic couplings between mutations in a PDZ domain and mutations in its peptide ligand allowed us to identify hundreds of mutations that alter the binding specificity of the domain. These mutations are concentrated in 20 major energetically coupled pairs of sites that are the primary determinants of binding specificity. Despite its distributed nature, specificity encoding is highly modular, with different residues in the PDZ domain largely encoding specificity for each residue in the ligand.

The largely independent encoding of specificity in six different modules for each of six different ligand sites means that specificity can be orthogonally re-programmed for each peptide site. The specificity for each peptide residue is encoded in a spatially clustered set of residues and by a mixture of direct and allosteric mechanisms of action.

Our results also show that N-terminal ligand residues that remain dynamic upon binding to the PDZ domain make an important contribution to affinity. However, mutations in the PDZ domain do not alter the effects of mutations in these dynamic residues, suggesting that they contribute little to the specificity of binding through the domain, at least in the immediate sequence space. The affinity contribution from these residues appears to be mostly via electrostatic interactions. It is possible therefore that multiple charge changes on the PDZ domain may be required to alter the specificity of recognition of these sites.

While dynamic residues do not enable widespread changes in specificity through the domain, we do find that they can do so by communicating with structured residues in the ligand. We also find points of communication between the dynamic residues that could propagate to the binding site. To our knowledge, our analysis is the first comprehensive quantification of energetic couplings between dynamic and structured portions of a disordered peptide and between both regions and a protein to which they bind (Fig. 7).

The rapid and high-throughput quantification of energetic couplings makes it possible to gain a comprehensive and mechanistic view of how specificity and affinity are defined in protein interaction domains and peptides. The approach can be applied to both structured and dynamic regions of proteins, allowing it to be used on the wide spectrum of protein forms that includes varying levels of protein disorder[36].

Looking forward, we propose that applying this approach to many different protein-peptide and protein-protein interactions will generate a dataset of sufficient size and diversity to train machine learning models to predict, understand and engineer specificity changes from sequence for all protein-protein interactions.

References

1. Diss, G. Towards attaining a quantitative and mechanistic model of a cell. *Nat. Rev. Mol. Cell Biol.* **21**, 301–302 (2020).
2. Van Der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
3. Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A Million peptide motifs for the molecular biologist. *Mol. Cell* **55**, 161–169 (2014).
4. Katuwawala, A., Peng, Z., Yang, J. & Kurgan, L. Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions. *Comput. Struct. Biotechnol. J.* **17**, 454–462 (2019).
5. Wright, P. E. & Dyson, H. J. Linking folding and binding. *Curr. Opin. Struct. Biol.* **19**, 31–38 (2009).
6. Alderson, T. R., Pritišanac, I., Kolarić, Đ., Moses, A. M. & Forman-Kay, J. D. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc. Natl. Acad. Sci.* **120**, e2304302120 (2023).
7. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).
8. Fuxreiter, M. Fuzzy protein theory for disordered proteins. *Biochem. Soc. Trans.* **48**, 2557–2564 (2020).
9. Mittag, T. *et al.* Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci.* **105**, 17772–17777 (2008).
10. Tuttle, L. M. *et al.* Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex. *Cell Rep.* **22**, 3251–3264 (2018).
11. Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61–66 (2018).

12. Skriver, K., Theisen, F. F. & Kragelund, B. B. Conformational entropy in molecular recognition of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **83**, 102697 (2023).
13. Teilum, K., Olsen, J. G. & Kragelund, B. B. On the specificity of protein-protein interactions in the context of disorder. *Biochem. J.* **478**, 2035–2050 (2021).
14. Brodsky, S. *et al.* Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol. Cell* 1–13 (2020) doi:10.1016/j.molcel.2020.05.032.
15. Baughman, H. E. R. *et al.* An intrinsically disordered transcription activation domain increases the DNA binding affinity and reduces the specificity of NFκB p50/RelA. *J. Biol. Chem.* **298**, 102349 (2022).
16. Nguyen Ba, A. N. *et al.* Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Sci. Signal.* **5**, rs1 (2012).
17. Davey, N. E. *et al.* Attributes of short linear motifs. *Mol BioSyst* **8**, 268–281 (2012).
18. Nourry, C., Grant, S. G. N. & Borg, J.-P. PDZ Domain Proteins: Plug and Play!
19. Amacher, J. F., Brooks, L., Hampton, T. H. & Madden, D. R. Specificity in PDZ-peptide interaction networks: Computational analysis and review. *J. Struct. Biol. X* **4**, 100022 (2020).
20. Laub, M. T. & Goulian, M. Specificity in Two-Component Signal Transduction Pathways. *Annu. Rev. Genet.* **41**, 121–145 (2007).
21. Tonikian, R. *et al.* A Specificity Map for the PDZ Domain Family. *PLoS Biol.* **6**, e239 (2008).
22. Raman, A. S., White, K. I. & Ranganathan, R. Origins of Allostery and Evolvability in Proteins: A Case Study. *Cell* **166**, 468–480 (2016).
23. Doyle, D. A. *et al.* Crystal structures of a complexed and peptide-free membrane protein-binding domain: Molecular basis of peptide recognition by PDZ. *Cell* **85**, 1067–1076 (1996).
24. Songyang, Z. *et al.* Recognition of Unique Carboxyl-Terminal Motifs by Distinct PDZ Domains. *Science* **275**, 73–77 (1997).
25. Mostarda, S., Gfeller, D. & Rao, F. Beyond the binding site: The role of the β2 - β3 loop and extra-domain structures in PDZ domains. *PLoS Comput. Biol.* **8**, (2012).
26. Saro, D. *et al.* A thermodynamic ligand binding study of the third PDZ domain (PDZ3) from the mammalian neuronal protein PSD-95. *Biochemistry* **46**, 6340–6352 (2007).
27. Ivarsson, Y. & Jemth, P. Affinity and specificity of motif-based protein–protein interactions. *Curr. Opin. Struct. Biol.* **54**, 26–33 (2019).
28. Niethammer, M. *et al.* CRIPT, a Novel Postsynaptic Protein that Binds to the Third PDZ Domain of PSD-95/SAP90. *Neuron* **20**, 693–707 (1998).
29. Skelton, N. J. *et al.* Origins of PDZ Domain Ligand Specificity. *J. Biol. Chem.* **278**, 7645–7654 (2003).
30. Tarassov, K. *et al.* An in vivo map of the yeast protein interactome. *Science* **320**, 1465–1470 (2008).
31. Faure, A. J. & Lehner, B. MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis and allostery from deep mutational scanning data. Preprint at https://doi.org/10.1101/2024.01.21.575681 (2024).
32. Zhou, H.-X. & Pang, X. Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. *Chem. Rev.* **118**, 1691–1741 (2018).
33. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).

34. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
35. Amacher, J. F., Cushing, P. R., Brooks, L., Boisguerin, P. & Madden, D. R. Stereochemical Preferences Modulate Affinity and Selectivity among Five PDZ Domains that Bind CFTR: Comparative Structural and Sequence Analyses. *Structure* **22**, 82–93 (2014).
36. Forman-Kay, J. D. & Mittag, T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Struct. Lond. Engl. 1993* **21**, 1492–9 (2013).
37. Grant, B. J., Skjærven, L. & Yao, X. The Bio3D packages for structural bioinformatics. *Protein Sci.* **30**, 20–30 (2021).
38. Weng, C., Faure, A. J., Escobedo, A. & Lehner, B. The energetic and allosteric landscape for KRAS inhibition. *Nature* (2023) doi:10.1038/s41586-023-06954-0.
39. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* **21**, 207 (2020).
40. Wagih, O. Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
41. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
42. Bolognesi, B. *et al.* The mutational landscape of a prion-like domain. *Nat. Commun.* **10**, (2019).
43. de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
44. Saldanha, A. J. Java Treeview — extensible visualization of microarray data. **20**, 3246–3248 (2004).
45. Meng, E. C. *et al.* UCSF ChimeraX : Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).

Acknowledgements

Funding: This work was funded by European Research Council (ERC) Advanced grant (883742), the Spanish Ministry of Science and Innovation (LCF/PR/HR21/52410004, EMBL Partnership, Severo Ochoa Centre of Excellence), the Bettencourt Schueller Foundation, the AXA Research Fund, Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322), and the CERCA Program/Generalitat de Catalunya. T.Z. was funded by a EMBO Long-term Postdoctoral Fellowship (ALTF 525-2021) and a Marie Skłodowska-Curie Postdoctoral Fellowship (GPIDR, 101068134).

We acknowledge all members of the Lehner lab for feedback during the project, particularly Toni Beltran, Albert Escobedo, Andre Faure, Cristina Hidalgo, Taylor Mighell, Michael Thompson, Magda Topolska and Chenchun Weng for feedback/advice with experiments and data analysis. We thank Júlia Domingo for sharing plasmid vectors pGJJ211 and pGJJ215. We acknowledge the CRG Genomics Unit for DNA sequencing services. Molecular graphics and analyses were performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization,

and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Author contributions

Conceptualization: T.Z. and B.L. Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization: T.Z. Funding acquisition: T.Z. and B.L. Resources, Supervision: B.L. Writing – original draft, review & editing: T.Z. and B.L.

Data availability

All raw DNA sequencing data and associated processed data files (binding fitness measurements for all datasets) have been deposited in the Gene Expression Omnibus under the accession number GSE265816. All other data required to reproduce analyses are available at https://zenodo.org/records/11048045.

Code availability

All custom scripts required to reproduce analyses are available at https://github.com/lehner-lab/fuzzy_specificity.

Competing interests

B.L. is a founder and shareholder of ALLOX.

List of supplementary materials

Materials and Methods
Figs. S1 to S9
Tables S1 to S4

Figures

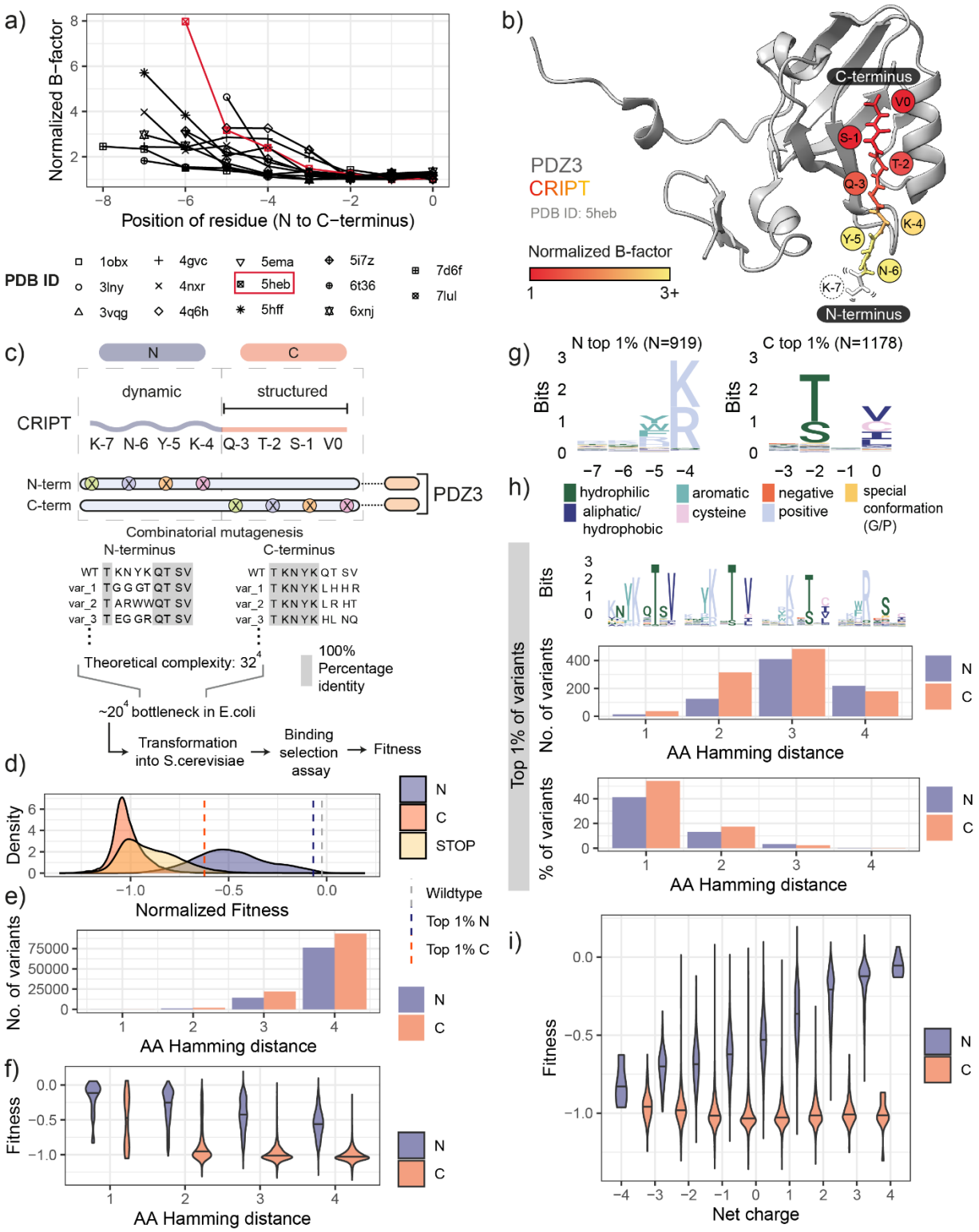

Figure 1. Combinatorial mutagenesis of a peptide ligand. a) Normalized b-factor vs position of residue for PDB structures of PDZ3 crystallized with its ligand b) PDZ3-CRIPT structure (PDB ID: 5heb[22]) where CRIPT is coloured by b-factor (normalized to be on the same scale as all other available PDZ3 structures with their ligands). K-7 is represented in white and dashed/moving lines since it is absent from the crystal structure. c) library design and pipeline to determine binding fitness for all possible mutants of dynamic N and structured C terminus of CRIPT d)

density distribution of binding fitness and e) AA hamming distance for all variants of N vs C f) Hamming distance vs. fitness of N vs C g) Position weight matrix (PWM) based on binding fitness scores for top 1% of variants in N vs C. h) Number and percent of top 1% of variants stratified by AA Hamming distance and their associated PWMs. i) Increasing net charge of residues in N results in incrementally increased fitness, but not in C.

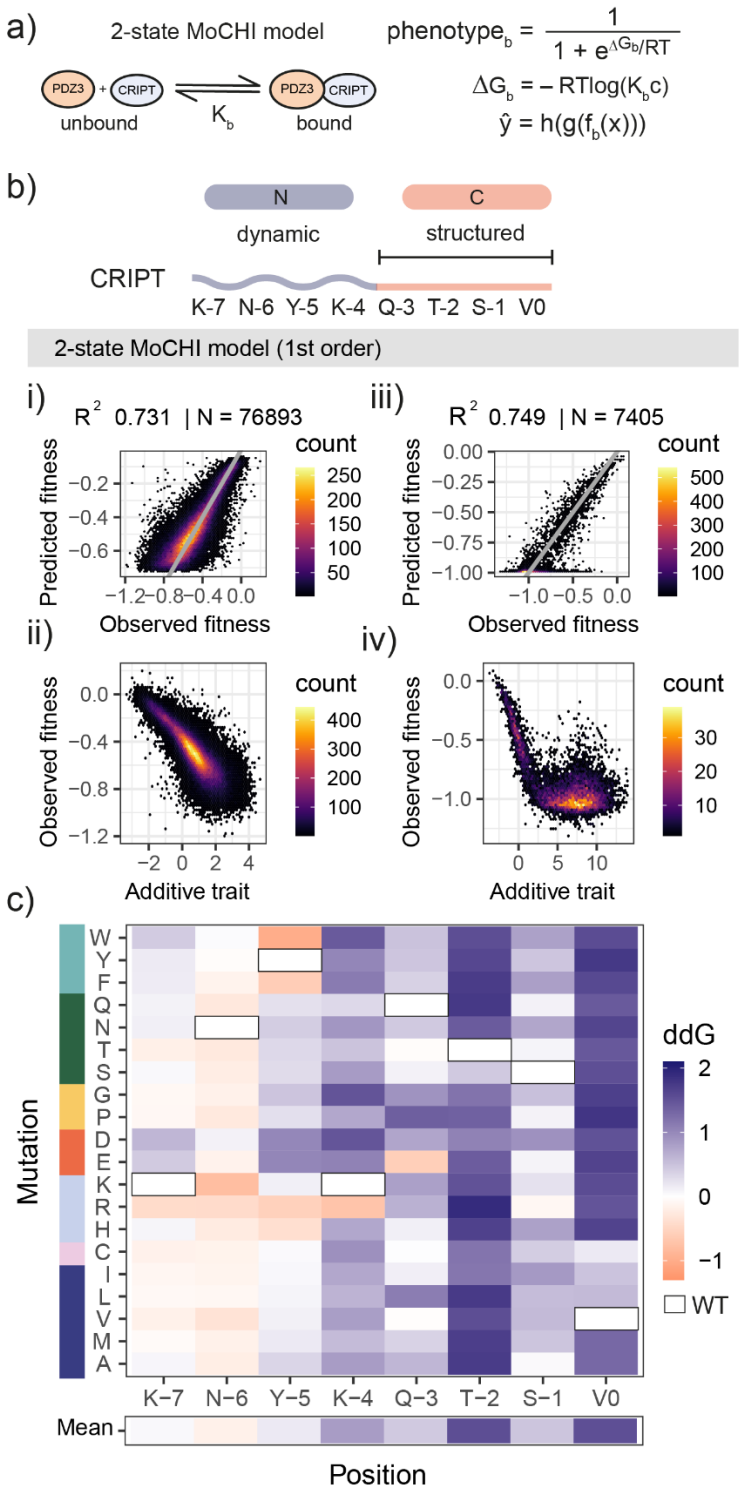

Figure 2. Energetic couplings in dynamic and structured portions of CRIPT. a) Two-state MoCHI thermodynamic model to transform binding fitness scores of mutations into energetic terms (ddG) of binding. b) Performance of model and additive trait coefficients of the dynamic N (left, panel i and ii) and structured C (right, panel iii and iv) portions of CRIPT. c) Heatmap of energetic binding terms for the dynamic N (left) and structured C (right) portions of CRIPT for all possible mutations (y-axis, coloured and ordered by physicochemical property).

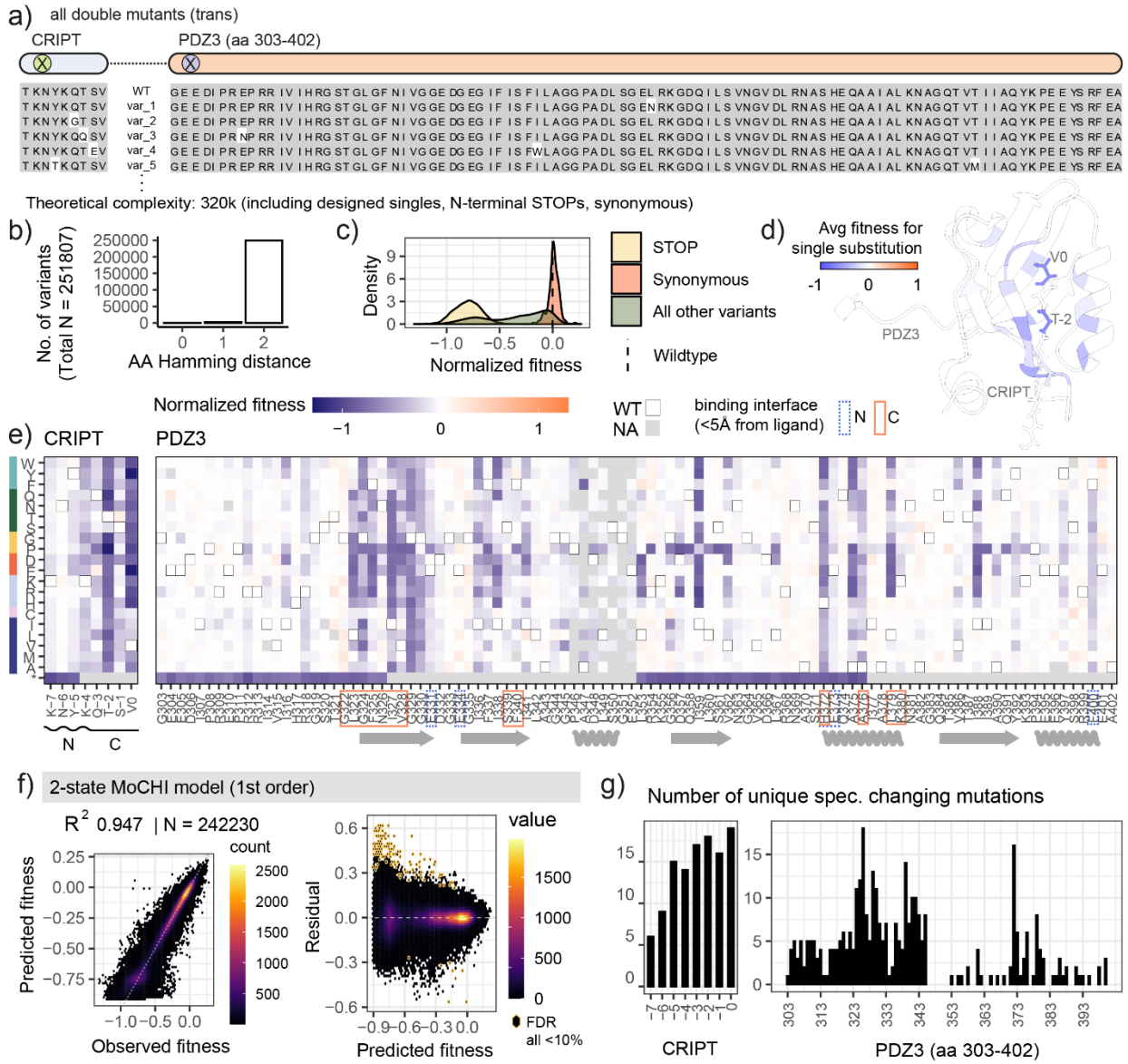

Figure 3. Measuring >200,000 energetic couplings between PDZ3 and CRIPT allows identification of specificity-encoding sites between the domain and peptide. a) Design of PDZ3-CRIPT trans library in which a library of all single substitutions in CRIPT is combined with a library of all single substitutions in PDZ3. Both libraries also contained designed STOP mutations and synonymous substitutions to comprise a total of 320k possible genotypes. b) AA Hamming distance and c) density distribution of binding fitness across the library for variants that pass quality thresholds. d) PDZ3-CRIPT structure (PDB ID: 5heb[22]) coloured by the average binding fitness effect for single substitutions. The canonical/constrained binding motif sites in CRIPT are labelled and have the strongest average binding fitness defects, as expected. e) heatmaps showing binding fitness of single substitutions in CRIPT (left) and PDZ3. STOP mutations are designed to only be in the N-terminus of CRIPT and two blocks of PDZ3. Most widely deleterious effects are at the binding interface (also seen in d)). f) Performance of first order 2-state MoCHI model on >240k variants (left) and residuals to the model (right) where those hexbins that pass FDR threshold of 10% (significantly different z-score from background) are outlined in yellow. g) Number of unique specificity-changing mutations (FDR<0.1) across CRIPT (left) and PDZ3 (right).

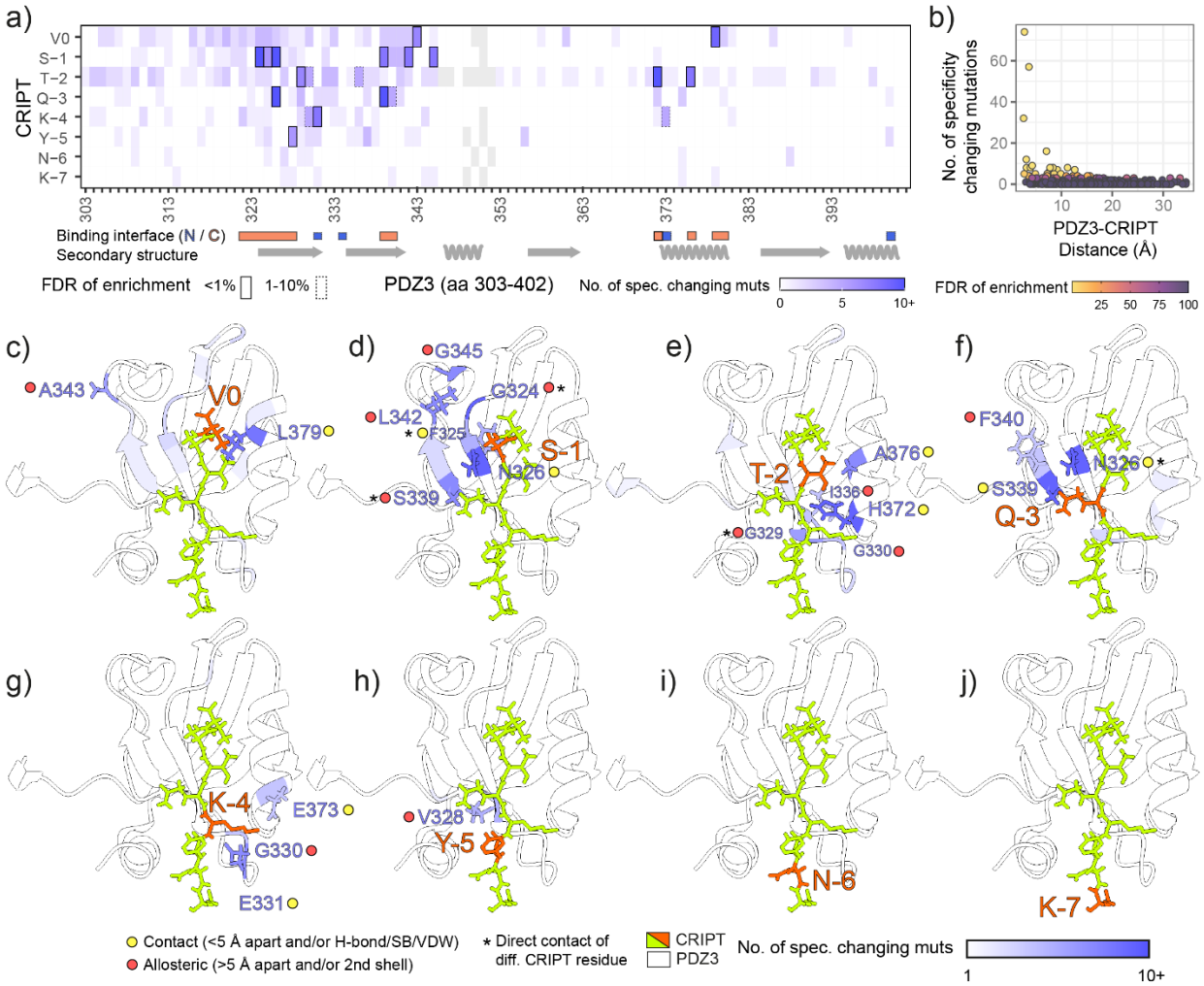

Figure 4. Major specificity encoding residues for each residue of a peptide. a) Heatmap showing number of specificity-changing mutations across every position-pair in PDZ3 and CRIPT. Outlined boxes mark those that pass FDR thresholds for enrichment as shown (solid FDR <0.01, dashed FDR 0.01-0.1) and comprise the major specificity encoding sites. PDZ annotations for binding interface (< 5Å) and secondary structure are shown along x axis). b) Number of specificity-changing mutations across each position pair versus distance between the position pair in question. Circles are coloured by FDR of enrichment, showing as in a) that most position pairs are not enriched in specificity-changing mutations and that position-pairs that are enriched in specificity-changing mutations tend to be closer to each other in the structure. c) – j) PDZ3-CRIPT (PDB ID: 5heb[22]) structure coloured by number of specificity-changing mutations, split by the identity of the CRIPT residue (highlighted in orange [V0, S-1, T-2, Q-3, K-4, Y-5, N-6, K-7] in separate panels). Number of specificity-changing mutations is thresholded between 1 and 10 to more clearly show range for significant (FDR<0.1) sites. Only positions in PDZ3 that pass the FDR threshold of 0.1 are labelled.

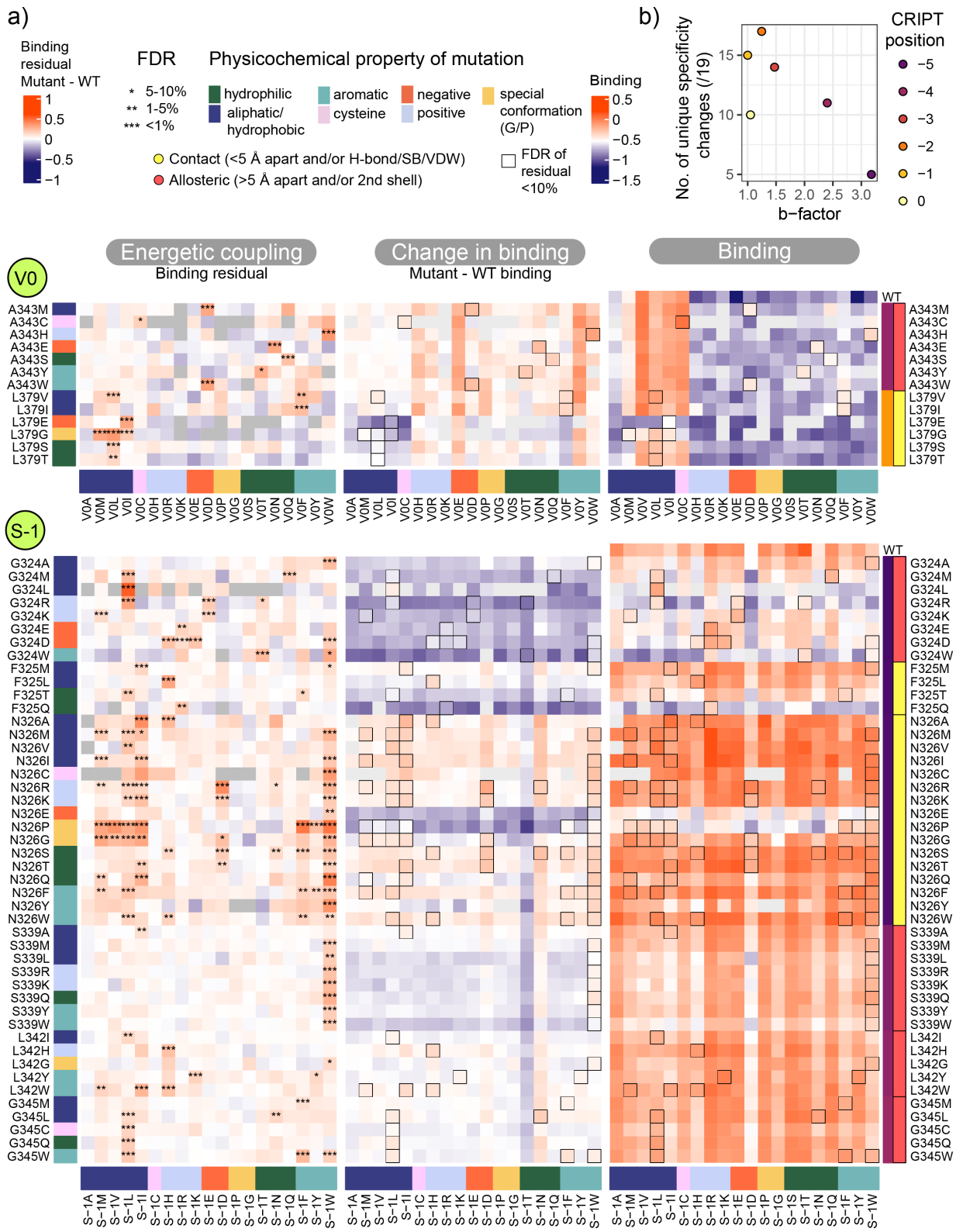

Figure 5. A complete map of all specificity-changing mutations across major specificity encoding sites in PDZ3-CRIPT a) Three columns of heatmaps show the energetic coupling (binding fitness residual), change in binding (binding fitness of mutant vs. wildtype), and raw binding fitness scores for all major specificity-changing PDZ3 mutations.

Mutations are colored by physicochemical property and positions in PDZ are colored by their position in PDZ (from N to C-terminus) as well as their contact/allosteric classification for the CRIPT residue that they are coupled to. b) Number of observed unique specificity changes in CRIPT vs b-factor for each residue.

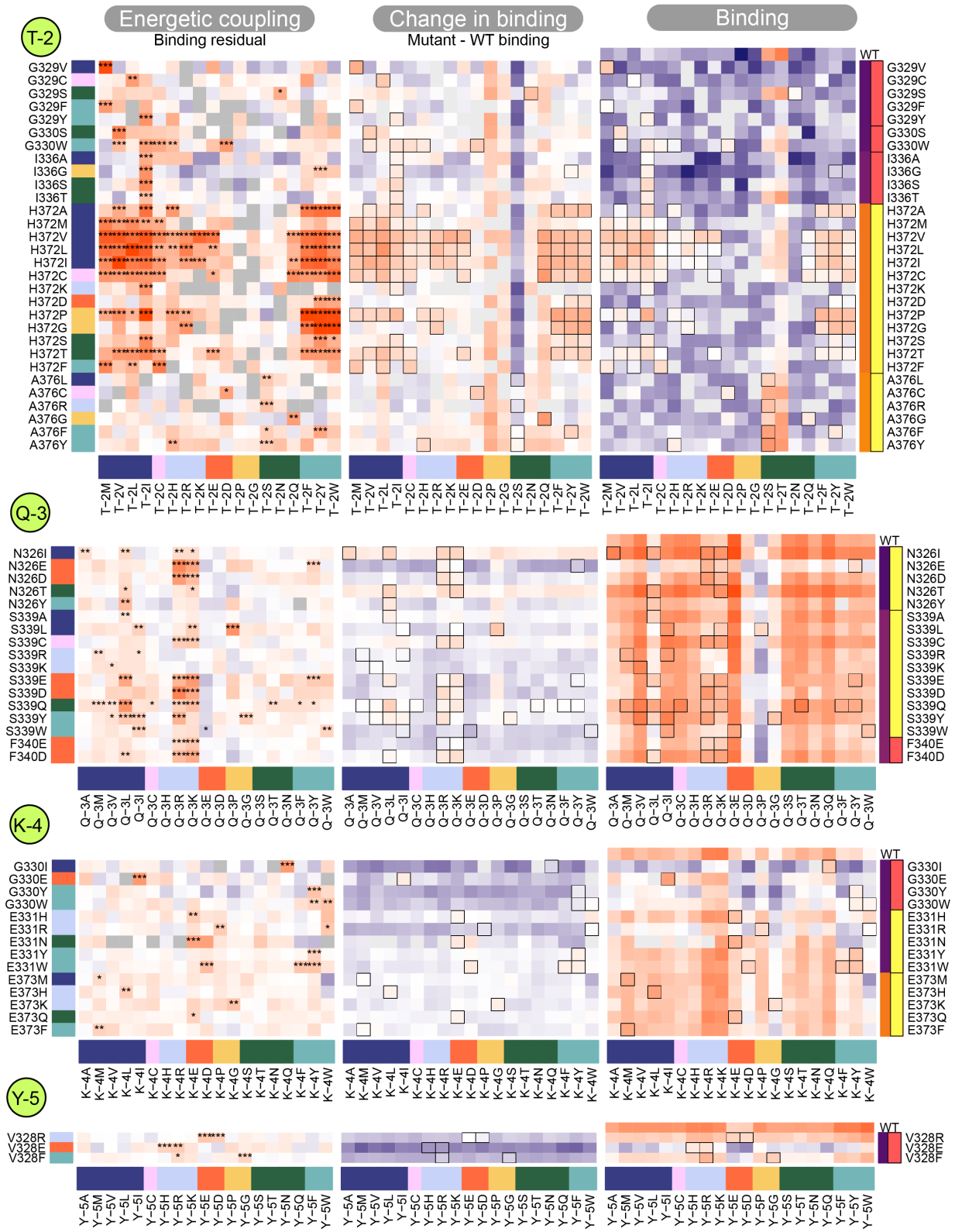

Figure 5. (continued)

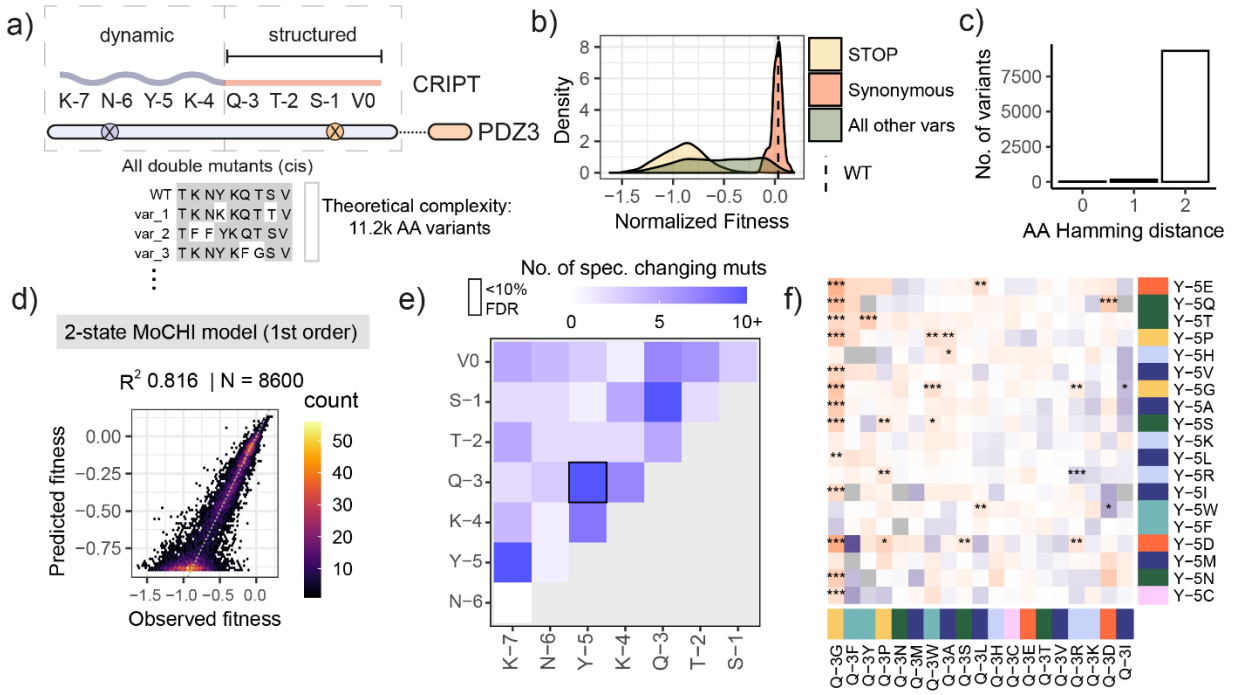

Figure 6. Measuring >8,000 energetic couplings between CRIPT residues reveals specificity-encoding positions within the dynamic peptide. a) Design of CRIPT cis double mutant library b) Density distributions of fitness and c) barplot of amino acid hamming distance for all variants d) Performance of first order MoCHI model fit on the CRIPT cis double mutant data e) Heatmap of the number of specificity-changing mutations per pair of CRIPT positions, with one significantly enriched coupled site outlined in black (FDR<0.1) f) Heatmap of residuals for each pair of mutations in the significantly coupled CRIPT sites Q-3 and Y-5.

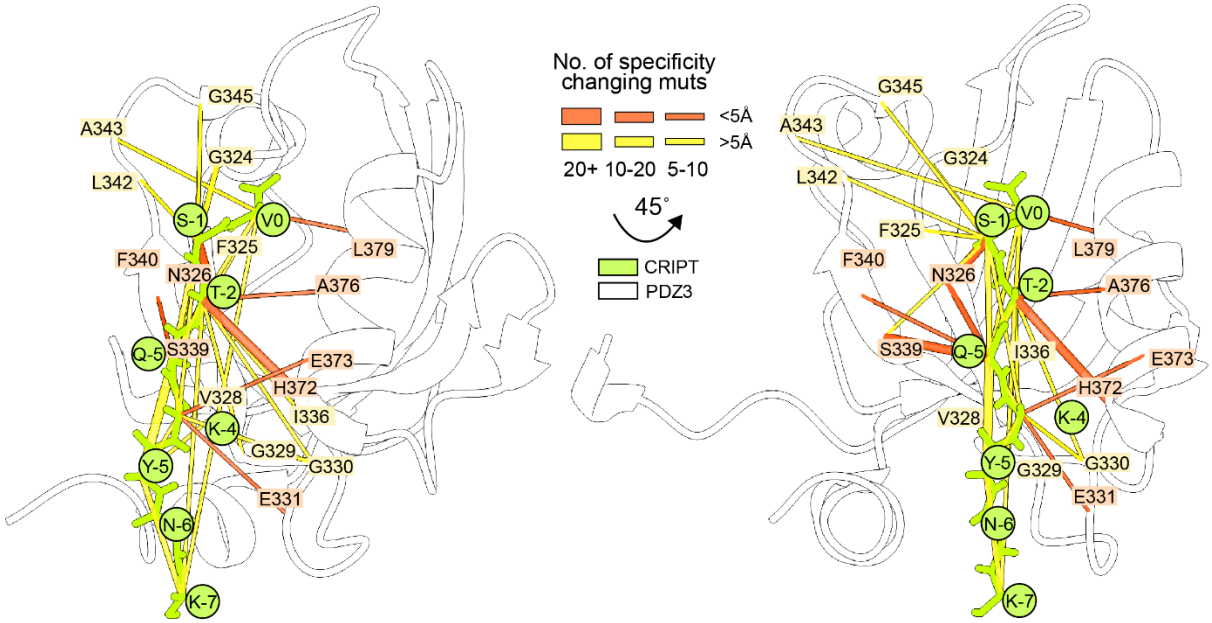

Figure 7. Summary of domain-peptide and peptide-peptide energetic couplings. PDZ3-CRIPT (PDB ID: 5heb[22]) structure where energetically coupled sites are indicated with coloured pseudobonds. Orange pseudobonds indicate couplings between structurally close residues (<5Å), yellow pseudobonds indicate couplings between structurally far residues (>5Å) or those for which distance cannot be determined in the crystal structure. Width of pseudobonds indicates effect size (i.e. number of specificity-changing mutations between that pair of residues).

Materials and Methods

Media

- LB: 10 g/L Bacto-tryptone, 5 g/L Yeast extract, 10 g/L NaCl. Autoclaved 20 min at 120ºC.
- YPD: 20 g/L glucose, 20 g/L Peptone, 10 g/L Yeast extract. Autoclaved 20 min at 120ºC.
- SORB: 1 M sorbitol, 100 mM LiOAc, 10 mM Tris pH 8.0, 1 mM EDTA.
- Filter sterilized (0.2 mm Nylon membrane, ThermoScientific).
- Plate mixture: 40% PEG3350, 100 mM LiOAc, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0. Filter sterilized.
- Recovery medium: YPD (20 g/L glucose, 20 g/L Peptone, 10 g/L Yeast extract) +0.5 M sorbitol. Filter sterilized.
- SD -URA: 6.7 g/L Yeast Nitrogen base without amino acid, 20 g/L glucose, 0.77 g/L complete supplement mixture drop-out without uracil. Filter sterilized.
- SD -URA/ADE: 6.7 g/L Yeast Nitrogen base without amino acid, 20 g/L glucose, 0.76 g/L complete supplement mixture drop-out without uracil, adenine and methionine. Filter sterilized.
- MTX competition medium: SD –URA/ADE + 200 ug/mL methotrexate (BioShop Canada Inc., Canada), 2% DMSO.
- DNA extraction buffer: 2% Triton-X, 1% SDS, 100mM NaCl, 10mM Tris-HCl pH8, 1mM EDTA pH8.

B-factor analysis

We obtained all 214 available PDZ domain family entries bound to a ligand from the PDB. We used the Bio3D[37] R package to read in the entries, filtering for those with 2 unique chains in the structure (n = 61), further filtered for those that had B-factors available (n = 58). We normalized the B-factors by the minimum value within that chain in order to be able to compare B-factors across crystal structures. We filtered the length of chain B to be in the first quartile of the dataset (in order to roughly filter for PDZ domains bound to a peptide ligand that was 10 aa or less) and present the 14 PDZs bound to a short ligand and their respective normalized B-factors in Fig. 1a).

Library construction

We designed four libraries to probe different questions about the PDZ3-CRIPT interaction (Fig. S1a). All libraries had the same backbone bindingPCA vector structure of DLG4-PDZ3 (aa 303-402) N-terminally fused to DHFR3 and CRIPT (0 to -8) N-terminally fused to DHFR1,2 as in[34]. All plasmids used in the study are described in table S2.

Combinatorial N- and C-terminal CRIPT libraries were each ordered as an NNK degenerate oligo from IDT and cloned into the bindingPCA vector containing PSD95-PDZ3 (aa 303-402) fused to DHFR3 and no bait fused to DHFR1,2 (hence “empty”). The library transformation was bottlenecked in *E. coli* such that the combinatorial space of variants was reduced from the possible 32^4 total variants (32 possible codons at 4 positions) to 20^4 in order to facilitate greater downstream sequencing depth for each variant.

The CRIPT “cis” double mutant library was ordered as an IDT pool of NNK oligos with 28 NNK degenerate oligos encoding all different combinations of double mutants across CRIPT

positions 0 to -8. The library was transformed into the same empty vector as the CRIPT N and CRIPT C libraries described above.

The PDZ3-CRIPT “trans” double mutant library was ordered as 3 separate TWIST oligo pools. PDZ3 (aa 303-402) was split into two non-overlapping blocks for mutagenesis: Block 1 encoding the first 50 aa in the domain and Block 2 encoding the latter 50 aa in the domain. We designed 1000 variants encoding the following for each block: wildtype, all possible single mutants, half of all possible synonymous mutants, and half of all possible single STOP codon variants starting at the N-terminus of each block. Variants were designed such that each possible single mutant codon was firstly scored by the number of nucleotide substitutions away from wildtype (such that 2-3 substitutions were preferred, followed by 1, followed by 0) and then by the most optimal yeast codon (based on the *S. cerevisiae* codon usage table). This design strategy enabled us to use non-overlapping reads to sequence the long amplicons that encoded PDZ3 at one end and CRIPT at the other (schematic in Fig. S4a, designed library and results in Fig. S4b). The CRIPT single mutant library was designed in the same way but ordered as a single block due to its short length. Each library was amplified using distinct primers and transformed separately into the same backbone/bindingPCA vector (PGJJ001 as in [34]). The PDZ3 portion of the Block 1 and Block 2 vectors were then cloned in two separate reactions into the vector with the CRIPT single mutant library to produce 2 double mutant library vectors (PDZ3 Block 1 single mutants + CRIPT single mutants, PDZ3 Block 2 single mutants + CRIPT single mutants).

Large-scale transformations of libraries into yeast and competition assays

We transformed each library of variants into *S. cerevisiae* in 3 replicates at a large volume scaled to the size of the library in order to ensure that all variants were present in multiple (at least 100) copies as in previous work[34,38] to prevent bottlenecking the library. We grew the cultures to saturation in synthetic complete media with 2% glucose as a carbon source. We harvested these cultures as the “input” replicates to our selection assay, and subjected the input library to selection. The selection experiment is based on a well-described protein complementation assay[30] wherein the yeast are grown in synthetic complete media with added methotrexate (MTX), a drug that requires dihydrofolate reductase (DHFR) for metabolization. In the presence of MTX, CRIPT (or the bait protein) must be bound to PDZ3 (the prey protein) to bring together the split fragments of DHFR and enable cell growth (Fig. S1a). We harvested these cultures as the “output” from the selection assay. We extracted the DNA from each replicate using a standard phenol chloroform procedure as in[34]. To prepare the DNA for Illumina sequencing, we used two PCR steps to obtain the amplicon for each library (table S3). In PCR1, we added frameshifting oligonucleotides (table S4) and amplified the regions of interest (amplicon with constant region and mutated region) from the extracted DNA with 5 cycles. In PCR2 we added Illumina sequencing barcodes with PCR using the minimum number of cycles necessary to reach amplification plateau for each sample based on a qPCR run.

Next Generation Sequencing and analysis of sequencing data (read counts to fitness scores)

We use sequencing as a quantitative readout for binding between PDZ3 and CRIPT (Fig. S1b). We obtained reliable sequencing data for a total of close to half a million variants across all four libraries, obtaining a fitness score (and associated error) for each variant using DimSum[39]. Parameters used to filter sequencing reads for DimSum required an input count of 10 reads in at least one replicate, and we filtered each dataset for the specific design of each library (i.e. libraries made with NNK degenerate oligos were filtered for NNK design and the custom pdz3-cript trans libraries were filtered to only keep designed variants).

Normalization of fitness scores across experiments

All 4 libraries contained overlapping variants that had highly correlated fitness (binding) scores when processed independently (Fig. S2f,h, Fig. S8 c). In order to make all scores comparable across the study, we used a linear transformation based on these highly correlated shared variants to normalize each library (Fig. S2g,i, Fig. S8d).

Position-weight matrices

All position-weight matrices were constructed using ggseqlogo[40] with a custom Zappo-based[41] color scheme to mark physicochemically related amino acids.

Calculation of physicochemical properties

We calculated features from a curated list of amino acid property scales (n=386) (http://www.genome.jp/aaindex/) as in[42] to quantify the correlation of these scores with the binding scores for the combinatorial N and C libraries in Fig. 1. We present those features that had a | Spearman's r | > 0.4 for either the N or C libraries in Fig. S2J.

Modeling phenotype to free energy with MoCHI

To translate the fitness scores, which capture the phenotypic effects of mutations, into free energy terms, we used the MoCHI package[31] to model the fitness with a two-state thermodynamic model for protein binding. Briefly, MoCHI takes as input amino acid sequences of each variant and predicts their fitness while correcting for global non-linearities (non-specific epistasis). Using the coefficients extracted from the model, we obtain the change in free energy associated with each mutation for the phenotype in question (in our case, binding). We used default parameters for a two-state model with one phenotype (binding) for all datasets. We used L1 and L2 regularization with a lambda of 10^{-6}. We evaluated the model using the held-out "fold" from the 10 times that the model was run on the dataset.

The CRIPT C combinatorial library contained an overabundance (>90%) of non-binding variants, so we balanced the dataset by sampling the distribution of >100k variants without replacement so that the variants with the largest distance from the peak of non-binding variants (specified by 2 s.d. away from mean of STOP codon binding distribution) had a higher probability of being sampled (Fig. S3a-b). We used a distance function based on a power-law distribution to weight the sampling probability as follows:

$$W_i = k^{1-\frac{x_i}{d}}$$

Where W_i is the sampling weight of each variant binding score x_i, d is the mean binding score for the non-binding/dead mode, and k is a constant (=10^{10}, chosen to balance the dataset towards a reasonable percentage of binding variants [~25%], Fig. S3b).

To test the reproducibility of the model results, we repeated the sampling procedure 10 times and performed the three implementations of MoCHI modeling (linear model with no global epistasis, two-state thermodynamic model with first order terms, two-state thermodynamic model with second order terms) (Fig. S3f). We found the modeling results to be extremely stable across 10 iterations and therefore show one representative iteration in Fig. 2, Fig. S3 and Fig. S4.

Analysis of residuals to quantify energetic couplings

For both PDZ3-CRIPT trans and cis CRIPT double mutant libraries we quantified the residuals from the observed vs. mean predicted binding fitness for each variant in the respective dataset. We converted the residuals to Z-scores (using the error derived from DimSum as the denominator) and performed a Z-test to derive p-values for each variant as the Z-scores were normally distributed. We corrected the p-values for multiple testing (Benjamini-Hochberg) and report the FDR values associated with residuals where relevant. Residuals (energetic couplings) that are significant by the multiple test-corrected Z-test, >0 and pertain to variants that pass the non-binding threshold (i.e. are binding) are classified as specificity-changing. To test for enrichment of specificity-changing mutations in PDZ3-CRIPT pairs, we did a test for enrichment (hypergeometric test) of specificity-changing mutations across all pairs of sites. We again performed multiple test correction (Benjamini-Hochberg) on these to identify the major specificity encoding residues/coupled sites as those that pass an FDR threshold of 0.1. We highlight these major energetically coupled sites in Fig. 4 and Fig. 5a, and list them in table S1.

Unsupervised clustering of binding residuals

We clustered the vector of binding residuals to all possible CRIPT variants for each PDZ3 variant that had at least one specificity-changing mutation (N=340), filtered to include those PDZ3 variants with >80% data present (N=290). We used Cluster 3.0[43] to perform unsupervised hierarchical clustering with a weighted cosine distance, average linkage and otherwise default parameters. We used JavaTreeview[44] to visualize the full cluster plot and several manually highlighted clusters in Fig. S7.

Protein contact determination

We used getContacts (https://getcontacts.github.io/) to predict contacting residues using get_static_contacts.py with the 5heb.pdb[22] structure, --itypes option set to “all”, and otherwise default parameters. We also used the Bio3D[37] package to calculate inter-residue distances between specific residues in PDZ3 and CRIPT based on the 5heb[22] structure from PDB.

Visualization of protein structures

All protein structures were based on the PDZ3-CRIPT structure with PDB ID 5heb[22]. In Fig. 1b we present the full crystalized structure, but since the K-7 position is missing, we added this residue using ChimeraX[45] v1.4 and therefore present it as a cartoon with motion lines around it and no associated B-factor. All other represented protein structures, e.g. in Fig. 3d, Fig. 4c-j, Fig. 7, and Fig. S7 include only those residues in PDZ3 and CRIPT that we had in our mutagenesis design, i.e. aa 303-402 in PDZ3 and 0 to -7 in CRIPT. All quantitative visualization on structures was performed with the color by attribute function in ChimeraX and links between energetically coupled residues (Fig. 7) were drawn via the pseudobonds function.

Supplementary Figures

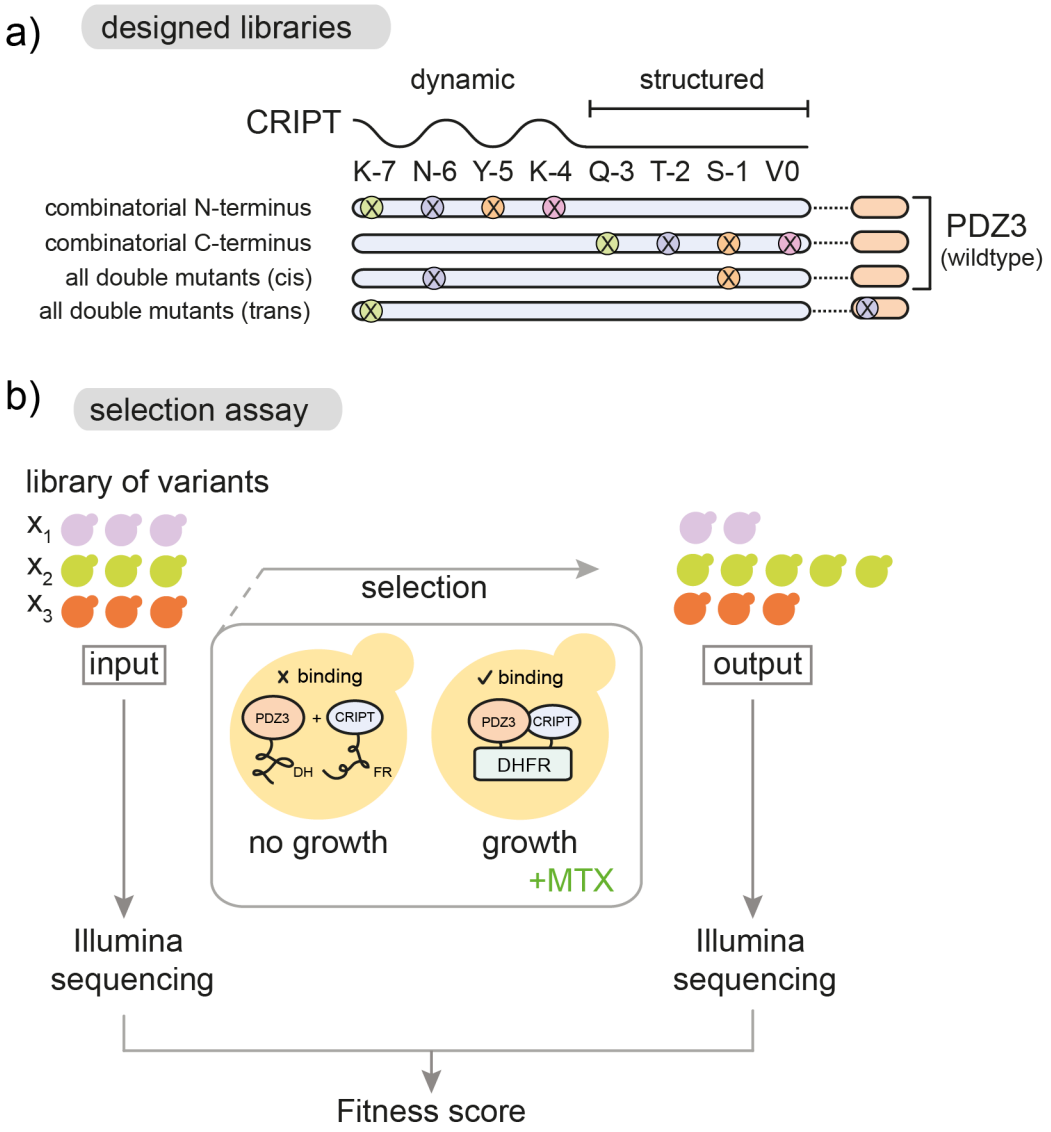

Figure S1. Overview of experimental design. a) designed libraries used to probe genetic encoding of PDZ3-CRIPT binding. b) libraries of variants are transformed into yeast and selected via a protein complementation assay (refs) that quantifies PDZ3-CRIPT binding.

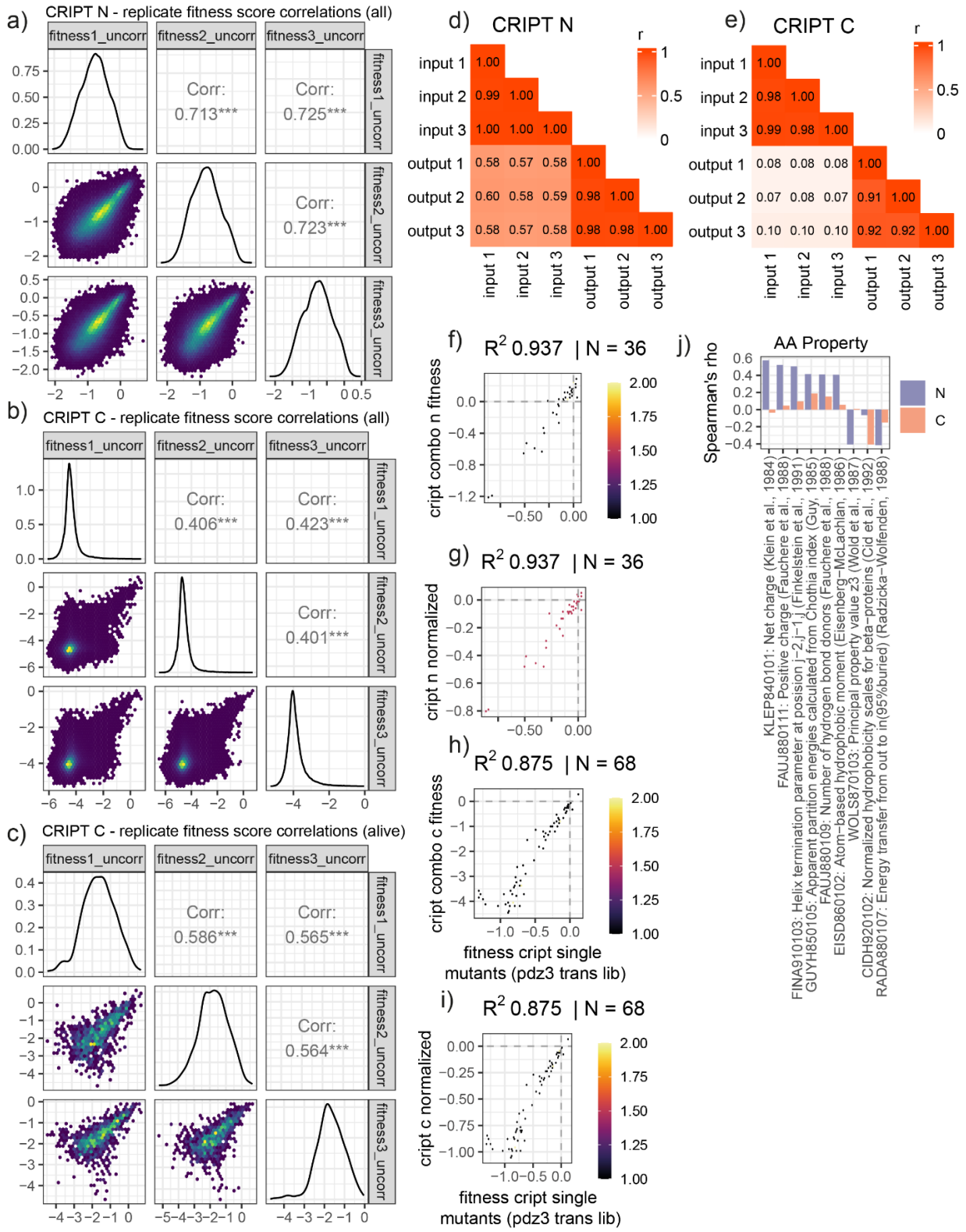

Figure S2. Overview of CRIPT combinatorial N and C data and quality. a) replicate fitness correlations (Pearson's r) for all N-terminal CRIPT variants b) replicate fitness correlations (Pearson's r) for all C-terminal CRIPT variants and c) top 1% of variants d) replicate count correlations (Pearson's r) for N-terminal CRIPT and e) C-terminal CRIPT. f-i)

normalization of fitness via a linear transformation using shared variants between CRIPT combinatorial libraries and the pdz3-cript trans double mutant library. j) Top physicochemical features that correlate (Spearman's rho > 0.4) with binding fitness scores in N or C dataset.

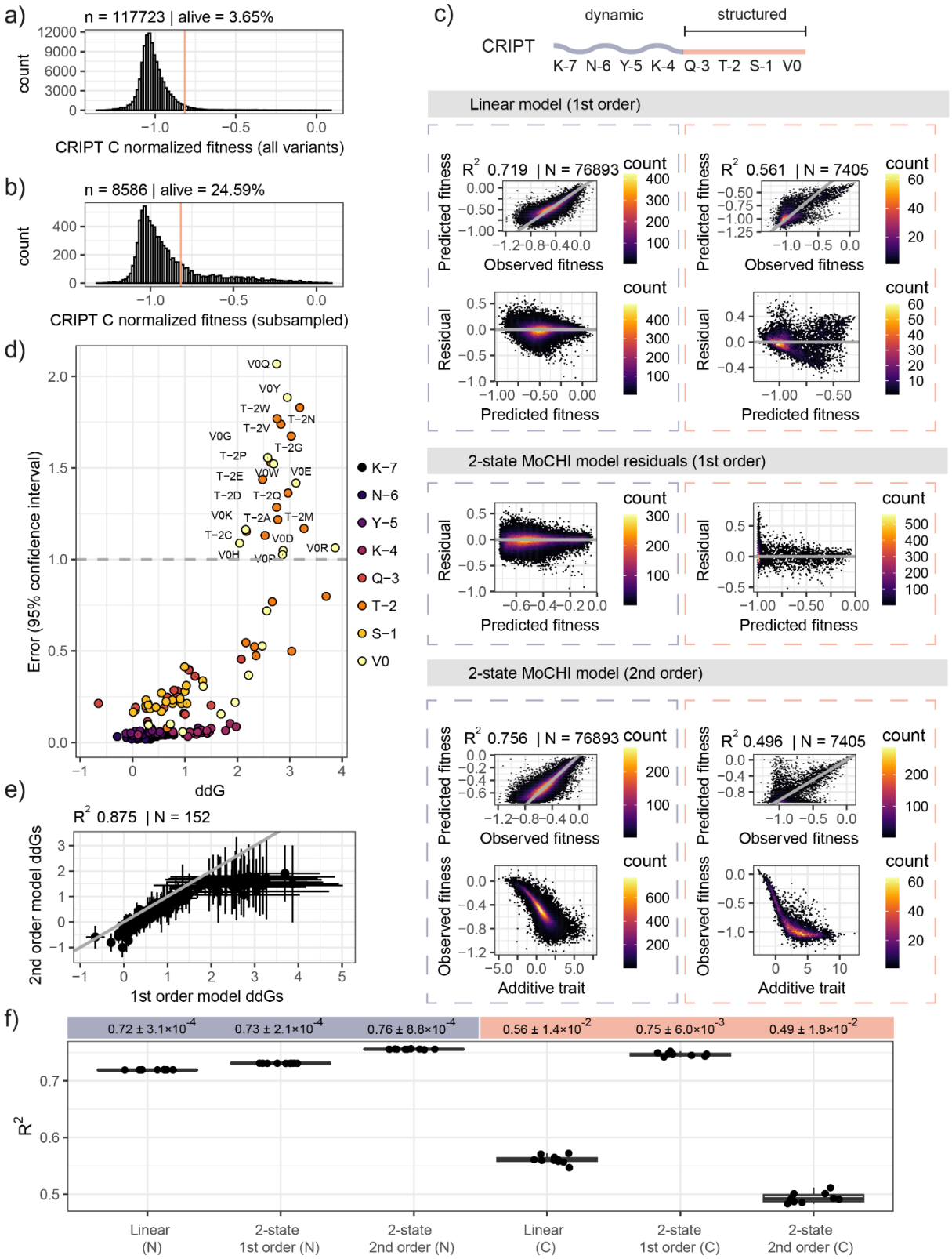

Figure S3. Additional MoCHI modeling results for combinatorial CRIPT N and CRIPT C datasets a) distribution of all variants in CRIPT C combinatorial library, marking percentage of variants that are “alive”/binding (i.e. 2 s.d. away from the mean of the distribution of STOP codons) b) distribution of subsampled variants that were used to balance

the dataset (see Methods) c) performance of several models on CRIPT N (left) and C (right) datasets, including a linear model that does not account for global epistasis (top), residuals of a 2-state MoCHI model shown in Fig. 2 (middle), and a 2-state MoCHI model that takes into account interactions between mutations (bottom). d) The relationship between ddG values and error, with cutoff of 1 kcal/mol indicated by dashed line to represent confident ddGs – all ddG values with high error values are in the V0 and T-2 positions that also have the highest ddG values. e) correlation between ddGs from 1st order vs 2nd order MoCHI model for CRIPT N and C datasets. Error bars for x and y axis indicate respective 95% confidence interval in kcal/mol. f) Results from running all models on 10 iterations of subsampling the CRIPT C dataset. Mean of model performance (i.e. explained variance, R^2 observed vs. predicted fitness) via 10-fold cross-validation for the 10 models is shown with 95% confidence intervals (2*s.d.) for each model.

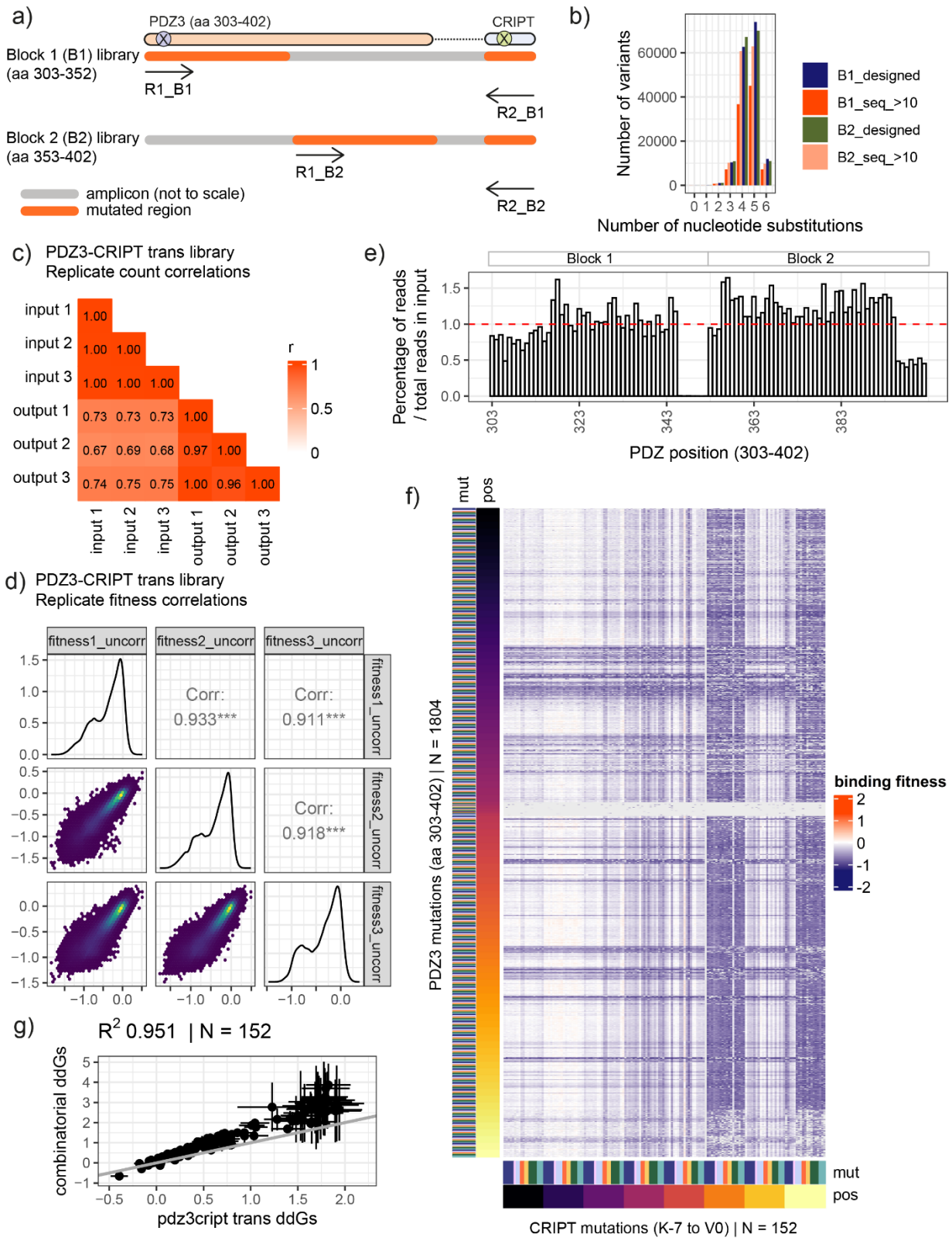

Figure S4. Overview of PDZ3-CRIPT trans library data and quality. a) sequencing strategy of PDZ3-CRIPT double mutant library where R1 and R2 denote forward and reverse sequencing reads, respectively. b) PDZ3-CRIPT library sequenced variants (_seq_>10 to reflect more than 10 input reads in the sample) reflect library design to maximize

number of nucleotide substitutions for each encoded variant along with the most optimal yeast codon (see methods) c) replicate count correlations for PDZ3-CRIPT library d) replicate fitness correlations for PDZ3-CRIPT library e) percent of reads in input sample at each position (out of total reads in input) for each of the 100 positions assayed across the PDZ domain – even sequencing expectation is 1%, denoted by dashed red line, showing that most positions are evenly sampled except for the missing tail end of block 2. f) all-by-all heatmap of binding fitness for PDZ3-CRIPT double mutants. Mutation physicochemical properties are represented by colours as shown in Fig. 1), positions are coloured from N to C (dark to light gradient). g) correlation between ddGs from two-state MoCHI model trained on combinatorial CRIPT libraries vs. PDZ3-CRIPT libraries with corresponding 95% confidence intervals as error bars on x and y axes.

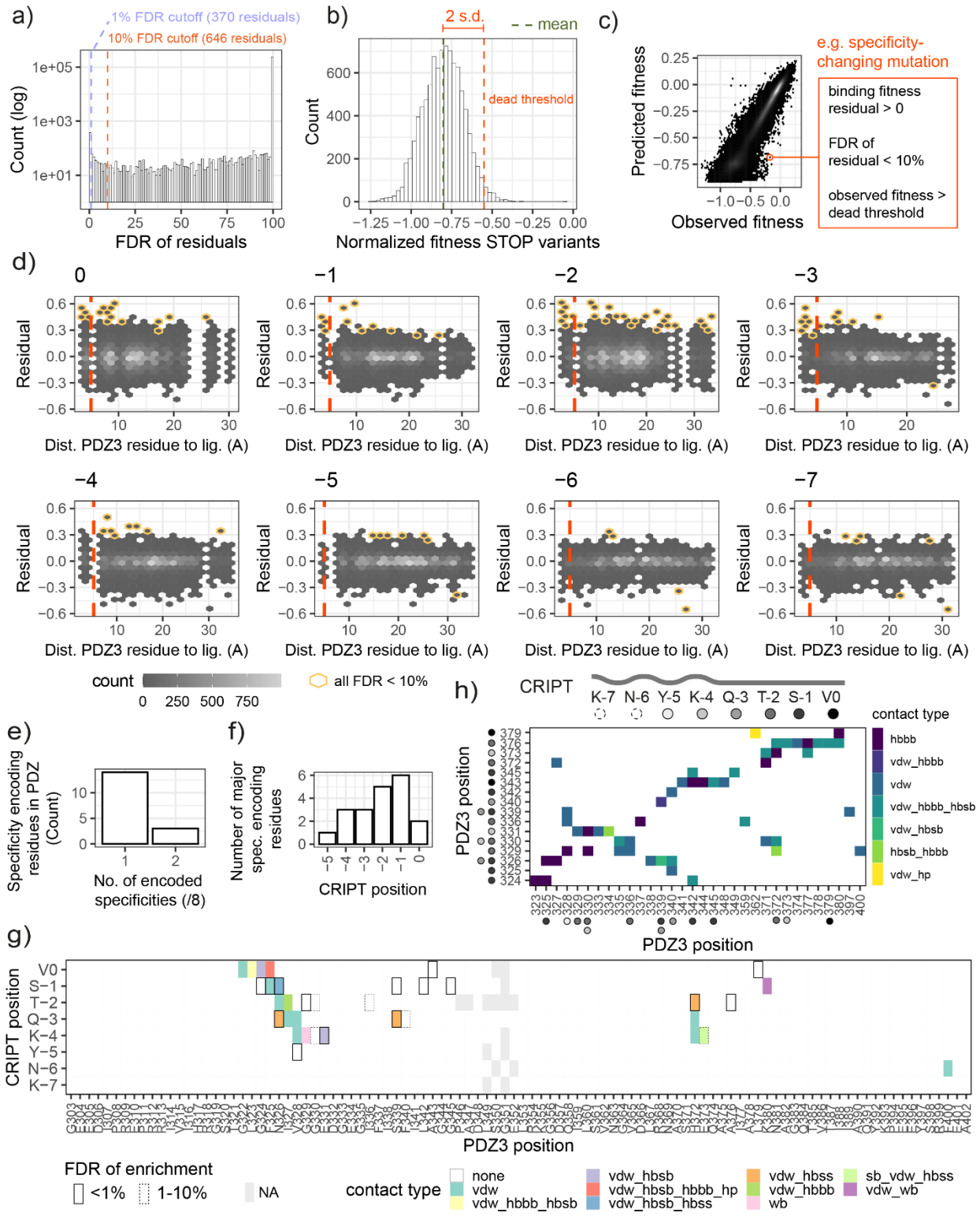

Figure S5. Determination of specificity-changing mutations and association with structural contacts. a) FDR of binding fitness residuals to MoCHI 2-state model for PDZ3-CRIPT mutants with different cutoffs b) distribution of STOP codon binding fitness and threshold for “dead”/non-binding threshold (2 s.d. away from STOP codon variant fitness

mean). c) definition of a specificity-changing mutation d) density plot of distance between PDZ3-CRIPT mutated position pairs vs. binding fitness residual (y axis), with those hexbins that have all residuals with FDR<0.1 outlined in yellow e) the number of CRIPT positions for which each major specificity encoding residue in PDZ encodes specificity, showing that the vast majority encode specificity for only one CRIPT position f) Number of major specificity encoding residues (in PDZ3) for each CRIPT position (-6 and -7 have zero) g) PDZ3 position vs. CRIPT position showing position pairs enriched in specificity-changing mutations (major specificity-encoding sites) outlined by FDR value (solid vs dashed line representing <1% vs 1-10% respectively) and the contact type of each pair as classified by getContacts (see Methods); vdw = van der waal's, hbbb = hydrogen bond backbone-backbone, hbsb = hydrogen bond sidechain-backbone, hbss = hydrogen bond sidechain-sidechain, sb = salt bridge, wb = water bond, hp = hydrophobics g) PDZ3 positions of major specificity-encoding residues (y axis) and their contacts in PDZ3 as determined by getContacts. CRIPT residue for which each PDZ3 position encodes specificity is represented by filled circles.

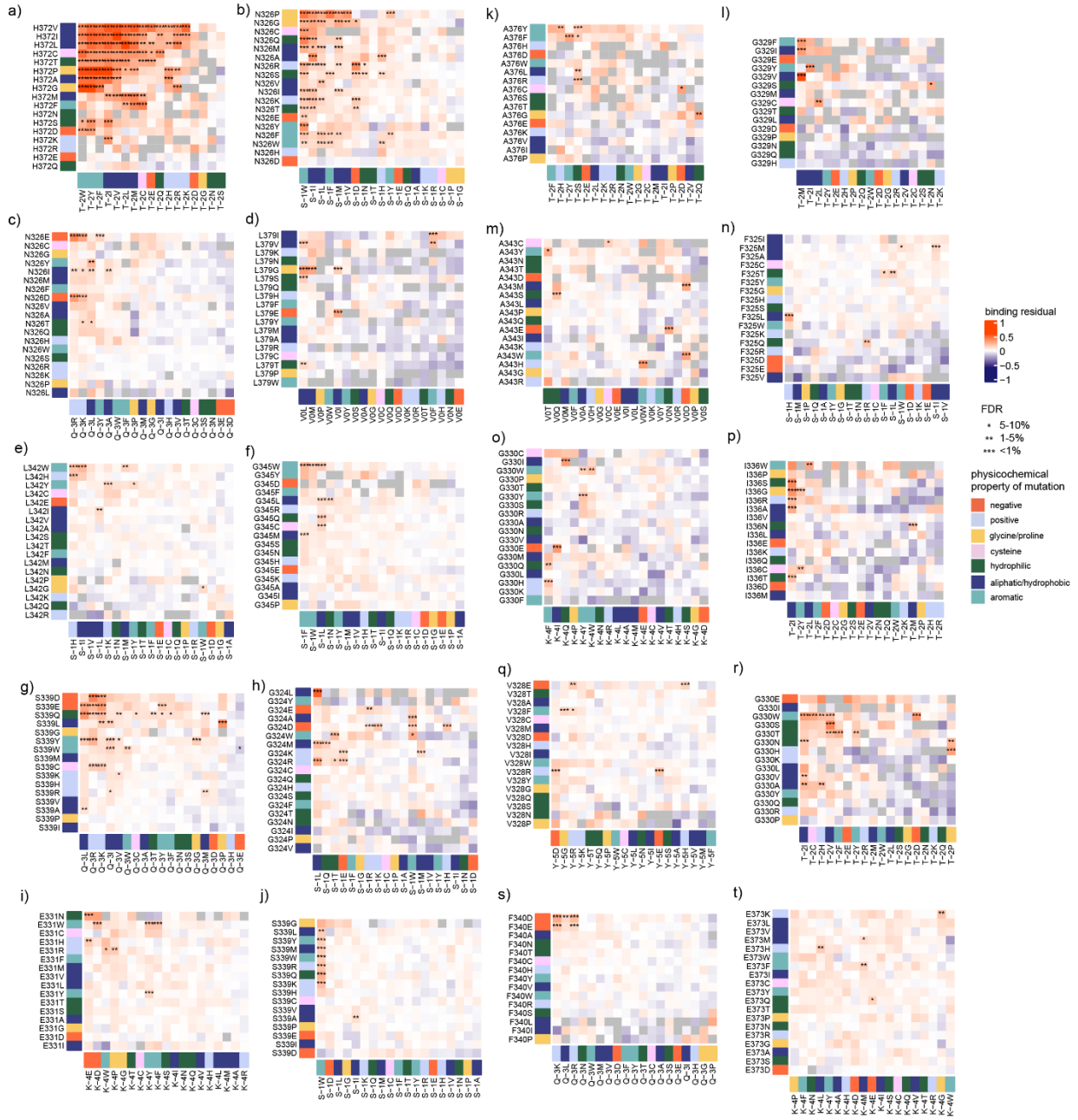

Figure S6. Clustered heatmaps of all 20 PDZ3-CRIPT major pairs of specificity encoding positions (i.e. significantly enriched FDR<0.1 for specificity-changing mutations).

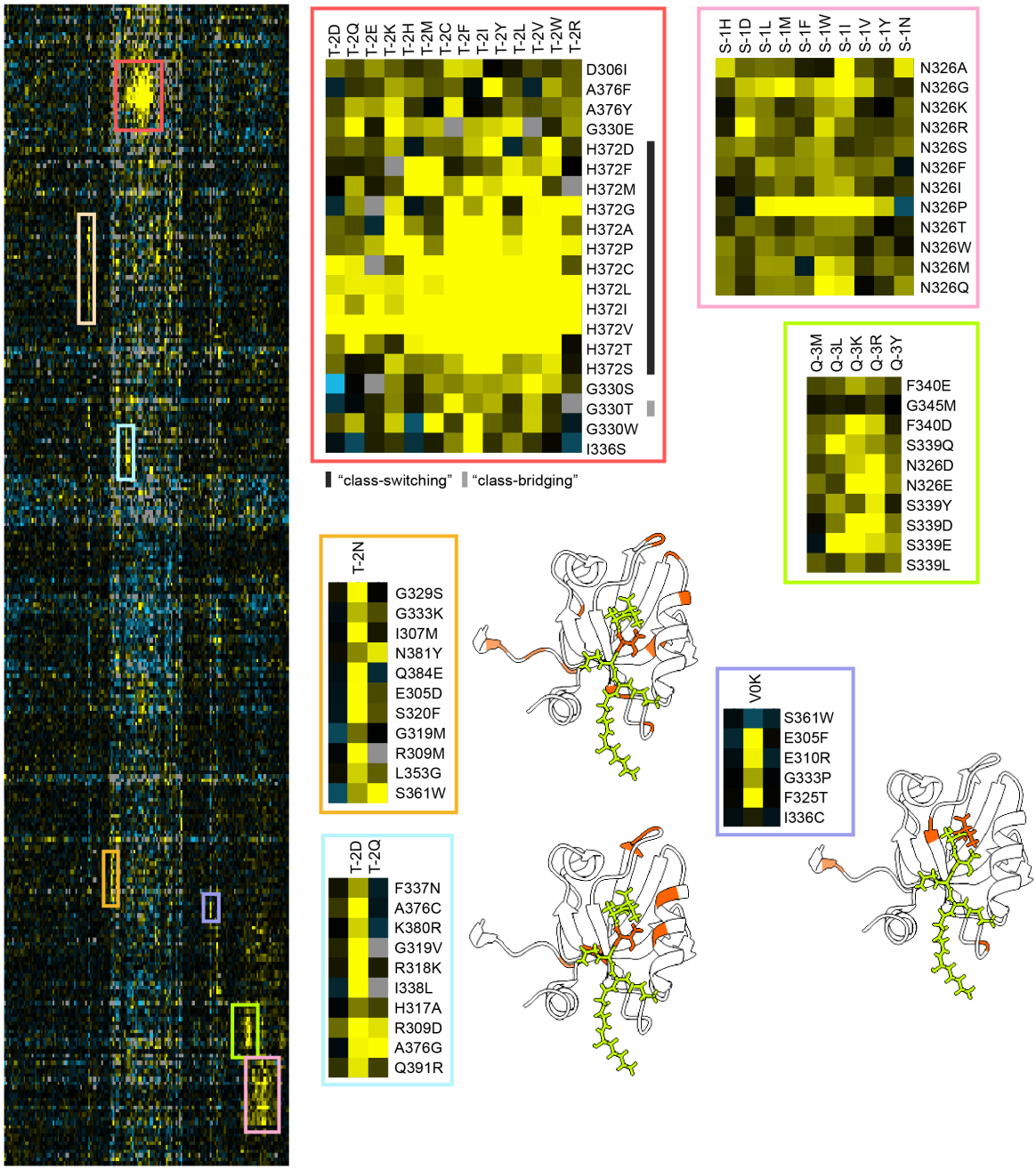

Figure S7. Clustered heatmap of binding residuals for all PDZ mutations (y-axis, N=290) that have at least one associated specificity-changing mutation in CRIPT (x axis, N=152). Outlined boxes represent manually-chosen clusters that point to PDZ positions enriched for specificity-changing mutations (H372, N326, S339) as shown in detail in Fig. 5 but also other sets of positions that are seemingly unrelated (orange, purple, blue clusters) with corresponding structural annotations of these sites to the right of each cluster.

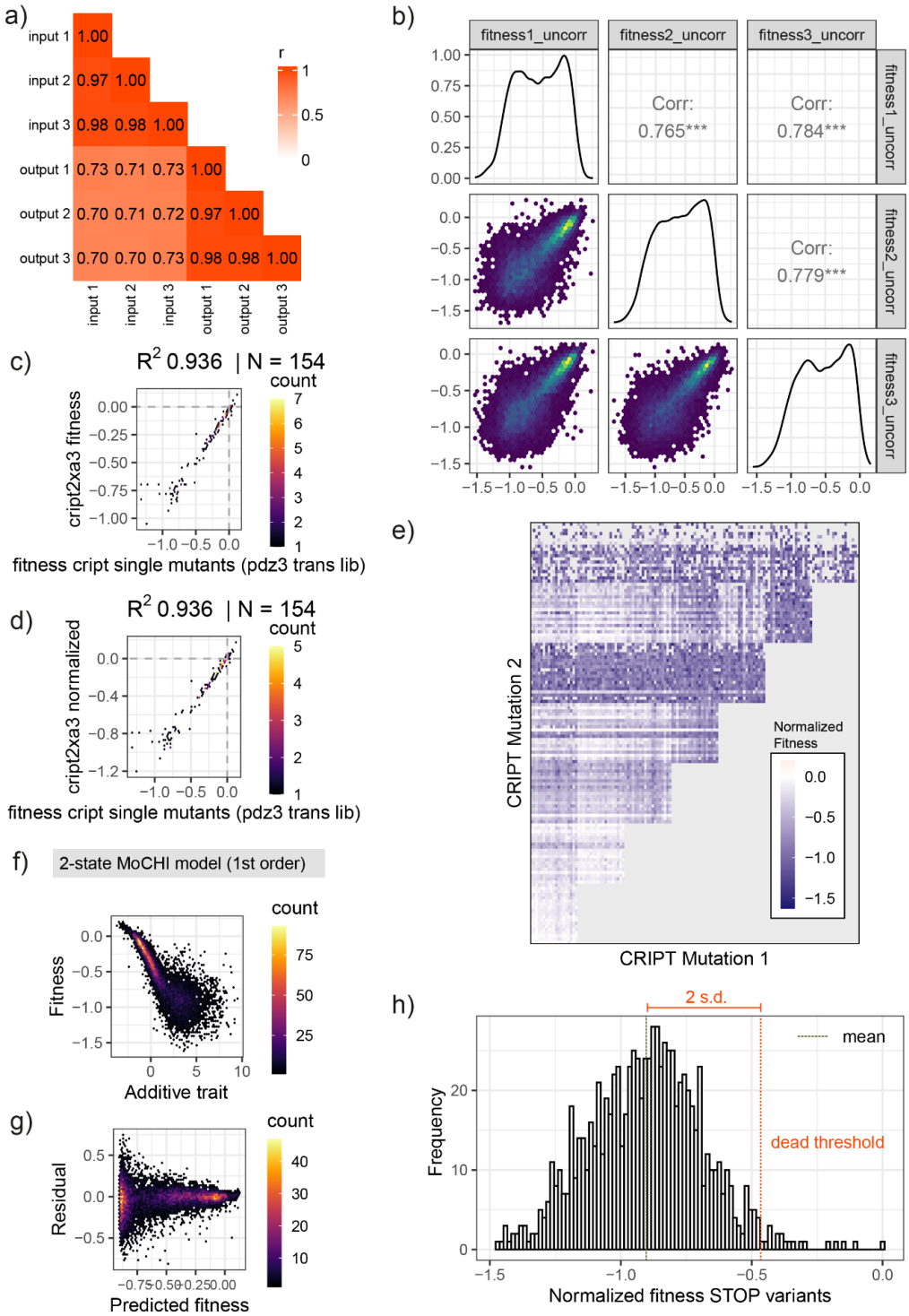

Figure S8. Overview of CRIPT cis double mutant library data and quality. a) replicate count correlations and b) replicate fitness correlations for all variants. c) correlation and d) normalization using linear transformation of shared variants across CRIPT cis double mutant library (y axis) and PDZ3-CRIPT double mutant trans library e) all-by-all binding fitness heatmap of all double mutants f) additive trait coefficients and g) residuals from fitting MoCHI two-state model h) distribution of variants with STOP codons and justification for non-binding/dead variants threshold (2 s.d. away from mean of STOP variants).

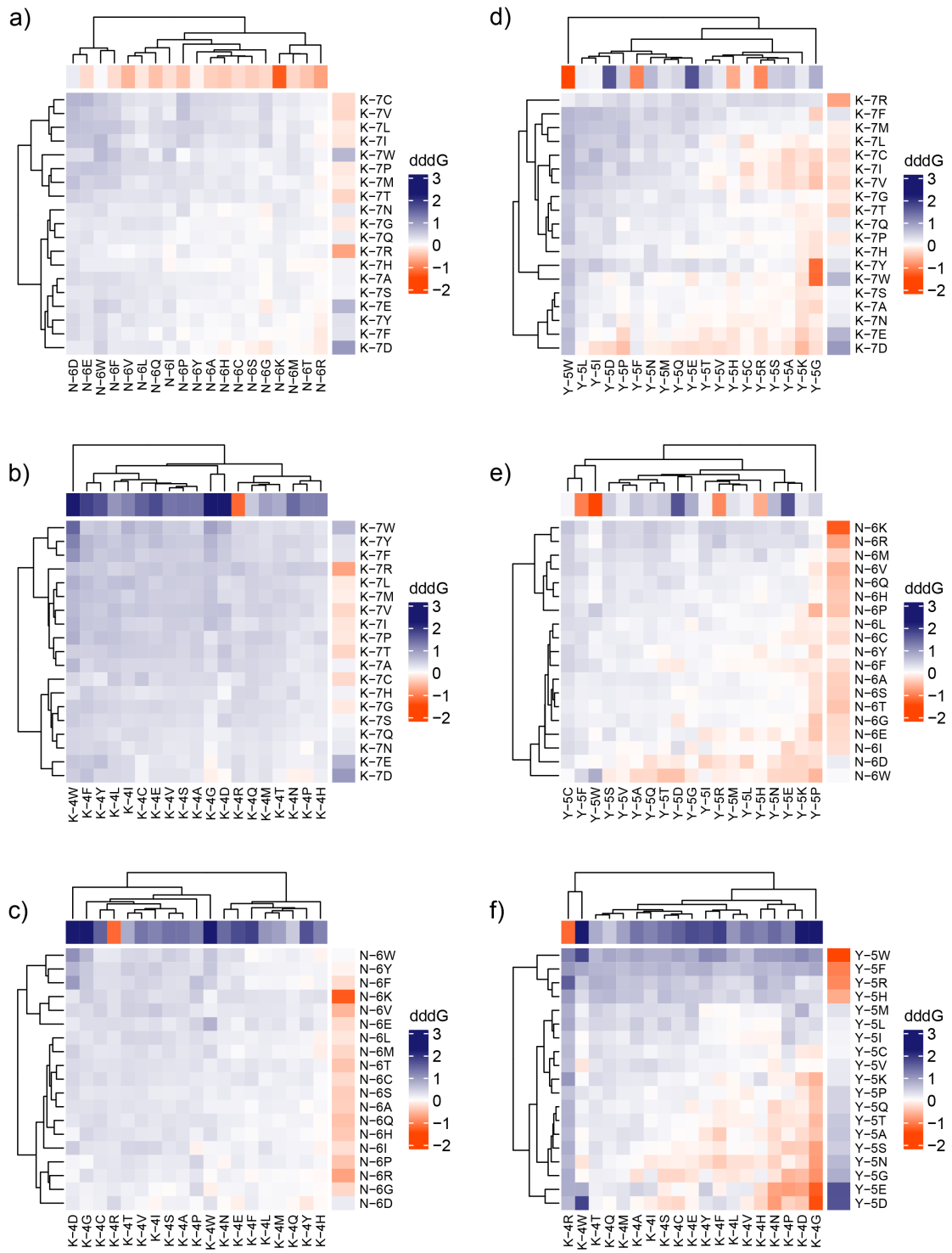

Figure S9. 1st and 2nd order interactions for CRIPT N as determined by MoCHI[31]. a)-f) Clustered heatmap of 2nd order interactions (dddGs) for CRIPT N combinatorial dataset with first order ddG terms for the individual mutations plotted on the x and y axes.

Supplementary tables:

Rank	Pair ID	CRIPT N vs. C	FDR (enrichment in number of spec. changing mutations across all pairs)	Number of spec. changing mutations	PDZ3 residue	CRIPT residue	Distance between residues (Å)
1	372_-2	C	2.03E-125	74	H372	T-2	2.715445
2	326_-1	C	1.49E-83	57	N326	S-1	3.593695
3	339_-3	C	1.22E-38	32	S339	Q-3	2.543977
4	324_-1	C	7.82E-14	16	G324	S-1	7.03372
5	326_-3	C	1.32E-08	12	N326	Q-3	3.08644
6	379_0	C	7.58E-06	9	L379	V0	4.245281
7	376_-2	C	3.74E-05	8	A376	T-2	3.924201
8	329_-2	C	0.000116	7	G329	T-2	9.165152
9	331_-4	N	0.000116	8	E331	K-4	3.126404
10	339_-1	C	0.000116	8	S339	S-1	8.3229
11	345_-1	C	0.000116	8	G345	S-1	11.18858
12	342_-1	C	0.000121	8	L342	S-1	7.746873
13	343_0	C	0.000607	7	A343	V0	12.59954
14	325_-1	C	0.007777	6	F325	S-1	7.572008
15	328_-5	N	0.007777	6	V328	Y-5	8.612266
16	330_-2	C	0.025456	5	G330	T-2	9.862768
17	336_-2	C	0.028273	5	I336	T-2	8.526768
18	340_-3	C	0.033814	5	F340	Q-3	4.862758
19	330_-4	N	0.057264	5	G330	K-4	7.177915
20	373_-4	N	0.076377	5	E373	K-4	2.670947

Table S1. Major (FDR<0.1, hypergeometric test) energetically coupled pairs of sites that control specificity

plasmid	description	link	use	source
pGJJ001	bindingPCA vector with DHFR3 and DHFR1,2 fused to nothing (empty bait and prey sites)	https://benchling.com/s/seq-FRMVd55qhXRSdnCerIB4?m=slm-HFeYB6ebmq6hKEVlXJ7S	Vector for CRIPT single mutant library combined with PDZ3 block 1 and block 2 libraries	Faure, Domingo, Schmiedel et al. Nature 2022[34]
pGJJ215	bindingPCA vectors with DHFR3 fused to wildtype PDZ3, empty "bait" for inserting CRIPT libraries	https://benchling.com/s/seq-YgJTHoCogB748M3lvrz1?m=slm-96kiUpsaHHKMnzOCiKvl	Vector for CRIPT N and C combinatorial libraries, CRIPT cis double mutant library	JDE/Lehner lab
pGJJ211	DHFR3 fused to wildtype PDZ3	https://benchling.com/s/seq-ZLKW8KZ0LU1BsoN88fBn?m=slm-pK7yOttj3EJwRkSJzK98	Vector for PDZ3 block 1 and block 2 libraries	JDE/Lehner lab

pGJJ518	bindingPCA vector with wildtype PDZ3, wildtype CRIPT	https://benchling.com/s/seq-5EQ0SO1NzUXy4soExGvc?m=slm-maaM0f6pwkJWq60vFxQR	Wildtype control (used to obtain empirical estimates of sequencing error, quantify amount of plasmid in genomic DNA extractions and test bindingPCA assay)	This work

Table S2. Plasmid sequences used in this study

amplicon	link	notes
CRIPT_N / CRIPT_C / CRIPT_CIS_DOUBLE	https://benchling.com/s/seq-ZjVNfyB4q0j7ghLbVvos?m=slm-tekXiMuhgKyP7qCWdIhJ	Reference amplicon for CRIPT N, CRIPT C combinatorial libraries and CRIPT CIS double mutant library
PDZB1_CRIPT	https://benchling.com/s/seq-m7a9OYYNBzp1HrSv0cgQ?m=slm-j3wlHLONBYjdpZy7uZVO	Reference amplicon for PDZ3-CRIPT trans double mutant library -- block 1 of PDZ3
PDZB2_CRIPT	https://benchling.com/s/seq-5krvszGJdAW94riW4Pmi?m=slm-b0Ku0YyzgXMeTFhjQ54N	Reference amplicon for PDZ3-CRIPT trans double mutant library -- block 2 of PDZ3

Table S3. Amplicon sequences used in this study

name	sequence	forward_reverse	library_1	library_2	library_3
oTZ132	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGTGGAGGCGGATCCACC	forward_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ134	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNAGGTGGAGGCGGATCCACC	forward_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ135	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNAGGTGGAGGCGGATCCACC	forward_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ136	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCNBAGGTGGAGGCGGATCCACC	forward_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ137	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTBHYAGGTGGAGGCGGATCCACC	forward_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ138	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAHYYAGGTGGAGGCGGATCCACC	forward_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ133	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTtcgaaggctttaatttgaCTAGTCTA	reverse_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ139	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNtcgaaggctttaatttgaCTAGTCTA	reverse_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ140	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNtcgaaggctttaatttgaCTAGTCTA	reverse_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE

oTZ141	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTRNWtcgaaggctttaatttgaCTAGTCTA	reverse_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ142	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCRSNtcgaaggctttaatttgaCTAGTCTA	reverse_CRIPT	CRIPT_N	CRIPT_C	CRIPT_CIS_DOUBLE
oTZ154	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGTGGAGGCGGATCCACC	reverse_CRIPT	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ155	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNAGGTGGAGGCGGATCCACC	reverse_CRIPT	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ156	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNAGGTGGAGGCGGATCCACC	reverse_CRIPT	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ157	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCNBAGGTGGAGGCGGATCCACC	reverse_CRIPT	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ158	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTBHYAGGTGGAGGCGGATCCACC	reverse_CRIPT	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ159	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAHYYAGGTGGAGGCGGATCCACC	reverse_CRIPT	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ160	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGGGAGGTGGAGCTAGC	forward_PDZ3B1	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ161	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNTCGGGAGGTGGAGCTAGC	forward_PDZ3B1	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ162	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNTCGGGAGGTGGAGCTAGC	forward_PDZ3B1	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ163	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCNWTCGGGAGGTGGAGCTAGC	forward_PDZ3B1	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ164	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAWBRTCGGGAGGTGGAGCTAGC	forward_PDZ3B1	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ165	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTBANWTCGGGAGGTGGAGCTAGC	forward_PDZ3B1	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ166	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCAGACCTCAGTGGGGAG	forward_PDZ3B2	PDZ3CRIPT_TRANS_DOUBLE	NA	NA

oTZ167	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNGCAGACCTCAGTGGGGAG	forward_PDZ3B2	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ168	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNWGCAGACCTCAGTGGGGAG	forward_PDZ3B2	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ169	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNWGCAGACCTCAGTGGGGAG	forward_PDZ3B2	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ170	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCSNWGCAGACCTCAGTGGGGAG	forward_PDZ3B2	PDZ3CRIPT_TRANS_DOUBLE	NA	NA
oTZ171	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTANNTGCAGACCTCAGTGGGGAG	forward_PDZ3B2	PDZ3CRIPT_TRANS_DOUBLE	NA	NA

Table S4. List of frameshifting oligonucleotides used for PCR1