Determining protein structures using genetics

Jörn M. Schmiedel[1], Ben Lehner[1-3*]

[1] Systems Biology Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain

[2] Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain

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

*email: ben.lehner@crg.eu

Summary

Determining the three dimensional structures of macromolecules is a major goal of biological research because of the close relationship between structure and function. Structure determination usually relies on physical techniques including x-ray crystallography, NMR spectroscopy and cryo-electron microscopy. Here we present a method that allows the high-resolution three-dimensional structure of a biological macromolecule to be determined only from measurements of the activity of mutant variants of the molecule. This genetic approach to structure determination relies on the quantification of genetic interactions (epistasis) between mutations and the discrimination of direct from indirect interactions. This provides a new experimental strategy for structure determination, with the potential to reveal functional and *in vivo* structural conformations at low cost and high throughput.

Introduction

Mutations within a protein or RNA can have non-independent effects on fitness [1-4]. Indeed, the effects of double mutants have long been used to probe the energetic couplings between positions in a protein to understand determinants of protein folding and stability [5,6]. Early work revealed that at least some strongly interacting positions within a protein are in direct structural contact [5-8]. Deep mutagenesis of proteins [9-12] and RNAs [13-16] has further confirmed this conclusion that some – but by no means all – genetic (or epistatic) interactions occur between structurally proximal mutations.

Further support for the idea that non-independence between mutations provides structural information comes from the analysis of amino acid and nucleotide sequence evolution. Here, correlated pairs of amino acids or nucleotides in multiple sequence alignments identify co-evolving positions within proteins and RNAs [17-19]. These patterns of co-evolution have been used to identify energetically-coupled positions and 'sectors' within proteins [20,21]. Moreover, when very large numbers of homologous proteins and RNAs are avaiable in sequence databases, the application of global statistical models has proven sufficient to discriminate direct structural contacts from patterns of co-evolution [22-24], allowing the prediction of macromolecular structures and interactions [25-34].

Could epistatic interactions quantified from deep mutational scanning experiments be used to determine macromolecular structures? If successful, structure determination by deep mutagenesis would offer a number of advantages over established techniques. First, it requires no specialized equipment or expertise beyond the ability to mutate a molecule, select functional variants, and quantify enrichments by sequencing. Appropriate *in vitro* and *in vivo* selection assays already exist for very many molecules of interest and generic assays based on folding, stability, and physical interactions have also been developed [9,35-38]. Second, it could be applied to molecules whose structures are difficult to determine by physical techniques such as intrinsically disordered and membrane proteins. Third, unlike evolutionary coupling analysis there is no requirement for large numbers of homologous sequences and so it could be applied to fast-evolving, recently-evolved and *de novo* designed proteins and RNAs [26,32,39]. Finally, and perhaps most importantly, it would provide a general strategy to determine the physiologically relevant structures of molecules whilst they are performing particular functions that can be selected for, including *in vivo* within cells. A cheap and straightforward approach for studying macromolecular structures *in vivo* would be a very exciting new frontier for cell biology.

Here we show that deep mutational scanning (DMS) of proteins can provide sufficient information to determine their high-resolution three-dimensional structures. Our statistical approach quantifies how often mutations between positions interact epistatically and how such epistatic interaction patterns correlate. These metrics accurately identify individual tertiary structure contacts as well as secondary structure elements within a protein. The same approach also identifies contacts between protein interaction partners. DMS data alone suffices to determine protein structures with accuracies down to 1.9Å backbone root mean square deviation (RMSD) compared to known reference structures. Moreover, we show that deep learning can further improve prediction performance, allowing the use of much sparser and lower quality DMS datasets for structure determination. This approach therefore provides a new experimental strategy for structure determination that can reveal functional and *in vivo* structural conformations at low cost and high throughput.

Results

Epistasis is enriched in but not exclusive to structural contacts

To investigate how genetic – or epistatic – interactions between mutations in a protein relate to structure we first used deep mutational scanning data for the immunoglobulin-binding protein G B1 domain (GB1) generated by Olson, et al. [11]. This dataset is the most complete double mutant deep mutagenesis of a protein domain reported to date and was generated by replacing each of 55 residues of the wild-type domain with all 19 alternative amino acids both individually and as double mutant pairwise combinations, resulting in a library of more than half a million variants (55*19 = 1,045 single mutants plus nearly 55*54/2*19*19 = 536,085 double mutants). mRNA display was used to combine an *in vitro* immunoglobulin G binding assay with a sequencing readout to determine protein fitness via changes in variant frequencies in the library before and after binding (Extended Data Figure 1, steps 1-4); resulting in a two orders of magnitude measurement range with a median relative error of fitness estimates of 2.8% (see Figure 2A, Table 1 and Methods).

We first computed which double mutant variants show epistatic fitness effects, i.e. non-independent fitness effects of the constituting single mutant variants (Figure 1B). Non-specific dependencies between mutants might be introduced by non-linearities in the fitness assay, systematic biases in error magnitudes as well as non-specific epistatic behavior, e.g. from thermodynamic stability effects [1,9]. We thus applied a non-parametric null model - the running median of double mutant fitness values given the constituting single mutant fitness values - for the independence of mutations. Equivalently, we calculated 5th and 95th percentile fitness surfaces; and classified double mutants with fitness lower than the 5th percentile as negative epistatic and double mutants with fitness higher than the 95th percentile as positive epistatic. We restricted the evaluation of positive or negative epistasis, however, to specific subsets of the data, where measurement errors do not impede epistasis classification (Extended Data Figure 2C, see Methods), which results in about 80% and 55% of double mutants being suitable for positive or negative epistasis classification, respectively, with a lot of variability across the position matrix (Extended Data Figures 2D-F and Table 1).

Consistent with previous observations [10-12], both positive and negative epistatic double mutants are enriched for proximal variants, for example, more than 2-fold at 8Å distance (side-chain heavy atom minimal distance, Figure 1C). However, about 75% of epistatic interactions are

between positions that are not in direct contact in the tertiary protein structure (as judged by an 8Å distance cutoff), suggesting that indirect effects often underlie epistatic interactions within a molecule [20,21]. The challenge for structure determination therefore becomes how to infer direct structural contacts from the mixture of direct and indirect effects that must underlie epistasis.

Figure 1

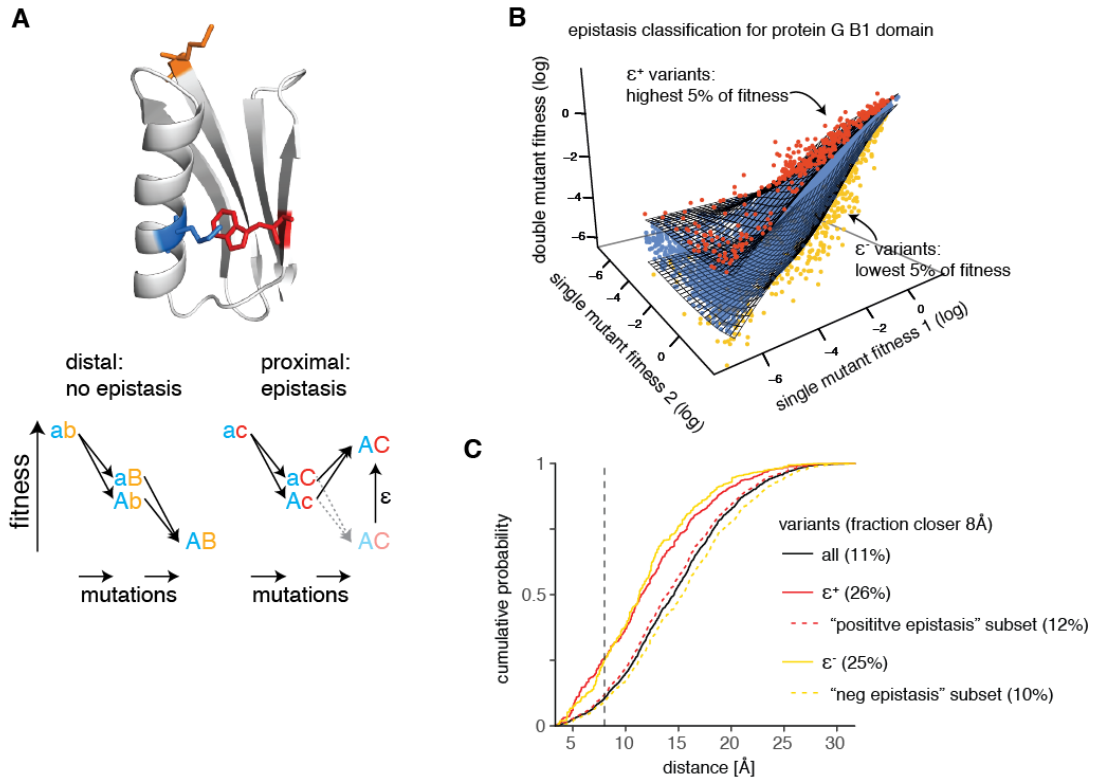


Figure 1: Extracting epistatic mutational effects from deep mutational scanning of a protein domain

A. Premise: If epistatic interactions relate to structural contacts then quantifying epistatic interactions should suffice to predict a molecule's structure. Structure: protein G B1 domain (PDB entry: 1pga) with residues a, b, and c colored.

B. Classifying epistatic variants based on deviations from expected fitness (based on quantile fitness surface approach). Variants above the 95th or below the 5th percentile of double mutant fitness given their single mutant fitness values were classified as positive (red, ε^{+}) or negative (yellow, ε^{-}) epistatic, respectively. Shown is a random sample of 10^4 variants in GB1 domain [11].

C. Distance distribution of epistatic variants separated by more than 5 amino acids in the linear sequence. (side-chain heavy atom distance in reference structure). Positive and negative epistasis subsets refer to the sets of variants applicable for epistasis analysis (see Extended Data Figure 2C). All variants, n = 400647; positive epistatic variants ε^{+}, n = 14127; positive epistasis subset, n = 315862; negative epistatic variants ε^{-}, n = 9837; negative epistasis subset, n = 208442.

Aggregated epistatic interactions predict tertiary structure contacts

To distill direct contacts from a list of thousands of epistatic double mutants we first aggregated epistatic information on the amino acid position-pair level by calculating the fraction of positive or negative epistatic double mutant variants per position pair (Figure 2A).

In the GB1 epistasis dataset even moderate enrichments for positive and negative epistatic variants are mutually exclusive (Extended Data Figure 3A). Moreover, the strongest positive and negative epistatic enrichments are separated in two clusters of proximal positions in the protein that exhibit mostly either positive or negative interactions among themselves, but hardly any epistatic interactions between clusters (Figures 2B,D and Extended Data Figure 3B), as also noted before by Olson, et al. [11].

Consistent with epistatic interaction clusters forming a dense network of proximal positions, we find that, of the top 55 epistatic pairs, 42% and 35% are direct contacts (connected by one edge smaller than 8 Ångström (Å), 3.9 and 3.2-fold over expectation) and another 45% and 55% share a common neighbor (connected via two edges < 8Å), for positive and negative epistatic interactions respectively, while interactions across more edges are depleted (Figure 2C; throughout the manuscript we only consider position pairs spaced by more than 5 amino acids in the linear sequence; closer positions are trivially also close in 3D space, and their proximity contributes little to successful structure prediction [28]).

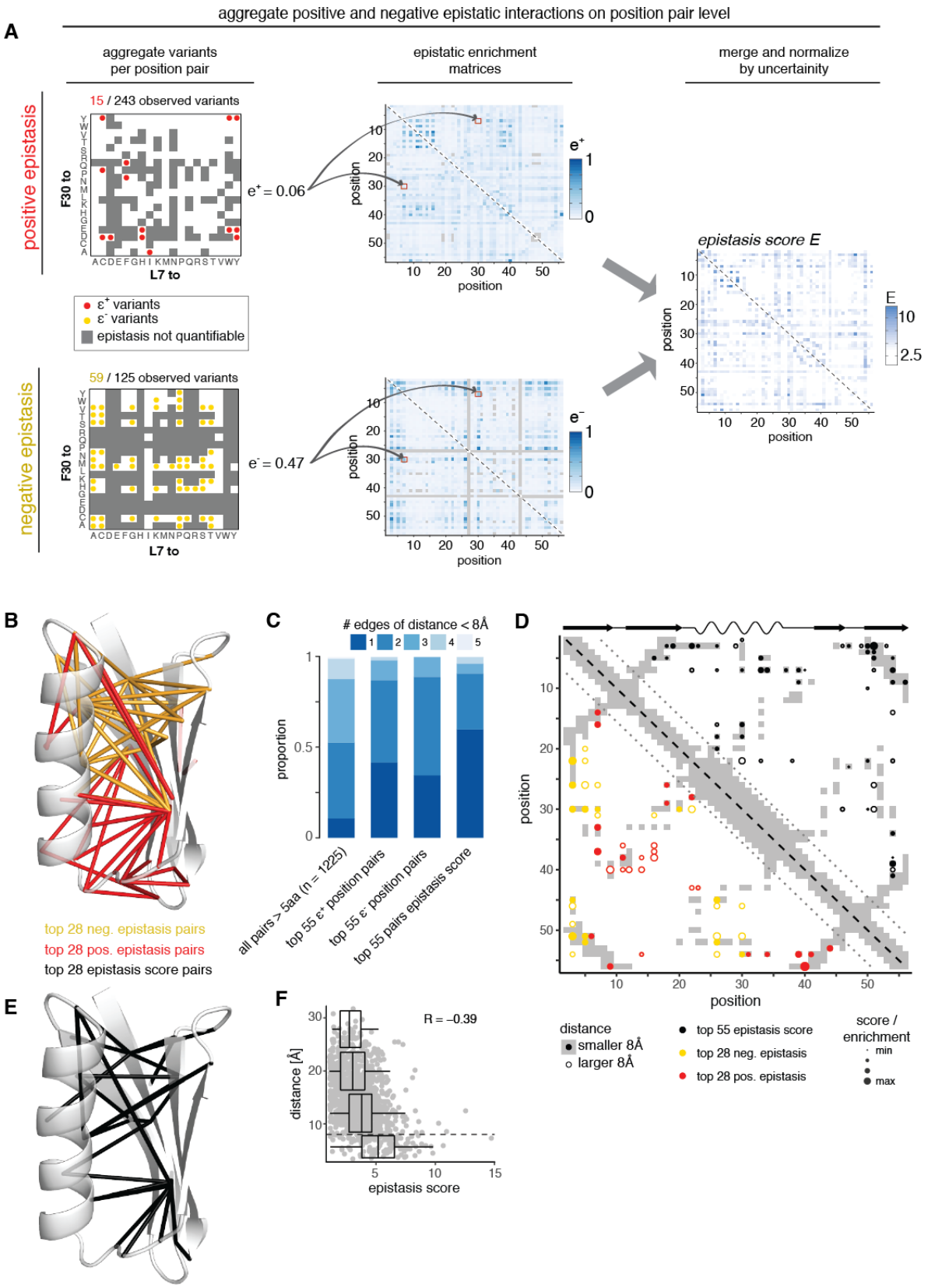
While aggregation of epistatic information between position pairs thus better discriminates structural contacts than individual epistatic interactions, positive and negative epistatic interactions still contain disparate structural information of the protein domain. We therefore merged positive and negative epistatic information by computing the weighted averages of

145 epistatic fractions per position pair given their uncertainty due to fitness measurement errors
146 and the finite number of observed double mutant variants via a resampling approach (Extended
147 Data Figure 1 and Methods). A final *epistasis score* per position pair was obtained by
148 normalizing these weighted averages by their uncertainty (a z-score), thus giving priority to
149 position pairs with high confidence enrichments.

150 The position pairs with highest *epistasis scores* are well distributed across the domain (Figures
151 2D,E), and the number of direct contacts (one edge < 8Å) among the top 55 *epistasis score*
152 pairs increases to 60% (Figure 2D), thus showing that the *epistasis score* successfully
153 incorporates information from both positive and negative epistasis to discriminate direct
154 contacts. Moreover, direct contacts as a whole are enriched for high *epistasis scores*, while
155 further away position pairs show a gradual decrease of *epistasis scores* (Pearson correlation
156 coefficient $R = -0.39$, $p < 10^{-6}$, Figure 2F).

157 Thus, although many interactions are indirect, physical contacts are an important determinant of
158 epistasis and aggregating information on position pairs and merging positive and negative
159 epistasis information better discriminates these direct structural contacts across the protein
160 domain.

Figure 2


A
aggregate positive and negative epistatic interactions on position pair level
aggregate variants per position pair
epistatic enrichment matrices
merge and normalize by uncertainity
positive epistasis
15 / 243 observed variants
F30 to
L7 to
e+ = 0.06
e+
1
0
position
ε+ variants
ε- variants
epistasis not quantifiable
negative epistasis
59 / 125 observed variants
e- = 0.47
e-
epistasis score E
E
10
2.5
B
top 28 neg. epistasis pairs
top 28 pos. epistasis pairs
top 28 epistasis score pairs
C
edges of distance < 8Å
proportion
all pairs > 5aa (n = 1225)
top 55 ε+ position pairs
top 55 ε- position pairs
top 55 pairs epistasis score
D
distance
smaller 8Å
larger 8Å
top 55 epistasis score
top 28 neg. epistasis
top 28 pos. epistasis
score / enrichment
min
max
E
F
R = −0.39
distance [Å]
epistasis score

161

Figure 2: Aggregated epistasis scores enrich for direct structural contacts

A. Workflow for aggregating positive and negative epistatic interactions on the position-pair level and merging them into a final *epistasis score*.

B. Top 28 position pairs (> 5 amino acids in linear sequence) each with highest positive (red) and negative (yellow) epistatic fractions marked on the reference structure (PDB entry 1pga).

C. Minimal number of edges (contact with distance < 8Å in reference structure) connecting position pairs. One edge – positions are direct contacts, two edges – positions have a common contact and so forth.

D. Interaction score map for top 28 position pairs each with highest positive (red) and negative (yellow) epistatic fractions (lower left triangle) and top 55 position pairs with highest *epistasis score* (upper right triangle). Dot size indicates relative epistatic enrichments or score; dot fill indicates distance below 8Å. Underlying in grey is the contact map of the reference structure (PDB entry 1pga, distance < 8Å) and shown on top its secondary structure elements (wave – alpha helix, arrow – beta strand).

E. Top 28 position pairs (> 5 amino acids in linear sequence) with highest *epistasis scores* marked on the crystal structure (PDB entry 1pga).

F. Distance of position pairs as a function of *epistasis scores*. Boxplots are spaced in distance intervals [0,8), [8,16), [16,24) and [24,32) Å. Dashed horizontal line indicates 8Å. Pearson correlation coefficient is indicated.

Tertiary structure neighborhood leads to correlated epistatic patterns

If epistasis arises mainly from structural interactions, a position's epistatic interaction profile with all other positions in the protein should provide a signature of its structural location (Figure 3A). Comparing these signatures between positions should thus reveal structurally close positions - similar to how correlated epistasis profiles in genetic interaction networks serve to identify physical and functional interaction partners [40].

190 To test the idea that pattern correlation should reveal structural proximity, we calculated the
191 correlations between the epistatic enrichment vectors for all position pairs (Figure 3B).
192 Consistently, pair-distances and similarity of epistasis patterns between positions are strongly
193 correlated (Pearson correlation coefficient = -0.43, $p < 10^{-6}$, n = 1225, Figure 3D). Top
194 correlated pairs from positive or negative interaction patterns do, however, form mutually
195 exclusive clusters within the protein domain that are nearly identical to the clusters observed for
196 direct positive and negative interactions (Figure 3C and E, c.f. Figure 2B). Thus, while
197 correlations of epistatic interaction patterns are a good indicator of distance within the protein
198 structure, they suffer from the same issues as epistatic enrichments, namely poor discrimination
199 of direct and indirect interactions and disparate structural information.

200 We reasoned that partial correlations - the association between two positions after accounting
201 for the global correlation structure - might provide the possibility to eliminate the dependencies
202 observed in the epistasis pattern structure and thus help to distinguish direct from indirect
203 contacts; similar to how mean-field approaches can help discriminate direct from indirect
204 evolutionary couplings in multiple sequence alignments [22,28,41]. We derived partial correlations
205 by inversion of the correlation matrices, merged values from positive and negative epistatic
206 patterns by their estimated uncertainty, and ranked these merged values by their z-scores,
207 which we refer to as *association scores* (Figure 3B and Methods).

208 In contrast to the correlation of epistasis patterns, partial correlation of epistasis patterns for
209 both positive and negative epistasis display no clustering but are well distributed across the
210 whole protein domain, consistent with partial correlations removing dependencies between
211 correlated pairs (Figure 3C). Moreover, the merged *association scores* are less well correlated
212 with pair-distance (Pearson correlation coefficient R = -0.26, $p < 10^{-6}$, n = 1225) and show a
213 more binary all-or-none response, with most distant position pairs having an *association score*
214 around 0 and only proximal pairs systematically deviating to higher values (Figure 3D).
215 Moreover, the top pairs involve many different individual positions and are well distributed
216 across the protein domain (Figure 3C and E). Thus, *association scores* are able to prioritize
217 direct over indirect structural contacts across the whole protein domain.

218

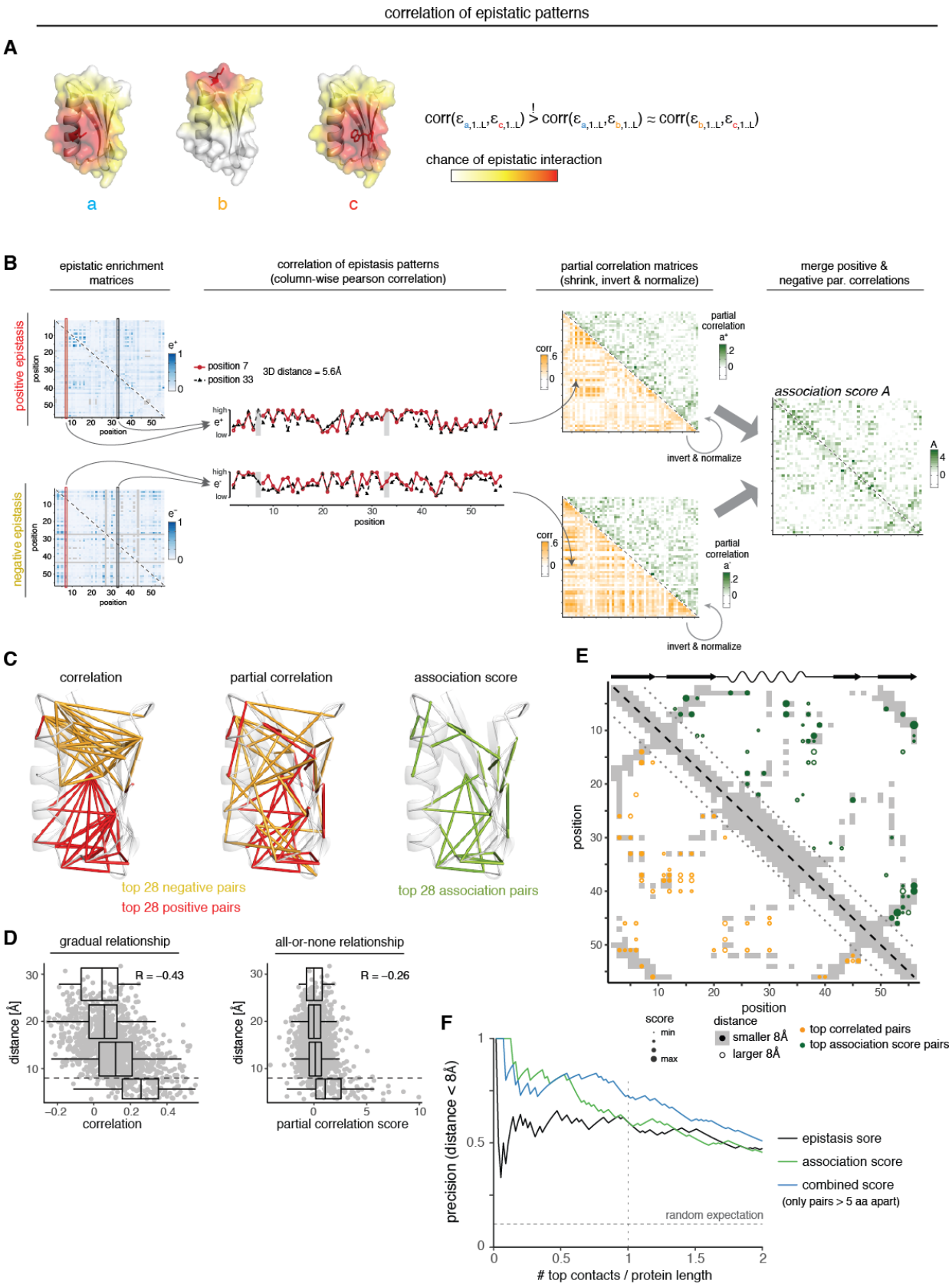
Combining *epistasis* and *association* scores better discriminates structural contacts

We derived a *combined score* by summing the standardized *epistasis* and *association scores*, to explore whether combining information from individual epistatic interactions and epistasis interaction patterns can improve proximity estimates; thereby prioritizing position pairs that are both enriched for direct epistatic interactions and have correlated epistasis patterns.

We evaluated the precision of the three interaction scores in predicting direct contacts in the protein domain. For all pairs separated by more than 5 amino acids in the linear sequence, the *epistasis score* has a roughly constant precision of around 60% across the first 2*L predicted contacts (L being the mutated length of the protein i.e. 55 amino acids). The *association score* has higher precision than the *epistasis score* up to the first L predicted contacts, with a precision of 79% at L/2 top contacts. Finally, the *combined score* has similar precision to the *association score* for the first L/2 contacts, but then remains at higher precision, with an improvement of about 10-15% over the individual scores at more predicted contacts (73% at L contacts); showing that combining information from epistatic interactions and interaction patterns further improves the discrimination of direct structural contacts.

Together, the derivation of the interaction scores demonstrates that it is possible to discriminate direct three-dimensional structural contacts from a mainly non-proximal set of epistatic interactions within a protein domain.

Figure 3


correlation of epistatic patterns
A
a
b
c
corr(ε_a,1..L, ε_c,1..L) > corr(ε_a,1..L, ε_b,1..L) ≈ corr(ε_b,1..L, ε_c,1..L)
chance of epistatic interaction
B
epistatic enrichment matrices
correlation of epistasis patterns (column-wise pearson correlation)
partial correlation matrices (shrink, invert & normalize)
merge positive & negative par. correlations
positive epistasis
negative epistasis
position 7
position 33
3D distance = 5.6Å
invert & normalize
association score A
C
correlation
partial correlation
association score
top 28 negative pairs
top 28 positive pairs
top 28 association pairs
D
gradual relationship
all-or-none relationship
R = −0.43
R = −0.26
distance [Å]
correlation
partial correlation score
E
position
score
min
max
distance
smaller 8Å
larger 8Å
top correlated pairs
top association score pairs
F
precision (distance < 8Å)
top contacts / protein length
epistasis sore
association score
combined score
(only pairs > 5 aa apart)
random expectation

238

Figure 3: Tertiary structure neighborhood leads to correlated epistatic patterns

A. Mutations in directly contacting residues should interact similarly with all other mutations in the protein. Thus, the similarity of epistasis patterns of two positions with all other positions in the protein should inform about their structural proximity.

B. Workflow for quantification of correlated epistasis patterns. Pairs of columns from epistatic enrichment matrices (here columns 7 and 33) are compared and their Pearson correlation coefficients are calculated, which constitute entries in the correlation matrix (here entries 7:33 and 33:7, due to matrix symmetry). Correlation matrices are inverted to yield the partial correlation matrices. Finally, entries of the positive and negative partial correlation matrices are merged (weighted average by uncertainty) and z-normalized to yield *association scores* (see Methods).

C. Top 28 position pairs (> 5 amino acids in linear sequence) marked on reference structure. Left: Top pairs from positive (red) or negative (yellow) epistasis pattern correlations. Middle: Top pairs after partial correlation transformation. Right: Top *association score* pairs (merge positive and negative partial correlations).

D. Distance of position pairs as a function of merged correlation (left) or *association scores*. Boxplots are spaced in intervals of 8Å. Dashed horizontal line indicates 8Å. Pearson correlation coefficient is indicated.

E. Interaction score map for top 55 position pairs with highest merge correlation (positive and negative correlations merged, lower left triangle, orange) and *association scores* (upper right triangle, green). Dot size indicates relative correlations or scores; dot fill indicates distance below 8Å. Underlying in grey is the contact map of the reference structure (PDB entry 1pga, distance < 8Å) and shown on top its secondary structure elements (wave – alpha helix, arrow – beta strand).

F. Precision of interaction scores to predict direct contacts (distance < 8Å in crystal structure 1pga) as a function of top scoring position pairs. Only position pairs with linear chain distance greater than 5 amino acids are considered (n = 1225 pairs, n = 131 direct contacts in reference structure). Horizontal dashed line indicates random expectation.

Periodic epistatic patterns reveal secondary structure arrangements

We investigated whether the periodic geometrical arrangement of amino acid residues in secondary structures results in periodic epistasis patterns [26,42]. Within an alpha helix with 3.6 residues per helical turn, a helical position would be predicted to interact epistatically with the third or fourth-over position along the linear amino acid chain (Figure 4A). Equivalently, within a beta strand, positions should interact epistatically with the next-but-one position.

We used a two-dimensional kernel smoothing approach to estimate the positions of alpha helices and beta strands from the deep mutational scanning data (Figure 4B). Here, the propensity of a position to belong to an alpha helix or a beta strand depends on whether it shows the expected periodicity in its interaction with neighboring positions, as well as whether neighboring positions display similar propensities for the same secondary structure element, and how strong these interactions are compared to those found in randomized data sets (see Methods).

We found that, while secondary structure element predictions derived from direct interaction-based *epistasis scores* are somewhat inaccurate and underpowered, predictions derived from correlation-based *association scores* (as well as *combined scores*) coincide very well with secondary structure elements in the reference structure (Figure 4C and Extended Data Figure 4C), with precision and recall values of about 90% (Extended Data Figure 4D). This suggests that the correlated profiles of epistatic interactions are informative about side chain orientations and also that eliminating transitive interactions is important for (secondary) structure prediction.

We further used two-dimensional kernel smoothing to detect parallel and anti-parallel beta sheet interactions, by applying beta strand kernels to off-diagonal entries on the interaction score matrices (Extended Data Figure 4A, see Methods). Several stretches of position pairs show the expected alternating interaction profiles for either parallel or anti-parallel beta sheets (Extended Data Figure 4B), with the top predictions corresponding to the known beta-sheet interactions in the reference structure (Figure 4D). Furthermore, updating beta strand predictions according to inferred beta sheet pairings can further improved beta strand prediction itself, notably introducing a correct split of beta strand 1 and 2 and adjusting the length of beta strands 3 and 4 (Figure 4C,D).

Together this shows that epistatic interaction data contains information on the periodic secondary structure of a protein domain and, *vice versa*, that secondary structure strongly influences genetic interactions.

Figure 4

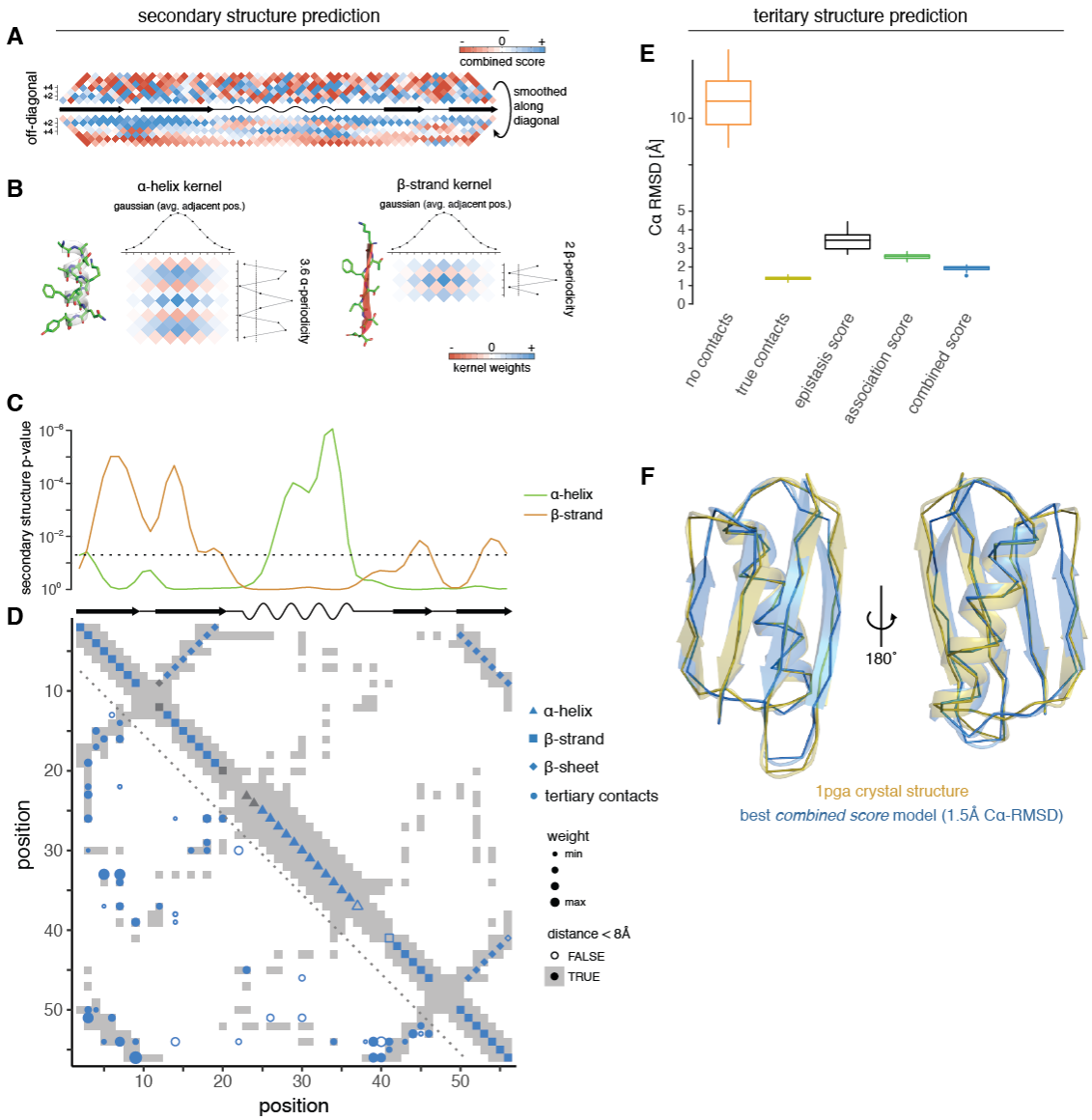


Figure 4: Secondary and tertiary structure prediction from deep mutational scanning data

A. Local interactions reveal signatures of secondary structure elements. Middle line is diagonal of interaction score map (rotated by 45 degree) and shows secondary structure elements of reference structure (PDB entry 1pga). Data above diagonal shows *combined score* data close to the diagonal, i.e. local interactions. Below the diagonal, the same data are smoothed with a Gaussian kernel along the direction of the diagonal (i.e. horizontally, length of Gaussian kernel as for kernels in panel b) to reveal periodicities in local interactions.

B. Two-dimensional kernels for alpha helix and beta strand detection. Kernel has a sinusoidal or alternating profile in the off-diagonal direction to detect alpha helices and beta strands propensities, respectively and a Gaussian profile along the diagonal, to average over propensities of adjacent positions.

C. Secondary structure propensity derived from kernel smoothing (orange – beta strand, green – alpha helix). P-values were derived by comparison to randomized datasets (see Methods). Dashed line indicates p = 0.05.

D. All structural predictions derived from *combined score* data. Lower left: Top 55 non-local (>5 aa in linear sequence) position pairs, i.e. tertiary contacts (circles); fill indicates correct prediction at 8Å, size of circles indicates relative score. Upper right: Predicted secondary structure elements (triangle – alpha helix, square – beta strand, diamond – beta sheet interaction). Fill indicates correct prediction. Note that beta strand predictions are derived by intersection of beta strand propensity (as shown in panel C) and results from beta sheet prediction (Extended Data Figure 4B, see Methods). Underlying in grey is the contact map of the reference structure (PDB entry 1pga, distance < 8Å) and shown on top are its secondary structure elements (wave – alpha helix, arrow – beta strand).

E. Accuracy ($C\alpha$ root-mean-square deviation) of top 5% structural models generated from deep mutational scanning data derived restraints compared to GB1 reference structure. Structural models were generated in XPLOR-NIH by simulated annealing with restraints derived from top 55 top scoring position pairs, secondary structure element prediction and beta sheet pairing predictions from the indicated interaction scores. No contacts – negative control with restraints only for secondary structure (predicted by PSIPRED)[43].

True contacts – positive control with 55 contacts (random subset), secondary structure elements and beta sheet interactions restraints derived from reference structure.

F. Overlay of top structural model generated with restraints from *combined score* (blue) and crystal structure (gold, PDB entry 1pga). Shown is backbone ribbon and secondary structure cartoon generated in PyMOL [44].

Protein structure determination by deep mutagenesis

Together, these findings show that deep mutagenesis data contain substantial information about a protein's secondary and tertiary structure. We therefore tested whether the data would suffice to determine *ab initio* the structure of protein G domain B1. We performed structural simulations by simulated annealing using the XPLOR-NIH modeling suite [45], with structural restraints derived from the deep mutational scanning data (see Methods). In particular, we defined distance restraints (distance < 8Å between Cβ atoms) for the top scoring position pairs; we found that using the top L (L = 55) contacts gave best results (Extended Data Figure 4F). Furthermore, we defined dihedral angle restraints for predicted secondary structure elements. Finally, we defined restrictive distance restraints (distance smaller than 2.1Å for N-H : C=O atom pairs) for beta sheet positions that form hydrogen bonds with each other.

We evaluated the top 5% of structural models (25/500, evaluation based on XPLOR internal energy terms) generated against the known crystal structure of protein G domain B1 (PDB entry 1pga) (Figures 4E and Extended Data Figure 4F). Models predicted from *combined score* data performed best, with an average Cα-root mean squared deviation of the top models ($\langle C\alpha - RMSD \rangle$) of 1.9Å and an average template modeling score of 0.71, which is very close to the optimum achievable with our simulation protocol (using contacts, secondary structure elements and beta sheet interactions from the reference structure, $\langle C\alpha - RMSD \rangle = 1.4\text{Å}$ and TM score = 0.8); and the top evaluated *combined score* structural model has a $C\alpha - RMSD$ of only 1.5Å (Figure 4F). Consistent with somewhat lower precision of contact and secondary structure predictions, models generated with restrains from *epistasis* or *association scores* have on average a lower accuracy ($\langle C\alpha - RMSD \rangle = 3.4\text{Å}$ and $\langle C\alpha - RMSD \rangle = 2.6\text{Å}$, respectively), with *association score* models performing consistently better (Figures 4E and Extended Data 4F).

Together, this shows that deep mutation scanning alone is sufficient to accurately determine the structure of a protein domain.

Contact prediction in additional protein domains

To test the generality of our approach, we analyzed two additional, incomplete deep mutational scanning datasets. First, a mutational scan of the 75 amino acid Pab1 RRM2 domain (Figure 5A), for which fitness was assessed in a complementation assay [10]. Second, a mutational scan of the hYAP65 WW domain (Figure 5C), in which 33 out of 50 amino acids were mutated and fitness was assayed by binding to a polyproline peptide ligand in a phage display assay [46]. Both datasets were created by 'doped' oligonucleotide synthesis and thus consist primarily of amino acid changes elicited by just one nucleotide change, which results in only 10% of possible double mutants being present. Additionally, their selection assays have smaller measurement ranges than that of the GB1 domain, which results in higher relative errors of fitness estimates as well as in negative epistasis being quantifiable for a smaller fraction of double mutants, as low as 0.8% in the case of the WW domain (Extended Data Figure 5A, see Table 1 for comparison of dataset properties).

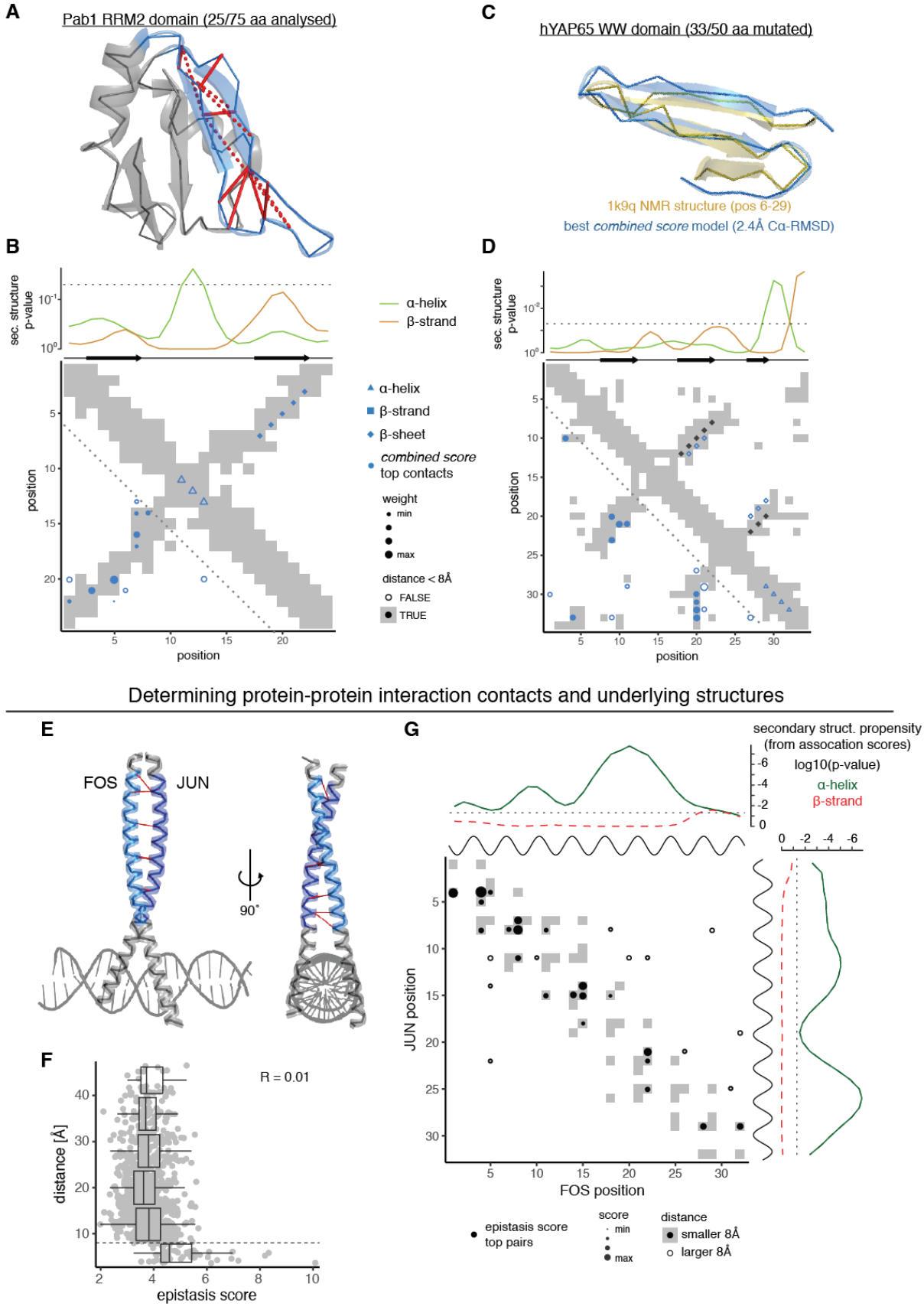
For the RRM domain, three 25 amino acid segments were mutated separately and we restricted our analysis to the central one, as it is the only segment that exhibits a reasonable number of intra-segment contacts in the reference structure (Figure 5A). We find that predicted tertiary contacts fall on or very close to known contacts in the region of the anti-parallel beta sheet and the intervening loop region (Figure 5B), with a precision of 57% for the top L/2 and 50% for the top L position pairs of the *combined score* (3-fold and 2.7-fold over expectation, respectively). Predicted beta strand propensities peak at the correct positions, albeit with low statistical significance; additionally, an alpha helical propensity is detected in the intervening loop region. Nonetheless, the correct anti-parallel beta sheet conformation at the correct position pairs is predicted.

For the WW domain, we find that top predicted tertiary contacts fall on or very close to known interactions between the beta strands and the N-terminal and C-terminal loop regions (Figure 5D), with a precision of 59% for the top L/2 and 38% for the top L position pairs of the *combined score* (3.9-fold and 2.5-fold over expectation, respectively). Secondary structure elements are not well predicted, but beta sheet interactions are predicted in the right anti-parallel conformation of $\beta 1 - \beta 2 - \beta 3$, though the exact pairing between positions is off by one to two positions. We determined the three dimensional structure of the secondary structure-rich central part of the domain (positions 6 to 29, 24 amino acids), using restraints derived from top *combined score* pairs and PSIPRED-predicted secondary structure elements (see Methods).

397 The top 5% of structural models have an average accuracy of 3.3Å $\langle C\alpha - RMSD \rangle$ compared to
398 the reference structure, which is on par with simulations using a set of 'true' contacts ($\langle C\alpha -$
399 $RMSD \rangle = 3.6$Å) (Extended Data Figure 5C). Moreover, the structural model with the best
400 XPLOR-NIH energy has an accuracy of 2.4Å $C\alpha - RMSD$ (or 2.0Å over 22 of the 24 residues)
401 (Figure 5C). Despite similar precision of predicted contacts, *association* and *combined score*-
402 derived WW domain structural models are more accurate than *epistasis score*-derived models
403 (Extended Data Figure 5C).

404 Together these results strongly support the generality of our approach for extracting structural
405 information from deep mutagenesis data, including from sparser and lower quality data.

Figure 5


A
Pab1 RRM2 domain (25/75 aa analysed)
C
hYAP65 WW domain (33/50 aa mutated)
1k9q NMR structure (pos 6-29)
best combined score model (2.4Å Cα-RMSD)
B
sec. structure p-value
α-helix
β-strand
position
α-helix
β-strand
β-sheet
combined score top contacts
weight
min
max
distance < 8Å
FALSE
TRUE
D
Determining protein-protein interaction contacts and underlying structures
E
FOS
JUN
90°
F
R = 0.01
distance [Å]
epistasis score
G
secondary struct. propensity (from assocation scores)
log10(p-value)
α-helix
β-strand
JUN position
FOS position
epistasis score top pairs
score
distance
smaller 8Å
larger 8Å

406

407 Figure 5 – Predicting structural contacts in two additional proteins and a
408 protein-protein interaction

409 A. Pab1 RRM2 domain structure (PDB entry 1cvj) with 25/75 positions analyzed here
410 highlighted in blue. Top 12 *combined score* position pairs are connected with red lines,
411 solid if distance < 8Å, dashed otherwise.

412 B. Structural predictions derived from *combined scores* in RRM domain. Upper plot shows
413 secondary structure propensities from kernel smoothing (p = 0.05 indicated as dashed
414 line). Just below are shown the secondary structure elements in the reference structure.
415 Map shows top 12 *combined score* position pairs in lower left and secondary structure
416 predictions in upper right triangle. Shape indicates type of prediction, fill indicates correct
417 prediction. Underlying is the contact map of the reference structure (grey if < 8Å).

418 C. Overlay of top structural model of hYAP65 WW domain (positions 6-29) generated with
419 restraints from *combined score* (blue) and solution NMR structure (gold, PDB entry
420 1k9q).

421 D. Structural predictions derived from *combined scores* in WW domain. Upper plot shows
422 secondary structure propensities from kernel smoothing (p = 0.05 indicated as dashed
423 line). Just below are shown the secondary structure elements in the reference structure.
424 Map shows top 17 *combined score* position pairs in lower left and secondary structure
425 predictions in upper right triangle. Shape indicates type of prediction, fill indicates correct
426 prediction. Underlying is the contact map of the reference structure. Black diamonds
427 indicate positions of beta sheet pairing in reference structure. Crystal structure of the
428 leucine zipper domains of FOS and JUN with a DNA strand (PDB entry 1fos). The
429 mutated regions (32 amino acids each) are highlighted in light blue (FOS) and dark blue
430 (JUN). Top 10 epistasis score pairs are shown with red dashes.

431 E. Distance of position pairs as a function of interaction scores. Boxplots are spaced in
432 distance intervals of 8Å. Dashed horizontal line indicates 8Å. Pearson correlation
433 coefficient is indicated.

434 F. FOS-JUN trans interaction score map for top 32 position pairs with highest *epistasis*
435 *scores*. Note that protein-protein interaction maps are not symmetric. Dot size indicates
436 relative score; dot fill indicates distance below 8Å; underlying in grey is the contact map
437 of the reference structure (PDB entry 1fos, distance < 8Å). Shown on top and to the right

438 of the contact map are the known alpha helices and secondary structure propensities
439 derived from *association scores* of FOS and JUN, respectively (black – known alpha
440 helix; green – predicted alpha helix propensity, orange - predicted beta strand
441 propensity; see Extended Data Figures 5F,G).

442 Contact prediction in a protein-protein interaction

443 Genetic interactions do not only occur between mutations within individual proteins but also
444 between molecules that physically interact [2]. We investigated a deep mutational scanning
445 dataset of the coiled-coil interaction between the proteins encoded by the proto-oncogenes *FOS*
446 and *JUN* (Figure 5E) [9]. In this experiment, all possible single amino acid changes were made in
447 each of 32 positions of each protein and the physical interaction of all single and (trans-)double
448 mutants was quantified using a deep sequencing-based protein complementation assay. After
449 filtering, the dataset contains 43% of all possible double mutants and has a median relative
450 error of fitness measurements of 3.6% (Table 1).

451 When assessing the enrichment of epistatic interactions between positions in the two interaction
452 partners we find a striking all-or-none relationship between *epistasis scores* and pair-distances
453 (Figure 5F), with all distant pairs contained in a low *epistasis score* peak and only proximal
454 interactions enriched for high *epistasis scores* (Pearson correlation coefficient R = 0.01, n =
455 1024). Indeed, the top 11 *epistasis score* pairs are all proximal interactions, and the precision of
456 contact prediction is 75% for the top L/2 contacts and 66% for the top L contacts (12-fold and
457 10.5-fold over expectation). Moreover, top *epistasis score* pairs are evenly distributed across
458 the interaction surface (Figures 5E and 5G).

459 When correlating epistatic patterns between columns of the epistatic enrichment matrices, one
460 is comparing the epistatic interactions that two positions in FOS have with all positions in JUN.
461 Therefore, the similarity of column-wise epistatic patterns reveals the *cis* relationships between
462 positions in FOS (Extended Data Figure 5F). Similarly, correlating epistasis patterns across
463 combinations of rows of the epistatic enrichment matrices reveals *cis* relationships between
464 positions in JUN. We find that cis-interaction maps from *association scores* for both FOS and
465 JUN are highly enriched for strong local interactions with an alpha helical periodicity; and
466 applying our secondary structure prediction algorithms to the *cis*-interaction maps reveals strong
467 alpha helix propensities across the full lengths of both FOS and JUN (Figures 5G and Extended
468 Data Figure 5G).

469 This shows that deep mutagenesis of protein interaction partners can accurately predict direct
470 contact across the interaction surface as well as the underlying structures of the interaction
471 partners themselves.

472

473 Deep learning improves contact prediction

474 Evolutionary coupling-based structural predictions have been successfully improved by machine
475 learning approaches that transform the two-dimensional interaction score maps after learning
476 the stereotypical patterns between evolutionary coupling-predicted contact maps and the actual
477 contact maps of the known structures [47,48].

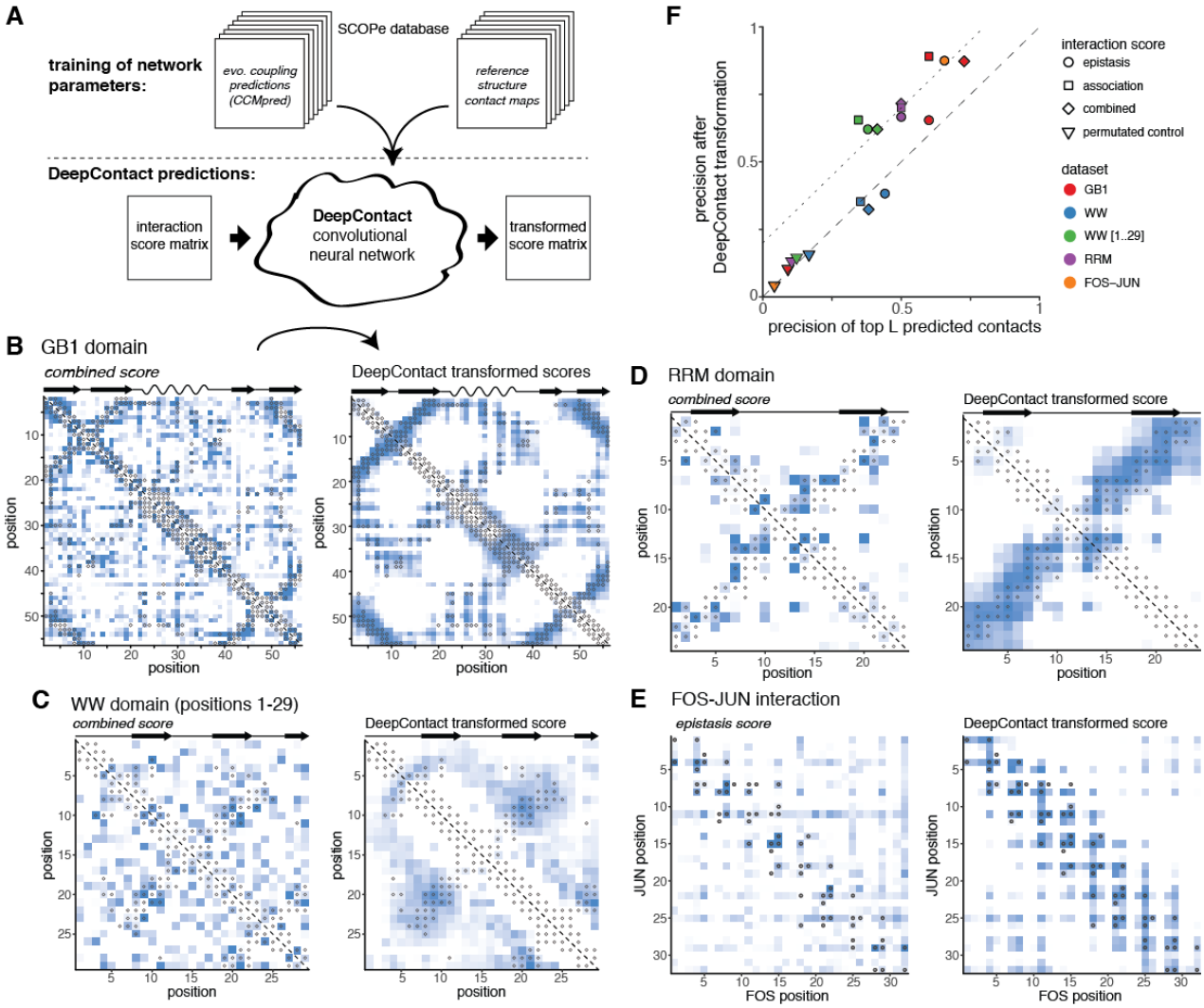
478 We tested whether such an approach can also improve deep mutagenesis-derived contact
479 predictions. We applied a convolutional neural network approach called *DeepContact*,
480 developed by Liu, et al. [47]. The basic *DeepContact* architecture takes as a sole input a two-
481 dimensional interaction score map that it then transforms based on the structural patterns it has
482 previously learned on evolutionary coupling-derived contact predictions for representative
483 families of the SCOPe database [49] (Figure 6A and Methods). When transforming evolutionary
484 coupling-derived contact predictions of proteins not contained in the training set, this basic
485 *DeepContact* architecture has been shown to improve contact prediction precision by about 10-
486 20% [47].

487 We first transformed the GB1 domain *combined score* interaction map with the DeepContact
488 network (Figure 6B). These transformations take as sole input our deep mutational scanning-
489 derived predictions and include no evolutionary coupling or otherwise-derived structural
490 predictors for GB1. The scores on the transformed map are much less noisy, with high scores
491 exclusively focused in areas of known contacts, especially those of secondary structure element
492 interactions, and areas devoid of contacts showing homogenously low scores. Moreover, the
493 transformed scores are highly correlated with pair-distances in the reference structure (Pearson
494 correlation coefficient $R = -0.68$, $p < 10^{-6}$, $n = 1225$, Extended Data Figure 6A). The precision of
495 top predicted contacts improves from 82% to 96% for L/2 and from 73% to 87% for L predicted
496 contacts (Figure 6F). *Epistasis score*-derived predictions improve by about 5%, while
497 a*ssociation score*-derived predictions improve by 29% at L predicted contacts. In contrast,
498 randomized interaction score maps show no changes in prediction performance over random
499 expectation after transformation with DeepContact.

Interaction score maps for the other datasets show similar improvements to GB1 both in terms of cleaner interaction score maps that resemble the reference contact maps as well as increases in contact prediction precision of up to 30% (Figure 6 C-F).

This shows that machine learning can substantially improve contact map prediction from deep mutational scanning data, allowing the use of sparser and lower quality data for accurate prediction.

Figure 6



508 Figure 6: Deep learning improves contact prediction from deep
509 mutagenesis data

510 A. DeepContact convolutional neural network transforms DMS-derived interaction score
511 maps based on learned structural patterns [47]. The particular DeepContact architecture
512 used here takes as only input the DMS-derived interaction score map and transforms it
513 based on structural patterns previously learned on an orthogonal and independent
514 training set (in which it compared evolutionary coupling-derived contact predictions with
515 contacts in known structures of representative protein families in the SCOPe database).

516 B. GB1 domain *combined score* interaction map before (left panel) and after (right panel)
517 transformation with *DeepContact* convolutional neural network. Heat maps show scores
518 (normalized to have similar range). Grey open circles show contacts (side-chain heavy
519 atom distance < 8Å) in reference structure.

520 C. WW domain *combined score* interaction map before (left) and after (right) DeepContact
521 transformation. Note that the maps shown here lack the 5 c-terminal positions (see
522 Extended Data Figure 6B for full map).

523 D. RRM domain *combined score* interaction map before (left) and after (right) *DeepContact*
524 transformation.

525 E. FOS-JUN trans-interaction *epistasis score* interaction map before (left) and after (right)
526 *DeepContact* transformation.

527 F. Precision of top L predicted contacts of different interaction scores before and after
528 *DeepContact* transformation for the four datasets. Color indicates dataset, shape
529 indicates interaction score. Permutated control score is average over three random
530 permutations of *combined score* matrices (in case of FOS-JUN *epistasis score*
531 matrices). Dashed diagonal line indicates no changes in precision, dotted diagonal line
532 shows precision improvement of 20% after DeepContact learning.

533

534

Minimal data quality requirements for successful protein structure prediction

We further investigated how robust our prediction strategy is to changes in data quality by artificially down-sampling the GB1 domain dataset, thus assessing the minimal requirements for deep mutational scanning datasets to be useful for structure prediction.

First, we considered the sequencing read coverage. The GB1 domain dataset consists of about 600 million sequencing reads [11]. We find that artificially down-sampling the sequencing read coverage of the dataset to 25% or 10% hardly affects the precision of predicted tertiary contacts (Figures 7A). Only when using just 2.5% of sequencing reads (15 million) does the precision of top L contacts drop below 50%.

Next, we simulated a 'doped' dataset, by only considering amino acid mutations that can be reached by one mutation in the nucleotide sequence - thus reducing the coverage of double mutants to ~10% (similar to RRM and WW domain datasets). The doped dataset with full sequencing read coverage exhibits a drop in precision of predicted tertiary contacts of about 20%. Moreover, the doped dataset shows an increased sensitivity to lower sequencing read coverage.

We also tested the effect of reducing the signal-to-noise ratio (i.e. the measurement range of selection assay relative to the median error of fitness estimates), which results in non-quantifiably of negative epistasis (Extended Data Figures 2D-F and 5A). We thus tested how our prediction strategy performs on the GB1 domain dataset when only positive epistasis information is available; and find that it results in a drop of precision of about 20%, comparable to that observed for a doped dataset. In contrast, only using negative epistasis information resulted in a drop to ~35% precision, as low as a doped dataset with 10% sequencing coverage.

We evaluated secondary structure prediction performance of the various down-sampled GB1 domain datasets. Beta strand and alpha helix predictions are hardly affected by lowered data quality or partial epistasis information (Extended Data Figure 7A). In contrast, precision and recall of beta sheet positional pairing is strongly affected by dataset quality, although often the correct overall conformation of beta sheets is still recovered (Extended Data Figure 7B).

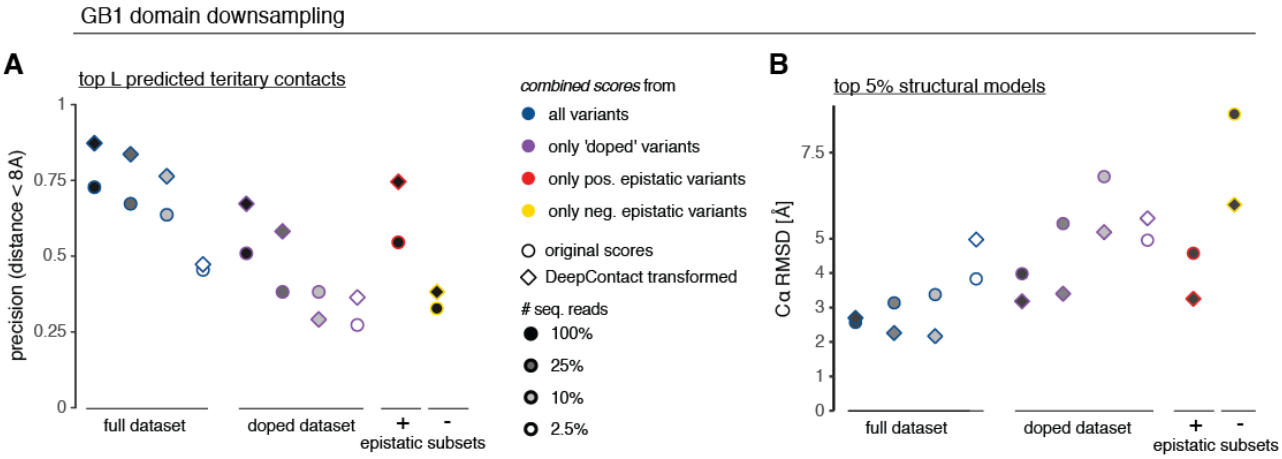
We next tested whether DeepContact could also improve prediction performance on these down-sampled datasets. Similar to interaction scores derived from full datasets, DeepContact

transformation of *combined scores* derived from down-sampled GB1 datasets improves the precision of predicted contacts by about 10-25% even for quite low quality datasets, i.e. the complete datasets with at least 10% read coverage, doped datasets with at least 25% read coverage or the dataset with only positive epistasis information (Figure 7A).

Finally, we evaluated how differences in prediction performance of tertiary contacts affect structural modeling. We find that changes in accuracy of the top structural models roughly scale with changes in contact prediction performance (Figure 7B). Down-sampling of sequencing reads in the complete dataset from 100% to 2.5% leads to a drop in accuracy from 2.5Å to 4Å $\langle C\alpha - RMSD \rangle$, which is roughly also the accuracy of top structural models from the doped dataset and the dataset using only positive epistasis information. Accuracies for lower quality datasets range from 5Å to 9Å $\langle C\alpha - RMSD \rangle$. DeepContact increases the accuracy of the top structural models by up to 2.6Å. For the complete datasets with only 25% or 10% of sequencing reads, the top structural models have better accuracy than those from the complete dataset with full sequencing read coverage but untransformed scores. Also, structural models based on DeepContact transformed scores from the doped dataset with full or 25% sequencing coverage and those from the dataset using only positive epistasis information reach average accuracies of 3.2Å $\langle C\alpha - RMSD \rangle$. Only for the two datasets with 2.5% sequencing read coverage do structural simulations based on DeepContact transformed scores not improve model accuracy.

Together these findings show that contact and structure prediction from deep mutational scanning data can also work for lower quality datasets and that the use of deep learning allows the use of much sparser and lower quality datasets.

Figure 7



589 Figure 7: Deep learning allows contact and structure prediction from
590 sparser and lower quality datasets

591 A. Precision of top L *combined score* position pairs for different down-sampled versions of
592 GB1 dataset. Color indicates type of dataset (blue – full dataset, purple – 'doped'
593 dataset, red – only positive epistasis information, yellow – only negative epistasis
594 information), fill indicates number of sequencing reads used in analysis, shape indicates
595 whether DeepContact has been used to transform the interaction score matrix.

596 B. Accuracy $\langle C\alpha - RMSD \rangle$ of top 5% structural models derived with tertiary contact
597 restraints from down-sampled GB1 datasets compared to reference structure. Note that
598 for better comparability, for these structural simulations only distance restraints were
599 derived from *combined scores* but the same secondary structure restraints predicted
600 from PSIPRED and no beta sheet pairing restraints were used for all simulations. Colors
601 and fills as in panel A.

Discussion

We have shown here that simply quantifying the activity of a large number of single and double mutant variants of a macromolecule can provide enough information to determine its high-resolution 3D structure.

We found that although most epistasis within a protein occurs between positions that are not direct structural contacts, aggregation on position pairs, merging of positive and negative epistasis information and partial correlation analysis of epistasis patterns can successfully discriminate direct from indirect structural contacts. Thus, mostly indirect epistatic couplings can be transformed to predict accurate structural contacts and elements. We have shown that this approach works robustly across multiple protein domains and a protein interaction. Moreover, we have demonstrated that the application of a convolutional neural network previously trained on patterns of co-evolution in proteins of known structure both improves structure prediction and allows the use of much lower quality deep mutation datasets.

Determining structures by mutagenesis requires an *in vitro* or *in vivo* selection assay. For many important molecules and drug targets, specific selection assays based on known functions or interaction partners already exist [11,14,15,46,50-55]. Additionally, many generic selection assays have recently been developed that should allow the stability or functional activity of many proteins to be assayed *in vivo* without the need for much prior knowledge about the protein under investigation [35-38]. Moreover, many molecules have known interaction partners – proteins, DNA, RNA, or small ligands – for which *cis*- and *trans*-epistasis can thus be assessed by binding assays [9,55]. We have shown here how *trans*-epistasis, for which library design is relatively easy, can lead to information about direct contacts in interaction surfaces as well as in the individual molecules.

Although structural information exists in the epistasis maps, our analyses and previous work [5-16] has shown that many epistatic interactions occur between positions that are not in direct structural contact. Indeed, in the GB1 domain the interactions are strikingly modular, with two mutually exclusive clusters of positive and negative epistatic interactions. This is consistent with many interactions being due to functional or energetic couplings between positions. The cluster of mostly positive epistatic interactions corresponds to a dynamic region involved in IgG binding [11]. In contrast, the cluster of negative epistatic interactions identifies positions important for the thermodynamic stability of the domain [56], and the periodicity of local negative epistatic

633 interactions provide evidence for a shift towards an alternative three-helical conformation that
634 has been previously reported for this sequence family (Extended Data Figure 4E) [57]. This
635 modular organization of epistasis is thus reminiscent of the concept of energetically-coupled
636 protein sectors identified from patterns of sequence co-evolution [20,21,58].

637 For macromolecules with very large numbers of homologs available in sequence databases,
638 correlated changes in sequence can provide sufficient information for structure determination [25-
639 34]. However, for many proteins and RNAs insufficient numbers of homologs are available, and
640 for fast evolving, recently-evolved or *de novo* designed molecules this is a fundamental
641 limitation [26,32,37,39]. Moreover, co-evolutionary analysis provides information on the average
642 structure across a large set of homologs, whereas it is easy to envisage how deep mutagenesis
643 could be used to directly determine alternative conformations of macromolecules when they are
644 performing particular selectable functions. The success of evolutionary coupling analysis for
645 predicting the structures of diverse folds and macromolecules does, however, strongly support
646 the generality of the approach outlined here. The demonstration that a deep learning approach
647 previously trained on evolutionary couplings dramatically improves the prediction of contacts
648 from deep mutagenesis data further supports this.

649 As a proof-of-principle we have shown that information from deep mutational scanning
650 experiments alone is sufficient for accurate structure prediction. In practice, however, integration
651 with other structural information is likely to further boost performance. As a first step, we used a
652 deep learning approach, DeepContact, that was trained on evolutionary couplings to learn
653 stereotypical structural patterns in contact maps [47]. DeepContact improved DMS-derived
654 contact prediction precision by up to 30% for individual proteins (GB1, WW and RRM domains).
655 Moreover, even though it had only been trained on data from individual proteins, it also
656 improved DMS-derived contact predictions for the FOS-JUN protein-protein interaction.
657 Integration with other structural predictors [47,59,60] and homology-driven structure modeling [61,62] is
658 likely to further improve accuracy and lower the data quality requirements for structure
659 determination by deep mutagenesis.

660 An analysis of incomplete and down-sampled variants of the GB1 dataset suggests that a high
661 signal-to-noise ratio (measurement range relative to experimental error of fitness estimates),
662 which allows both positive and negative epistasis to be quantified, is an important factor for
663 generating datasets with a quality sufficient for protein structure prediction. For datasets with
664 complete epistasis information, however, sequencing coverage hardly affected prediction

665 performance. In contrast, prediction performance on the incomplete, 'doped' dataset was
666 sensitive to sequencing coverage. Down to 25% sequencing coverage, however, performance
667 could be recovered by deep learning. Together, these analyses suggest that our approach
668 should be easily applicable to longer molecules. For example, with the experimental effort
669 undertaken to create and sequence the 55 amino acid GB1 library, a protein of length ~350
670 amino acids should be assayable at similar prediction performance (using a doped library with
671 25% sequencing coverage, i.e. 2.5% of the data: $55aa/\sqrt{0.025} \approx 350aa$). Such libraries for
672 longer proteins could be created via fragment-based ligation approaches [52] or via random
673 mutagenesis and barcode-variant linking [35].

674 Taken together, the results presented here establish deep mutagenesis as a new experimental
675 strategy for structure determination. The approach that we have outlined is not the only one
676 that can be envisaged to predict direct structural contacts from deep mutagenesis data, and
677 other related approaches are also likely to work [63]. The determination of macromolecular
678 structures by physical techniques requires access to very expensive scientific infrastructure. In
679 contrast, deep mutagenesis only requires techniques familiar to many molecular biologists and
680 access to sequencing that is increasingly low cost and available to all. Most importantly,
681 however, deep mutagenesis allows the structures of macromolecules to be studied whilst they
682 are performing particular functions *in vitro* as well as *in vivo* in the cell. As such, deep
683 mutagenesis opens up the possibility of low cost and high throughput determination of *in vivo*
684 macromolecular structures by many molecular biology and genomics labs. A large-scale project
685 to systematically determine the structures of proteins and protein domains should therefore be
686 possible using the existing infrastructure of genomics institutes.

687

Table 1: Dataset properties

Dataset	Mutated positions	% double mutants[$]	% doubles quantifiable[#]		# input reads per double mutant (median)[*]	measurement range (log fitness units)[+]	relative error (median)[&]
			positive epistasis	negative epistasis			
Protein G B1 domain	55	97	80	55	248	5.8	2.8%
hYAP WW domain	33	10	8.3	0.8	73	0.8	8.6%
Pab1 RRM2 domain	25	11	8.3	3.9	209	3.1	3.7%
FOS-JUN	2 x 32	43	37	31	124	8.6	3.6%

[$] median percentage of all possible double mutants (361 per position pair) that passed read quality thresholds per position pair

[#] median percentage of all possible double mutants (361 per position pair) that passed read quality thresholds and are deemed suitable for epistasis quantification per position pair

[*] summed number of reads across all input replicates for double mutants that passed read quality thresholds

[+] measurement range of selection assay: log fitness range between peak of lethal mutants and the wild-type variant

[&] median error of fitness estimates of double mutant variants relative to measurement range of selection assay

Acknowledgements

We are grateful to Yang Liu and Jian Peng for making their DeepContact code available and for their advice. We thank members of the Lehner lab, T. Gross, G. Mönke, M. Bolognesi and C. Camilloni for discussions and feedback. This work was supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2011-26206 and SEV-2012-0208), the AXA Research Fund, the Bettencourt Schueller Foundation, Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR, SGR-831), the EMBL-CRG Systems Biology Program, and the CERCA Program/Generalitat de Catalunya. J.M.S. was supported by an EMBO Long-Term Fellowship (ALTF 857-2016). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 752809 (J.M.S).

Author Contributions

Conceptualization, B.L. and J.M.S.; Methodology, J.M.S.; Investigation, J.M.S.; Writing, J.M.S. and B.L.; Supervision, B.L.

Competing Interests

The authors declare no competing interests.

Additional Information

Correspondence and requests for materials should be addressed to B.L.

References

1 Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596-604, doi:10.1016/j.sbi.2009.08.003 (2009).

2 Lehner, B. Molecular mechanisms of epistasis within and between genes. *Trends in genetics : TIG* **27**, 323-331, doi:10.1016/j.tig.2011.05.007 (2011).

3 Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods* **11**, 801-807, doi:10.1038/nmeth.3027 (2014).

4 Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein science : a publication of the Protein Society* **25**, 1204-1218, doi:10.1002/pro.2897 (2016).

5 Horovitz, A. & Fersht, A. R. Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *J Mol Biol* **214**, 613-617, doi:10.1016/0022-2836(90)90275-Q (1990).

6 Carter, P. J., Winter, G., Wilkinson, A. J. & Fersht, A. R. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). *Cell* **38**, 835-840 (1984).

7 Ackermann, E. J., Ang, E. T., Kanter, J. R., Tsigelny, I. & Taylor, P. Identification of pairwise interactions in the alpha-neurotoxin-nicotinic acetylcholine receptor complex through double mutant cycles. *J Biol Chem* **273**, 10958-10964 (1998).

8 Chen, J. & Stites, W. E. Energetics of side chain packing in staphylococcal nuclease assessed by systematic double mutant cycles. *Biochemistry* **40**, 14004-14011 (2001).

9 Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, 594, doi:10.7554/eLife.32472 (2018).

10 Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19**, 1537-1551, doi:10.1261/rna.040709.113 (2013).

11 Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology : CB* **24**, 2643-2651, doi:10.1016/j.cub.2014.09.072 (2014).

12 Sahoo, A., Khare, S., Devanarayanan, S., Jain, P. C. & Varadarajan, R. Residue proximity information and protein model discrimination using saturation-suppressor mutagenesis. *eLife* **4**, 371, doi:10.7554/eLife.09532 (2015).

13 Li, C. & Zhang, J. Multi-environment fitness landscapes of a tRNA gene. *Nature Ecology & Evolution* **15**, 1, doi:10.1038/s41559-018-0549-8 (2018).

14 Li, C., Qian, W., Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837-840, doi:10.1126/science.aae0568 (2016).

15 Domingo, J., Diss, G. & Lehner, B. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature* **558**, 117-121, doi:10.1038/s41586-018-0170-7 (2018).

16 Puchta, O. *et al.* Network of epistatic interactions within a yeast snoRNA. *Science* **352**, 840-844, doi:10.1126/science.aaf0965 (2016).

17 Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317, doi:10.1002/prot.340180402 (1994).

18 Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of molecular biology* **193**, 693-707 (1987).

19 Gloor, G. B., Martin, L. C., Wahl, L. M. & Dunn, S. D. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* **44**, 7156-7165, doi:10.1021/bi050293e (2005).

20 Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774-786, doi:10.1016/j.cell.2009.07.038 (2009).

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

22 Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108**, E1293-1301, doi:10.1073/pnas.1111471108 (2011).

23 Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106**, 67-72, doi:10.1073/pnas.0805923106 (2009).

24 Burger, L. & van Nimwegen, E. Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. *PLoS Computational Biology* **6**, e1000633, doi:10.1371/journal.pcbi.1000633 (2010).

25 Weinreb, C. *et al.* 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell* **165**, 963-975, doi:10.1016/j.cell.2016.03.030 (2016).

26 Tóth-Petróczy, A. *et al.* Structured States of Disordered Proteins from Genomic Sequences. *Cell* **167**, 158-170.e112, doi:10.1016/j.cell.2016.09.010 (2016).

27 Hopf, T. A. *et al.* Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **149**, 1607-1621, doi:10.1016/j.cell.2012.04.012 (2012).

28 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).

29 Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184-190, doi:10.1093/bioinformatics/btr638 (2012).

30 De Leonardis, E. *et al.* Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Research* **43**, 10444-10455, doi:10.1093/nar/gkv932 (2015).

31 Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences* **109**, 10340-10345, doi:10.1073/pnas.1207864109 (2012).

32 Ovchinnikov, S. *et al.* Protein structure determination using metagenome sequence data. *Science* **355**, 294-298, doi:10.1126/science.aah4043 (2017).

33 Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248, doi:10.7554/eLife.09248 (2015).

34 Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030, doi:10.7554/eLife.02030 (2014).

35 Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics* **50**, 874-882, doi:10.1038/s41588-018-0122-z (2018).

36 Weile, J. *et al.* A framework for exhaustively mapping functional missense variants. *Molecular Systems Biology* **13**, 957, doi:10.15252/msb.20177908 (2017).

37 Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168-175, doi:10.1126/science.aan0693 (2017).

38 Kim, I., Miller, C. R., Young, D. L. & Fields, S. High-throughput analysis of in vivo protein stability. *Molecular & cellular proteomics : MCP* **12**, 3370-3378, doi:10.1074/mcp.O113.031708 (2013).

39 Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nature Biotechnology* **30**, 1072-1080, doi:10.1038/nbt.2419 (2012).

40 Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425-431, doi:10.1126/science.1180823 (2010).

41 Stein, R. R., Marks, D. S. & Sander, C. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Computational Biology* **11**, e1004182, doi:10.1371/journal.pcbi.1004182 (2015).

42 Andreani, J. & Söding, J. bbcontacts: prediction of β-strand pairing from direct coupling patterns. *Bioinformatics* **31**, 1729-1737, doi:10.1093/bioinformatics/btv041 (2015).

43 Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* **292**, 195-202, doi:10.1006/jmbi.1999.3091 (1999).

44 The PyMOL Molecular Graphics System, V. S., LLC. *The PyMOL Molecular Graphics System, Version 1.8* (2015).

45 Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *Journal of magnetic resonance (San Diego, Calif. : 1997)* **160**, 65-73 (2003).

46 Araya, C. L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences* **109**, 16858-16863, doi:10.1073/pnas.1209751109 (2012).

47 Liu, Y., Palmedo, P., Ye, Q., Berger, B. & Peng, J. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell systems*, doi:10.1016/j.cels.2017.11.014 (2017).

48 Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. & Bonvin, A. M. J. J. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins* **86 Suppl 1**, 51-66, doi:10.1002/prot.25407 (2018).

49 Fox, N. K., Brenner, S. E. & Chandonia, J. M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* **42**, D304-309, doi:10.1093/nar/gkt1240 (2014).

50 Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**, 413-422, doi:10.1534/genetics.115.175802 (2015).

51 Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences* **110**, E1263-1272, doi:10.1073/pnas.1303309110 (2013).

52 Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *bioRxiv*, doi:10.1101/213835 (2017).

53 Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **5**, e16965, doi:10.1038/nature23902 (2017).

54 Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature Methods* **7**, 741-746, doi:10.1038/nmeth.1492 (2010).

55 McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138-142, doi:10.1038/nature11500 (2012).

56 Otwinowski, J. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol Biol Evol*, doi:10.1093/molbev/msy141 (2018).

57 Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences* **106**, 21149-21154, doi:10.1073/pnas.0906408106 (2009).

58 Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *eLife* **7**, 41, doi:10.7554/eLife.34300 (2018).

59 Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999-1006, doi:10.1093/bioinformatics/btu791 (2015).

60 Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Computational Biology* **13**, e1005324, doi:10.1371/journal.pcbi.1005324 (2017).

61 Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93, doi:10.1016/S0076-6879(04)83004-0 (2004).

62 Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**, 7-8, doi:10.1038/nmeth.3213 (2015).

63 Rollins, N. J. *et al.* 3D protein structure from genetic epistasis experiments. doi:10.1101/320721 (2018).

64 Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome biology* **18**, 741, doi:10.1186/s13059-017-1272-5 (2017).

65 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461, doi:10.1093/bioinformatics/btq461 (2010).

66 Barlow, R. *Statistics : a guide to the use of statistical methods in the physical sciences*. (Wiley, 1989).

67 Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology* **4**, Article32, doi:10.2202/1544-6115.1175 (2005).

68 Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* **33**, 4721-4729 (1994).

69 Pires, J. R. *et al.* Solution structures of the YAP65 WW domain and the variant L30 K in complex with the peptides GTPPPPYTVG, N-(n-octyl)-GPPPY and PLPPY and the application of peptide libraries reveal a minimal binding epitope. *Journal of molecular biology* **314**, 1147-1156, doi:10.1006/jmbi.2000.5199 (2001).

70 Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98**, 835-845 (1999).

71 Glover, J. N. & Harrison, S. C. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **373**, 257-261, doi:10.1038/373257a0 (1995).

72 Adhikari, B., Bhattacharya, D., Cao, R. & Cheng, J. CONFOLD: Residue-residue contact-guided ab initio protein folding. *Proteins* **83**, 1436-1449, doi:10.1002/prot.24829 (2015).

73 Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710, doi:10.1002/prot.20264 (2004).

74 Seemayer, S., Gruber, M. & Söding, J. CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130, doi:10.1093/bioinformatics/btu500 (2014).

Methods:

Datasets and preprocessing

Protein G B1 domain

Protein G B1 domain (GB1) deep mutational scanning data was obtained from the supplementary information of Olson, et al. [11]. The data consists of summed read counts of three replicate experiments assaying the binding affinity of GB1 variants to immunoglobulin G (IgG).

Read frequencies of each single or double mutant variant in input library and output library (after binding affinity assay) were calculated as variant read counts relative to wild-type variant read counts. A variant's fitness was calculated as the natural logarithm of the ratio of output to input read frequency, i.e. $f_i = log(\frac{n_i^{out}/n_{wt}^{out}}{n_i^{in}/n_{wt}^{in}})$, with n as read counts, superscripts denoting input or output sequencing library and subscripts denoting variant *i* or wild-type variant.

The standard error of fitness estimates was calculated from read counts under Poissonian assumptions, i.e. $\sigma_i = \sqrt{\frac{1}{n_i^{in}} + \frac{1}{n_i^{out}} + \frac{1}{n_{wt}^{in}} + \frac{1}{n_{wt}^{out}}}$ [64]. We note that this is a lower bound estimate of the actual error, due to the lack of replicate information.

Each measurement assay has a lower measurement limit due to unspecific background effects (Extended Data Figure 2A). In the case of the IgG-binding assay for GB1, this is presumably mainly due to unspecific carry-over on beads [11]. The fitness values derived from the measurement are therefore a convolution of the actual binding affinities to IgG and nonspecific carryover, i.e. $exp(f_i^{measured}) = exp(f_i^{binding}) + exp(f^{carryover})$. Fitness values of variants close to the lower measurement limit of the assay are therefore dominated by unspecific carryover effects. The lower measurement limit of the assay was estimated by two approaches that yielded identical estimates. One, from a kernel density estimate of the single mutant fitness distribution (R function *density* with parameter *bw* set to 0.15), where the position of the lower mode of the data corresponded to $f^{carryover} = -5.85$ (~0.3% on linear scale). Two, from examining the fitness distribution of double mutants with expected fitness lower than -8 log-units, i.e. double mutants resulting from two lethal or nearly lethal single mutant variants, whose fitness values are thus expected to be dominated by background effects. The median of this background fitness distribution yielded an estimate of $f^{carryover} = -5.85$.

937 7% of double mutant variants were discarded due to too low sequencing coverage in input or
938 output libraries (Extended Data Figure 2B). That is, variants that had less than 200 input reads
939 and no output reads were discarded, because it is not possible to determine their fitness. Above
940 200 input reads, variants without output reads are certain to be dominated by nonspecific
941 carryover effects. These variants were retained and their fitness was calculated by setting their
942 output read count to 0.5.

943 GB1 down-sampling

944 Down-sampling of the full GB1 dataset was performed in three different ways.

945 For the 'doped' datasets, we only allowed amino acid changes created by one nucleotide
946 mutation from the wild-type sequence (ENA entry M12825). To down-sample the sequencing
947 read coverage, for each variant we picked as a down-sampled read count the draw of
948 successes from a binomial distribution with the number of sequencing reads in the full datasets
949 as trials and the target down-sampling rate (25%, 10% or 2.5%) as chance of success. For the
950 read down-sampled and doped datasets (and combinations of both), the analysis workflow for
951 the full dataset was repeated.

952 For the down-sampled datasets taking only positive or negative epistatic information into
953 account, we calculated *epistasis* and *association scores* from epistatic enrichment matrices and
954 partial correlation matrices of only positive or negative epistasis information. Instead of merging
955 positive and negative matrices and then calculating z-scores, we only calculated z-scores with
956 the individual errors from positive or negative epistasis information. The *combined scores* (for
957 which results are reported) for each set was then calculated as for the full dataset by summing
958 standardized *epistasis* and *association scores*.

959 hYAP WW domain

960 hYAP WW domain data was obtained from Sequence Read Archive (SRA) entry SRP015751 [46].
961 Paired-end reads were merged with USearch [65] and merged reads with any base having a
962 Phred base quality score below 20 were discarded. Read counts from the two technical
963 sequencing replicates were merged and read counts for the same amino acid variants with at
964 most one synonymous mutation in another codon were summed up. The dataset consists of an
965 input library and three output libraries after consecutive rounds of selection in a phage display
966 assay. Fitness was estimated as the slope of log frequency (variant counts divided by wild-type
967 counts) changes over the rounds of selection experiment [46]. For each variant at each selection

968 step a Poissonian error of $\sigma_{i,x} = \sqrt{\frac{1}{n_i^x} + \frac{1}{n_{wt}^x}}$ was calculated, with x denoting the selection step.
969 Slopes were calculated as weighted straight line least square fits [66]. Comparison of library-wide
970 changes in variant frequencies between selection rounds suggested differential selection
971 pressures across the rounds. We thus applied a non-equidistant spacing of 0.6, 1.17 and 1.22
972 between selection rounds when calculating slopes. Only variants that have more than 10 reads
973 in the input library and at least one read after the first selection were retained for further analysis
974 (45% of constructed double mutants). The lower fitness limit was calculated as the weighted
975 mean fitness of all variants containing STOP codons (-0.78 in log-fitness units).

976 Pab1 RRM2 domain

977 Pab1 RRM2 domain data was obtained from the Supplementary Table 5 of Melamed, et al. [10].
978 Reported enrichment scores were log-transformed to obtain fitness values. Output reads per
979 variant were deduced from the number of input reads times the enrichment score and used to
980 calculate a Poissonian error of the fitness estimate. Single mutant count data is not provided
981 and we thus estimated the error of single mutant fitness estimates to be 0.01. Lower bound of
982 fitness assay was estimated as weighted mean fitness of all double mutant variants containing
983 STOP codons (-3.1 log-fitness units).

984 FOS-JUN interaction

985 Raw count tables were provided by Guillaume Diss [9]. The dataset consists of input and output
986 sequencing libraries after selection for physical interaction between the two proteins in a protein
987 complementation assay in three biological replicates. Per sequencing library, read counts from
988 all synonymous variants were summed up. Only variants that had more than 10 reads in each of
989 the three input libraries were used for further analysis (43% of double mutants). Per input/output
990 replicate, fitness of each variant was calculated as the log change in frequency compared to the
991 wild-type variant (as for GB1). A Poissonian error for each variant's fitness estimate was
992 derived.

993 The dataset has a large dynamic range, thus many low-fitness variants with low input read
994 coverage have very low or no output read counts (per replicate ~1/3 of variants have below 3
995 output counts, ~15% of variants have zero output counts), effectively reducing the dynamic
996 range of the assay for low input variants and distorting the estimate of the overall fitness
997 distribution (see Extended Data Figure 5E). To overcome this, we implemented a Bayesian
998 estimator of fitness. For each double mutant variant, we first identified the 1000 nearest

neighbors in single mutant fitness space (i.e. those double mutants whose respective single mutant fitness values are similar to the single mutant fitness values of the variant under consideration) with sufficient input coverage (more than 100 reads in the input library). From this set of 1000 nearest neighbors we calculated the expected distribution of double mutant fitness values, which served as a prior distribution. For the variant under consideration we calculated the likelihood distribution of fitness values given its input and output read counts under Poissonian assumptions. Fitness was then estimated as the mean of the distribution resulting from the multiplication of prior and likelihood distributions. Error of fitness estimate was estimated as the standard deviation of the resulting distribution. Estimated fitness from the three replicate experiments were merged by weighted averaging.

Epistasis classification

Epistasis was calculated from a non-parametric null model (Figure 1B) in order to account for nonlinearities close to the lower limit of the fitness assay measurement range, non-specific epistatic behavior resulting from e.g. thermodynamic stability thresholds as well as differential uncertainty of fitness measurements across the fitness landscape, due to lower read counts in the output for low fitness variants.

First, double mutant fitness values were corrected by subtracting the average local fitness computed using a two-dimensional local polynomial regression (using R function *loess* with span = 0.2). This was necessary to avoid boundary effects of quantile-based fits in boundary regions with non-zero slopes. 5th and 95th percentile surfaces were then fit to these corrected double mutant fitness values, by computing for each double mutant variant the 5th and 95th percentile of the fitness distribution made up of the 1% closest neighbors in single mutant fitness space. Double mutant variants with fitness values below the 5th or above the 95th percentile were categorized as negative or positive epistatic, respectively (Figure 1B).

The evaluation of positive or negative epistasis was, however, restricted to specific subsets of the data where measurement errors do not impede epistasis classification (Extended Data Figure 2C). The data subset deemed suitable for positive epistasis classification is limited to regions where

- the 95th percentile fitness surface is below wild-type fitness

1029 ● at least one single mutant fitness value is significantly smaller than wild-type fitness at
1030 two standard errors

1031 ● the expected fitness (sum of both single mutant fitness values) is not significantly higher
1032 than wild-type at two standard errors

1033 The rationale for these criteria is to avoid double mutants from two neutral single mutants,
1034 because these are dominated by measurement noise of overabundant wild-type like variants.
1035 No restrictions were instead applied to the lower limits of the measurement range, because
1036 otherwise no/little epistasis quantification would have been available for several positions with
1037 very strong detrimental effects as well as because strong positive epistatic effects are observed
1038 in these regions, despite the dominance of background measurement effects.

1039 The data subset in which variants were potentially classified as negative epistatic is limited to
1040 data regions where

1041 ● the 5th percentile fitness surface is above the 95th percentile of the background effect
1042 distribution; this value is derived from the 95th percentile of double mutant fitness values
1043 with expected fitness below -8 (analogous to lower fitness limit estimation, see above).

1044 ● both single mutant fitness values are significantly higher than the lower limit of the
1045 fitness assay measurement range at two standard errors

1046 ● the expected fitness (sum of both single mutant fitness values) is not significantly higher
1047 than wild-type at two standard errors

1048 The rationale for criteria 1 and 2 is to avoid background measurement effects that make
1049 negative epistasis quantification unreliable.

1050 As a result of these restrictions as well as differences in initial coverage, the number of double
1051 mutant variants that can be used to assess positive and negative epistasis varies substantially
1052 across position pairs and datasets (see Table 1, Extended Data Figures 2D-F and 5A&D).

1053

1054 Epistatic interactions (*epistasis scores*)

1055 We derived several interaction scores to estimate which position pairs are in close contact in the
1056 tertiary structure (see Figures 2A and 3A,B and Extended Data Figure 1). These scores are
1057 based on summarizing epistasis information on the position pair-level and accounting for the

uncertainty inherent in the summarized estimates due to differential error of fitness estimates across the measurement range as well as varying numbers of double mutants amenable to epistasis classification (see Table 1, Extended Data Figures 2D-F and 5A&D).

To summarize epistasis information on the position pair-level, we calculated the fraction of positive or negative epistatic variants per position pair. The fraction of positive epistatic variants per position pair is the number of positive epistatic variants divided by the number of all variants that lie in the double mutant space amenable to positive epistasis classification (Extended Data Figure 1, step 5b, equivalent calculation for negative epistasis fraction). Because enrichments with positive and negative epistatic variants per position are anti-correlated (Extended Data Figure 3A), we treated both separately and only aggregated them to derive the final interaction scores.

To estimate the uncertainty in epistatic fractions per position pair for downstream analyses we implemented a re-sampling approach (Extended Data Figure 1, step 5, described here for positive epistatic variants, but equivalent for negative epistatic variants). In each of 10.000 re-sampling runs:

- each variant's fitness was drawn from a normal distribution with the measured fitness as mean and the uncertainty due to sequencing coverage as standard deviation $f_{ij}^{sampled} = \mathcal{N}(f_{ij}, \sqrt{\sigma_{ij}^2 + s_i^2 * \sigma_i^2 + s_j^2 * \sigma_j^2})$, with s_x as the local slope of the median fitness landscape in direction of the respective single mutant (step 5a)
- positive epistasis of variants was re-classified given the drawn fitness values (also step 5a)
- each position pair's fraction of positive epistatic variants was drawn from the posterior probability distribution of how likely an underlying true fraction of epistatic variants E_{xy}^{+} is to generate the observed fraction of epistatic variants given the finite number of overall variants, i.e. $\mathrm{e}_{xy}^{+} \sim \mathrm{p}(\mathrm{E}_{xy}^{+} \mid \#\, \varepsilon_{xy}^{+}, \#\, \mathrm{variants}_{xy})$ (step 5b). The posterior probability distribution is the product of a prior probability distribution – the kernel density estimate of the expected epistatic fractions across all position pairs (calculated using R function *density* with parameter *bw* set to 0.05) – and the likelihood function for the underlying true fraction of epistatic variants given the observed fraction of epistatic variants and the overall number of variants under binomial sampling assumptions

To derive an interaction score from the epistatic fractions per position pair, mean positive and negative epistatic fractions across resampling runs were combined by weighted averaging, with

weights as the inverse variances of epistatic fractions across resampling runs, i.e. $\langle e_{xy} \rangle = \frac{\langle e_{xy}^{+} \rangle * \sigma_{e_{xy}^{+}}{}^{-2} + \langle e_{xy}^{-} \rangle * \sigma_{e_{xy}^{-}}{}^{-2}}{\sigma_{e_{xy}^{+}}{}^{-2} + \sigma_{e_{xy}^{-}}{}^{-2}}$.

To arrive at the final *epistasis score,* the mean epistatic fractions were further normalized by their combined uncertainty, $E_{xy} = \langle e_{xy} \rangle / \sigma_{xy}$, with $\sigma_{xy} = \left(\sigma_{e_{xy}^{+}}{}^{-2} + \sigma_{e_{xy}^{-}}{}^{-2}\right)^{-1/2}$ (step 6).

Epistasis pattern correlations (*association scores*)

In addition to the *epistasis score* we derived an interaction score from the partial correlation of epistasis patterns between position pairs, termed *association score*. The rationale behind this score is that proximal positions in the protein should have similar distances and geometrical arrangements towards all other positions in the protein and should therefore also have similar patterns of epistatic interactions with all other positions.

In each re-sampling run we constructed a symmetric matrix of the drawn positive epistatic fractions e_{xy}^{+} with number of rows and columns as the number of mutated positions (Extended Data Figure 1, step 5c, equivalent for negative epistatic fractions). Missing values (positions pairs without observed variants) were imputed by drawing a random value from the overall distribution of epistatic fractions. A pseudo count equal to the first quartile of the epistatic fraction distribution was added to each epistatic fraction. Diagonal elements (epistatic fractions of a position with itself) were set to 1. The matrix values were transformed by the natural logarithm (to make distributions more symmetric, thus correlations are not dominated by few position pairs with large epistatic fractions) and for each pair of columns the Pearson correlation coefficient was calculated to arrive at the correlation matrix (step 5d).

A shrinkage approach was used to improve the estimate of the correlation matrix [67]. In brief, the empirical correlation matrix is shrunk towards the identity matrix in order to minimize the mean-squared error between estimated and true correlation matrix. Additionally, this yields a positive definite and well-conditioned correlation matrix, suitable for inversion. All computations on correlation matrices, shrinkage and matrix inversion were performed with the R package *corpcor*, functions *cor.shrink* and *pcor.shrink* [67].

Partial correlations of epistatic patterns between each position pair were calculated by inverting the shrunk correlation matrix and normalizing each off-diagonal entry of the inverted matrix by

the geometric mean of the two respective diagonal entries, i.e. $a_{xy}^{+} = \frac{r_{xy}^{-1}}{\sqrt{r_{xx}^{-1} * r_{yy}^{-1}}}$, with r_{xy}^{-1} as the (x,y)-entry of the inverted correlation matrix (Extended Data Figure 1, step 5d). Equivalent to the *epistasis score,* positive and negative partial correlation estimates were merged by calculating weighted averages of their mean estimates across re-sampling runs, with weights as the inverse variances across resampling runs, i.e. $\langle a_{xy} \rangle = \frac{\langle a_{xy}^{+} \rangle * \sigma_{a_{xy}^{+}}{}^{-2} + \langle a_{xy}^{-} \rangle * \sigma_{a_{xy}^{-}}{}^{-2}}{\sigma_{a_{xy}^{+}}{}^{-2} + \sigma_{a_{xy}^{-}}{}^{-2}}$, and the final *association score* normalized by the combined uncertainty, $A_{xy} = {\langle a_{xy} \rangle}/{\sigma_{xy}}$, with $\sigma_{xy} = \left(\sigma_{a_{xy}^{+}}{}^{-2} + \sigma_{a_{xy}^{-}}{}^{-2}\right)^{-1/2}$ (step 6).

Aggregating *epistasis* and *association scores* (*combined scores*)

We further derived a *combined score* by summing the standardized *epistasis* and *association scores*, i.e. $C_{xy} = \frac{E_{xy} - \langle E \rangle}{\sigma_E} + \frac{A_{xy} - \langle A \rangle}{\sigma_A}$. We note that this is a naïve approach to combining the information from these two complementary sources, and surely more sophisticated approaches that further improve proximity estimates can be developed.

Secondary structure prediction

We used a two-dimensional kernel smoothing approach to predict secondary structure elements from interaction score matrices (Figure 4A-C). For a given position along the linear chain (on the diagonal of the interaction score matrix), interaction scores (off the diagonal) are integrated with distance-specific weighting according to the kernel, which reflects the known geometry of secondary structures.

The alpha kernel takes on a sinusoidal profile perpendicular to the diagonal to weight interactions according to whether the position pair considered should have congruent side-chain orientations (see diagonal and perpendicular profiles in Figure 4B). The kernel was defined as $k_{\alpha}(d,p) = \left(\cos\left(p * \frac{2\pi}{3.6}\right) + {}^{1}/_{3}\right) * e^{-\frac{d^2}{c^2}}$, with $d = |2x - i - j|$ as the diagonal distance of the interaction *ij* (off the diagonal) to the reference position *x* (on the diagonal) and $p = |i - j|$ as the perpendicular distance of the interaction off the diagonal. The kernel weight for positions with p

> 5 was set to 0, thus only including interactions across little more than the first helical turn. Finally, $c = 4$ is the integration scale for the Gaussian kernel along the diagonal. While smaller integration scales do yield nosier results and longer integration scales can lead to non-detection of shorter secondary structure stretches, we found that in practice our whole approach (including the actual detection algorithm described below) is robust to alterations of the integration length.

The kernel smoothed alpha value for a given position x on the diagonal is then calculated as the sum over all interaction scores times their kernel weights $K_{\alpha,x} = \sum_i \sum_j k_\alpha(d,p) * S_{ij}$, where S_{ij} is one of the interaction scores (*epistasis, association* or *combined score*) at position pair *ij*.

The beta kernel takes an alternating profile perpendicular to the diagonal to weight interactions according to alternating side-chain orientations in a beta strand and was defined as $k_\beta(d,p) = \left((p+1)\, mod\, 2 - {}^{1}\!/_{3}\right) * e^{-\frac{d^2}{c^2}}$, with $c = 4$. Only interactions with perpendicular distances equal or smaller than two (i.e. $k_\beta(d, p > 3) = 0$) were included.

To calculate whether kernel-weighted interaction scores of a specific position are larger than expected, they were compared to kernel-weighted scores obtained from 10^4 randomized control datasets. Randomization was performed by shuffling all interaction scores, while preserving matrix symmetry, and kernel-weighted interaction scores from randomized control datasets were calculated for each position independently to control for possible boundary effects in positions close to the borders of the protein chain. A p-value for each position was calculated as the fraction of random controls with kernel smoothed values above that of the real data.

Secondary structure elements were identified by searching for continuous stretches of positions with high propensities to belong to either alpha helices or beta strands. The following workflow was implemented:

1. calculate a combined p-value for seeds of length 3 by combining position-wise p-values using Fisher's method for both alpha and beta kernel smoothed interaction scores
2. separate positions according to whether combined p-values of seeds from alpha or beta kernels are more significant, i.e.
    2.1. for alpha helical propensity calculations only consider stretches of at least 5 consecutive positions for which the combined p-value of seeds for alpha kernel smoothing is smaller than that from beta kernel smoothing (thus setting the lower size limit of alpha helical elements to five)

2.2. for beta strand propensity calculations only consider stretches of at least 3 consecutive positions for which the combined p-value of seeds for beta kernel smoothing is smaller than that from alpha kernel smoothing (thus setting the lower size limit of beta strands to three)

For alpha helices and beta strands separately and while combined p-values of seeds < 0.05

3. select the most significant seed
4. test whether extension to any side yields lower combined p-value
   4.1. if yes: extend seed in this direction and repeat step 4
   4.2. else: repeat step 4 once to see whether further extension in same direction yields lower combined p-value
      4.2.1. if yes: extend and repeat step 4
      4.2.2. else: proceed to step 5
5. fix as secondary structure element and delete all 'used' p-values (and combined seed p-values), such that other elements cannot incorporate them
6. check whether other already fixed elements are adjacent or at most one position away
   6.1. if yes: merge both elements
7. repeat steps 3-6 until no more seeds with combined p-value < 0.05 are left

This yields a list of predicted locations of secondary structure elements. We note that the secondary structure elements predicted from deep mutational scanning data could be compared to and combined with predictions derived from other tools, such as PSIPRED (Jones, 1999), to further improve reliability.

To detect beta sheet interactions a modified beta strand kernel was used. In contrast to beta strand detection, the beta sheet interaction kernel is centered on each off-diagonal position. For beta sheet kernels diagonal and perpendicular distances are therefore modified as $d = |x + y - i - j|$ and $p = |x - i - (y - j)|$. The kernels to detect parallel and anti-parallel beta sheets differ in which is their 'diagonal' direction, i.e. the direction at which consecutive position pairs interact in the beta sheet (Extended Data Figure 4A). Therefore, parameters *d* and *p* were swapped for the anti-parallel beta sheet kernel. Also, because these positions can be deemed the most crucial for deciding whether a position participates in a beta sheet interaction or not, we up-weighted these positions (those with $p = 0$) in the kernel by a factor of two, i.e. $K_\beta(d, 0) = 4/3 * e^{-\frac{d^2}{c^2}}$.

Beta sheet interactions were identified by searching for the most significant stretches of parallel and anti-parallel interactions (similar to workflow for alpha helices and beta strands), then identifying the set of most significant interactions that is consistent with previously predicted secondary structure elements.

In particular, step 1 & 3-7 from the above-described workflow were performed for the parallel beta sheet kernel on each sub-diagonal (parallel to the main diagonal) of the interaction score matrix separately; and for the anti-parallel beta sheet kernel on each perpendicular diagonal of the interaction score matrix separately.

The steps were modified as follows:

- for anti-parallel beta sheet interactions, only positions with a distance greater than 1 from the main diagonal were used to calculate seed p-values (assuming anti-parallel beta sheet interactions need a turn of at least length two to be connected)
- for parallel beta sheet interactions, only sub-diagonals with a distance greater than 4 from the main diagonal were considered (assuming parallel beta sheet interactions of two adjacent beta strands need a linker region)

We extended the workflow with the following steps to predict beta sheet interactions within the protein domain:

8. compute association of seeds with known beta strands (e.g. seed positions overlap strand 1 on one side and coincide with strand 3 on the other side)
9. while there are seeds with $p < 0.05$: pick most significant seed from either the parallel or anti-parallel sheet subset
10. check consistency with secondary structure elements
    - 10.1. discard the seed and jump back to step 9 if
        - 10.1.1. it is overlapping or too close to an alpha helix or the linker region between two beta strands that interact (minimal distance smaller one)
        - 10.1.2. at least one of the two strands it is associated with already has two other beta sheet interactions or the total number of beta sheet interactions exceeds 2*(#beta strands – 1)
    - 10.2. modify secondary structure elements and start anew from step 3 if
        - 10.2.1. one side of the seed is not associated to a known beta strand: create this beta strand
        - 10.2.2. if both sides of the seed are associated with the same known strand: split the strand and create a linker region in-between the strands

10.3. else fix the beta sheet interaction and delete all other interactions that are associated with the same strands and haven't been fixed yet, jump back to step 9

11. if no more seeds with $p<0.001$, finish

12. update beta strands: keep only those positions that are part of a beta sheet interaction

For beta sheet pairing detection in the GB1 domain (as reported in Figure 4D and Extended Data Figures 4A-C and 7B) we used as input the secondary structure element predictions derived from the deep mutational scanning data (as shown in Figure 4A-C and Extended Data Figure 4D). For the RRM and WW domains, we used as input PSIPRED predicted secondary structure elements, due to the insufficient signal from secondary structure element predictions from deep mutational scanning data.

Protein distance metrics

The minimal side chain heavy atom distance, i.e. the minimal distance between any two side-chain heavy atoms of a position pair (in case of glycine, $C\alpha$), was used as the general distance measure. A direct contact was defined as minimal side-chain heavy atom distance < 8Å. For all evaluations of predicted contact precision we only considered position pairs with non-trivial tertiary contacts (those with a linear sequence separation of greater than 5 positions).

We do find that, while using all heavy atoms to calculate distances increase the true positive rates of predicted contacts by about 10%, side-chain heavy atom distances display much higher true positive rates over random expectation, thus suggesting that side-chain interactions are more informative for epistatic interactions (Extended Data Figure 7C).

The Floyd-Warshall algorithm (implemented as custom script in R) was used to calculate the minimal number of edges <8Å that connect any two positions in the protein.

Reference structures used as comparison were

- GB1 domain: PDB entry 1pga, X-ray diffraction structure [68]
- WW domain: PDB entry 1k9q, solution NMR structure [69]
- RRM domain: PDB entry 1cvj (chain A), X-ray diffraction structure of human Pab1 [70]; note that the central section of the yeast RRM domain analysed is one nucleotide longer than the corresponding homologous region in the human RRM domain. We thus

arbitrarily removed position 14 (in the loop region) when comparing the DMS-derived predictions to the human Pab1 structure.

- FOS-JUN interaction: PDB entry 1fos (chains E and F), X-ray diffraction structure [71]

We found that precision or accuracy calculated against other reference structures differed only marginally, thus we have limited reporting to the aforementioned PDB entries.

Protein folding

To *ab initio* determine protein structures, we performed simulated annealing using the XPLOR-NIH modeling suite [45] with structural restraints derived from the deep mutational scanning data. Simulations were performed in three stages, in each of which 500 structural models were generated. Stages 1 and 2 served to identify inconsistencies among defined structural restraints. Additionally, in stage 2 an average structure of the best 10% of models was calculated. Stage 3 served to refine this average structure to obtain a final set of best models.

Restraints from top predicted contacts (position pairs with highest interaction scores and linear chain separation greater than 5 positions) were implemented by setting Cβ-Cβ atom distances (Cα in case of Glycine) between positions to range between 0 and 8Å and weighting the restraints according to their relative interaction score (interaction score divided by mean interaction score of all predicted contacts used).

Restrains from secondary structures elements were implemented as dihedral angle restraints. Dihedral angles of both beta strands and alpha helices were set to range between values commonly observed in crystal structures [72], for alpha helices $\Phi_\alpha = -63.5° \pm 4.5°$ and $\Psi_\alpha = -41.5° \pm 5°$ and for beta strands $\Phi_\beta = -118° \pm 10.7°$ and $\Psi_\beta = 134° \pm 8.6°$.

Restraints for beta sheet interactions were implemented by setting H-N:O=C hydrogen bond distances between interacting positions to range between 1.8 and 2.1Å [72], with weight one. Predictions of beta sheet interactions derived from deep mutational scanning data yield a string of interacting positions, but hydrogen bonding in beta sheets occurs in specific non-continuous patterns between position pairs (between alternating positions off the interaction diagonal in parallel beta sheets and between every second set of position pairs in anti-parallel beta sheets). Specifically, for each set of interacting positions there are two alternative patterns of hydrogen bonding possible. These alternative possibilities of pairing were implemented as mutually exclusive selection pairs with the “assign … or” syntax in Xplor-NIH.

1301 Distance restraints were implemented in XPLOR-NIH as NOE (nuclear Overhausser effect)
1302 potential, with potential type set to “soft” for stages 1 and 2 and “hard” for the final simulation
1303 stage. Dihedral angle restraints were implemented via the CDIH potential.

1304 After simulation stages 1 and 2 restraints were checked for their consistency with predicted
1305 structural models. First, structural models were clustered according to their violations of
1306 distance and dihedral angle restraints (k-means clustering, k = 4). Clusters were ranked by the
1307 mean total energy (from all energy potentials used) of their 50 models with lowest total energy
1308 (or all of their structures if clusters are smaller 50 models). From the 50 models with the lowest
1309 total energy from the top-ranked cluster (or however many top-ranked clusters were necessary
1310 to arrive at 50 models) the fraction of models that violate specific restraints was recorded. For
1311 the subsequent simulation stage, distance restraints were down-weighted according to the
1312 fraction of models that violated them, $w_{x,i} = w_{x,i-1} * (1 - f_x)^2$, and distance restraints with a
1313 weight below 0.1 were discarded. There is no option to weight dihedral angle restraints, thus
1314 instead dihedral angle restraints with a ‘weight’ below 1/3 were discarded for the subsequent
1315 simulation stages.

1316 The top 5% structural models from simulation stage 3 were evaluated against the reference
1317 structure. The TM-score program (update 2016/03/23) was used to calculate the Cα root mean
1318 squared deviation and the template modeling score [73].

1319 Several types of control simulations were performed to judge the predictive power of restraints
1320 derived from deep mutational scanning data. As a negative control we performed simulations
1321 without restraints from predicted contacts and beta sheet interactions, but with restraints from
1322 secondary structure elements predicted by PSIPRED (version 3.3, [43]). As a positive control we
1323 performed simulations with restraints derived from the reference structure. Here, L true contacts
1324 of position pairs with linear chain distance greater than 5 amino acids were randomly sampled
1325 and beta sheet interactions were determined by PyMOL [44]. These simulations serve as a
1326 positive control and give the maximally achievable accuracy of our Xplor-NIH workflow.

1327 For the WW domain, simulations on the full mutated 33aa section gave mediocre results, both
1328 when using combined scores with PSIPRED predicted secondary structure (5.8Å Calpha-
1329 RMSD), as well as when using perfect information from the reference structure (4.1Å Calpha-
1330 RMSD). Upon inspection, this seemed to be an issue of the unstructured tail regions. We thus
1331 conducted structural simulations for a truncated version of the WW domain using only mutated
1332 positions 6-29 (the core region including the three beta strands).

For structural simulations of down-sampled GB1 datasets (and DeepContact transformed versions thereof) we used distance restraints derived from top predicted contacts and secondary structure restraints derived from PSIPRED predictions, but no restraints for beta sheet pairing. This was done to avoid skewed results due to false beta sheet pairing predictions in low quality datasets (Extended Data Figure 7B). For structural simulations from DeepContact-transformed predictions, we find that using more tertiary contacts results in better models. We conclude that this is because the deep learning algorithm focuses many strong predictions in few structural features (such as interactions of secondary structure elements), which are therefore the top contacts. Restraints in other regions of the protein are therefore only included if more predicted contacts are used for restraint calculations, therefore improving structural predictions. Because of this, when comparing structural simulations from scores derived before and after deep learning, we compare the top 5% of structural models derived with the top L predicted contacts from original scores with those derived with the top 1.5*L predicted contacts from DeepContact transformed scores.

DeepContact learning

DeepContact software was obtained from GitHub (https://github.com/largelymfs/deepcontact) [47]. We are grateful to Yang Liu and Jian Peng for also making - without any hesitation - their basic DeepContact network architecture available on their GitHub repository and helping us with the implementation. The DeepContact architecture used here only takes one 2D input of predicted contact scores and returns a 2D map of transformed scores (denoted as “DeepContact CCMPred only” in [47] and described in the first paragraph of the result section therein). The DeepContact architecture employed came with a pre-trained network model that had been trained on solved structures of the 40% homology filtered ASTRAL SCOPe 2.06 database (see GitHub repository and Liu, et al. [47]), which were filtered to avoid structure and sequence redundancy of the training data. Because CCMpred scores [74] are distributed in the range of 0 to 1, we pre-normalized our deep mutational scanning derived interaction scores to this range (such that the minimum score on the interaction score matrix was 0 and the maximum score was 1) before providing them as an input to DeepContact. As negative control, we created for each dataset three random permutations of *combined score* matrices (while preserving matrix symmetry; in case of FOS-JUN dataset non-symmetric *epistasis score* matrices were permutated), which were transformed by the DeepContact algorithm. These control datasets show no increased precision of random expectation (Figure 6F).

Code availability

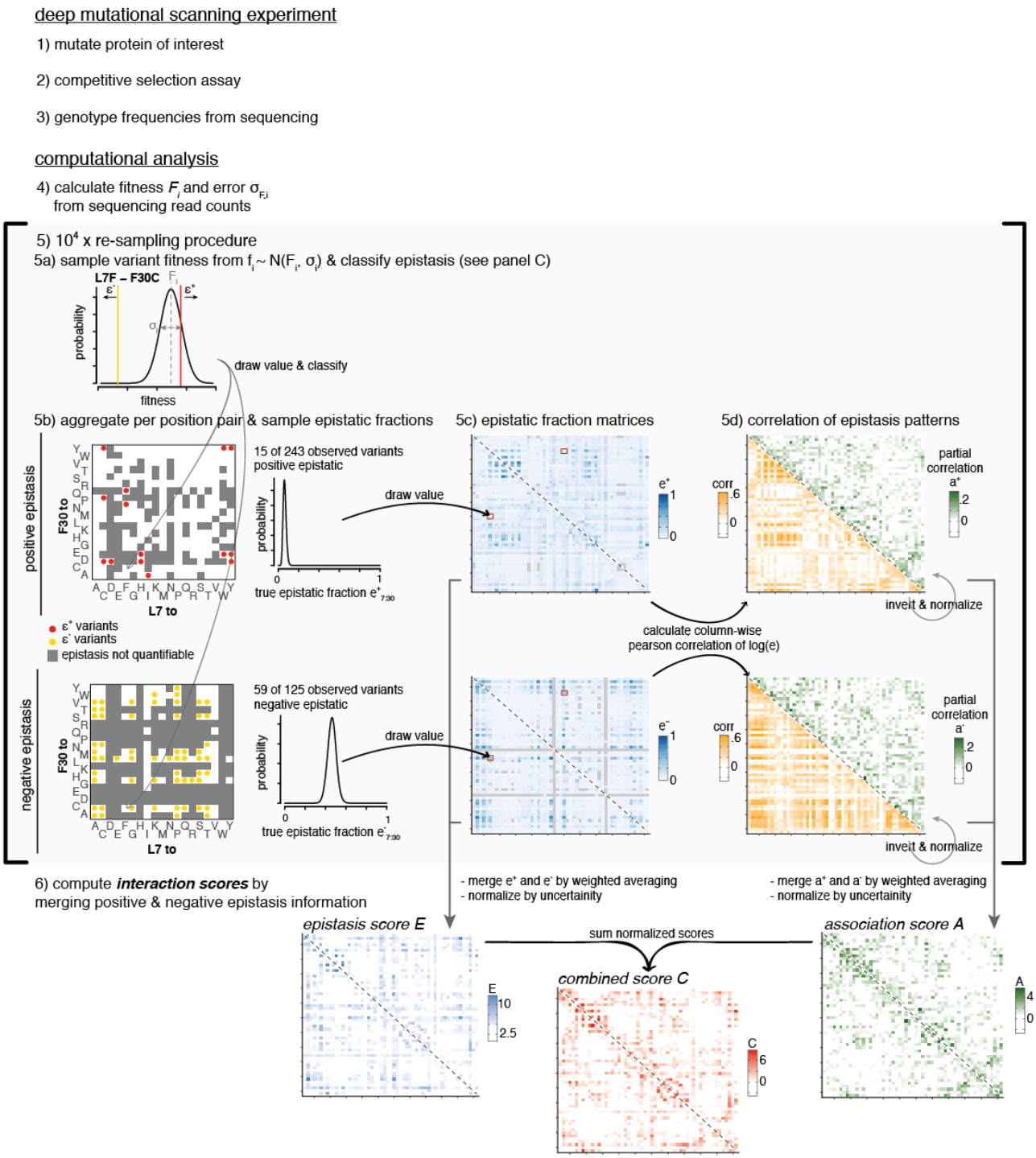
Data was analyzed with custom scripts written and executed in R programming language, version 3.4.3. Structural simulations were performed with Xplor-NIH modeling suite version 2.46. Analysis scripts are available at https://github.com/lehner-lab/DMS2structure.

Data availability

No primary data was generated in this study. Processed interaction scores for all datasets are included in Supplementary Table 1. All intermediate steps of data processing can be recapitulated with the scripts at https://github.com/lehner-lab/DMS2structure.

Extended Data Figures

Extended Data Figure 1

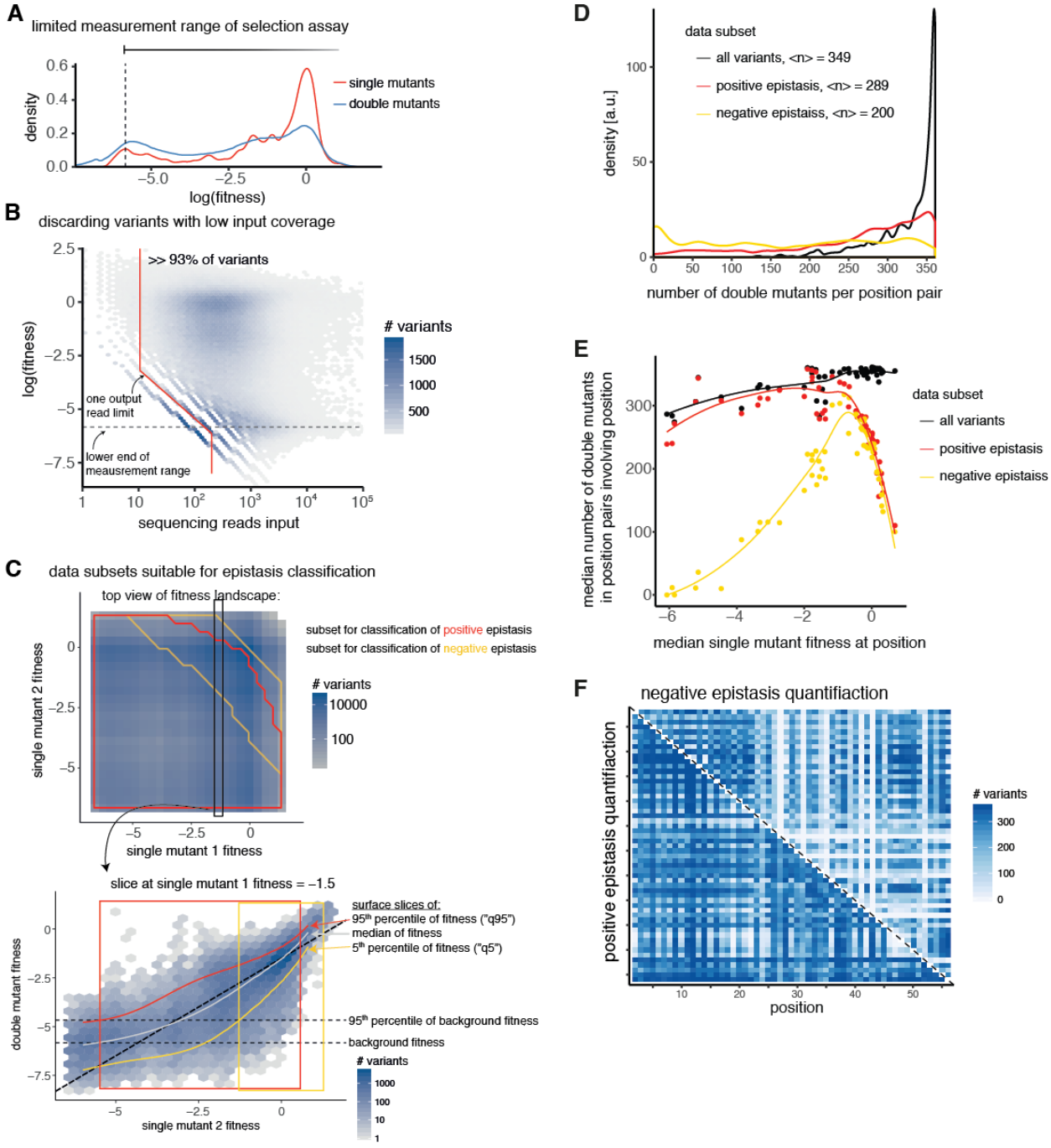


1378 # Extended Data Figure 1: Deep mutational sequencing data to contact
1379 prediction workflow

1380 Overview of workflow to predict interacting position pairs from deep mutational scanning
1381 datasets (see Methods and Results).

1382

Extended Data Figure 2



Extended Data Figure 2: GB1 deep mutational scanning data processing

A. Distribution of fitness values for single and double mutant variants. Lower peak in distributions indicates lower limit of fitness assay measurement range (see Methods).

1387 B. Two-dimensional variant density showing dependency of fitness values on sequencing
1388 read counts in input library. For variants with very low coverage in the input library low
1389 fitness values cannot be accurately estimated. Red line shows sequence read cutoff
1390 used for variant inclusion (93% of variants included for downstream analysis). Horizontal
1391 dashed line indicates lower limit of fitness assay measurement range.

1392 C. Reliable quantification of positive and negative epistasis is limited to subsets of the data.
1393 Upper plot shows two-dimensional double mutant variant density in single mutant fitness
1394 space. Red outline shows single mutant fitness space that enables positive epistasis
1395 quantification. Yellow outline shows single mutant fitness space that enables negative
1396 epistasis quantification. Lower plot shows an example slice through fitness landscape at
1397 single mutant fitness = -1.5. Red, grey and yellow curves show slices through quantile
1398 fitness surfaces of 95th percentile, median and 5th percentile, respectively (see Figure
1399 1B). Diagonal dashed line shows $double\ mutant\ fitness =$
1400 $single\ mutant\ 2\ fitness - 1.5$ (expected fitness = observed fitness). Horizontal dashed
1401 lines give lower limit of fitness assay measurement range and the 95th percentile of
1402 fitness values of variants dominated by background fitness effects. Red and yellow
1403 boxes indicate the range that includes 99% of variants within the slice that are suitable
1404 for positive or negative epistasis quantification, respectively.

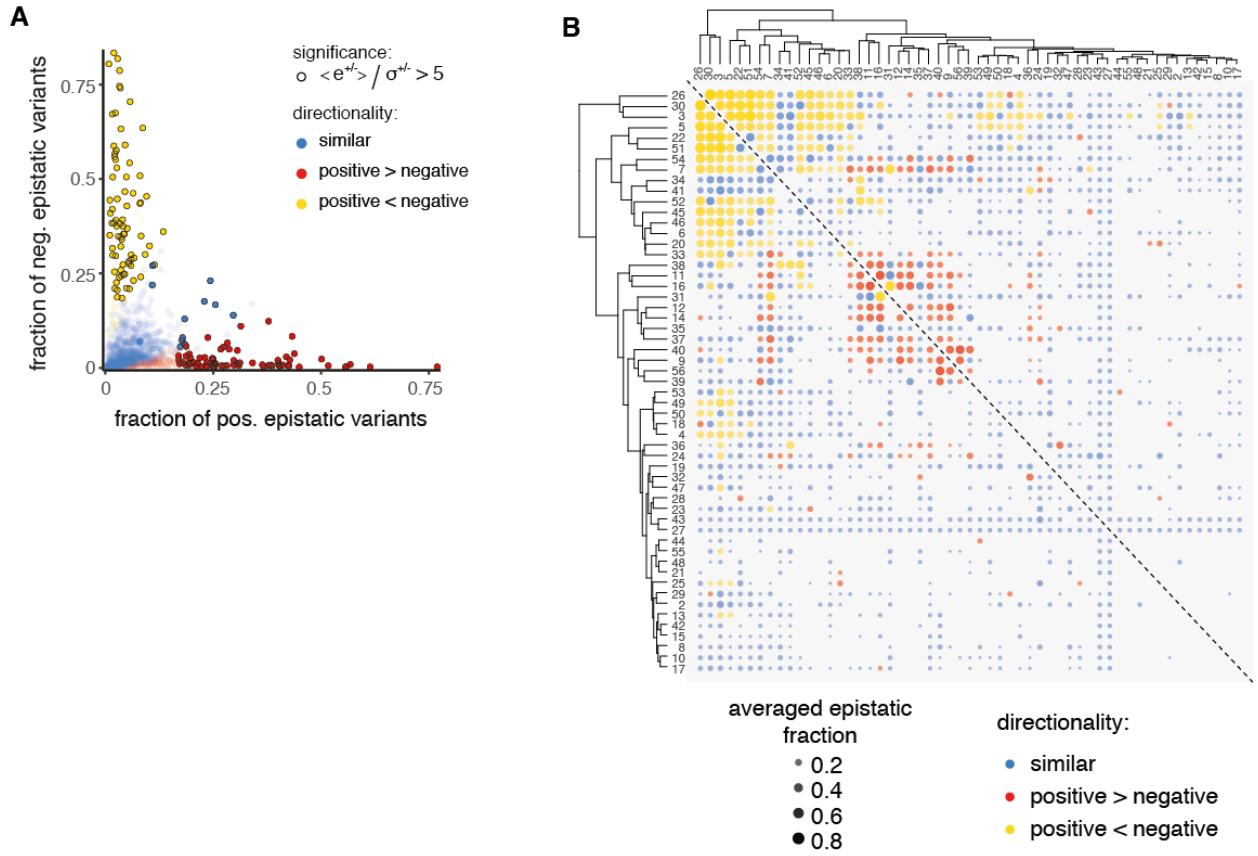
1405 D. Distribution of number of double mutant variants across all position pair. Legend gives
1406 median number of double mutants per position pair for different data subsets.

1407 E. Relationship between median single mutant fitness at a position and the median number
1408 of double mutants observed in position pairs the position is involved in. Curves are loess
1409 smoothed. Across all variants, positions with stronger fitness effects show lower
1410 coverage of double mutants. Restrictions for quantification of positive epistasis
1411 additionally reduce coverage for positions with mostly neutral or positive effects. Finally,
1412 restrictions for quantification of negative epistasis strongly reduce coverage for positions
1413 with strong fitness effects, due to the lower measurement limit of the fitness assay.

1414 F. Number of double mutants for which positive (lower left triangle) or negative (upper right
1415 triangle) epistasis can be quantified per position pair plotted on the interaction matrix.

1416

Extended Data Figure 3



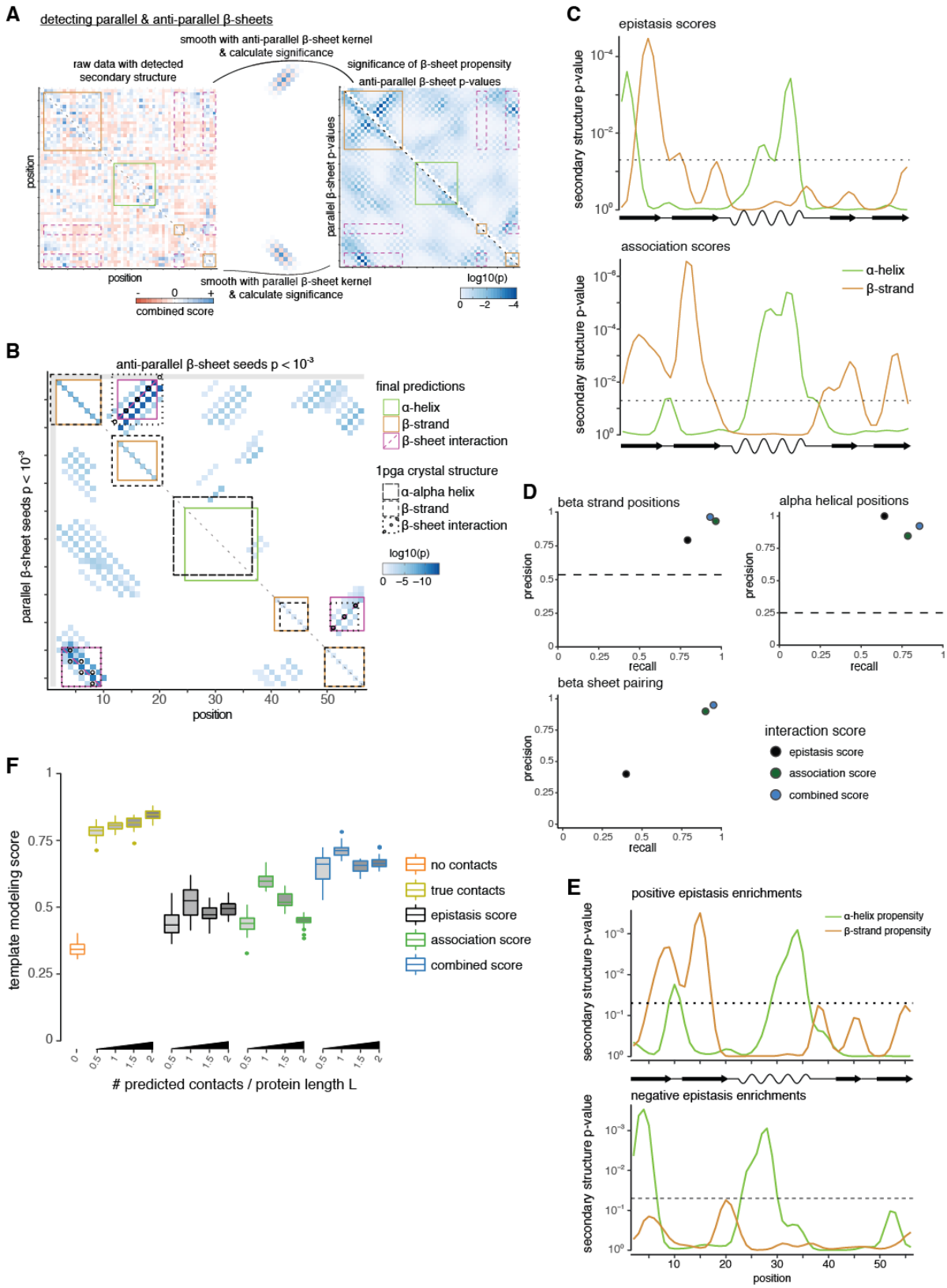
Extended Data Figure 3: Positive and negative epistasis enrichments across position pairs

A. Position pair-wise fractions of positive and negative epistatic variants. Black circles mark position pairs with highly significant fractions (either positive or negative, or both); red dots: positive epistatic fraction significantly larger than negative epistatic fraction; yellow dots: negative epistatic fraction significantly larger than positive epistatic fraction; blue dots: no significant differences.

B. Hierarchical cluster analysis of epistatic fraction patterns. Positions are clustered according to the Euclidean distance of their mean epistatic fractions (weighted average of positive and negative epistatic fractions, weights are inverse variances of fractions in resampling runs) to all other positions. Note that directionality of interactions (positive or negative fractions more significant) was not used for clustering but only marked post-analysis. Clustering shows two highly interconnected clusters of positions that interact mostly positively or

negatively within each cluster but hardly any strong interactions are observed between the two clusters (with exception of positions 7 and 54).

Extended Data Figure 4



1435

1436 Extended Data Figure 4: Secondary and tertiary structure prediction for
1437 GB1 domain

1438 A. Detecting beta sheet pairing with two-dimensional kernel smoothing. Left plot shows raw
1439 combined score interaction matrix, with secondary structure element predictions (see
1440 Figure 4A,B) marked as squares along the diagonal (red – beta strand, green – alpha
1441 helix). Off-diagonal orange rectangles show potential regions of beta sheet pairing. Right
1442 plot: calculation of beta sheet pairing propensity with beta sheet kernels. Upper right
1443 triangle shows anti-parallel beta sheet propensity. Lower left shows parallel beta sheet
1444 propensity.

1445 B. Matrix of aggregated propensities of beta sheet pairing stretches (upper right – anti-
1446 parallel, lower left – parallel, $p < 10^{-3}$) and the predictions for beta sheet pairing and
1447 secondary structure elements derived from them. In brief, predictions are performed by
1448 picking the highest propensity stretch that is consistent with predicted beta strands, if
1449 necessary modifying beta strand predictions (e.g. introducing an initially not predicted
1450 split between beta strands 1 and 2), then disregarding all stretches that conflict with the
1451 picked top-stretch. This procedure is repeated until no more beta sheet stretches with
1452 propensity $P < 10^{-3}$ are left. Finally, beta strand predictions are updated such that only
1453 positions involved in a beta sheet interaction are retained. Reference elements from
1454 crystal structure are shown as comparison (lower triangle – parallel beta sheets, upper
1455 triangle – anti-parallel beta sheet, diagonal –secondary structure elements).

1456 C. Secondary structure propensity derived from kernel smoothing (red – beta strand, green
1457 – alpha helix) for *epistasis* (upper*)* and *association scores* (lower). P-values were
1458 derived by comparison to randomized datasets (see Methods). Dashed line indicates p =
1459 0.05.

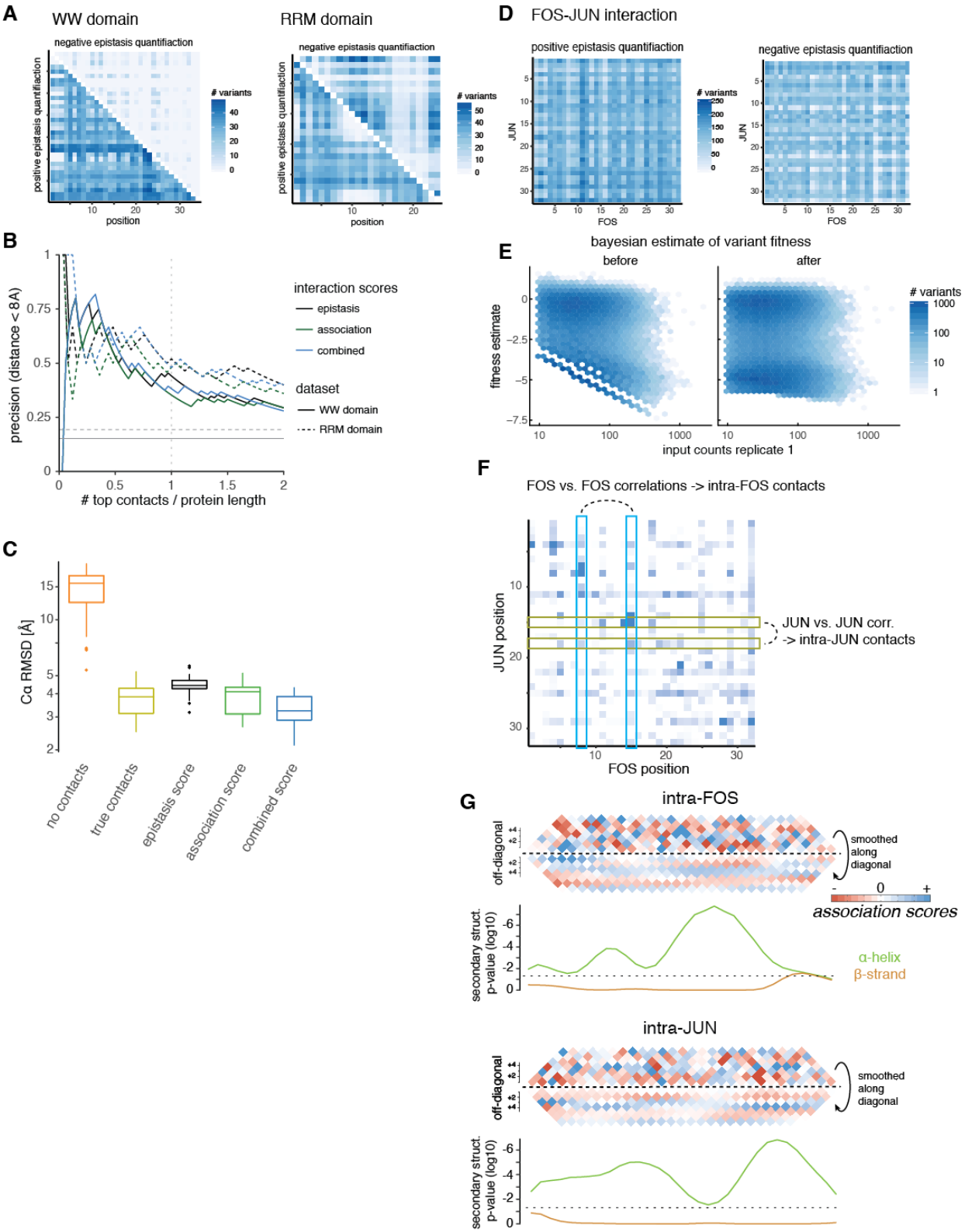
1460 D. Precision and recall for beta strand, alpha helix and beta sheet predictions from
1461 *epistasis*, *association* and *combined scores* (in comparison to crystal structure). Dashed
1462 lines for beta strand and alpha helical positions give random expectation. Random
1463 expectation for beta sheet pairing precision is below 1%.

1464 E. Secondary structure propensities derived from local positive or negative epistatic
1465 enrichments. The upper panel shows secondary structure propensity derived from
1466 positive epistatic interactions, which are in line with secondary structure elements in the
1467 GB1 crystal structure (PDB entry 1pga). The lower panel shows secondary structure
1468 propensity derived from negative epistatic interactions, which are devoid of beta strand

signals and instead show a three-helical pattern, which is reminiscent of the three-helical structure of the protein G A domain that binds albumin [57].

F. Template modeling score of top 5% structural models compared to crystal structure 1pga and the dependency on number of predicted contacts used. “No contacts” – only restraints for secondary structure predicted by PSIPRED. “True contacts’’ – restraints derived from 0.5-2*L contacts (linear sequence separation greater than 5 positions, random subset), secondary structure elements and beta sheet interactions from crystal structure. All other: restraints derived from top 0.5-2*L contacts, secondary structure element and beta sheet interaction predictions from the three interaction scores, as indicated by color.

Extended Data Figure 5



1480

1481 # Extended Data Figure 5: WW and RRM domain and FOS-JUN
1482 interaction dataset properties

1483 A. Number of double mutants for which positive (lower left triangle) or negative (upper right
1484 triangle) epistasis can be quantified per position pair plotted on the interaction matrix for
1485 WW (left) and RRM (right) domains.

1486 B. Precision of interaction scores to predict direct contacts (distance < 8Å in reference
1487 structure) as a function of top scoring position pairs for WW and RRM domain interaction
1488 scores. Color denotes interaction scores, solid lines for WW domain, dashed lines for
1489 RRM domain. Grey horizontal lines give random expectation. Only position pairs with
1490 linear sequence separation greater than 5 amino acids are considered.

1491 C. Accuracy ($C\alpha$ root-mean-square deviation) of top 5% structural models of the WW
1492 domain (core positions 6-29) generated from deep mutational scanning data derived
1493 restraints compared to reference structure (PDB entry 1k9q). Structural models were
1494 generated in XPLOR-NIH by simulated annealing with restraints derived from top 17
1495 tertiary contacts and secondary structure elements predicted by PSIPRED. No beta
1496 sheet pairing information was used.

1497 D. Number of double mutants for which positive (left) or negative (right) epistasis can be
1498 quantified per position pair plotted on the trans-interaction matrix of the FOS-JUN
1499 interaction.

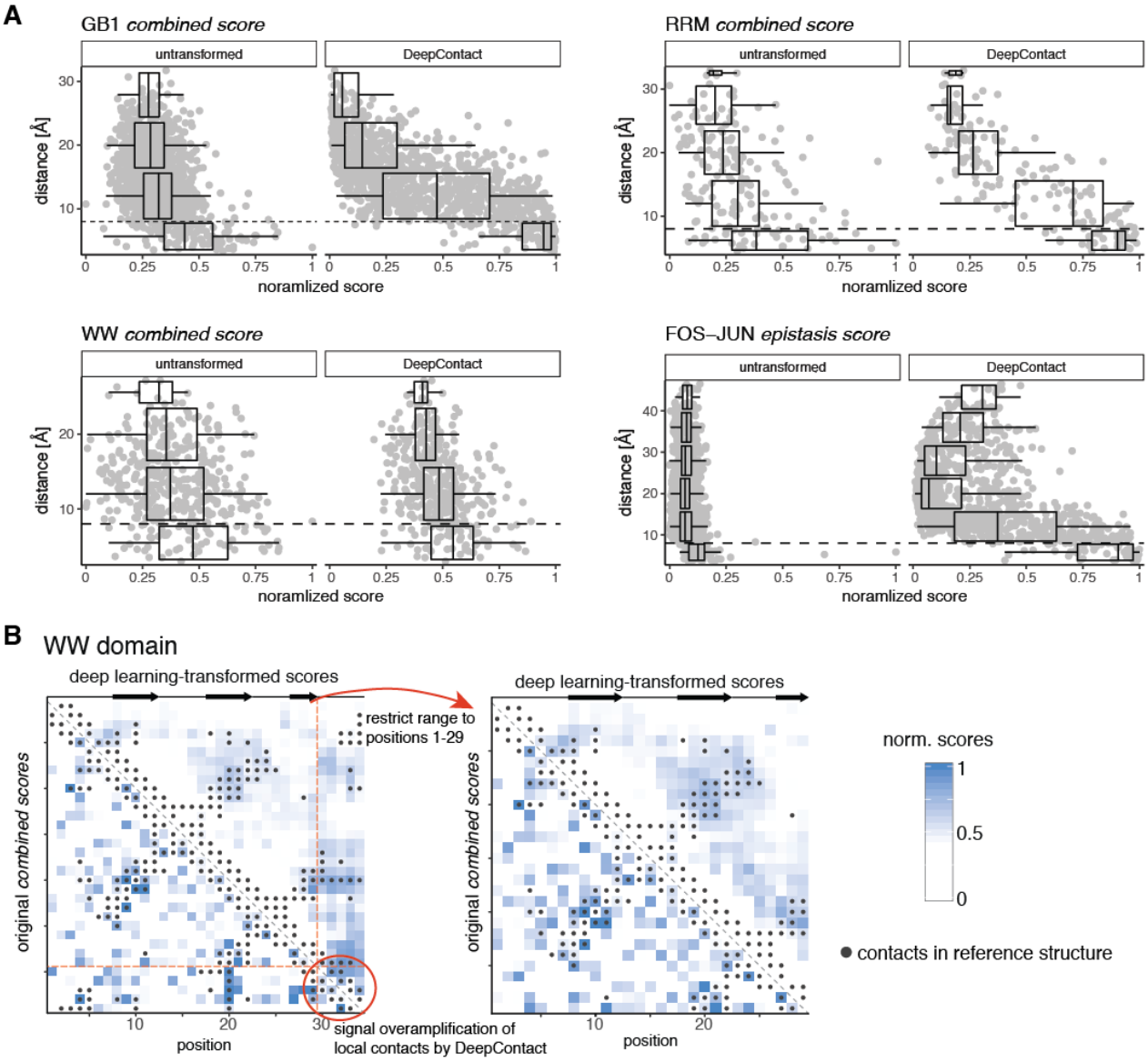
1500 E. Bayesian estimation of fitness values in FOS-JUN interaction data. Mutants with low
1501 input sequencing coverage display limited measurement range and many dropouts
1502 (~15% of variants without reads in output). Left panel shows original fitness distribution
1503 as function of input coverage in replicate 1, right panel shows Bayesian estimates of
1504 fitness as function of input coverage in replicate 1.

1505 F. Learning about intra-molecular contacts in FOS or JUN from epistatic pattern
1506 correlations. Column-wise correlation of epistatic patterns of the trans interaction score
1507 map serve to calculate intra-FOS *association scores* and thus reveal relationships
1508 between positions in FOS. Likewise, row-wise correlation of epistatic patterns reveal
1509 relationships between positions in JUN.

1510 G. Local interactions in intra-molecular *association scores* reveal secondary structures of
1511 protein interaction partners. Upper panels: Data above diagonal shows *association score*
1512 data close to the diagonal, i.e. local interactions. Data below the diagonal is smoothed
1513 with a Gaussian kernel to reveal interaction periodicity. Lower panels: Secondary
1514 structure propensities derived from kernel smoothing (see Figure 4A-C). Green indicates
1515 alpha helical propensity, orange indicates beta sheet propensity, $p = 0.05$ is indicated by
1516 dashed line.

1517

Extended Data Figure 6



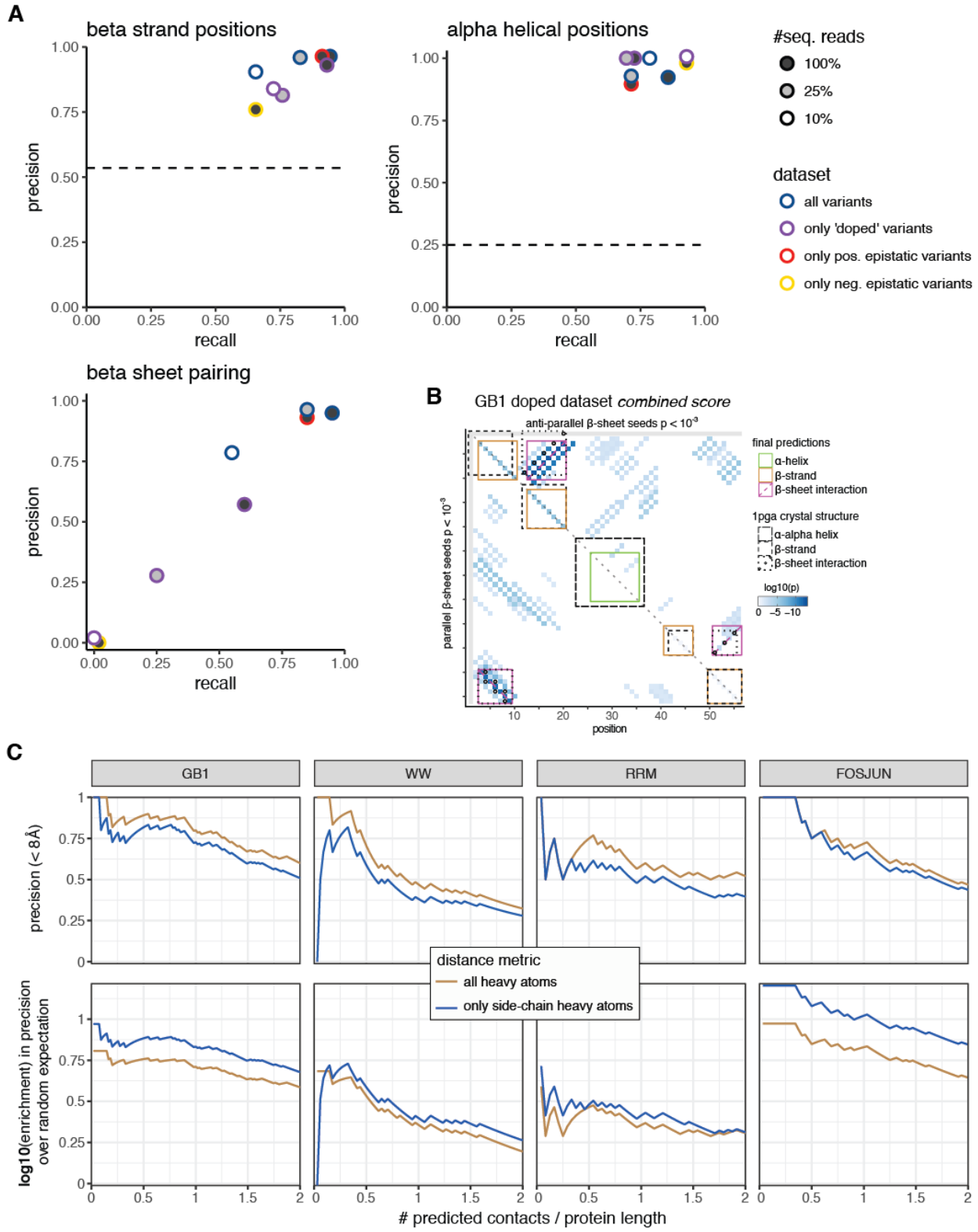
Extended Data Figure 6: Deep learning improves contact prediction from deep mutagenesis data

A. Distance of position pairs as a function of interactions scores before (left panel, scores normalized to interval [0,1]) and after (right panel) transformation with *DeepContact* for the four datasets. Boxplots are spaced in distance intervals [0,8), [8,16), [16,24), [24,32), [32,40) and [40,48) Å. Dashed horizontal line indicates 8Å.

B. Left: Full WW domain *combined score* interaction map before (lower left) and after (upper right) DeepContact transformation. DeepContact amplifies a signal from local

contacts in the C-terminal region of the domain, thus concentrating the strongest transformed signal in this region. Removing positions 30-34 removes this artefact (right plot). Heat maps show interaction scores that have been normalized to have similar range. Grey dots show contacts (distance < 8Å) in reference structure.

Extended Data Figure 7



Extended Data Figure 7: Distance metric comparison and secondary structure prediction for lower data-quality GB1 datasets

A. Precision and recall for beta strand, alpha helix and beta sheet predictions derived from combined scores of down-sampled GB1 datasets (in comparison to reference structure). Dashed lines for beta strand and alpha helical positions give random expectation. Random expectation for beta sheet pairing precision is below 1%. Note that some coinciding data points were slightly moved for better identifiability.

B. Beta sheet pairing predictions for the doped GB1 dataset with 100% sequencing read coverage (cf. Extended Data Figure 4B). Beta sheet pairing between beta strands 1 and 2 is predicted in correct anti-parallel direction, but exact pairing of positions are off by 2; thus precision and recall of beta sheet pairing for doped GB1 dataset drops to ~60% (see panel B).

C. Differences in precision and enrichment over random expectation for all heavy atom or side-chain heavy atom distance metrics. As expected, using all heavy atoms (including backbone heavy atoms) increases precision of predicted contacts by about 10%. Restricting distance measurements to side-chain heavy atoms, however, increases precision over random expectation, often by more than 2-fold (note the log10-scale), indicating that side-chain distances are more informative for epistatic interactions. For these calculations, only position pairs with linear sequence separation greater than 5 amino acids were considered.