Determining protein structures using genetics

Jörn M. Schmiedel[1,2], 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

*Corresponding Author and Lead Contact: 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 should be widely applicable, including to macromolecules whose structures are difficult to determine by physical techniques.

Introduction

Mutations within a protein or RNA can have non-independent effects on folding, stability and activity (Fowler and Fields, 2014; Lehner, 2011; Starr and Thornton, 2016; Tokuriki and Tawfik, 2009). Consistently, quantifying the evolutionary couplings between amino acid and nucleotide occurrences in protein and RNA families yields structural and functional information tha has been used to predict macromolecular structures, physical interactions and mutational effect sizes (Altschuh et al., 1987; Burger and van Nimwegen, 2010; Dunn et al., 2008; Ekeberg et al., 2013; Gloor et al., 2005; Göbel et al., 1994; Halabi et al., 2009; Hopf et al., 2012; Hopf et al., 2017; Jones et al., 2012; Kamisetty et al., 2013; Lockless and Ranganathan, 1999; Marks et al., 2011; McLaughlin Jr et al., 2012; Michel et al., 2014; Morcos et al., 2011; Ovchinnikov et al., 2015; Ovchinnikov et al., 2017; Seemayer et al., 2014; Tóth-Petróczy et al., 2016; Weigt et al., 2009; Weinreb et al., 2016). Current state of the art structure prediction algorithms routinely incorporate evolutionary sequence information to predict tertiary protein contacts (Schaarschmidt et al., 2018). However, the success of this strategy depends on the depth and diversity of sequence alignments, with the approach failing for many proteins because of insufficient number of homologs (Marks et al., 2012; Ovchinnikov et al., 2017; Tóth-Petróczy et al., 2016).

An alternative approach to structure prediction would be to perform experiments that directly measure the dependencies between mutations in a molecule (Fowler and Fields, 2014) (Figure 1A). Such an approach, if successful, would provide a general strategy to determine macromolecular structures, including for molecules without sufficient numbers of homologs in sequence databases and for molecules whose structures are difficult to determine by physical techniques. All that would be required would be a selection assay that allows the molecular activity of a sufficiently large number of individual mutations and combinations of mutations to be quantified. However, whether such an approach is actually feasible is unknown.

Here we show that deep mutagenesis of a molecule can indeed provide sufficient information to determine its macromolecular structure. Our method quantifies how often residue pairs interact epistatically – that is non-independently – in experimental data and how such epistatic interaction patterns correlate. These data accurately identify individual tertiary structure contacts as well as secondary structure elements and their interactions in a protein domain. Moreover, we show that these data can be used to predict the structure of the protein domain at 1.9Å root mean squared deviation (RMSD) compared to a reference crystal structure.

Results

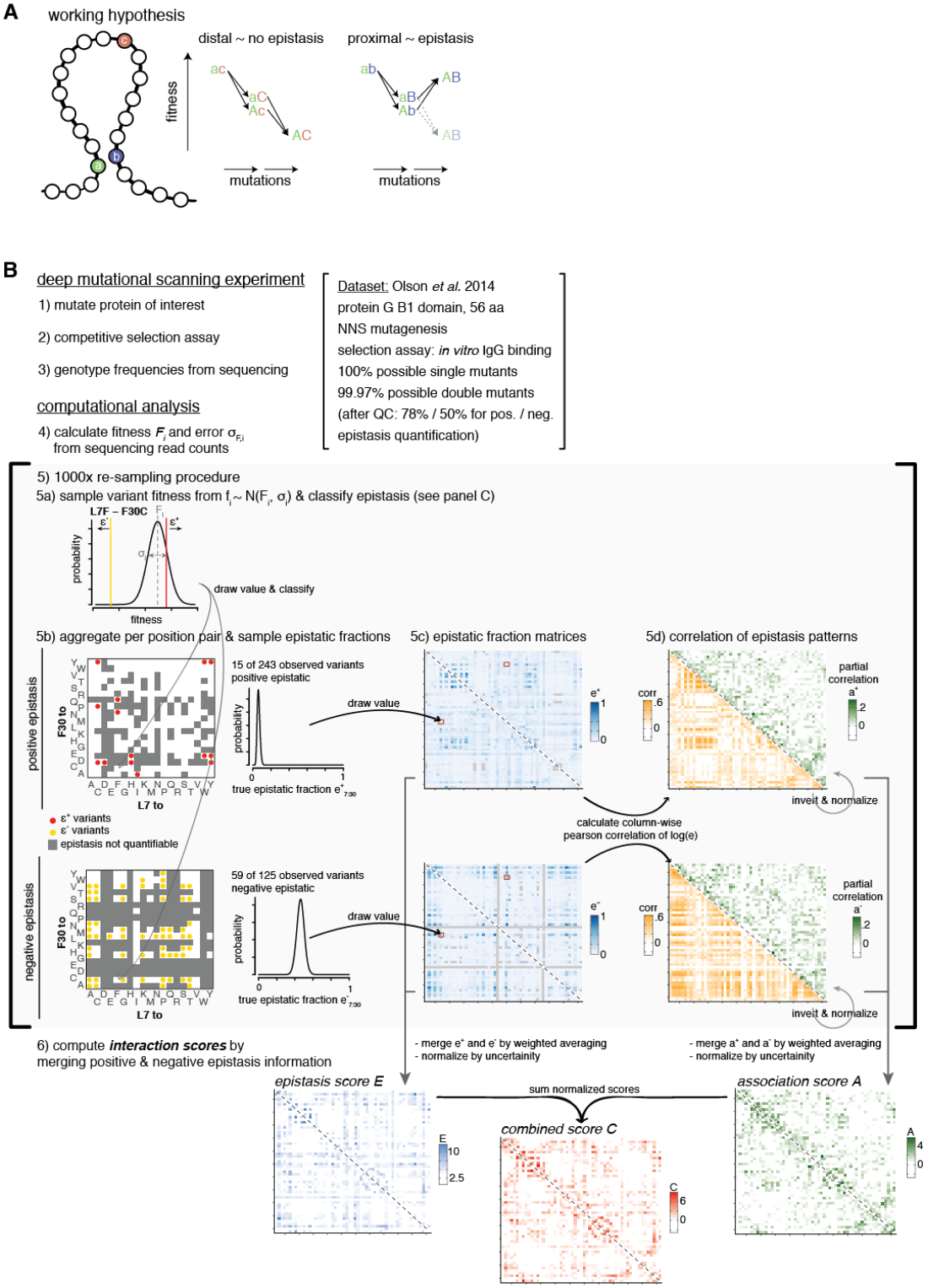
Epistatic variants within a protein domain are enriched for structural proximity

We used deep mutational scanning data for the immunoglobulin-binding protein G domain B1 (Olson et al., 2014) to investigate how genetic – or epistatic – interactions between mutations in a protein relate to structure. 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 (Figure 1B, step 1). The resulting library consists of more than half a million variants (55*19 = 1,045 single mutants plus 55*54/2*19*19 = 536,085 double mutants). mRNA display was then used to combine an in vitro immunoglobulin G binding assay with a sequencing readout of protein fitness (here defined as propensity to bind immunoglobulin, Figure 1B, step 2). The protein fitness of each mutant was calculated as the change in read frequencies in the library due to selection relative to the wild-type variant (Figure 1B, steps 3 & 4). The assay provides a measurement range of more than two orders of magnitude, before background effects dominate the measured fitness values (Figure S1A), and a median error relative to the measurement range of 3.2%, as judged from count-based error estimates (see 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 1). We 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 (Figure 2A). We then similarly calculated 5th and 95th percentile fitness surfaces; and classified double mutants as positive epistatic if they were fitter than the 95th percentile fitness surface and as negative epistatic if they were less fit than the 5th percentile fitness surface (Figure 2A). Basing the assessment of epistasis on this non-parametric null model accounts for nonlinearities introduced by the finite measurement range of the fitness assay, variable magnitudes of measurement error across the fitness landscape (Figure S1B), as well as non-specific epistatic behavior resulting from, for example, thermodynamic stability threshold effects (Diss and Lehner, 2018;

84 Tokuriki and Tawfik, 2009). We restricted the evaluation of positive or negative epistasis to
85 specific subsets of the data, where measurement errors do not impede epistasis classification
86 (Figure S1D, see Methods). Most importantly, we classified negative epistasis only for double
87 mutants whose expected fitness is significantly above the lower measurement limit of the fitness
88 assay. As a result of these restrictions as well as differences in initial coverage, the number of
89 double mutant variants that can be used to assess positive and negative epistasis varies
90 substantially across position pairs (with an average of 252 and 187 variants per position pair in
91 the positive and negative epistasis subsets, respectively; Figure S2). Consistent with our initial
92 hypothesis and previous observations (Melamed et al., 2013; Olson et al., 2014; Sahoo et al.,
93 2015), both positive and negative epistatic double mutants are enriched for proximal variants;
94 and this is not due to the specific subsets of the data being analyzed (Figure 2B).

Figure 1


A
working hypothesis
fitness
distal ~ no epistasis
proximal ~ epistasis
ac
aC
Ac
AC
ab
aB
Ab
AB
AB
mutations
mutations
B
deep mutational scanning experiment
1) mutate protein of interest
2) competitive selection assay
3) genotype frequencies from sequencing
computational analysis
4) calculate fitness F_i and error $\sigma_{F,i}$ from sequencing read counts
Dataset: Olson et al. 2014
protein G B1 domain, 56 aa
NNS mutagenesis
selection assay: in vitro IgG binding
100% possible single mutants
99.97% possible double mutants
(after QC: 78% / 50% for pos. / neg. epistasis quantification)
5) 1000x re-sampling procedure
5a) sample variant fitness from $f_i \sim N(F_i, \sigma_i)$ & classify epistasis (see panel C)
L7F – F30C
probability
fitness
draw value & classify
5b) aggregate per position pair & sample epistatic fractions
positive epistasis
negative epistasis
F30 to
L7 to
ε⁺ variants
ε⁻ variants
epistasis not quantifiable
15 of 243 observed variants positive epistatic
59 of 125 observed variants negative epistatic
draw value
true epistatic fraction $e^+_{7,30}$
true epistatic fraction $e^-_{7,30}$
5c) epistatic fraction matrices
5d) correlation of epistasis patterns
corr
partial correlation a⁺
partial correlation a⁻
inveit & normalize
calculate column-wise pearson correlation of log(e)
6) compute interaction scores by merging positive & negative epistasis information
- merge e⁺ and e⁻ by weighted averaging
- normalize by uncertainity
- merge a⁺ and a⁻ by weighted averaging
- normalize by uncertainity
epistasis score E
sum normalized scores
combined score C
association score A

95

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

A. We hypothesize that epistatic interactions mainly result from structural interactions. Epistasis is the non-independence of single mutant fitness effects, i.e. fitness of the double mutant deviates from (multiplicative) expectation.

B. Workflow to predict interacting position pairs from deep mutational scanning datasets (see Methods and Results).

QC: quality control filtering of double mutant variants.

Figure 2

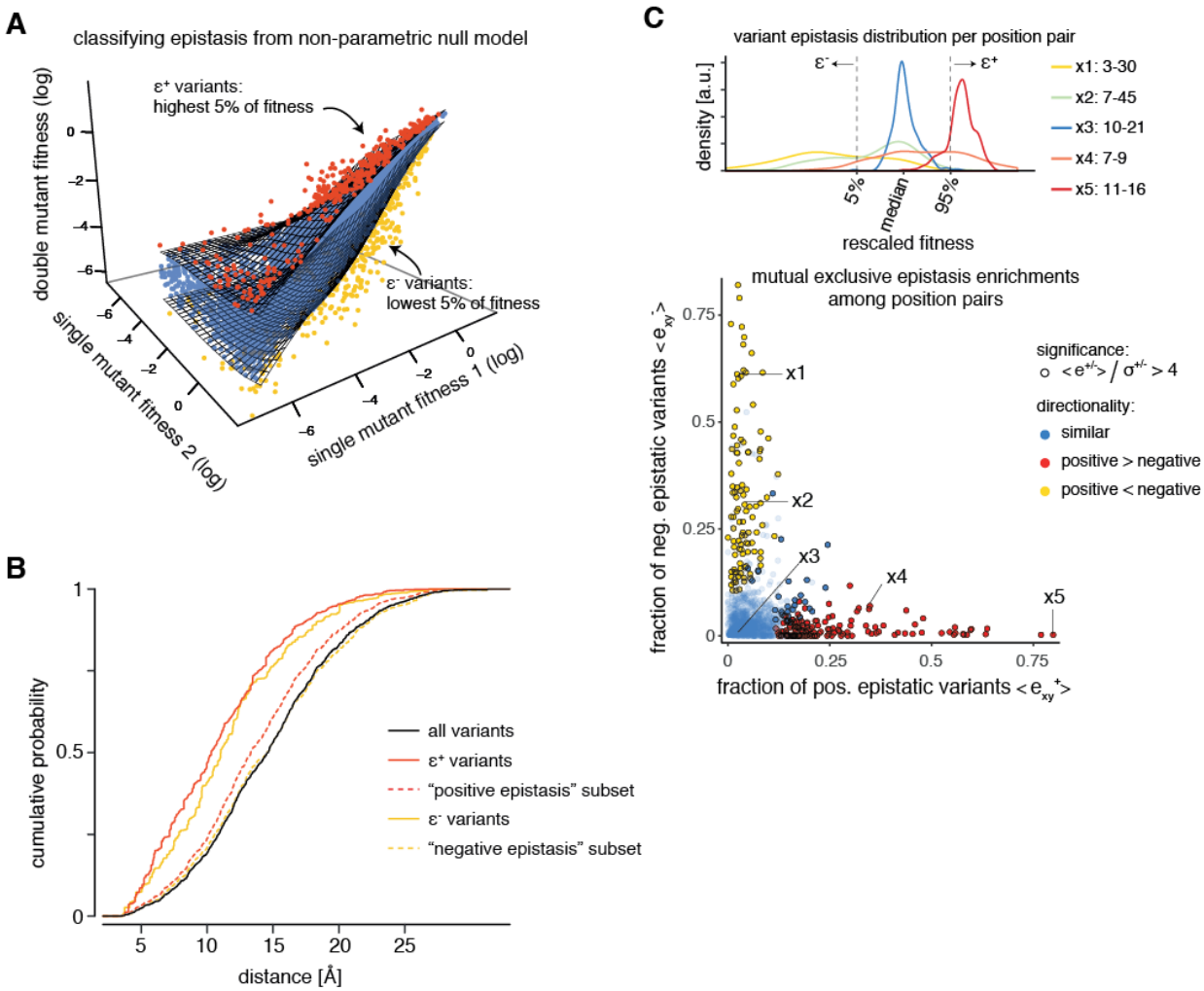


Figure 2: Epistatic variants are enriched for structural proximity

A. Non-parametric null model of epistasis, based on running quantile surfaces, accounts for non-linear dependencies due to lower fitness limit of assay and unspecific epistatic effects. 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.

B. Cumulative probability of minimal distances between side chain heavy atoms of double mutant variant pairs. Positive and negative epistasis subsets refer to the sets of variants applicable for epistasis analysis, see Figure S1D. All variants, n = 497011; positive

116 epistasis subset, n = 378884; positive epistatic variants ε^{+}, n = 16907; negative
117 epistasis subset, n = 262768; negative epistatic variants ε^{-}, n = 10708.

118 C. Upper panel: Examples of double mutant variant distribution of position pairs in rescaled
119 fitness space (fitness of variants below median rescaled by 5th percentile ("5%"), fitness
120 of variants above median fitness rescaled by 95th percentile ("95%")). Lower panel:
121 Position pair-wise fractions of positive and negative epistatic variants. Shown are mean
122 epistatic fractions across 1.000 resampling runs. Black circles mark position pairs with
123 highly significant fractions (mean fraction (either positive or negative epistatic)
124 normalized by standard deviation of fraction over resampling runs greater than 4). Red
125 marks position pairs with significantly larger positive than negative epistatic fractions.
126 Yellow marks opposite position pairs. Blue marks positions pairs with non-significant
127 differences. "x1-5" mark example pairs shown in upper panel.

128

129 Position pairs enriched for epistatic interactions predict tertiary 130 structure contacts

131 For each position pair, we next calculated the fraction of positive or negative epistatic double
132 mutant variants (Figure 1B, step 5). Even moderate enrichments for positive and negative
133 epistatic variants appear to be mutually exclusive (Figure 2C). We find that the reason for this
134 are two clusters of positions that exhibit mostly positive or negative interactions among
135 themselves as well as towards subsets of non-cluster positions (Figure S3; as also noted before
136 by Olson et al. (2014)). Both clusters of positive and negative interactions form blocks on the
137 contact map in regions that are enriched for teritary contacts (Figure 3A). We calculated the
138 number of edges of distance smaller than 8 Ångström (Å) that connect position pairs. Consistent
139 with interaction clusters forming a dense network of proximal positions, we find that 45% of the
140 top 100 positions pairs with high fractions of positive or negative epistatic interactions are direct
141 contacts (one edge, 2.4-fold over expectation) and another 41% or 45% share a common
142 neighbour (two edges), while interactions across more edges are depleted (Figure 3C).

143 We attempted to derive a single *epistasis score* per position pair from positive and negative
144 epistasis interations. We therefore first estimated the uncertainty in the fraction of positive or
145 negative epistatic variants per position pair due to measurement errors and the finite number of
146 observed double mutant variants (Figure 1B, step 5). We implemented a resampling approach

147 where, in each of 1000 runs, each variant's fitness is drawn from a normal distribution with the
148 measured fitness as mean and the uncertainty due to sequencing coverage as standard
149 deviation (step 5a). In each resampling run we re-classified positive and negative epistasis of
150 variants given the drawn fitness values. Finally, the fraction of positive (or negative) epistatic
151 variants for each position pair was drawn from the posterior probability distribution of epistatic
152 fractions (step 5b). The posterior probability distribution is the product of the prior probability
153 distribution - the distribution of epistatic fractions over all position pairs - and the likelihood
154 function – how likely a given fraction is to generate the observed fraction of epistatic variants.
155 The uncertainty of the fraction of epistatic variants per position pair is then the standard
156 deviation of drawn fractions over all resampling runs.

157 We then combined positive and negative epistasis information by computing the weighted
158 average of mean epistatic fractions over all resampling runs given their uncertainty. The final
159 *epistasis score* is a z-score, which is obtained by further normalizing this weighted average by
160 its uncertainty, to give priority to high confidence enrichments (step 6).

161 The position pairs with highest *epistasis scores* are well distributed across the contact map
162 (Figure 3B), and the number of direct contacts (one edge < 8Å) among the top 100 *epistasis*
163 *score* pairs increases to 57% (Figure 3C), thus showing that the *epistasis score* succesfully
164 incorporates information from both positive and negative epistasis. Additionally, *epistasis scores*
165 generally correlate with proximity in the tertiary structure (minimal distance between side-chain
166 heavy atoms, rank correlation coefficient = -0.39, $p < 10^{-6}$, n = 1485; Figure 3D).

167 We examined the precision to predict direct contacts of amino acids spaced more than five
168 positions apart (commonly used as restraints for structural simulations, Fig. 3E). Precision is
169 82% for the top L/5 (n=11, L as length of the protein, 56aa) *epistasis score* pairs, 71% for the
170 top L/2 pairs and 61% for the top L pairs. Of note, the precision to predict direct contacts from
171 *epistasis scores* is 10-20% higher than that of the non-normalized epistasis score, thus showing
172 the need to account for experimental uncertainity.

173 Together, our analyses suggest that physical interactions are a main determinant of epistasis in
174 this protein domain. Nonetheless, the large fractions of epistatically interacting position pairs
175 separated by a common neighbor (two edges), as well as their connectivity observed on the
176 contact map, highlight the fact that strong epistatic interactions do also arise from indirect
177 interactions. Nonetheless, epistatic interaction data suffice to predict tertiary contacts with high
178 precision.

Figure 3

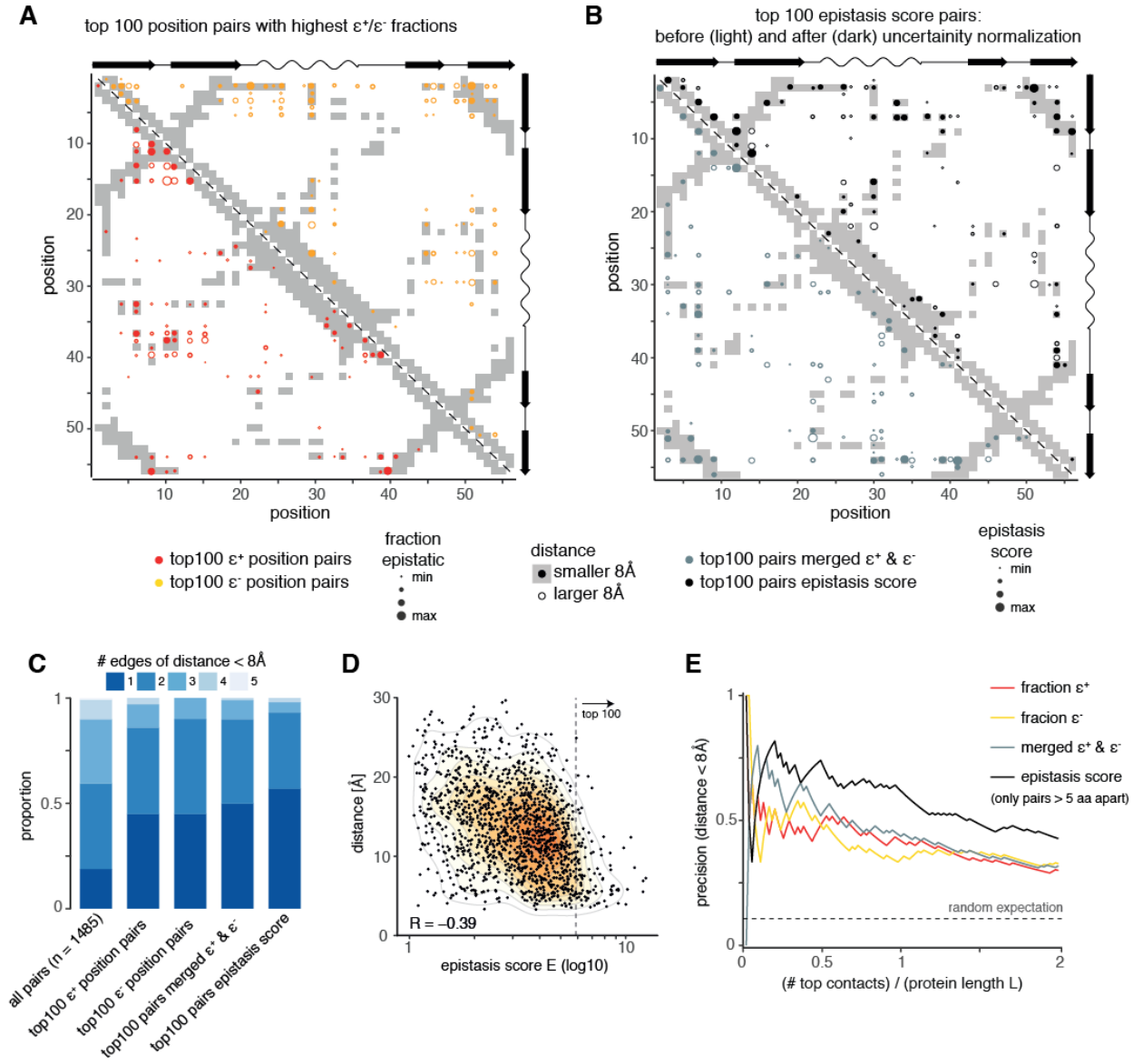


Figure 3: Position pairs enriched for epistatic interactions predict tertiary structure contacts

A. Interaction score map for top 100 position pairs with highest positive (lower triangle, red) and negative (upper triangle, yellow) epistatic fractions. Dot size indicates mean epistatic fractions across 1000 re-sampling runs relative to average over top 100 pairs, dot fill indicates distance below 8Å. Underlying in grey is the contact map of the crystal

structure (PDB entry 1pga, distance < 8Å) and shown on top and to the right are its secondary structure elements (wave – alpha helix, arrow – beta strand).

B. Interaction score map for top 100 pairs with highest merged epistatic enrichments before normalization by uncertainty (light blue, lower triangle) and highest *epistasis scores* (black, upper triangle). Dot size indicates interaction score relative to mean of top 100 pairs, everything else as in panel A.

C. Minimal number of edges (contact with distance < 8Å in crystal structure 1pga) connecting a position pair.

D. Correlation of pair-wise *epistasis score* and minimal heavy atom distance. Fill of contour map indicates density (red – dense, white – sparse) of position pairs. Spearman correlation coefficient is indicated.

E. Precision of interaction scores to predict direct contacts (distance < 8Å in crystal structure 1pga) as a function of the number 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). Horizontal dashed line indicates random expectation. Protein length is 56 amino acids.

Tertiary structure neighborhood leads to correlated epistatic patterns

We reasoned that, if epistasis arises mainly from physical interactions within the protein domain, proximal positions should also exhibit similar epistasis interactions with all other positions in the protein (Figure 4A). To test this, we estimated the correlation matrix of positive (or negative) epistatic fractions across all position pairs in each run of our resampling procedure (Figures 1B (step 5c) and 4A, see Methods).

Indeed, the top 100 position pairs with the highest epistasis pattern correlation are enriched for being direct neighbors (51%) and for sharing a common neighbor (Figure 4B) and, conversely, more proximal position pairs have more highly correlated epistasis patterns (rank correlation coefficient = -0.47, $p < 10^{-6}$, n = 1485, Figure 4D). Nonetheless, correlation patterns do show a strong position-dependency, with positions close in the linear sequence forming clusters of lower or higher correlations (Figures 1B (step 5d lower triangles in the heat maps) and 4C (lower triangle)). Thus, similar to individual epistatic interactions, correlations of epistatic

218 interaction patterns are highly enriched for proximal pairs but also contain information on
219 indirect structural interactions.

220 We tested whether partial correlations of epistasis patterns, that is dependencies of epistasis
221 patterns between position pairs after accounting for the epistasis patterns of all other positions,
222 would minimize indirect interactions and reveal more direct interactions. We derived partial
223 correlations by inversion of the correlation matrix, estimated the uncertainty of individual values
224 from resampling, merged values from positive and negative epistatic patterns, and ranked these
225 by their z-scores, which we refer to as *association scores* (Figure 1B, steps 5c and 6).

226 In contrast to the correlation of epistatic patterns, *association scores* are slightly less correlated
227 with proximity (rank correlation coefficient $R = -0.36$, $p < 10^{-6}$, $n = 1485$), with most position pairs
228 having an *association score* around 0 and only few pairs deviating to higher values (Figure 4D).
229 These pairs with highest *association scores*, however, are strongly enriched for being direct
230 neighbors in the structure (71%, Figure 4B). Furthermore, top pairs are well distributed across
231 the contact map and are enriched for pairs close in the linear chain (Figure 4C). These results
232 are consistent with a more binary classification of being distant (*association score* around zero)
233 or being in direct contact (positive *association score*).

234 We further derived a *combined score* by summing the standardized *epistasis* and *association*
235 *scores*, to explore whether combining information from individual epistatic interactions and
236 epistasis interaction patterns can improve proximity estimates.

237 We evaluated the precision of the three interaction scores in predicting direct contacts in the
238 protein domain. For all pairs at least 5 amino acids apart, the *association score* predicts direct
239 contacts with similar precision to the *epistasis score* (Figure 4F). The *combined score* is, apart
240 from the very top contacts, consistently more precise than the individual scores (L/5 – 72%, L/2
241 – 79%, L – 68%, 2*L – 53% precision), showing that information from epistasis interactions and
242 their interaction patterns across the protein yield complementary results. Finally, we also find
243 that the interaction scores precisely predict direct contacts over long ranges in the linear chain.
244 For pairs that are at least 24 amino acids apart, precision is as high as 91% for the top L/5
245 contacts, 64% for the top L/2 contacts and 48% for the top L contacts (Figure 4G).

246

Figure 4

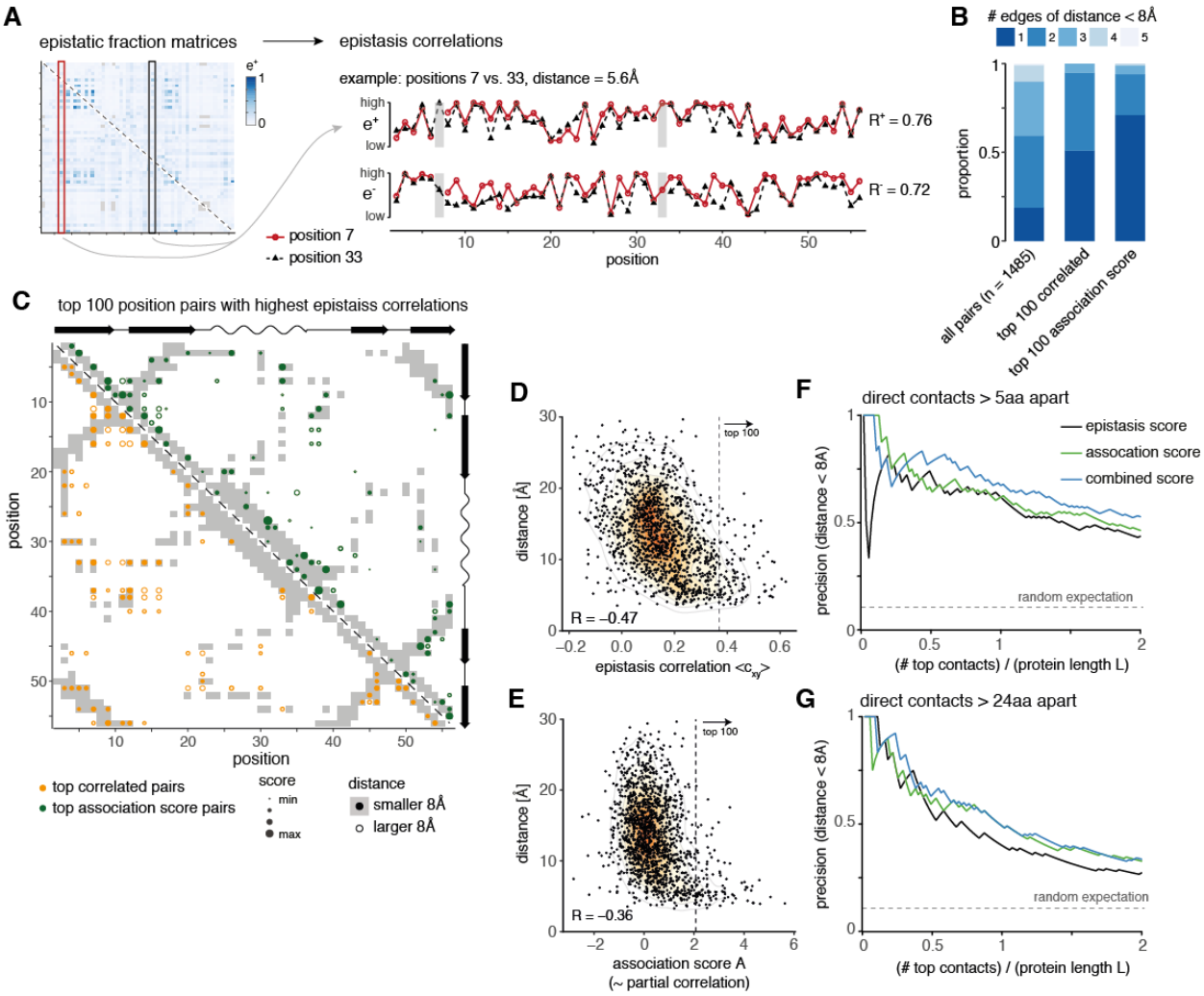


Figure 4: Tertiary structure neighborhood leads to correlated epistatic patterns

A. Example of epistatic interaction patterns of the two proximal positions 7 and 33. Matrix with fraction of positive epistatic variants is shown on the right. Epistatic interaction patterns across the protein correspond to columns of the matrix. Epistatic interaction patterns for positive (upper) and negative (lower) epistatic fractions of positions 7 and 33 are shown on the right. Pearson correlation coefficients for epistatic interaction patterns are given on the right, note that these do not allow a direct comparison to epistasis correlation values given in panel D. Grey vertical bars mark the positions under consideration.

B. Minimal number of edges (contact with distance < 8Å in crystal structure 1pga) connecting a position pair.

C. Interaction score map for top 100 position pairs with highest correlation (lower triangle, orange) and association scores (upper triangle, green). Dot size indicates score across 1000 re-sampling runs relative to average over top 100 pairs, dot fill indicates distance below 8Å. Underlying in grey is the contact map of the crystal structure (PDB entry 1pga, distance < 8Å) and shown on top and to the right are its secondary structure elements (wave – alpha helix, arrow – beta strand).

D. Correlation of pair-wise epistasis pattern correlation to minimal heavy atom distance. Shown are weighted averages of shrunk correlation coefficients for positive and negative epistatic patterns (means across re-sampling runs), with weights as the inverse variances across resampling runs (equivalent to epistatic fraction aggregation). Fill of contour map indicates density (red – dense, white – sparse) of position pairs. Spearman correlation coefficient is indicated.

E. Correlation of pair-wise *association scores* to minimal heavy atom distance. As in panel D.

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). Horizontal dashed line indicates random expectation.

G. Precision of interaction scores to predict direct contacts (distance < 8Å in crystal structure 1pga) greater than 24 amino acids apart (n = 465 pairs, n = 53 direct contacts). 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. This might be expected both locally as well as across the full length of the protein if structural interactions are a primary determinant of epistasis (Andreani and Söding, 2015; Tóth-Petróczy et al., 2016).

289 Within an alpha helix with 3.6 residues per helical turn, a helical position would be predicted to
290 interact epistatically with the third or fourth neighbor along the linear amino acid chain. Moreover
291 these positions should also have similar epistasis patterns towards all other positions in the
292 protein due to similar side chains orientations (Figure 5A). Similarly, within a beta strand,
293 positions should interact epistatically with the next-but-one neighbor with whom they should also
294 share similar epistasis patterns towards all other positions in the protein.

295 We used a two dimensional kernel smoothing approach to estimate the positions of alpha
296 helices and beta strands from the deep mutational scanning data (Figure 5B). Here, the
297 propensity of a position to belong to an alpha helix or a beta strand is calculated by taking into
298 account how the position interacts with neighboring positions as well as how neighboring
299 positions interact with each other. For alpha helices, the kernel has a sinusoidal profile with a
300 period of 3.6 perpendicular to the diagonal, to weight interactions with neighboring positions,
301 and a Gaussian profile along the diagonal to integrate over multiple adjacent positions. For beta
302 strands, the kernel has an alternating profile perpendicular to the diagonal, down-weighting
303 interactions with adjacent positions and up-weighting interactions with the next-but-one position,
304 and a Gaussian profile along the diagonal to integrate over multiple adjacent positions.

305 We calculated a p-value for each position to belong to an alpha helix or a beta strand by
306 comparing the value obtained from kernel smoothing the interaction score data to values
307 obtained when applying the kernels to 10.000 randomized data sets. To determine the range of
308 secondary structure elements we combined the p-values of individual positions by Fisher's
309 method, extending stretches of positions with highly significant scores for alpha helices or beta
310 strands until the combined p-values drop off when adding additional adjacent positions (see
311 Methods).

312 Predicting secondary structures using *combined scores* yielded secondary structure elements
313 that coincide well with secondary structure elements in the crystal structure (Figure 5C). Using
314 *epistasis* and *association scores* to predict secondary structures elements gave similar results,
315 although the epistasis score analysis failed to detect both beta strands at the end of the domain
316 due to low statistical power (Figure S4A).

317 We further used two-dimensional kernel smoothing to detect parallel and anti-parallel beta sheet
318 interactions, by applying beta strand kernels to off-diagonal entries on the interaction score
319 matrices (Figure 5D). Using *combined scores* we detected several putative stretches of parallel
320 and anti-parallel beta sheet interactions throughout the protein domain (Figure 5E). We defined
321 a set of rules that iteratively fix the most significant stretches of beta sheet interaction when they

322 are consistent with previously determined secondary structure elements and already fixed beta
323 sheet interactions (see Methods). This yielded an accurate prediction of beta sheet interactions,
324 in terms of the number and orientation of sheet interactions as well as the exact pairwise
325 interactions of all but two position pairs (Figure 5E). Furthermore, updating beta strands
326 according to inferred beta sheet pairings improved beta strand prediction itself, notably
327 introducing a correct split of beta strand 1 and 2 and adjusting the length of beta strands 3 and
328 4. Together, predictions of secondary structures and beta sheet interactions from *combined*
329 *scores* have a precision greater than 95% and a recall of greater than 85% (Figure 5F). Similar,
330 but slightly worse, results are obtained when applying this workflow to *epistasis* and *association*
331 *score* data instead (Figures S4B,C).

332

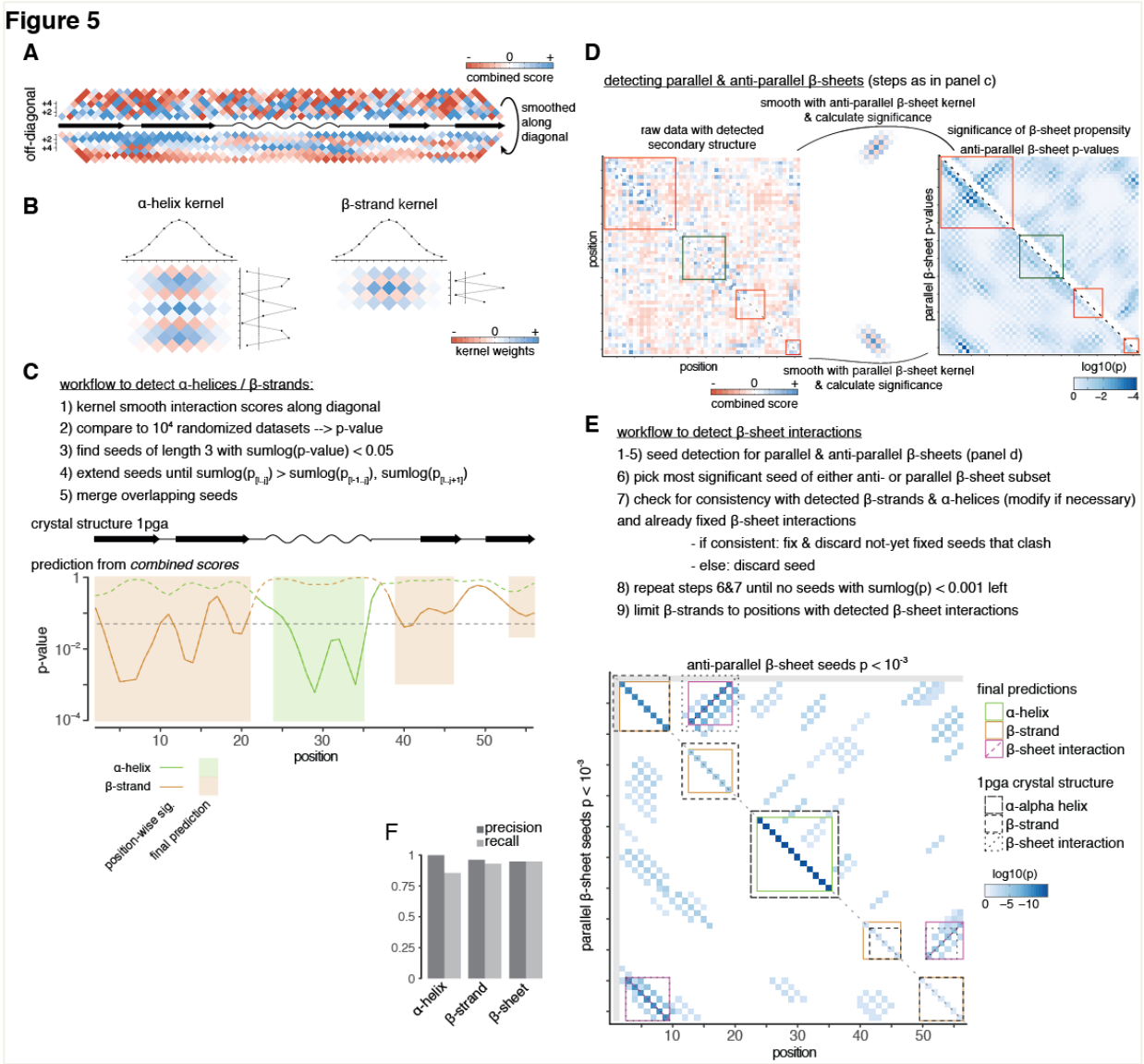


Figure 5: Periodic epistatic patterns reveal secondary structure arrangements

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 crystal 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 smoothed with a Gaussian kernel along the direction of the diagonal (i.e. horizontally, length of Gaussian kernel as for kernels in panel b) is shown.

342 b) Two-dimensional kernels for alpha helix and beta strand detection. Kernel has a
343 Gaussian profile in the direction of the diagonal, and a sinusoidal or alternating profile in
344 the off-diagonal direction for alpha helices and beta strands, respectively.
345 c) Workflow for secondary structure element prediction and results from *combined score*
346 data compared to known secondary structure from crystal structure 1pga (see Figure S4
347 for other interaction score data). In workflow, "sumlog(p)" referes to p-value combination
348 by Fisher's method (see Methods for more detail on workflow). In plot, lines give
349 position-wise p-value of secondary structure propensity; dashed lines indicate data was
350 not used (p-value for other secondary structure type lower); boxes indicate final
351 predictions, lower limit of box indicates p-value of final predictions (limited to minimum of
352 10^{-4}). Note that sumlog(p) of seeds length 3 is not shown.
353 d) Left: *Combined score* matrix. Right: Position-wise p-values of beta sheet propensity after
354 kernel smoothing off-diagonal elements with beta sheet kernels (shown in the middle)
355 and comparing to 10^4 random datasets (lower triangle - parallel beta sheet propensity,
356 upper triangle – anti-parallel beta sheet propensity). Secondary structure element
357 predictions from panel c) indicated on both matrices (green – alpha helix, red – beta
358 sheet).
359 e) Workflow to detect beta sheet interactions. Plot shows matrix of combined p-value for
360 extended (beta sheet) seeds, the final predictions and the reference elements in the
361 crystal structure (lower triangle – parallel beta sheets, upper triangle – anti-parallel beta
362 sheet, diagonal –secondary structure elements).
363 f) Precision and recall for position-wise secondary structure element and beta sheet
364 interaction predictions as shown in panel e).

365

366

367 Folding a protein domain from epistatic mutational patterns

368 Together, these findings show that genetic interaction data contain information on a protein's
369 secondary and tertiary structure. We therefore tested whether the information extracted from
370 the genetic data would suffice to predict *ab initio* the structure of protein G domain B1. We
371 performed simulated annealing with structural restraints derived from the deep mutational
372 scanning data. Structural restraints were defined for tertiary contacts, secondary structure
373 elements, and beta sheet interactions. In particular, 56 distance restraints (distance < 8Å

374 between Cβ atoms) for tertiary contacts were defined from the top scoring position pairs with a
375 linear chain distance greater than five amino acids. We found that using fewer or more contacts
376 generally decreases the accuracy of the predicted models (Figures S5A,B). Furthermore, we
377 defined dihedral angle restraints for predicted secondary structure elements. Finally, we defined
378 restrictive distance restraints (distance smaller than 2.1Å for N-H : C=O atom pairs) for beta
379 sheet positions that form hydrogen bonds with each other.

380 We performed structural simulations by simulated annealing using the Xplor-NIH modeling suite
381 (Schwieters et al., 2003). We set up a three-stage approach, with 500 structural models
382 generated per stage. Stages 1 and 2 served to identify restraints inconsistent with the most
383 plausible structural models. These restraints were subsequently removed, thus improving the
384 precision of the utilized restraints, especially those for tertiary contacts, for which precision
385 improved from ~65% up to 85% for the top 56 contacts (Figures 6A,B). Stage 3 then served to
386 refine an average structure for the top 10% of models from stage 2 to obtain a final set of
387 structural models.

388 We evaluated the top 5% structural models generated in the third simulation stage against the
389 known crystal structure of protein G domain B1 (PDB entry 1pga) (Figures 6C and S5A,B).
390 Models predicted from *combined score* data performed best, with an average Cα-root mean
391 squared deviation of the top models ($\langle C\alpha - RMSD \rangle$) of 1.9Å and an average template modeling
392 score of 0.75 (Figure 6D). This accuracy is similar to the resolution of the crystal structure
393 (2.07Å) and very close to the optimum achievable with our simulation protocol (using true
394 contacts, secondary structure elements and beta sheet interactions, $\langle C\alpha - RMSD \rangle = 1.4Å$ and
395 TM score = 0.79). Consistent with lower precision of contact and secondary structure
396 predictions, models generated with restrains from *epistasis* or *association scores* have on
397 average a somewhat lower accuracy ($\langle C\alpha - RMSD \rangle = 2.9Å$ and $\langle C\alpha - RMSD \rangle = 2.6Å$,
398 respectively), with association score models performing slightly but consistently better (Figures
399 6C and S5A,B). Defining restraints for beta sheet interactions results in an improvement of
400 about 50% average accuracy for the top structural models (Figure 6C, $\langle C\alpha - RMSD \rangle = 5.7Å$ for
401 *epistasis score*, $\langle C\alpha - RMSD \rangle = 5.4Å$ for *association score*, and $\langle C\alpha - RMSD \rangle = 3.8Å$ for
402 *combined score* without including restraints for beta sheet interactions).

403 Together, this shows that genetic interaction data alone is sufficient to accurately predict the
404 structure of a protein domain.

405

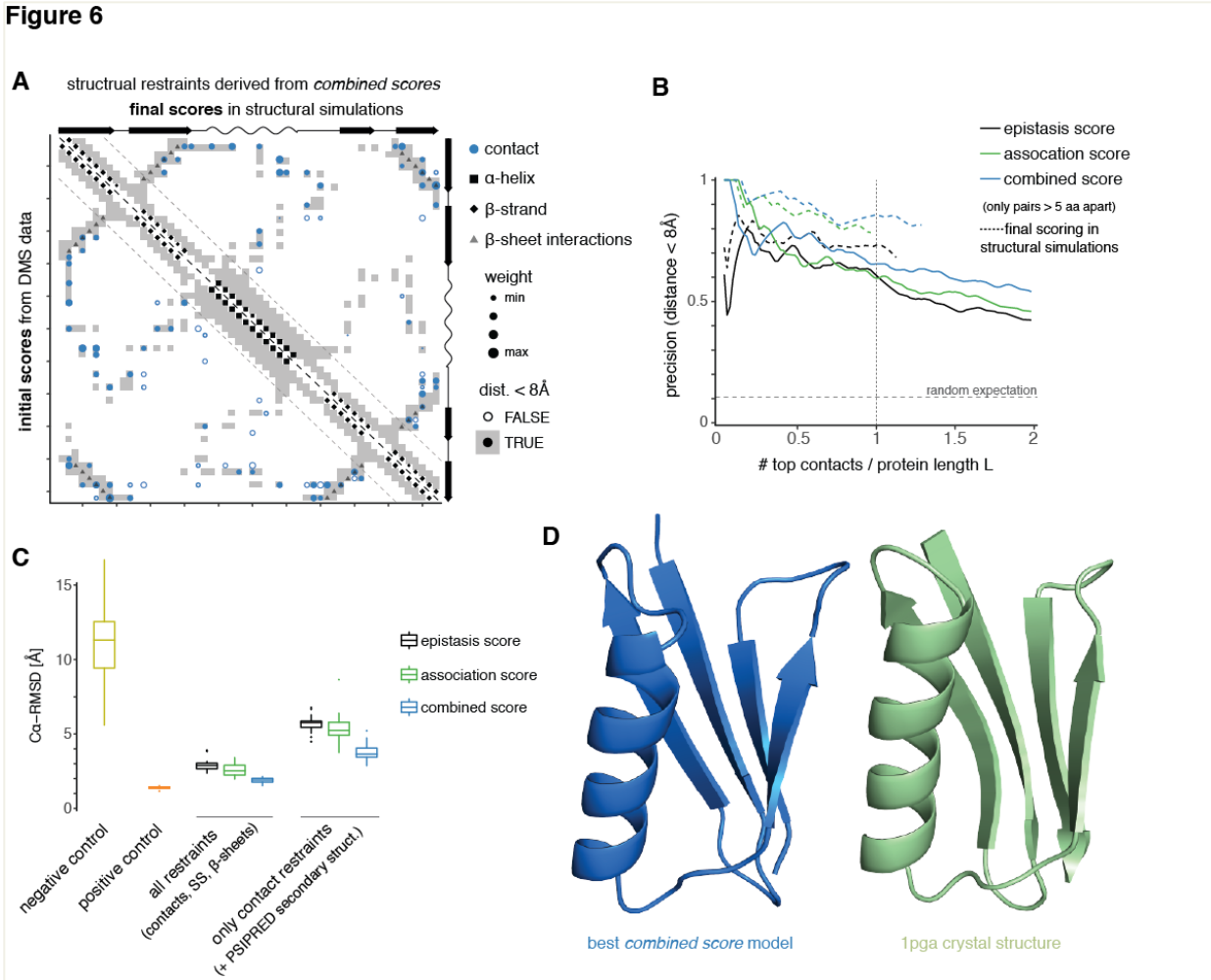


Figure 6: *Ab initio* folding of protein G domain B1

A. Map of structural restraints derived from *combined scores* (lower triangle - initial scores, upper triangle – scores after reweighting in structural simulations, scores below 0.1*mean(scores) were discarded). Top 56 tertiary contacts (greater than 5 amino acid linear chain distance) are shown as blue circles, alpha helices as squares (score is average weight of dihedral angle restraints), beta strands as diamonds (score is average weight of dihedral angle restraints) and beta sheet interactions as triangles. Underlying in grey is the contact map of the crystal structure (PDB entry 1pga, distance < 8Å) and shown on top and to the right are its secondary structure elements (wave – alpha helix, arrow – beta strand)

B. Improvements in precision of contact predictions due to consistency checking in structural simulations. Solid lines give precision as initially determined from interaction

419 scores (same as shown in Figure 4F). Dashed lines give precision after simulation stage
420 2 rescoring of contacts based on consistency with generated structural models for
421 simulations that were started with top 2*L contacts.

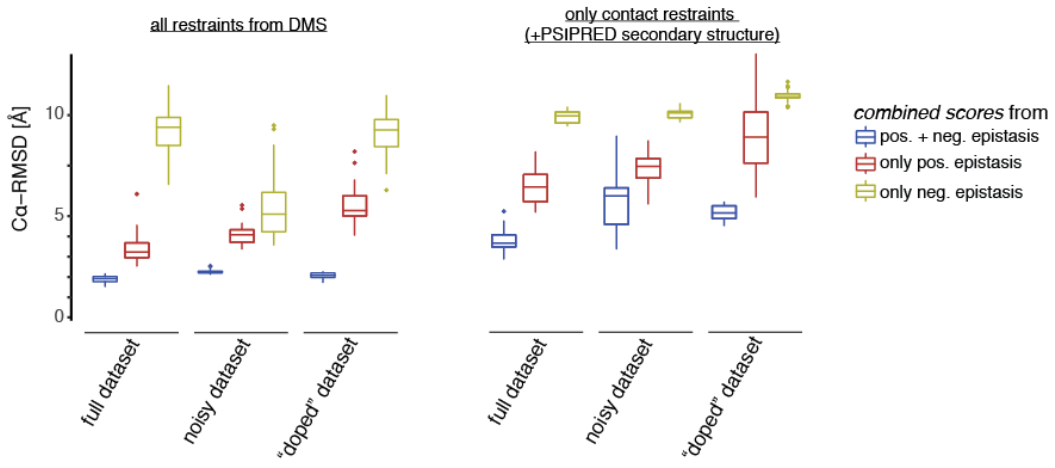
422 C. Accuracy ($C\alpha$ root mean squared deviation) of top 5% structural models generated from
423 different sets of restraints compared to crystal structure 1pga. Negative control – only
424 restraints for secondary structure predicted by PSIPRED (Jones, 1999). Positive control
425 – 56 true contacts (random subset), secondary structure elements and beta sheet
426 interactions from crystal structure. All restraints – top 56 contacts, secondary structure
427 element and beta sheet interaction predictions from the three interaction scores, as
428 indicated by color. Only contact restraints – top 56 contacts from interaction scores and
429 secondary structure elements as predicted by PSIPRED.

430 D. Best structural model generated with all restraints derived from *combined score* (left)
431 and crystal structure (PDB entry 1pga) (right). Visualization generated in PyMOL (The
432 PyMOL Molecular Graphics System, 2015).

433

434

Figure 7



435

436 Figure 7: Requirements for structure determination from deep mutational
437 scanning data

438 Down-sampling analysis to test data quality requirements for accurate prediction of
439 protein structures from deep mutational scanning datasets (see also Supplementary
440 Text). Results from structural simulations when using restraints for tertiary contacts,
441 secondary structure elements and interactions (left panel) or only tertiary contacts

442 supplemented with secondary structure elements predicted by PSIPRED (right panel)
443 which were derived from *combined scores*. Additionally we also varied the calculation of
444 combined scores, either calculated from full epistasis information (blue) or only positive
445 (red) or only negative (yellow) epistatic information. Full dataset refers to dataset used
446 throughout the study. Noisy dataset is down-sampled to 25% of sequencing read
447 coverage. “Doped” datasets contains only amino acid changes achieved by single
448 nucleotide mutations (31% of possible single mutants and 10% of possible double
449 mutants).
450 Boxplots show: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x
451 interquartile range; points, outliers.
452
453

454

Discussion

We have shown here that the structure of a protein domain can be accurately predicted simply by quantifying the activity of a large number of single and double mutant variants.

For macromolecules with very large numbers of homologs available in sequence databases, correlated changes in sequence can provide sufficient information for structure determination (Hopf et al., 2012; Kosciolek and Jones, 2014; Marks et al., 2011; Michel et al., 2014; Ovchinnikov et al., 2015; Ovchinnikov et al., 2017; Tóth-Petróczy et al., 2016; Weinreb et al., 2016). However, for many proteins and RNAs this is not the case, and for evolutionary novel folds and families this may never be so (Marks et al., 2012; Ovchinnikov et al., 2017; Tóth-Petróczy et al., 2016). In contrast, our approach should work for any molecule that can be subjected to deep mutagenesis and selection *in vitro* or *in vivo* (Araya et al., 2012; Diss and Lehner, 2018; Fowler et al., 2010; Li et al., 2016; Matreyek et al., 2017; Poelwijk et al., 2017; Rocklin et al., 2017; Starita et al., 2013; Starita et al., 2015; Starr et al., 2017; Weile et al., 2017). However, the same fundamental principle underlies both approaches – that mutations affect each other because of specific structural interactions (Poelwijk and Ranganathan, 2017). The successful application of evolutionary coupling analysis to many different molecules and problems (Hopf et al., 2012; Kosciolek and Jones, 2014; Marks et al., 2011; Michel et al., 2014; Ovchinnikov et al., 2015; Ovchinnikov et al., 2017; Tóth-Petróczy et al., 2016; Weinreb et al., 2016) therefore provides very strong evidence that the approach that we have outlined here will work for many different macromolecules. We thus envisage broad applicability of our framework, including to macromolecules whose structures are difficult to determine by physical methods, such as membrane proteins and intrinsically disordered proteins that only fold upon binding or in particular conditions.

As a proof-of-principle we used only information extracted from a deep mutational scanning experiment to predict the structure of protein G domain B1. However, taking advantage of recent advances in contact predictions, such as machine learning of sequence-derived protein features, (Schaarschmidt et al., 2018) and homology-driven structure modeling (Rohl et al., 2004) is likely to further improve structure prediction from genetic data, especially for more complex folds than the one tested here.

An analysis of incomplete and down-sampled variants of the deep mutational scanning dataset used here suggests that a high signal-to-noise ratio (measurement range relative to

experimental error of fitness estimates), which allows both positive and negative epistasis to be quantified, is an important factor for generating datasets with a quality sufficient for protein structure prediction (Figure 7 and Supplementary Text). At sufficient signal-to-noise ratios, however, even deep mutational scanning datasets containing an incomplete (random) subset of double mutants or sequenced at lower coverage might be amenable to protein structure prediction (Figure 7), thus suggesting that our framework should be easily applicable to longer molecules.

Finally, we note that our results strongly support the hypothesis (Diss and Lehner, 2018; Fowler and Fields, 2014; Melamed et al., 2013; Olson et al., 2014; Sahoo et al., 2015; Starr and Thornton, 2016) that structural interactions are the main cause of specific genetic interactions (epistasis) within macromolecules.

Acknowledgements

We thank members of the Lehner lab, Torsten Gross and Gregor Mönke 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).

Author Contributions

Conceptualization, B.L. and J.M.S.; Methodology, J.M.S.; Investigation, J.M.S.; Writing - Original Draft, J.M.S.; Writing - Review & Editing, J.M.S. and B.L.; Supervision, B.L.

Declaration of Interests

The authors declare no competing interests.

References

Adhikari, B., Bhattacharya, D., Cao, R., and Cheng, J. (2015). CONFOLD: Residue-residue contact-guided ab initio protein folding. Proteins *83*, 1436-1449.

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

Andreani, J., and Söding, J. (2015). bbcontacts: prediction of β-strand pairing from direct coupling patterns. Bioinformatics *31*, 1729-1737.

Araya, C.L., Fowler, D.M., Chen, W., Muniez, I., Kelly, J.W., and Fields, S. (2012). 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.

526 Burger, L., and van Nimwegen, E. (2010). Disentangling Direct from Indirect Co-Evolution of
527 Residues in Protein Alignments. PLoS Computational Biology *6*, e1000633.

528 Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. Elife *7*,
529 e32472.

530 Dunn, S.D., Wahl, L.M., and Gloor, G.B. (2008). Mutual information without the influence of
531 phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics *24*, 333-
532 340.

533 Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction
534 in proteins: Using pseudolikelihoods to infer Potts models. Physical Review E *87*, 012707.

535 Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields,
536 S. (2010). High-resolution mapping of protein sequence-function relationships. Nature Methods
537 *7*, 741-746.

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

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

543 Gloor, G.B., Martin, L.C., Wahl, L.M., and Dunn, S.D. (2005). Mutual information in protein
544 multiple sequence alignments reveals two classes of coevolving positions. Biochemistry *44*,
545 7156-7165.

546 Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue
547 contacts in proteins. Proteins *18*, 309-317.

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

550 Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Three-
551 Dimensional Structures of Membrane Proteins from Genomic Sequencing. Cell *149*, 1607-1621.

552 Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., and Marks,
553 D.S. (2017). Mutation effects predicted from sequence co-variation. Nature Biotechnology.

554 Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring
555 matrices. Journal of molecular biology *292*, 195-202.

556 Jones, D.T., Buchan, D.W.A., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural
557 contact prediction using sparse inverse covariance estimation on large multiple sequence
558 alignments. Bioinformatics *28*, 184-190.

559 Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based
560 residue-residue contact predictions in a sequence- and structure-rich era. Proceedings of the
561 National Academy of Sciences *110*, 15674-15679.

562 Kosciolek, T., and Jones, D.T. (2014). De novo structure prediction of globular proteins aided by
563 sequence variation-derived contacts. PLoS ONE *9*, e92197.

564 Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. Trends Genet
565 *27*, 323-331.

566 Li, C., Qian, W., Maclean, C.J., and Zhang, J. (2016). The fitness landscape of a tRNA gene.
567 Science *352*, 837-840.

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

570 Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C.
571 (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS ONE *6*,
572 e28766.

573 Marks, D.S., Hopf, T.A., and Sander, C. (2012). Protein structure prediction from sequence
574 variation. Nature Biotechnology *30*, 1072-1080.

575 Matreyek, K.A., Stephany, J.J., and Fowler, D.M. (2017). A platform for functional assessment
576 of large variant libraries in mammalian cells. Nucleic Acids Research *45*, e102.

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

579 Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R., and Fields, S. (2013). Deep mutational
580 scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA *19*,
581 1537-1551.

582 Michel, M., Hayat, S., Skwark, M.J., Sander, C., Marks, D.S., and Elofsson, A. (2014).
583 PconsFold: improved contact predictions improve protein models. Bioinformatics *30*, i482-488.

584 Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic,
585 J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures
586 native contacts across many protein families. Proceedings of the National Academy of Sciences
587 *108*, E1293-1301.

588 Olson, C.A., Wu, N.C., and Sun, R. (2014). A comprehensive biophysical description of pairwise
589 epistasis throughout an entire protein domain. Current biology : CB *24*, 2643-2651.

590 Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D.E., Kamisetty, H., Grishin, N.V.,
591 and Baker, D. (2015). Large-scale determination of previously unsolved protein structures using
592 evolutionary information. eLife *4*, e09248.

593 Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G.A., Kim, D.E., Kamisetty,
594 H., Kyrpides, N.C., and Baker, D. (2017). Protein structure determination using metagenome
595 sequence data. Science *355*, 294-298.

596 Poelwijk, F.J., and Ranganathan, R. (2017). The relation between alignment covariance and
597 background-averaged epistasis arXiv, 1-7.

598 Poelwijk, F.J., Socolich, M., and Ranganathan, R. (2017). Learning the pattern of epistasis
599 linking genotype and phenotype in a protein. bioRxiv, 213835.

600 Rocklin, G.J., Chidyausiku, T.M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L.,
601 Ravichandran, R., Mulligan, V.K., Chevalier, A.*, et al.* (2017). Global analysis of protein folding
602 using massively parallel design, synthesis, and testing. Science *357*, 168-175.

603 Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. (2004). Protein structure prediction using
604 Rosetta. Methods Enzymol *383*, 66-93.

605 Sahoo, A., Khare, S., Devanarayanan, S., Jain, P.C., and Varadarajan, R. (2015). Residue
606 proximity information and protein model discrimination using saturation-suppressor
607 mutagenesis. eLife *4*, 371.

608 Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., and Bonvin, A. (2018). Assessment of
609 contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins *86*
610 *Suppl 1*, 51-66.

611 Schäfer, J., and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix
612 estimation and implications for functional genomics. Statistical applications in genetics and
613 molecular biology *4*, Article32.

614 Schwieters, C.D., Kuszewski, J.J., Tjandra, N., and Clore, G.M. (2003). The Xplor-NIH NMR
615 molecular structure determination package. Journal of Magnetic Resonance *160*, 65-73.

616 Seemayer, S., Gruber, M., and Söding, J. (2014). CCMpred--fast and precise prediction of
617 protein residue-residue contacts from correlated mutations. Bioinformatics *30*, 3128-3130.

618 Starita, L.M., Pruneda, J.N., Lo, R.S., Fowler, D.M., Kim, H.J., Hiatt, J.B., Shendure, J., Brzovic,
619 P.S., Fields, S., and Klevit, R.E. (2013). Activity-enhancing mutations in an E3 ubiquitin ligase
620 identified by high-throughput mutagenesis. Proceedings of the National Academy of Sciences
621 *110*, E1263-1272.

622 Starita, L.M., Young, D.L., Islam, M., Kitzman, J.O., Gullingsrud, J., Hause, R.J., Fowler, D.M.,
623 Parvin, J.D., Shendure, J., and Fields, S. (2015). Massively Parallel Functional Analysis of
624 BRCA1 RING Domain Variants. Genetics *200*, 413-422.

625 Starr, T.N., Picton, L.K., and Thornton, J.W. (2017). Alternative evolutionary histories in the
626 sequence space of an ancient protein. Nature *5*, e16965.

627 Starr, T.N., and Thornton, J.W. (2016). Epistasis in protein evolution. Protein Science *25*, 1204-
628 1218.

629 The PyMOL Molecular Graphics System, V.S., LLC. (2015). The PyMOL Molecular Graphics
630 System, Version 1.8.

631 Tokuriki, N., and Tawfik, D.S. (2009). Stability effects of mutations and protein evolvability.
632 Current Opinion in Structural Biology *19*, 596-604.

633 Tóth-Petróczy, A., Palmedo, P., Ingraham, J., Hopf, T.A., Berger, B., Sander, C., and Marks,
634 D.S. (2016). Structured States of Disordered Proteins from Genomic Sequences. Cell *167*, 158-
635 170.e112.

636 Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct
637 residue contacts in protein-protein interaction by message passing. Proceedings of the National
638 Academy of Sciences *106*, 67-72.

639 Weile, J., Sun, S., Cote, A.G., Knapp, J., Verby, M., Mellor, J.C., Wu, Y., Pons, C., Wong, C.,
640 van Lieshout, N.*, et al.* (2017). A framework for exhaustively mapping functional missense
641 variants. Molecular Systems Biology *13*, 957.

642 Weinreb, C., Riesselman, A.J., Ingraham, J.B., Gross, T., Sander, C., and Marks, D.S. (2016).
643 3D RNA and Functional Interactions from Evolutionary Couplings. Cell *165*, 963-975.

644 Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein
645 structure template quality. Proteins *57*, 702-710.

646

Methods:

Datasets and preprocessing

Protein G domain B1 deep mutational scanning data was obtained from the supplementary information of Olson et al. (2014). The data consists of summed read counts of three replicate experiments assaying the binding affinity of protein G domain B1 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.

Additionally, a correction factor for wild-type fitness was derived, based on the observation that the upper modes of single and double mutant variant fitness distributions were not centered on the wild-type fitness value and that the epistasis distribution was highly non-symmetric. The kernel density estimate of the single mutant fitness distribution was calculated (using the R function *density* with parameter *bw* set to 0.15) and the correction factor was derived as the position of the upper fitness mode as $cf = 0.198$. Each fitness value was corrected by this factor, which centered the upper modes of single and double mutant fitness distributions on the wild-type fitness value and resulted in a symmetric epistasis distribution. We note, however, that this correction has little effect on the downstream analysis, because of the use of quantile-based estimates to call positive or negative epistasis. This would be different if the fitness deviation from expectation would have been used to calculate epistasis.

The standard error of fitness estimates was calculated from read counts under Poissonian assumptions, i.e. $\sigma_i = \sqrt{\frac{1}{n_i^{in}} + \frac{1}{n_i^{out}} + \frac{1}{n_{wt}^{in}} + \frac{1}{n_{wt}^{out}}}$. 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 (Figure S1A). In the case of the IgG-binding assay for protein G domain B1, this is presumably mainly due to unspecific carry-over on beads (Olson et al., 2014). The fitness values derived from the measurement are therefore a convolution of the actual binding affinities to IgG and

676 nonspecific carryover, i.e. $exp(f_i^{measured}) = exp(f_i^{binding}) + exp(f^{carryover})$. Fitness values of
677 variants close to the lower measurement limit of the assay are therefore dominated by
678 unspecific carryover effects. The lower measurement limit of the assay was estimated by two
679 approaches that yielded very similar estimates. One, from a kernel density estimate of the single
680 mutant fitness distribution (R function *density* with parameter *bw* set to 0.15), where the position
681 of the lower mode of the data corresponded to $f^{carryover} = -5.85$ (~0.3% on linear scale). Two,
682 from examining the fitness distribution of double mutants with expected fitness lower than -8, i.e.
683 double mutants resulting from two lethal or nearly lethal single mutant variants, whose fitness
684 values are thus expected to be dominated by background effects. The median of this
685 background fitness distribution yielded an estimate of $f^{carryover} = -5.89$. The mean of the two
686 estimates was used in downstream analyses.

687 7% of double mutant variants were discarded due to too low sequencing coverage in input or
688 output libraries (Figure S1C). That is, variants that had less than 200 input reads and no output
689 reads were discarded, because it is not possible to determine their fitness. Above 200 input
690 reads, variants without output reads are certain to be dominated by nonspecific carryover
691 effects. These variants were retained and their fitness was calculated by setting their output
692 read count to 0.5.

693

694 Epistasis classification

695 Epistasis was calculated from a non-parametric null model (Figure 2A) in order to account for
696 nonlinearities close to the lower limit of the fitness assay measurement range, non-specific
697 epistatic behavior resulting from e.g. thermodynamic stability thresholds as well as differential
698 uncertainty of fitness measurements across the fitness landscape (Figure S1B).

699 First, double mutant fitness values were corrected by subtracting the average local fitness
700 computed using a two-dimensional generalized additive model (using R function *gam* with
701 tensor product smooths, with the two constituting single mutant fitness values as independent
702 variables and the resulting double mutant fitness value as dependent variable). This was
703 necessary to avoid boundary effects of quantile-based fits in boundary regions with non-zero
704 slopes. 5th and 95th percentile surfaces were then fit to these corrected double mutant fitness
705 values, by computing for each double mutant variant the 5th and 95th percentile of the fitness
706 distribution made up of the 1% closest neighbors in single mutant fitness space. Double mutant

707 variants with fitness values below the 5th or above the 95th percentile were categorized as
708 negative or positive epistatic, respectively (Figure 2A).

709 The evaluation of positive or negative epistasis was, however, restricted to specific subsets of
710 the data where measurement errors do not impede epistasis classification (Figure S1D). The
711 data subset in which variants were potentially classified as positive epistatic is limited to data
712 regions where

713 - the 95th percentile fitness surface is below wild-type fitness
714 - at least one single mutant fitness value is significantly smaller than wild-type fitness at
715 five standard errors
716 - the expected fitness (sum of both single mutant fitness values) is not significantly higher
717 than wild-type at five standard errors
718 - the expected fitness is significantly higher than twice the lower fitness limit of the assay

719 The rationale for criteria 1-3 is to avoid double mutants from two neutral single mutants,
720 because these are dominated by measurement noise of wild-type like variants. Minimal
721 restrictions were instead applied to the lower limits of the measurement range (criterion 4),
722 because otherwise no/little epistasis quantification would have been available for several
723 positions with very strong detrimental effects as well as because strong positive epistatic effects
724 are observed in these regions, despite the dominance of background measurement effects.

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

727 - the 5th percentile fitness surface is above the 95th percentile of the background effect
728 distribution; this value is derived from the 95th percentile of double mutant fitness values
729 with expected fitness below -8 (analogous to lower fitness limit estimation, see above).
730 - both single mutant fitness values are significantly higher than the lower limit of the
731 fitness assay measurement range at five standard errors
732 - the expected fitness (sum of both single mutant fitness values) is not significantly higher
733 than wild-type at five standard errors

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

736 As a result of these restrictions as well as differences in initial coverage, the number of double
737 mutant variants that can be used to assess positive and negative epistasis varies substantially
738 across position pairs (with an average of 252 and 187 variants per position pair in the positive
739 and negative epistasis subsets, respectively; Figure S2).

740

741 # Epistatic interactions

742 We derived several interaction scores to estimate which position pairs are in close contact in the
743 tertiary structure (see Figure 1B). These scores are based on summarizing epistasis information
744 on the position pair-level and accounting for the uncertainty inherent in the summarized
745 estimates due to differential error of fitness estimates across the measurement range (Figure
746 S1B) as well as varying numbers of double mutants amenable to epistasis classification
747 (Figures S1D and S2).

748 To summarize epistasis information on the position pair-level, we calculated the fraction of
749 positive or negative epistatic variants per position pair. The fraction of positive epistatic variants
750 per position pair is the number of positive epistatic variants divided by the number of all variants
751 that lie in the double mutant space amenable to positive epistasis classification (Figure 1B, step
752 5b, equivalent calculation for negative epistasis fraction). Because enrichments with positive
753 and negative epistatic variants per position appear mutually exclusive (Figure S3B), we treated
754 both separately and only aggregated them to derive the final interaction scores.

755 To estimate the uncertainty in epistatic fractions per position pair for downstream analyses we
756 implemented a re-sampling approach (Figure 1B, step 5, described here for positive epistatic
757 variants, but equivalent for negative epistatic variants). In each of 1000 re-sampling runs:

758 - each variant's fitness was drawn from a normal distribution with the measured fitness as
759 mean and the uncertainty due to sequencing coverage as standard deviation $f_{ij}^{sampled} =$
760 $\mathcal{N}(f_{ij}, \sigma_{ij} + \sigma_i + \sigma_j)$ (step 5a)
761 - positive epistasis of variants was re-classified given the drawn fitness values (also step
762 5a, see also Figure 2A)
763 - each position pair's fraction of positive epistatic variants was drawn from the posterior
764 probability distribution of how likely an underlying true fraction of epistatic variants E_{xy}^{+} is
765 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 patterns

In addition to the *epistasis score* we derived an interaction score, termed *association score,* from the partial correlation of epistasis patterns between position pairs. 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 (Figure 1B, step 5c, equivalent for negative epistatic fractions). Missing values (positions pairs without observed variants) were supplemented by drawing a random value from the overall distribution of epistatic fractions. A pseudo count equal to the first percentile 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).

795 A shrinkage approach was used to improve the estimate of the correlation matrix (Schäfer and
796 Strimmer, 2005). In brief, the empirical correlation matrix is shrunk towards the identity matrix in
797 order to minimize the mean-squared error between estimated and true correlation matrix.
798 Additionally, this yields a positive definite and well-conditioned correlation matrix, suitable for
799 inversion. All computations on correlation matrices, shrinkage and matrix inversion were
800 performed with the R package *corpcor*, functions *cor.shrink* and *pcor.shrink* (Schäfer and
801 Strimmer, 2005).

802 Partial correlations of epistatic patterns between each position pair were calculated by inverting
803 the shrunk correlation matrix and normalizing each off-diagonal entry of the inverted matrix by
804 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
805 (x,y)-entry of the inverted correlation matrix (Figure 1B, step 5d). Equivalent to the *epistasis*
806 *score,* positive and negative partial correlation estimates were merged by calculating weighted
807 averages of their mean estimates across re-sampling runs, with weights as the inverse
808 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*
809 *score* normalized by the combined uncertainty, $A_{xy} = {\langle a_{xy} \rangle}/{\sigma_{xy}}$, with
810 $\sigma_{xy} = \left(\sigma_{a_{xy}^{+}}{}^{-2} + \sigma_{a_{xy}^{-}}{}^{-2}\right)^{-1/2}$ (step 6).

811 We further derived a *combined score* by summing the standardized *epistasis* and *association*
812 *scores*, i.e. $C_{xy} = \frac{E_{xy} - \langle E \rangle}{\sigma_E} + \frac{A_{xy} - \langle A \rangle}{\sigma_A}$, to explore whether combining information from individual
813 epistatic interactions and epistasis interaction patterns can further improve proximity estimates.
814 We note that this is a naïve approach to combining the information from these two sources, and
815 surely more sophisticated approaches that further improve proximity estimates can be
816 developed. Additionally, we note that it should be possible to also solve the inverse problem for
817 epistatic fraction matrix, thereby potentially improving proximity estimates from epistatic
818 interactions.

819

820 Secondary structure prediction

821 We used a two dimensional kernel smoothing approach to predict secondary structure elements
822 from deep mutational scanning data (Figure 5). For a given position along the linear chain (on

823 the diagonal of the interaction score matrix), interaction scores (off the diagonal) are integrated
824 with distance-specific weighting according to the kernel, which reflects the known geometry of
825 secondary structures.

826 The alpha kernel takes on a sinusoidal profile perpendicular to the diagonal to weight
827 interactions according to whether the position pair considered should have congruent side-chain
828 orientations (see diagonal and perpendicular profiles in Figure 5B). The kernel was defined as
829 $k_\alpha(d,p) = \left(\cos\left(p * \frac{2\pi}{3.6}\right) + {}^{1}/_{3}\right) * e^{-\frac{d^2}{c^2}}$, with $d = |2x - i - j|$ as the diagonal distance of the
830 interaction *ij* (off the diagonal) to the reference position *x* (on the diagonal) and $p = |i - j|$ as the
831 perpendicular distance of the interaction off the diagonal. The kernel weight for positions with p
832 > 5 was set to 0, thus only including interactions across little more than the first helical turn.
833 Finally, $c = 4$ is the integration scale for the Gaussian kernel along the diagonal. While in
834 principle smaller integration scales do yield nosier results and longer integration scales can lead
835 to non-detection of shorter secondary structure stretches, we found that in practice our whole
836 approach (including the actual detection algorithm described below) is robust to alterations of
837 the integration length.

838 The kernel smoothed alpha value for a given position x on the diagonal is then calculated as the
839 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
840 one of the interaction scores (*epistasis, association* or *combined score*) at position pair *ij*.

841 The beta kernel takes an alternating profile perpendicular to the diagonal to weight interactions
842 according to alternating side-chain orientations in a beta strand and was defined as $k_\beta(d,p) =$
843 $\left((p+1)\ mod\ 2 - {}^{1}/_{3}\right) * e^{-\frac{d^2}{c^2}}$, with $c = 4$. Only interactions with perpendicular distances equal
844 or smaller than two (i.e. $k_\beta(d, p > 3) = 0$) were included.

845 To calculate whether kernel-weighted interaction scores of a specific position are larger than
846 expected, they were compared to those obtained from 10^4 randomized control datasets.
847 Randomization was performed by shuffling all interaction scores, while preserving matrix
848 symmetry, and kernel-weighted interaction scores from randomized control datasets were
849 calculated for each position independently to control for possible boundary effects in positions
850 close to the borders of the protein chain. A p-value for each position was calculated as the
851 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 positions where 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 positions where 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 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 (Figure 5D). 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:

- the minimal combined p-value to consider a seed was lowered to 0.001
- for anti-parallel beta sheet interactions, only positions with a distance greater than two 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 six 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 < 0.001: 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 sides 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

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, Cα), was used as the general distance measure. The reference structure used was PDB entry 1pga, which was determined by X-ray diffraction (Gallagher et al., 1994). A direct contact or edge was defined as minimal side chain heavy atom distance < 8Å. The Floyd-Warshall algorithm (implemented as custom script in R) was used to calculate the minimal number of edges that connect any two positions in the protein.

Protein folding

To *ab initio* predict protein structures, we performed simulated annealing using the Xplor-NIH modeling suite (Schwieters et al., 2003) 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

946 models was calculated. Stage 3 served to refine this average structure to obtain a final set of
947 best models.

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

952 Restrains from secondary structures elements were implemented as dihedral angle restraints.
953 Dihedral angles of both beta strands and alpha helices were set to range between values
954 commonly observed in crystal structures (Adhikari et al., 2015), for alpha helices $\Phi_\alpha =$
955 $-63.5° \pm 4.5°$ and $\Psi_\alpha = -41.5° \pm 5°$ and for beta strands $\Phi_\beta = -118° \pm 10.7°$ and $\Psi_\beta = 134° \pm$
956 $8.6°$.

957 Restraints for beta sheet interactions were implemented by setting H-N:O=C hydrogen bond
958 distances between interacting positions to range between 1.8 and 2.1Å (Adhikari et al., 2015),
959 with weight one. Predictions of beta sheet interactions derived from deep mutational scanning
960 data yield a string of interacting positions, but hydrogen bonding in beta sheets occurs in
961 specific non-continuous patterns between position pairs (between alternating positions off the
962 interaction diagonal in parallel beta sheets and between every second set of position pairs in
963 anti-parallel beta sheets). Specifically, for each set of interacting positions there are two
964 alternative patterns of hydrogen bonding possible. These alternative possibilities of pairing were
965 implemented as mutually exclusive selection pairs with the "assign … or" syntax in Xplor-NIH.

966 Distance restraints were implemented in Xplor-NIH as NOE (nuclear Overhausser effect)
967 potential, with potential type set to "soft" for stages 1 and 2 and "hard" for the final simulation
968 stage. Dihedral angle restraints were implemented via the CDIH potential.

969 After simulation stages 1 and 2 restraints were checked for their consistency with predicted
970 structural models. Frist, structural models were clustered according to their violations of
971 distance and dihedral angle restraints (k-means clustering, k = 4). Clusters were ranked by the
972 mean total energy (from all energy potentials used) of their 50 models with lowest total energy
973 (or all of their structures if clusters are smaller 50 models). From the 50 models with the lowest
974 total energy from the top-ranked cluster (or however many top-ranked clusters were necessary
975 to arrive at 50 models) the fraction of models that violate specific restraints was recorded. For
976 the subsequent simulation stage, distance restraints were down-weighted according to the
977 fraction of models that violated them, $w_{x,i} = w_{x,i-1} * (1 - f_x)^2$, and distance restraints with a

weight below 0.1 were discarded. There is no option to weight dihedral angle restraints, thus instead dihedral angle restraints with a 'weight' below 1/3 were discarded for the subsequent simulation stages.

The top 5% structural models from simulation stage 3 were evaluated against the known crystal structure of protein G domain B1. The TM-score program (update 2016/03/23) was used to calculate the Cα root mean squared deviation and the template modeling score (Zhang and Skolnick, 2004).

Several types of control simulations were performed to judge the predictive power of restraints derived from deep mutational scanning data. Simulations without restraints from predicted contacts and beta sheet interactions, but with restraints from secondary structure elements predicted by PSIPRED (version 2.6, (Jones, 1999)), which serve as a negative control. Simulations with restraints derived from the known crystal structure of protein G domain B1 (PDB entry 1pga). Here, 56 true contacts of position pairs with linear chain distance greater than 5 amino acids were randomly sampled and beta sheet interactions were determined by PyMOL (The PyMOL Molecular Graphics System, 2015). These simulations serve as a positive control and give the maximally achievable accuracy of our Xplor-NIH workflow. Finally, a set of simulations that only used contact restraints derived from top predicted interaction sores and secondary structure restraints derived from PSIPRED predictions but no restraints for beta sheet interactions. These simulations serve to judge the power of predicting secondary structure elements and beta sheet interactions from deep mutational scanning data.

Code availability

Data was analysed 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. Custom R code is available upon request from the corresponding author.

Data availability

No primary data was generated in this study. Processed data is available upon request from the corresponding author.

Supplementary Text

Down sampling analysis

We evaluated how robust our prediction strategy is to lower data quality (Figure 7), thus assessing what might be the minimal requirements for deep mutational scanning datasets to be useful for structure prediction.

The protein G domain B1 dataset used here has a high signal-to-noise ratio of about 35x (measurement range relative to the median experimental error of fitness estimates). Assays based on cellular growth or phage display typically have smaller measurement ranges (Araya et al., 2012; Fowler et al., 2010; Li et al., 2016; Matreyek et al., 2017; Starita et al., 2013; Starita et al., 2015; Weile et al., 2017), thus in tendency smaller signal-to-noise ratios, which might render the majority of variants unsuitable for negative epistasis classification. We thus tested how our prediction strategy performs when only positive epistasis information is available. Using *combined scores* we find that accurate prediction of protein structure is still feasible with positive epistatic information only $\langle \mathrm{C\alpha - RMSD} \rangle = 3.4\mathring{A}$). This is mainly due to the prediction of beta sheet interactions, as the accuracy drop is larger if only tertiary contacts and secondary structure element prediction are used to model the domain ($\langle \mathrm{C\alpha - RMSD} \rangle = 6.4\mathring{A}$). Interestingly, using only negative epistasis information results in a complete loss of predictive power for the structure of the protein domain.

Furthermore, we tested the effect of a lower sequencing coverage. The dataset used here was produced from 600 million reads, roughly the equivalent of four flow cell lanes on an Illumina HiSeq sequencer (Olson et al., 2014). Because the number of double mutant variants scales quadratically with the protein length assayed, it is important to understand the effect of sequencing depth on prediction accuracy in order to assess whether scaling the approach to longer proteins is feasible. We artificially down-sampled the sequencing read coverage of the datasets by a factor four, thus doubling the error of fitness estimates ("noisy dataset"). The effect on accuracy was minor if using both positive and negative epistatic values as well as predicting beta sheet interactions $\langle \mathrm{C\alpha - RMSD} \rangle = 2.3\mathring{A}$), while the negative impact on accuracy increased if using only positive epistasis information or omitting restraints for beta sheet interactions ($\langle \mathrm{C\alpha - RMSD} \rangle = 4.1\mathring{A}$ and $\langle \mathrm{C\alpha - RMSD} \rangle = 5.6\mathring{A}$, respectively). We note that the reason for minor effects of down-sampling sequencing read coverage might come from the fact

1037 that for most variants the error from the fitness measurements itself might be the dominant
1038 source of uncertainty. Unfortunately, for the present dataset we cannot assess this due to the
1039 lack of replicate information. This does not affect the conclusions about the effects of lower
1040 sequencing coverage, but highlights the importance of performing the fitness experiments in
1041 replicate.

1042 Finally, we tested the effect of a lower coverage of double mutants in the dataset. Fewer double
1043 mutants will increase the uncertainty in estimating epistatic fractions per position pair, but
1044 limiting the number of double mutants constructed in an experiment would also allow to assay
1045 longer proteins. The present dataset was constructed by NNS-mutagenesis, i.e. replacing each
1046 wild-type amino acid with all other 19 amino acids. Synthesis strategies relying on doping or
1047 error-prone template replication can effectively only give rise to a fraction of all possible amino
1048 acid changes (Araya et al., 2012; Fowler et al., 2010; Starita et al., 2013; Starita et al., 2015;
1049 Weile et al., 2017). We thus considered only those single amino acid mutations in the dataset
1050 that can be reached by one mutation in the nucleotide sequence, thereby lowering the coverage
1051 of single mutants to 31% and the coverage of double mutants to 10% ("doped dataset"). Again,
1052 we find only a minor effect on accuracy if using the full epistatic information and predicted beta
1053 sheet interactions ($\langle \mathrm{C\alpha - RMSD} \rangle = 2.1\text{Å}$), but greater negative effects if using only positive
1054 epistasis information or omitting restraints for beta sheet interactions ($\langle \mathrm{C\alpha - RMSD} \rangle = 5.6\text{Å}$ and
1055 $\langle \mathrm{C\alpha - RMSD} \rangle = 5.1\text{Å}$, respectively).

Supplemental Figures

Supplemental Figure 1

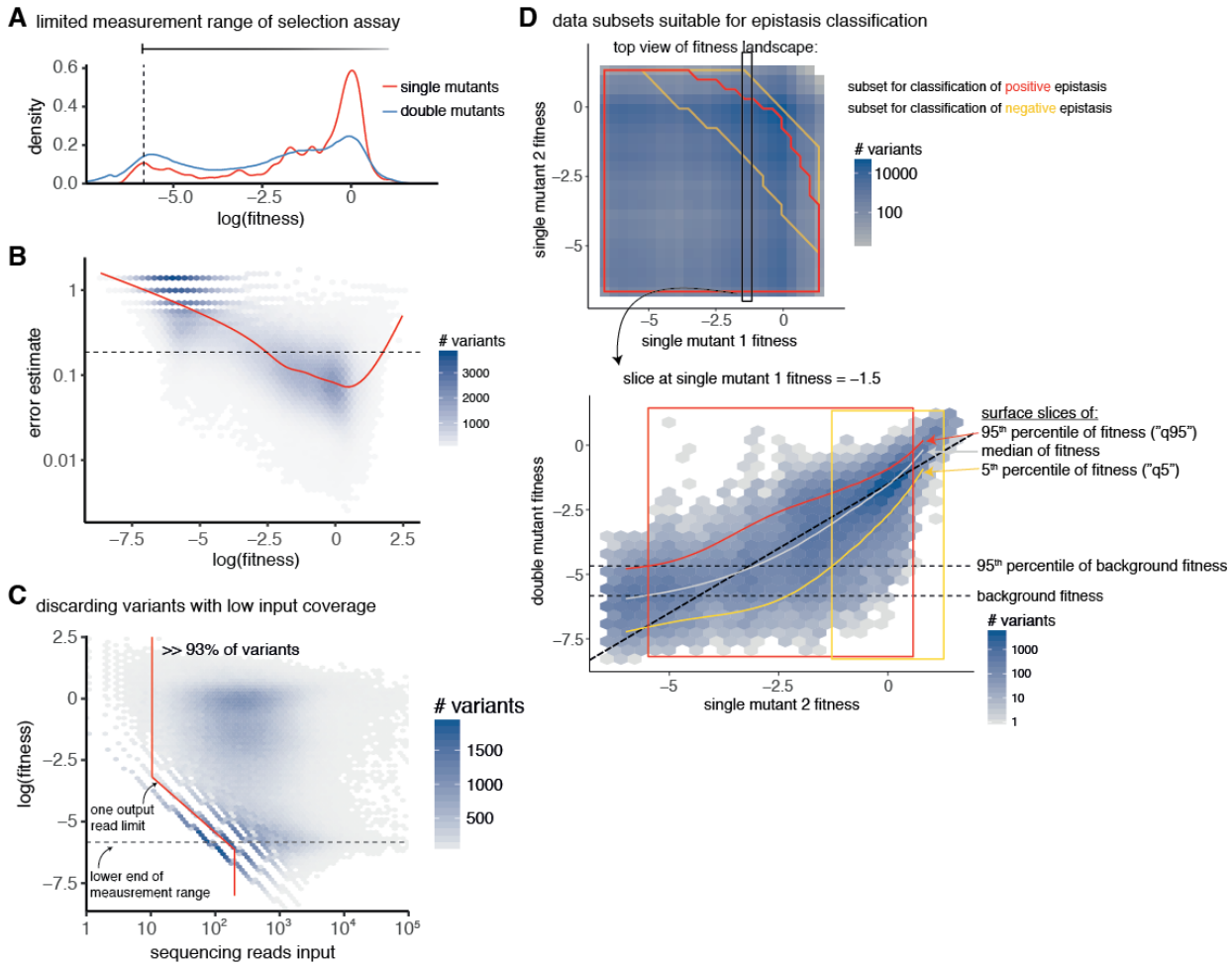


Figure S1: Deep mutational sequencing data processing

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).

B. Two-dimensional variant density showing relationship between variant fitness and their error estimates. Red line shows *loess* fit of data. Dashed horizontal line indicates median error estimate. Fitness values and their error estimates are linked due to common sequencing read count dependencies.

C. Two-dimensional variant density showing dependency of fitness values on sequencing read counts in input library. For variants with too low coverage in the input library low

1067 fitness values cannot be accurately estimated. Red line shows sequence read cutoff
1068 used for variant inclusion (93% of variants included for downstream analysis). Horizontal
1069 dashed line indicates lower limit of fitness assay measurement range.

1070 D. Reliable quantification of positive and negative epistasis is limited to subsets of the data.
1071 Upper plot shows two-dimensional double mutant variant density in single mutant fitness
1072 space. Red outline shows single mutant fitness space that enables positive epistasis
1073 quantification. Yellow outline shows single mutant fitness space that enables negative
1074 epistasis quantification. Lower plot shows an example slice through fitness landscape at
1075 single mutant fitness = -1.5. Red, grey and yellow curves show slices through quantile
1076 fitness surfaces of 95th percentile, median and 5th percentile, respectively (see Figure
1077 2A). Diagonal dashed line shows
1078 $double\ mutant\ fitness\ =\ single\ mutant\ 2\ fitness - 1.5$ (expected fitness = observed
1079 fitness). Horizontal dashed lines give lower limit of fitness assay measurement range
1080 and the 95th percentile of fitness values of variants dominated by background fitness
1081 effects. Red and yellow boxes indicate the range that includes 99% of variants within the
1082 slice that are suitable for positive or negative epistasis quantification, respectively.

1083

Supplemental Figure 2

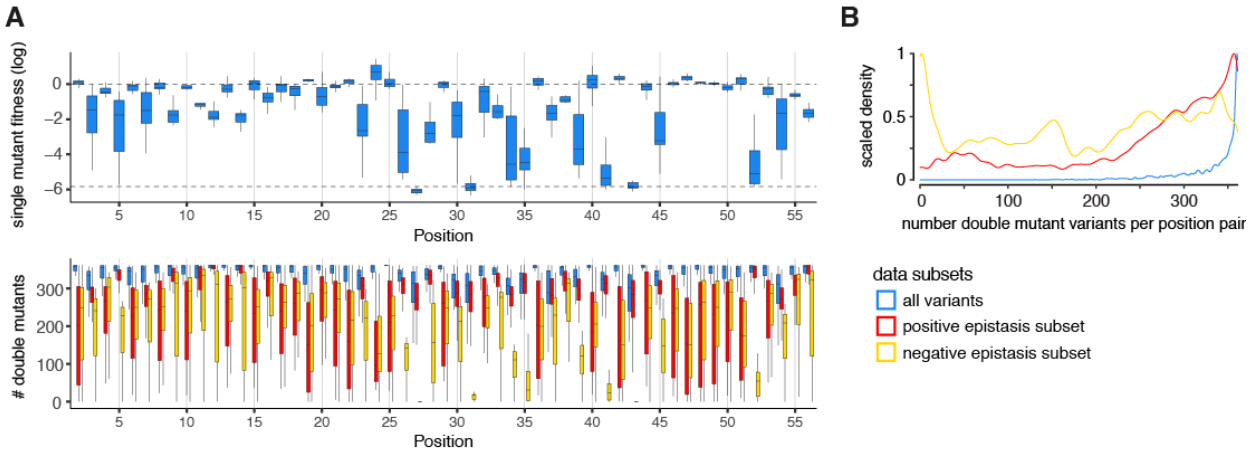


Figure S2: Variant number distributions across position pairs

A. Upper panel: Position-wise distributions of single mutant fitness values. Lower panel: Distributions of the number of double mutant variants for each position pair a position is involved in. Shown are distributions for all variants (blue), i.e. coverage after quality control (93% of double mutants), variants suitable to positive epistasis quantification (red) and variants suitable to negative epistasis quantification (yellow). Note how number of double mutant variants suitable to epistasis quantification relates to single mutant fitness distributions shown in upper panel.

B. Distribution of number of double mutant variants across all position pair. Colors as in panel A.

Supplemental Figure 3

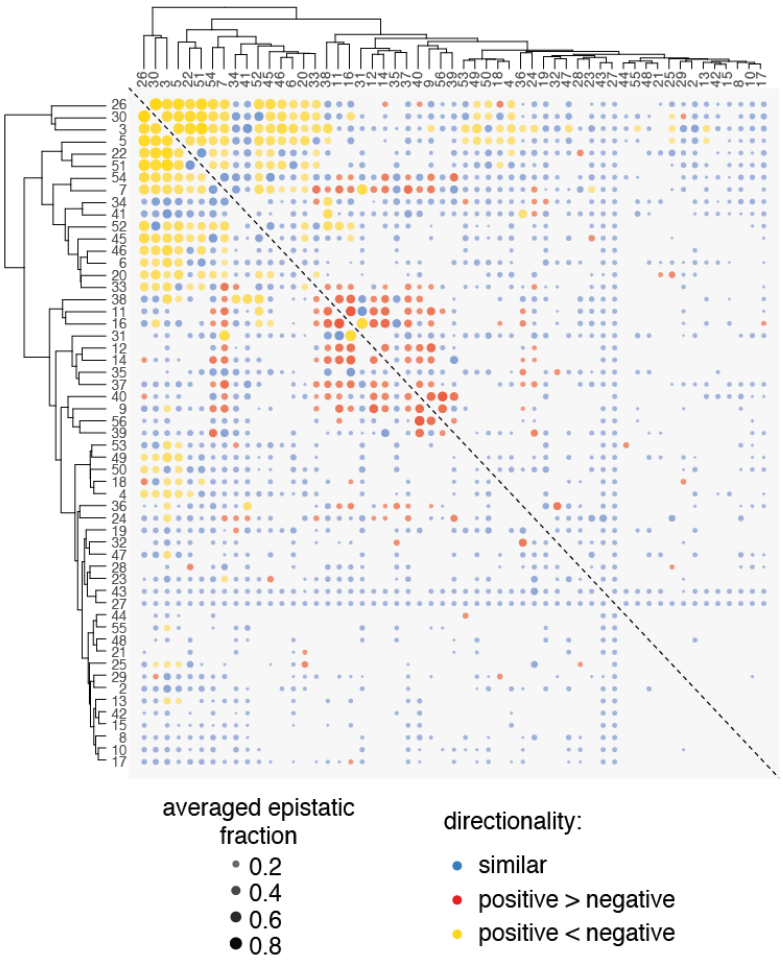


Figure S3: Positive and negative epistasis enrichments across position pairs

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 subsets of positions that interact mostly positively or negative with each other and other non-subset positions.

Supplemental Figure 4

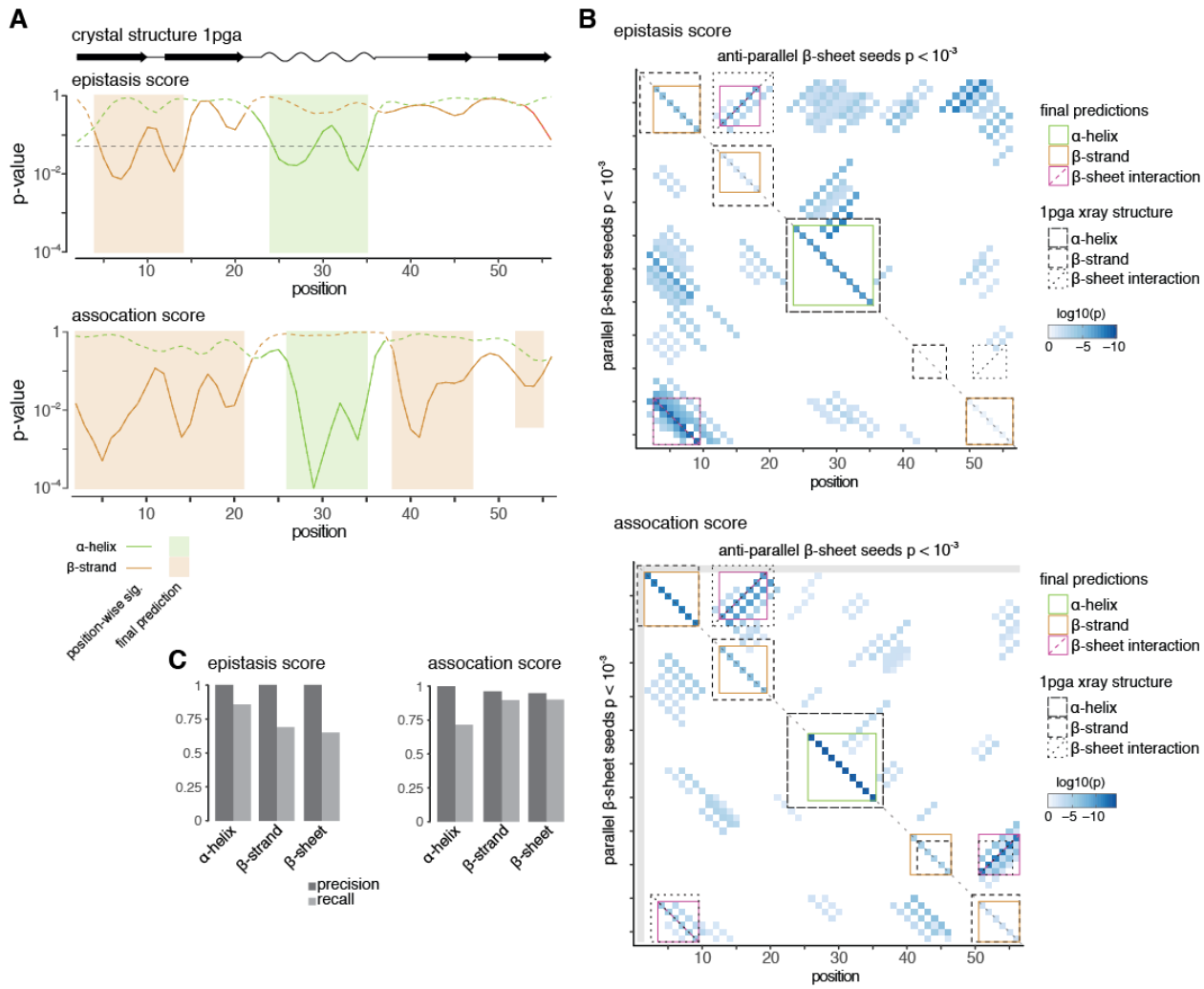


Figure S4: Secondary structure and beta sheet prediction from *epistasis* and *association scores*

A. Secondary structure element detection from *epistasis* (upper panel) and *association scores* (lower panel). Lines give position-wise p-value of secondary structure propensity; dashed lines indicate data was not used (p-value for other structure element lower); boxes indicate final predictions, lower limit of box indicates p-value of final predictions (limited to minimum of 10^{-4}). Secondary structure elements in crystal structure 1pga is shown for comparison on top.

B. Beta sheet interaction predictions from *epistasis* (upper panel) and *association scores* (lower panel). Plots show matrix of combined p-values for extended (beta sheet) seeds,

1118 the final predictions and the reference elements in the crystal structure (lower triangle –
1119 parallel beta sheets, upper triangle – anti-parallel beta sheets, diagonal – secondary
1120 structure elements).

1121 C. Precision and recall for position-wise secondary structure element and beta sheet
1122 interaction predictions from panel B.

1123

Supplemental Figure 5

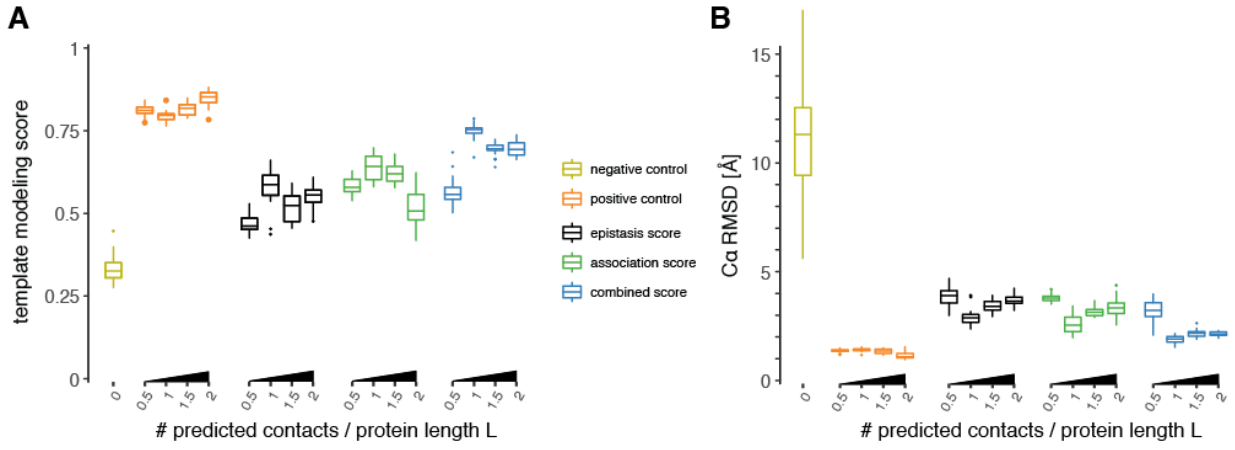


Figure S5: *Ab initio* folding of protein G domain B1

A. Template modeling score of top 5% structural models compared to crystal structure 1pga and the dependency on number of long-range contacts used. Negative control – only restraints for secondary structure predicted by PSIPRED. Positive control – 0.5-2*L ture contacts (random subset), secondary structure elements and beta sheet interactions from crystal structure. All other: all restraints – top 0.5-2*L contacts, secondary structure element and beta sheet interaction predictions from the three interaction scores, as indicated by color.

B. Dependency of accuracy ($C\alpha$ root mean squared deviation) of top 5% of structural models on number of contacts used. Colors as in panel A.