nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0431-x

Determining protein structures using deep mutagenesis

Jörn M. Schmiedel[1] and Ben Lehner[1,2,3]*

Determining the three-dimensional structures of macromolecules is a major goal of biological research, because of the close relationship between structure and function; however, thousands of protein domains still have unknown structures. 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 backbone structure of a biological macromolecule to be determined only from measurements of the activity of mutant variants of the molecule. This genetic approach to structure determination relies on the quantification of genetic interactions (epistasis) between mutations and the discrimination of direct from indirect interactions. This provides an alternative experimental strategy for structure determination, with the potential to reveal functional and in vivo structures.

Despite years of effort and technological developments, thousands of protein domains still have unknown three-dimensional structures[1]. Mutations within a protein or RNA can have non-independent effects on fitness (called genetic interactions or epistasis)[2–5] and double mutants have been used to probe the energetic couplings between positions in a protein to understand determinants of protein folding and stability[6,7]. Early work revealed that at least some strongly interacting mutations within a protein are in direct structural contact[6–10] (Fig. 1a). Deep mutational scanning (DMS) of proteins[11–14] and RNAs[15–18] has further revealed that some—but by no means all—epistatic interactions occur between structurally proximal mutations.

Support for the idea that non-independence between mutations provides structural information comes from the analysis of amino acid and nucleotide sequence evolution. Here, correlated pairs of amino acids or nucleotides in multiple sequence alignments identify co-evolving positions within proteins and RNAs[19–21]. These patterns of co-evolution have been used to identify energetically coupled positions and independently evolving 'sectors' within proteins[22,23]. Moreover, when very large numbers of homologous proteins and RNAs are available in sequence databases, the application of global statistical models can reveal direct structural contacts from patterns of co-evolution[24–26], allowing the prediction of macromolecular structures and interactions[1,27–35].

The question then becomes whether epistatic interactions quantified from DMS experiments can be used to determine macromolecular structures. If successful, structure determination by DMS would offer a number of advantages over established techniques. First, it requires no specialized equipment or expertise beyond the ability to mutate a molecule, select functional variants, and quantify enrichments by sequencing. Appropriate in vitro and in vivo selection assays already exist for many molecules of interest and generic assays based on folding, stability and physical interactions have also been developed[11,36–39]. Second, it could be applied to molecules for which structures are difficult to determine by physical techniques such as intrinsically disordered and membrane proteins. Third, unlike evolutionary coupling analysis there is no requirement for large numbers of homologous sequences[1,28,40] and so it could be applied to fast-evolving, recently evolved and de novo designed proteins and RNAs. Finally, and perhaps most importantly, it would provide a general strategy to determine the physiologically relevant structures of molecules while they are performing particular functions that can be selected for, including in vivo in cells. A potentially cheap and straightforward approach for studying macromolecular structures in vivo would be an exciting new frontier for cell and molecular biology.

Here we show that DMS of proteins can provide sufficient information to determine their three-dimensional backbone structures. Our statistical approach quantifies how often mutations between positions interact epistatically and how these epistatic interaction patterns correlate. These metrics accurately identify individual tertiary structure contacts as well as secondary structure elements within a protein. The same approach also identifies contacts between protein-interaction partners. DMS data alone suffice to determine backbone structures with accuracies of 1.9 Å C_α root mean square deviation (r.m.s.d.) compared to known reference structures. Moreover, we show that deep learning can further improve prediction performance, allowing the use of sparser and lower quality DMS datasets for structure determination. Our approach therefore provides an experimental strategy for structure determination that can reveal functional and in vivo structures.

Results

Epistasis is enriched in, but not exclusive to, structural contacts. We first investigated the relationship between epistasis and structure for more than half a million mutant variants ($55\times19=1{,}045$ single mutants plus nearly $55\times54/2\times19\times19=536{,}085$ double mutants) of the protein G B1 (GB1) domain[13]. For these variants, protein fitness was quantified using binding to an immunoglobulin G (IgG) fragment as a selection assay, resulting in a two-orders-of-magnitude measurement range with a median relative error of fitness estimates of 2.8% (Table 1 and Supplementary Fig. 1a).

We used a running median surface approach as a null model for the independence of the effects of double mutations (Fig. 1b) to account for nonspecific dependencies between mutants introduced by the fitness assay or nonspecific epistatic behavior from

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [3]ICREA, Barcelona, Spain. *e-mail: ben.lehner@crg.eu

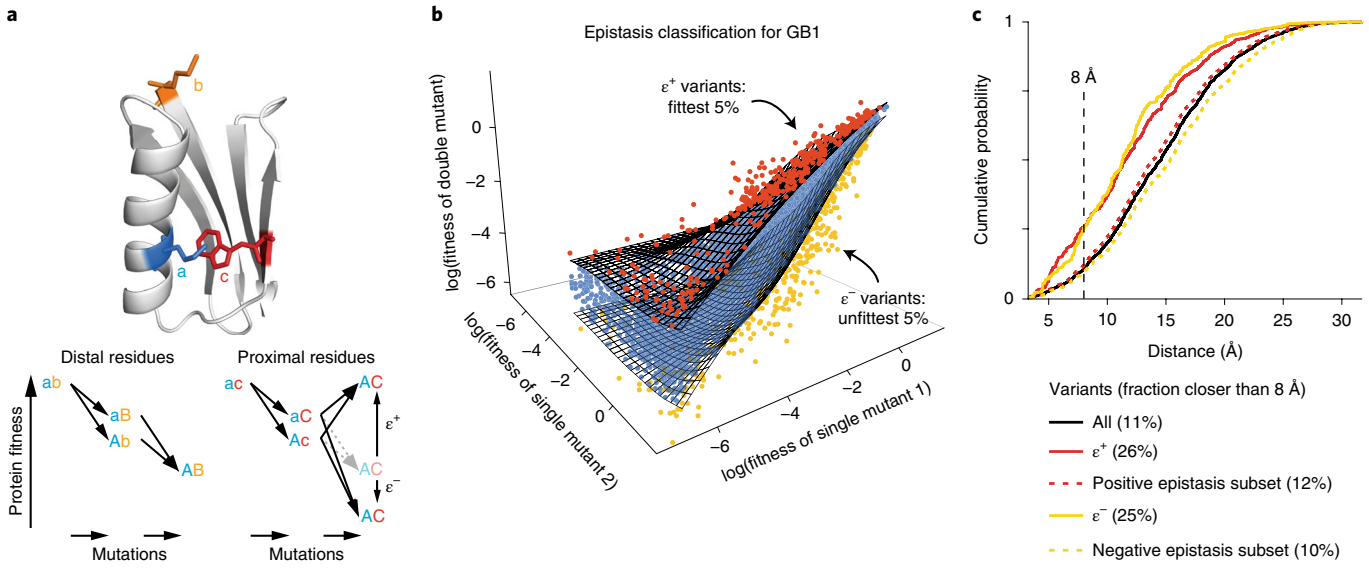

Fig. 1 | Extracting epistatic mutational effects from DMS of a protein domain. a, The premise of our study was that if genetic interactions (or epistasis) are mostly caused by structural interactions, then comprehensively quantifying epistatic interactions should suffice to predict the structure of a molecule. The structure of GB1 is shown (Protein Data Bank (PDB) entry 1PGA)[61] with residues a, b and c colored. **b**, Classifying epistatic variants based on deviations from expected fitness (quantile fitness surface approach). Variants with the 5% most extreme fitness values given fitness of their respective single mutants were classified as positive (red, ε^+) or negative (yellow, ε^-) epistatic. A random sample of 10^4 variants in the GB1 domain[13] is shown. **c**, Distance distribution of epistatic variants separated by more than 5 amino acids in the linear sequence (minimal side-chain heavy atom distance). Positive and negative epistasis subsets refer to the sets of variants suitable for epistasis analysis (see Supplementary Fig. 1c). All variants, $n = 400{,}647$; positive epistatic variants ε^+, $n = 14{,}127$; positive epistasis subset, $n = 315{,}862$; negative epistatic variants ε^-, $n = 9{,}837$; negative epistasis subset, $n = 208{,}442$.

Table 1 | Dataset properties

Dataset	Mutated amino acid positions	Double mutants (%)[a]	Quantifiable double mutants (%)[b]		Median number of input reads per double mutant[c]	Measurement range (log units)[d]	Median relative error[e]
			Positive epistasis	Negative epistasis			
GB1 domain[13]	55	97	80	55	248	6	2.8%
Human YAP WW domain[43]	33	10	8.3	0.8	73	0.8	8.6%
PAB1 RRM2 domain[12]	25	11	8.3	3.9	209	3.1	3.7%
FOS–JUN interaction[11]	2×32	43	37	31	124	8.6	3.6%

[a]The median percentage of all possible double mutants (361 per position pair) that passed read quality thresholds per position pair. [b]Median percentage of all possible double mutants (361 per position pair) that passed read quality thresholds and are deemed suitable for epistasis quantification per position pair. [c]Summed number of reads across all input replicates for double mutants that passed read quality thresholds. [d]Measurement range of selection assay: log-transformed fitness range between peak of lethal mutants and the wild-type variant. [e]Median error of fitness estimates of double-mutant variants relative to measurement range of selection assay.

thermodynamic stability effects[2,11]. Double mutants were classified as positive or negative epistatic if they have more extreme fitness than the 95th or 5th percentile fitness surfaces, respectively. Restricting the classification of epistasis to variants not impeded by measurement errors resulted in 80% and 55% of double mutants being suitable for positive or negative epistasis classification, respectively, with substantial variability across the position matrix (Table 1 and Supplementary Fig. 1b–f).

Consistent with previous observations[12–14], both positive and negative epistatic double mutants are enriched for proximal variants, for example, more than twofold at a distance of 8 Å (Fig. 1c, only considering position pairs separated by more than 5 amino acids in the linear sequence; closer positions are trivially also close in three-dimensional space and their proximity contributes little to successful structure prediction[30]). However, about 75% of epistatic interactions are between positions that are not in direct contact in the protein (as judged by an 8 Å distance cut-off), suggesting that indirect effects often underlie specific epistatic interactions within a molecule[22,23]. The challenge for structure determination therefore becomes how to infer direct structural contacts from the mixture of direct and indirect effects that underlie epistasis.

Likelihood of epistatic interactions and correlated interaction profiles predict tertiary structure contacts. To discriminate direct structural contacts from a list of thousands of epistatic double mutants, we used two measures.

The first, which we refer to as the enrichment score, quantifies how often double mutants between each pair of positions interact with positive or negative epistasis (Fig. 2a). Calculating the fraction of epistatic interactions separately for either positive or negative interactions enriches for structural contacts, but for different regions of the domain (Fig. 2b and Supplementary Fig. 2). Combining the positive and negative epistatic fractions, while taking into account quantification errors, further enriches for direct contacts (precision

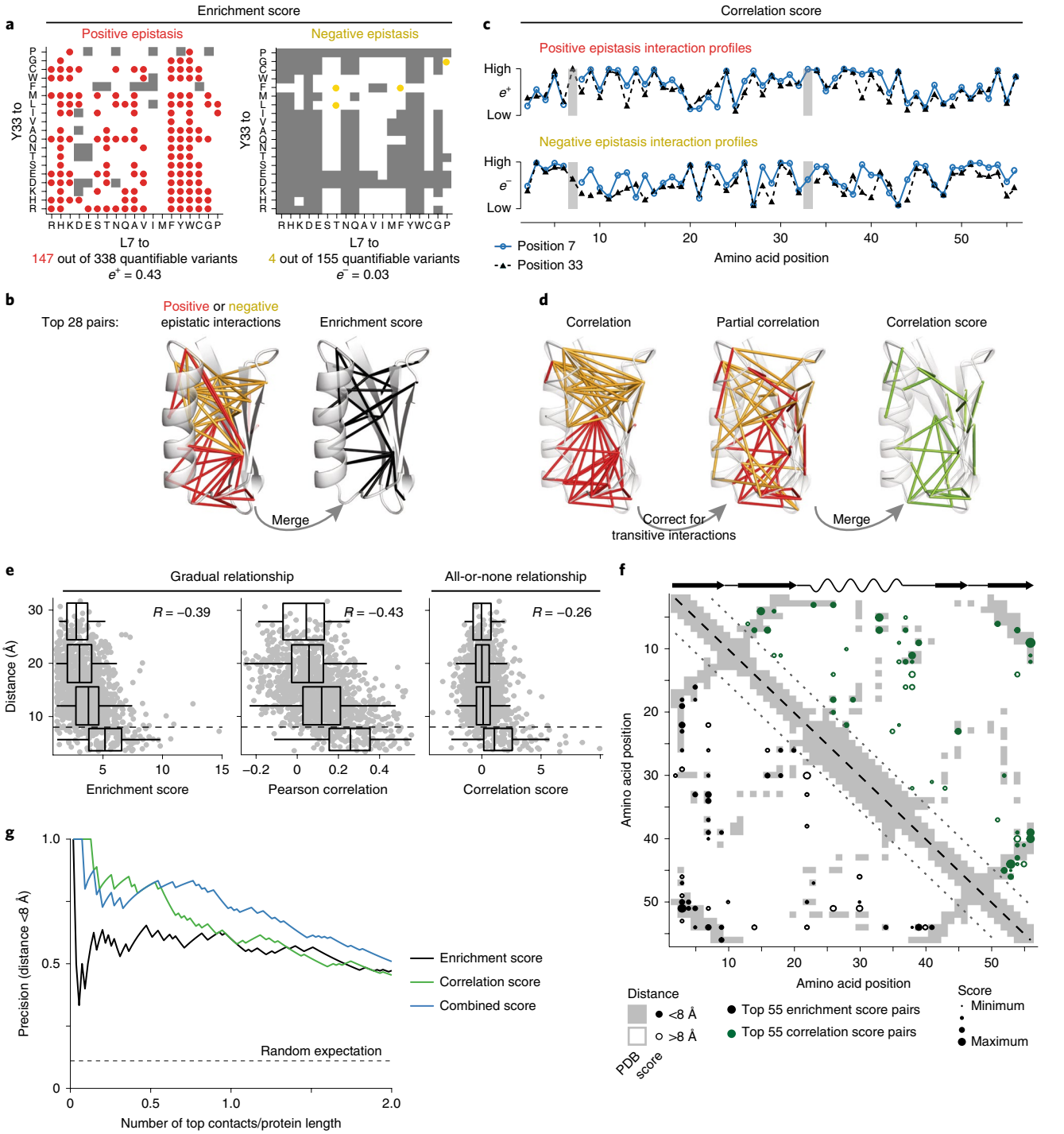

Fig. 2 | Likelihood of epistatic interactions and correlated interaction profiles predict tertiary structure contacts. a, The enrichment score quantifies the fraction of positive and negative epistatic interactions per position pair (here positions 7 and 33). Gray shading indicates epistatic interactions that are not quantifiable (see Supplementary Fig. 1c–f). **b**, Structural distribution of top 28 epistatic interaction pairs (PDB entry 1PGA). Left, pairs with the highest positive (red) and negative (yellow) epistatic enrichments. Right, pairs with highest enrichment scores. **c**, The correlation score quantifies the similarity of epistatic interaction profiles between position pairs. Example of positive (top) and negative (bottom) epistatic interaction profiles for positions 7 and 33 (marked by gray horizontal bars) is shown. **d**, Structural distribution of top 28 pairs with highest positive (red) or negative (yellow) Pearson correlations (left), partial correlations (middle) or correlation scores (right) of interaction profiles. **e**, Distance of position pairs (>5 amino acids in linear sequence, n = 1,225) as a function of enrichment scores, merged Pearson correlation of epistasis interaction profiles or correlation scores. Box plots are spaced in intervals of 8 Å; boxes cover the first to third quartile of the data, with the middle bar indicating the median, whiskers extend to 1.5× the interquartile range away from the box. Dashed horizontal line indicates an 8 Å threshold. Pearson correlation coefficients are indicated. **f**, Distribution of top 55 position pairs (>5 amino acids in linear sequence, indicated by dotted lines) with highest enrichment score (black, bottom left triangle) or correlation scores (green, top right triangle) on contact map of the reference structure (gray shading). Reference secondary structure elements (wave, α-helix; arrow, β-strand) are shown on top. **g**, Precision of interaction scores to predict direct contacts (distance < 8 Å) as a function of top scoring position pairs. There are 131 direct contacts out of 1,225 pairs (>5 amino acids in linear sequence), horizontal dashed line indicates random expectation.

(PRC) for top $L/2$ contacts $PRC_{L/2}=61\%$, $PRC_L=60\%$, with $L=55$ amino acids as the length of the mutated sequence, Fig. 2g), with these contacts evenly distributed across the domain (Fig. 2b,f).

The second score, which we refer to as the correlation score, quantifies the similarities of epistasis interaction profiles—how a position interacts with all other positions in the protein—between each pair of positions. The assumption that underlies this score is that positions close in space in a structure should interact similarly with all other positions (Fig. 2c). We used partial correlations—thus correcting correlations for transitive signals—to better distinguish direct from indirect contacts and again calculated scores separately for positive and negative interactions before merging them while taking into account quantification errors (Fig. 2d). The final correlation scores show a more binary all-or-none relationship with distance than the enrichment scores or when using simple correlations to quantify similarity (Fig. 2e), thus better prioritizing the top direct structural contacts across the whole domain (Fig. 2f,g, $PRC_{L/2}=79\%$, $PRC_L=60\%$).

Finally, combining the enrichment and correlation scores into a combined score by simply summing normalized scores further improves contact predictions, especially when considering lower ranked predictions ($PRC_{L/2}=82\%$, $PRC_L=73\%$; Fig. 2g).

Identification of secondary structure elements. We hypothesized that the periodic geometrical arrangement of amino acid residues in secondary structures (3.6 residues per α-helical turn and alternating side-chain directions in β-strands) might result in periodic epistasis patterns in DMS data[28,41] (Fig. 3a). We used a two-dimensional kernel smoothing approach to detect α-helical and β-strand periodicities (Fig. 3b) and found significant periodicities for one α-helix and four β-strands that coincide very well with secondary structure elements in the reference structure (Fig. 3c and Supplementary Fig. 3a). Moreover, stretches of off-diagonal, long-distance interactions show the expected alternating patterns for either parallel or anti-parallel β-sheets, with the top predictions corresponding to the known anti-parallel interactions of β1−β2 and β3−β4 as well as the parallel interaction of β1−β4 (Fig. 3d and Supplementary Fig. 3b,c). Furthermore, updating β-strand predictions according to inferred β-sheet pairings led to improved β-strand prediction, enforcing a split between β1 and β2 and correcting the length of β3 and β4 (Fig. 3c,d). Overall, these secondary structure element predictions achieve precision and recall values of about 90% when derived from correlation scores (or combined scores; Supplementary Fig. 3d). Predictions from enrichment scores are less precise, thus suggesting that eliminating transitive, indirect interactions is important for secondary structure prediction.

Tertiary structure prediction. We next tested whether the DMS data alone could be used to determine the structure of the protein domain. We performed structural simulations by simulated annealing using the XPLOR-NIH modeling suite[42], with the top L scoring position pairs as distance constraints as well as dihedral angle constraints for predicted secondary structure elements and restrictive distance constraints for predicted β-sheet positions that form hydrogen bonds (Fig. 3e).

Comparing the structural models against the experimentally determined crystal structure of GB1 revealed that the combined scores provided the best predictions, with the top 5% of models (25 out of 500, evaluated on internal energy terms and constraint satisfaction) having an average C_α r.m.s.d. of 1.9 Å and an average template modeling score of 0.71 (Fig. 3f,g and Supplementary Fig. 3f), which is very close to the optimum achievable with our simulation protocol (using contacts, secondary structure elements and β-sheet interactions from the reference structure, C_α r.m.s.d. = 1.4 Å and template modeling score = 0.8). Consistent with the slightly lower precision of tertiary contact and secondary structure predictions, models generated with constraints from enrichment or correlation scores have—on average—a lower accuracy (C_α r.m.s.d. = 3.4 Å and C_α r.m.s.d. = 2.6 Å, respectively), although correlation score models performed consistently better.

Together, this shows that DMS alone is sufficient to accurately determine the backbone structure of a protein domain.

Deep mutagenesis identifies protein interaction contacts and structures. Epistatic interactions can also occur between different proteins, for example between physical interaction partners[3]. We tested whether epistasis between two proteins quantified using our metrics could predict their structural interactions. We used a dataset[11] in which we had made all possible amino acid mutations at 32 positions in the products of the *FOS* and *JUN* proto-oncogenes and quantified the physical interaction of all single and (*trans*-)double mutants using a deep sequencing-based protein complementation assay (Fig. 4a and Table 1). Notably, enrichment scores show a binary all-or-none relationship with distance similar to the correlation scores in GB1 (Fig. 4b), with distant position pairs across the interaction surface contained in a low enrichment score peak and proximal interactions enriched for high enrichment scores. Indeed, the top 11 enrichment score pairs are all proximal interactions, and the precision of contact prediction is $PRC_{L/2}=75\%$ and $PRC_L=66\%$ (12-fold and 10.5-fold increase over expectation). Moreover, top enrichment score pairs are evenly distributed across the interaction surface (Fig. 4a,c).

Correlating the epistatic interaction profiles between columns of the epistatic enrichment matrices compares the epistatic interactions that two positions in FOS have with all positions in JUN. Therefore, the similarity of column-wise epistatic profiles identifies the *cis* relationships between positions in FOS, while row-wise interaction profiles identify *cis* relationships between positions in JUN (Supplementary Fig. 4a). The *cis*-interaction maps from correlation scores for both FOS and JUN are highly enriched in local interactions and applying our secondary structure prediction algorithms reveals strong α-helix propensities across the full lengths of both FOS and JUN, consistent with the coiled-coil structure of the complex (Fig. 4c and Supplementary Fig. 4b).

This shows that DMS of protein-interaction partners can accurately predict direct contacts across the interaction surface as well as reveal the underlying structural conformations of the interaction partners themselves.

Generality and data requirements for successful protein structure prediction. To test the generality of our approach, we analyzed two additional DMS of individual protein domains, the PAB1 RRM2 domain[12] and the human YAP65 WW domain[43] (Fig. 5a,b). These datasets contain only incomplete sets of double mutants (approximately 10%), were sequenced less deeply and have up to six times smaller measurement ranges, resulting in up to three times higher relative measurement errors and fewer double mutants that were suitable for quantification of epistasis (especially negative epistasis) (Table 1 and Supplementary Fig. 5a). Nonetheless, tertiary contacts can be predicted with good precision (combined score $PRC_{L/2}=57\%$ (threefold higher than random expectation) and $PRC_{L/2}=59\%$ (3.9-fold increase over expectation) for the RRM2 and WW domain, respectively; Fig. 5c,d and Supplementary Fig. 5b). Secondary structure predictions were inaccurate and underpowered (0% precision), but β-sheet pairing was inferred correctly (100% precision and recall for RRM2 domain), albeit off by one and two positions for the two anti-parallel sheet interactions in the WW domain (Fig. 5c,d).

We used the top $L/2$ predicted combined score contacts to model the structure of the secondary structure-rich central part of the WW domain (positions 6–29, 24 amino acids, see Methods). The top 5% of structural models have an average accuracy of 3.3 Å C_α r.m.s.d. compared to the reference structure (Fig. 5a), which is on

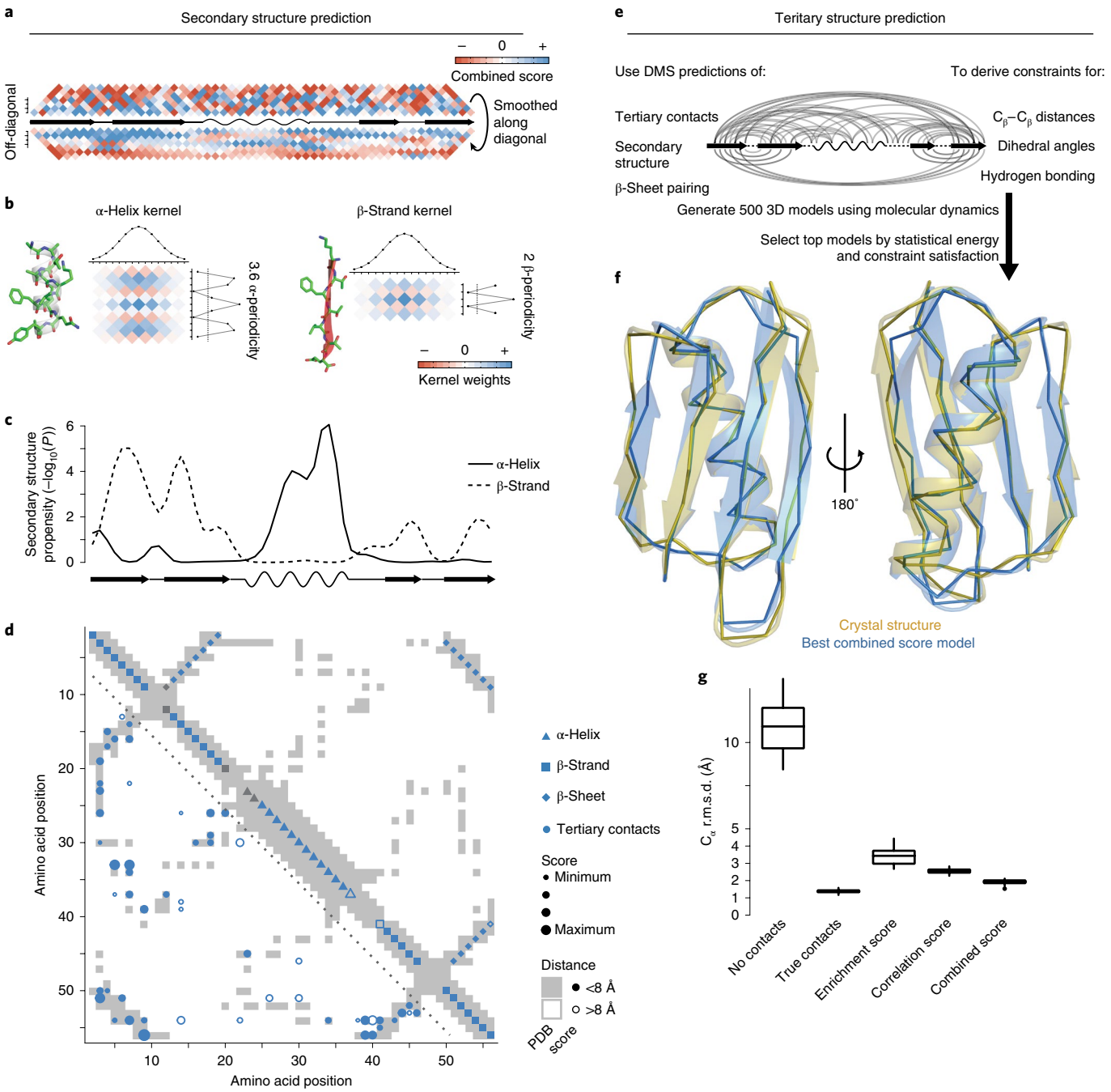

Fig. 3 | Secondary and tertiary structure prediction from DMS data. a, Local interactions (above diagonal, raw combined scores up to 7-amino-acid distance in linear sequence; below diagonal, scores smoothed with a Gaussian kernel) reveal signatures of secondary structure. Middle line is diagonal of interaction score map (rotated by 45°) and shows secondary structure elements of reference structure. **b**, Two-dimensional kernels with sinusoidal profile to detect stereotypical α-helical (left, period of 3.6) and β-strand (right, period of 2) interactions and perpendicular Gaussian profile to average over similar interaction patterns in adjacent positions. **c**, Secondary structure propensity *P* values derived from kernel smoothing (one-sided permutation test, see Methods) in comparison to reference structure secondary structures (wave, α-helix; arrow, β-strand). **d**, Structural predictions derived from combined score data compared to reference structure contact map (gray shading). Bottom left, top 55 non-local (>5 amino acids in linear sequence) tertiary contacts. Top right, predicted secondary structure elements. Fill indicates correct prediction. The β-strand predictions are derived by intersection of β-strand propensities (**c**) and β-sheet pairing predictions (Supplementary Fig. 3b,c). **e**, Schematic for the generation of three-dimensional structural models (see Methods for details). **f**, Overlay of top structural model of the GB1 domain generated with constraints from combined score (blue) and crystal structure (gold, PDB entry 1PGA). **g**, Accuracy (C_α r.m.s.d.) of top 5% structural models ($n = 25$) generated from interaction score-derived constraints (three right-most columns) compared to reference structure. Left, 'no contacts' indicates negative control with constraints only for secondary structure (predicted by PSIPRED)[62]. 'True contacts' indicates a positive control with constraints derived from 55 random tertiary contacts, secondary structure elements and β-sheet interactions of the reference structure. Box plots show boxes that cover the first to third quartile of the data, with the middle bar indicating the median, and whiskers that extend to 1.5× the interquartile range away from the box.

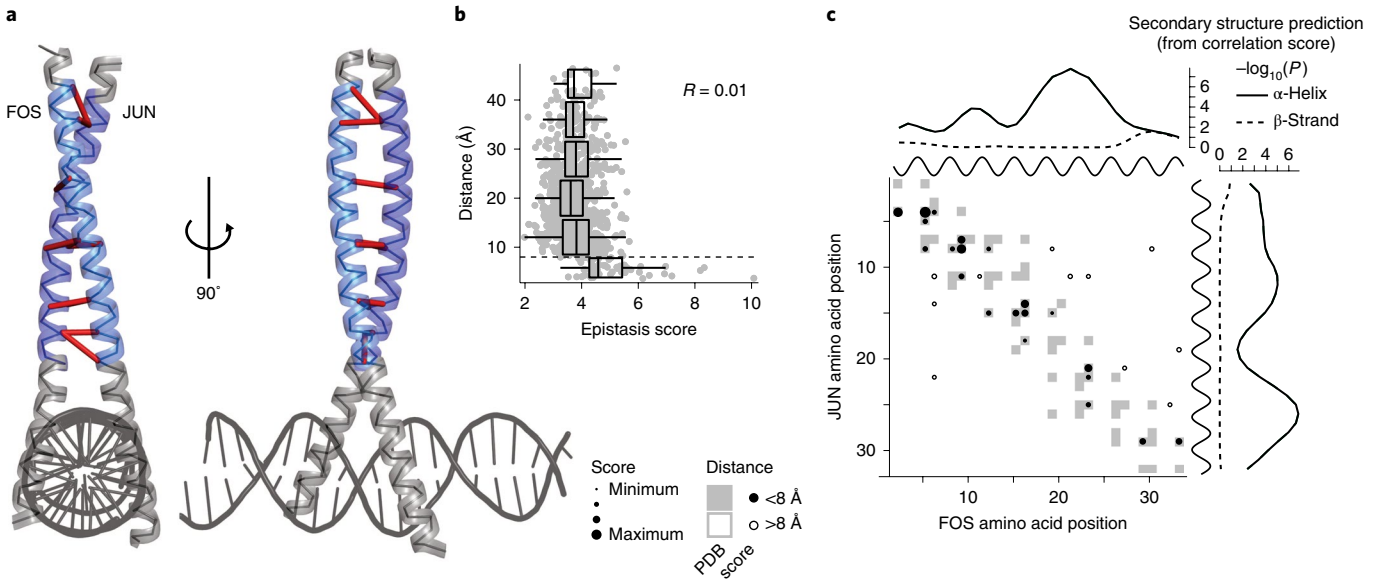

Fig. 4 | Deep mutagenesis identifies protein-interaction contacts. a, Crystal structure of the leucine zipper domains of FOS and JUN with a DNA strand (PDB entry 1FOS). The mutated regions (32 amino acids each) are highlighted in light blue (FOS) and dark blue (JUN)[11]. Top ten enrichment score pairs are shown with red dashes, note that two interactions between position 8 in FOS and positions 7 and 8 in JUN, as well as three interactions between positions 14 and 15 in FOS and positions 14 and 15 in JUN are hard to distinguish. **b**, Distance of position pairs as a function of enrichment scores ($n = 1{,}024$). Box plots are spaced in intervals of 8 Å; boxes cover the first to third quartile of the data, with the middle bar indicating the median, whiskers extend to 1.5× the interquartile range away from the box. Dashed horizontal line indicates an 8 Å threshold. Pearson correlation coefficient is indicated. **c**, FOS–JUN *trans*-interaction score map for top 32 position pairs with highest enrichment scores, compared to contact map of known interaction structure (1FOS, shown as an underlying structure in gray). Note that protein–protein interaction maps are not symmetric. Shown on top and to the right of the contact map are the known α-helices (black) as well as the secondary structure propensities derived from correlation scores of FOS and JUN (one-sided permutation test, see also Supplementary Fig. 4a,b).

par with simulations using a set of true contacts (C_α r.m.s.d. = 3.6 Å; Supplementary Fig. 5c). We could not make structural predictions for the RRM2 domain because it was mutagenized in three independent segments.

To estimate the minimal requirements for DMS datasets to be useful for structure prediction, we investigated how robust our prediction strategy is to changes in data quality by artificially downsampling the GB1 domain dataset.

First, we considered the sequencing read coverage and found that even using only 10% of the 600 million sequencing reads in the full GB1 dataset hardly affects the precision of predicted tertiary contacts ($PRC_L = 64\%$, a drop by 9% compared to the full dataset; Fig. 5e). Only when using just 2.5% of sequencing reads (15 million) does the precision of the top L contacts drop below 50% ($PRC_L = 45\%$).

Second, we simulated a 'doped' mutagenesis dataset, by only considering amino acid mutations that can be reached by one mutation in the nucleotide sequence—thus reducing the coverage of double mutants to approximately 10% (similar to the RRM2 and WW domain datasets). The doped dataset exhibits a decrease in precision of predicted tertiary contacts of about 20% ($PRC_L = 51\%$; Fig. 5e). Moreover, the doped dataset shows an increased sensitivity to lower sequencing read coverage.

Third, we tested the effect of small signal-to-noise ratios (that is, the measurement range of the selection assay relative to the median error of fitness estimates, which results in unquantifiability of negative epistasis, see Supplementary Figs. 1d–f and 5a), by using only positive epistasis information to calculate interaction scores. This also results in a drop of precision of about 20% ($PRC_L = 55\%$). By contrast, only using negative epistasis information resulted in a drop to 33% precision, as low as a doped dataset with low sequencing coverage.

Finally, we evaluated how differences in prediction performance of tertiary contacts affect structural modeling. Changes in accuracy of the top structural models scale with changes in contact prediction performance (Fig. 5f). Downsampling of sequencing reads in the complete dataset from 100% to 2.5% leads to a decrease in average accuracy from 2.5 Å to 4 Å C_α r.m.s.d., which is roughly also the accuracy of top structural models from the doped dataset and the dataset using only positive epistasis information.

Together, these results support the generality of our approach for extracting structural information from DMS data, including from sparser and lower quality datasets.

Deep learning improves contact prediction. Evolutionary coupling-based structural predictions have been successfully improved by machine-learning approaches that transform the two-dimensional interaction score maps after learning the stereotypical patterns between evolutionary coupling-predicted contact maps and experimentally determined contact maps[44,45].

We tested whether machine learning could also improve DMS-derived contact predictions. We applied a convolutional neural network called DeepContact[44], which transforms a two-dimensional interaction score map based on the structural patterns it has previously learned on evolutionary coupling-derived contact predictions for representative families of the SCOPe database[46] (Fig. 6a and Methods).

We first transformed the combined score interaction map of the GB1 domain using the DeepContact network. These transformations take as sole input our DMS-derived predictions and include no evolutionary coupling or otherwise-derived structural predictors for GB1. The scores on the transformed map are much less noisy, with high scores exclusively focused in areas of structural contacts, especially those of secondary structure element interactions, and areas

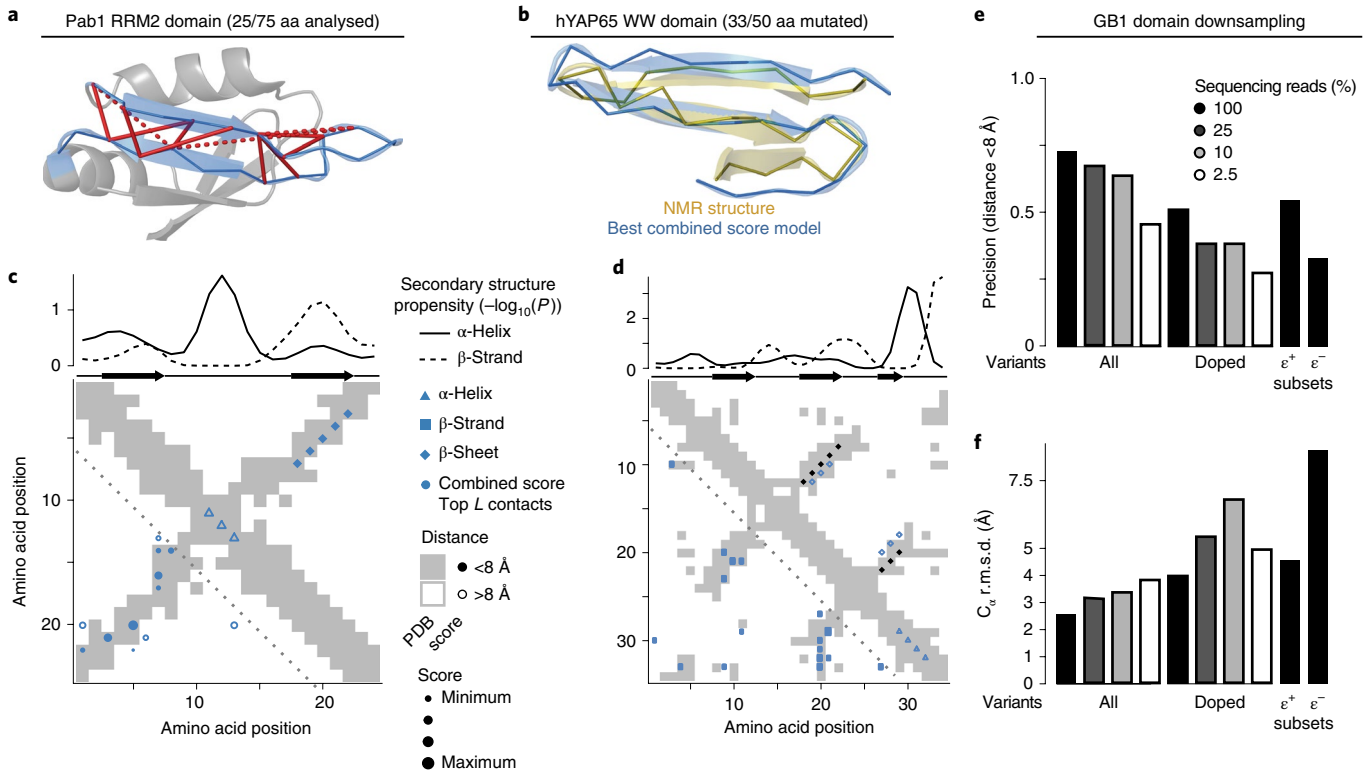

Fig. 5 | Generality and data requirements for successful protein structure prediction from DMS data. a, PAB1 RRM2 domain (PDB entry 1CVJ). The analyzed 25-amino acid segment is highlighted in blue. The top 12 combined score position pairs are connected with red lines, solid if distance <8 Å, dashed otherwise. **b**, Overlay of top structural model of the human YAP65 WW domain (positions 6–29) generated with constraints from combined score (blue) and solution NMR structure (gold, PDB entry 1K9Q). **c**, Structural predictions derived from combined scores in RRM2 domain. The top plot shows secondary structure propensities from kernel smoothing (one-sided permutation test) in comparison to secondary structures in the reference structure. Map shows top 12 combined score position pairs in bottom left and secondary structure predictions in top right triangle, in comparison to reference contact map (gray shading). **d**, Structural predictions derived from combined scores in the WW domain. The top plot shows secondary structure propensities from kernel smoothing (one-sided permutation test) in comparison to secondary structures in reference. Map shows top 17 combined score position pairs in bottom left and secondary structure predictions in top right triangles, in comparison to the reference contact map (gray shading). Black diamonds indicate positions of β-sheet pairing in the reference map. **e**, Precision of top *L* combined score position pairs for different downsampled versions of the GB1 dataset (in terms of type of variants analyzed or sequencing coverage). **f**, Accuracy (average C_α r.m.s.d.) of top 5% structural models ($n = 25$) derived with tertiary contact constraints from downsampled GB1 datasets compared to the reference structure.

devoid of structural contacts showing homogenously low scores (Fig. 6b). The precision of top predicted contacts improves from 82% to 96% for *L*/2 and from 73% to 87% for *L* predicted contacts (Fig. 6c).

Predictions derived from the two other GB1 interaction scores (enrichment and correlation scores) as well as the interaction score maps for the other datasets (downsampled GB1, FOS–JUN, RRM2 and WW data) show similar improvements both in terms of cleaner interaction score maps that better resemble the reference contact maps as well as increases in contact prediction performance of up to 30% (Fig. 6c and Supplementary Fig. 6). By contrast, randomized interaction score maps show no changes in prediction performance over random expectation after transformation with DeepContact.

Finally, we tested whether DeepContact-transformed contact predictions could also improve structural modeling. On downsampled GB1 datasets, DeepContact-transformed predictions increased the accuracy of structural models by up to 2.6 Å (Fig. 6d). For the complete datasets with only 25% or 10% of sequencing reads, the top structural models have better accuracy than those from the complete dataset with full sequencing read coverage but untransformed scores. In addition, structural models based on DeepContact-transformed scores from the doped dataset with full or 25% sequencing coverage and those from the dataset using only positive epistasis information reach average accuracies of 3.2 Å C_α r.m.s.d. Only for the two datasets with 2.5% sequencing read coverage do structural simulations based on DeepContact-transformed scores not improve model accuracy.

This shows that machine learning can substantially improve contact map prediction from DMS data, thus allowing the use of even sparser and lower quality data for accurate structure prediction.

Discussion

We have shown here that simply quantifying the activity of a large number of single- and double-mutant variants of a macromolecule can provide enough information to reliably determine its three-dimensional fold.

Our analyses and previous work[6–9,11–18] have shown that many epistatic interactions occur between positions that are not in direct structural contact. Indeed, in the GB1 domain, the interactions are strikingly modular, with two mutually exclusive clusters of positive and negative epistatic interactions arising potentially from differential energetic couplings to protein stability and binding (Fig. 2b,d and Supplementary Fig. 2c), reminiscent of the concept of semi-independent energetically coupled protein sectors that have been identified from patterns of sequence co-evolution[22,23].

Nonetheless, aggregating epistatic interactions on position pairs, merging of positive and negative epistasis information and partial

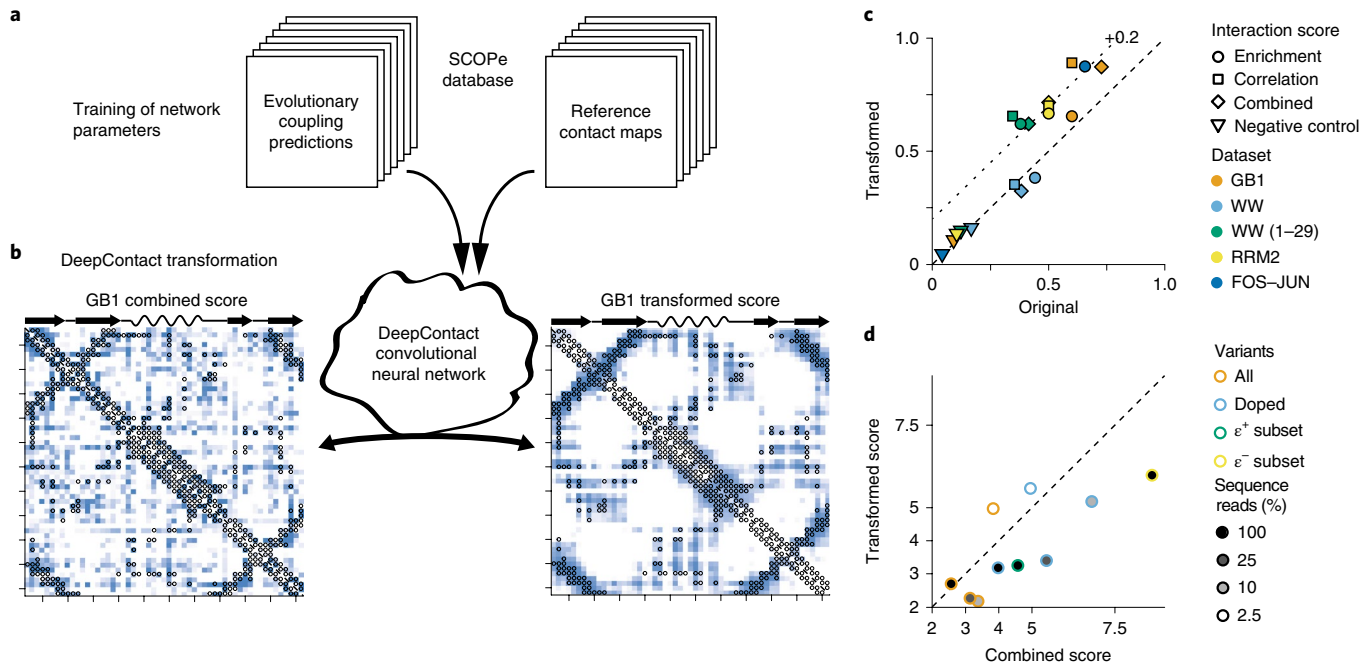

Fig. 6 | Deep learning improves contact prediction and structural models from deep mutagenesis data. a, DeepContact convolutional neural network transforms DMS-derived interaction score maps based on learned structural patterns[44]. The basic DeepContact architecture used here takes as the only input the DMS-derived interaction score map and transforms it based on structural patterns previously learned on an orthogonal and independent training set (in which it compared evolutionary coupling-derived contact predictions with contacts in known structures of representative protein families in the SCOPe database). **b**, The combined score interaction map of the GB1 domain before (left) and after (right) transformation with the DeepContact convolutional neural network. Heat maps show scores (low, white; high, blue). Gray open circles show contacts (distance < 8 Å) in the reference structure. **c**, Precision of top *L* predicted contacts before and after DeepContact transformation. Negative control is the average over three random permutations of combined score matrices (or in the case of the FOS–JUN dataset, enrichment score matrices). WW (1–29) refers to a truncated version of the WW domain; see Supplementary Fig. 6). **d**, Comparison of accuracy (average C_α r.m.s.d.) of top 5% GB1 structural models ($n = 25$ each) with constraints derived either from combined scores or from DeepContact-transformed combined scores for different (downsampled) GB1 DMS datasets.

correlation analysis of epistasis interaction profiles can successfully discriminate direct from indirect structural contacts. Thus, mostly indirect epistatic couplings can be transformed to accurately predict secondary structure elements and tertiary contacts to reveal the protein fold.

We have shown that our approach works robustly across multiple protein domains and a protein interaction. Moreover, we have demonstrated that the application of a convolutional neural network previously trained on patterns of co-evolution in proteins of known structure both improves structure prediction and allows the use of much lower quality DMS datasets. We note that our approach is likely to be only one of several that could work[47].

We expect that development of the computational approach (consideration of the underlying physico-chemistry, better scoring methods and extracting side-chain information) as well as integration with other structural predictors[44,48,49] and homology-driven structure modeling[50,51] is likely to further improve accuracy and lower the data-quality requirements for structure determination by deep mutagenesis.

Whether it will it be possible to determine the structures of larger molecules by deep mutagenesis remains to be investigated. It is currently unclear how the requirements for variant coverage scale with protein length or the complexity of folds. However, the fact that sparse double-mutant datasets can suffice for structure prediction, and the rapid development of DNA synthesis and sequencing technologies suggest that similar approaches may work for larger structures. Currently, DMS libraries for larger proteins could be created via fragment-based ligation[52] or programmed mutagenesis[53,54] and sequenced by linking variants to short barcodes[36,37] to overcome the current size limitations of short-read sequencers.

A limitation of the current approach is that, similar to methods based on evolutionary couplings of residues[24,30], it identifies tertiary contacts but does not provide atom-level structural information. However, our finding that epistatic interactions contain information on the periodic arrangement of side-chain orientations in secondary structure elements and that tertiary contacts are better described by side-chain than backbone atom distances (Supplementary Fig. 7) suggests that genetic interactions are mostly mediated by structural interactions of amino acid side chains and that it might be possible to extract additional information about their orientations to improve structural modeling.

Determining structures by DMS offers several practical advantages. The approach does not require the expensive scientific infrastructure that most physical techniques require and uses methods familiar to molecular biologists. Selection assays based on known functions or interaction partners already exist for many proteins[13,16,17,43,52,55–59] and the development of generic assays for stability and activity[36–39] should allow it to be applied to molecules of unknown function. The approach also potentially brings the power of high-throughput genomics to structural biology. For example, using the existing infrastructure of genomics institutes, a large-scale project to systematically determine the structures of all protein domains of unknown structure is a plausible endeavor. Finally, and perhaps most interestingly, DMS allows the structures of macromolecules to be studied in vivo in the cell[60]. Ultimately, it is the structures of macromolecules as they perform a particular function in vivo that are of most interest. Deep mutagenesis, selection and sequencing provide a generic approach for in vivo structural biology.

In summary, DMS provides an experimental strategy for structure determination and opens up the possibility of low-cost and high-throughput determination of in vivo macromolecular structures, both by individual laboratories and by large-scale genomics projects.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0431-x.

Received: 9 October 2018; Accepted: 29 April 2019;
Published online: 17 June 2019

References

1. Ovchinnikov, S. et al. Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017).
2. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
3. Lehner, B. Molecular mechanisms of epistasis within and between genes. *Trends Genet.* **27**, 323–331 (2011).
4. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
5. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).
6. Horovitz, A. & Fersht, A. R. Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *J. Mol. Biol.* **214**, 613–617 (1990).
7. Carter, P. J., Winter, G., Wilkinson, A. J. & Fersht, A. R. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (*Bacillus stearothermophilus*). *Cell* **38**, 835–840 (1984).
8. Ackermann, E. J., Ang, E. T., Kanter, J. R., Tsigelny, I. & Taylor, P. Identification of pairwise interactions in the α-neurotoxin–nicotinic acetylcholine receptor complex through double mutant cycles. *J. Biol. Chem.* **273**, 10958–10964 (1998).
9. Chen, J. & Stites, W. E. Energetics of side chain packing in staphylococcal nuclease assessed by systematic double mutant cycles. *Biochemistry* **40**, 14004–14011 (2001).
10. Roisman, L. C., Piehler, J., Trosset, J. Y., Scheraga, H. A. & Schreiber, G. Structure of the interferon–receptor complex determined by distance constraints from double-mutant cycles and flexible docking. *Proc. Natl Acad. Sci. USA* **98**, 13231–13236 (2001).
11. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, e32472 (2018).
12. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).
13. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).
14. Sahoo, A., Khare, S., Devanarayanan, S., Jain, P. C. & Varadarajan, R. Residue proximity information and protein model discrimination using saturation-suppressor mutagenesis. *eLife* **4**, e09532 (2015).
15. Li, C. & Zhang, J. Multi-environment fitness landscapes of a tRNA gene. *Nat. Ecol. Evol.* **2**, 1025–1032 (2018).
16. Li, C., Qian, W., Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837–840 (2016).
17. Domingo, J., Diss, G. & Lehner, B. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature* **558**, 117–121 (2018).
18. Puchta, O. et al. Network of epistatic interactions within a yeast snoRNA. *Science* **352**, 840–844 (2016).
19. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309–317 (1994).
20. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**, 693–707 (1987).
21. Gloor, G. B., Martin, L. C., Wahl, L. M. & Dunn, S. D. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* **44**, 7156–7165 (2005).
22. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
23. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).
24. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–E1301 (2011).
25. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
26. Burger, L. & van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.* **6**, e1000633 (2010).
27. Weinreb, C. et al. 3D RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963–975 (2016).
28. Tóth-Petróczy, A. et al. Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170 (2016).
29. Hopf, T. A. et al. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
30. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
31. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
32. De Leonardis, E. et al. Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.* **43**, 10444–10455 (2015).
33. Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *Proc. Natl Acad. Sci. USA* **109**, 10340–10345 (2012).
34. Ovchinnikov, S. et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248 (2015).
35. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
36. Matreyek, K. A. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
37. Weile, J. et al. A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).
38. Rocklin, G. J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
39. Kim, I., Miller, C. R., Young, D. L. & Fields, S. High-throughput analysis of in vivo protein stability. *Mol. Cell Proteomics* **12**, 3370–3378 (2013).
40. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
41. Andreani, J. & Söding, J. bbcontacts: prediction of β-strand pairing from direct coupling patterns. *Bioinformatics* **31**, 1729–1737 (2015).
42. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–73 (2003).
43. Araya, C. L. et al. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl Acad. Sci. USA* **109**, 16858–16863 (2012).
44. Liu, Y., Palmedo, P., Ye, Q., Berger, B. & Peng, J. Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst.* **6**, 65–74 (2018).
45. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. & Bonvin, A. M. J. J. Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins* **86**, 51–66 (2018).
46. Fox, N. K., Brenner, S. E. & Chandonia, J. M. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* **42**, D304–D309 (2014).
47. Rollins, N. J. et al. Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0432-9 (2019).
48. Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999–1006 (2015).
49. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).
50. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzym.* **383**, 66–93 (2004).
51. Yang, J. et al. The I-TASSER suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
52. Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. Preprint at *bioRxiv* https://doi.org/10.1101/213835 (2017).
53. Firnberg, E. & Ostermeier, M. PFunkel: efficient, expansive, user-defined mutagenesis. *PLoS ONE* **7**, e52031 (2012).
54. Wrenbeck, E. E. et al. Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* **13**, 928–930 (2016).
55. Starita, L. M. et al. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).
56. Starita, L. M. et al. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl Acad. Sci. USA* **110**, E1263–E1272 (2013).

57. Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).
58. Fowler, D. M. et al. High-resolution mapping of protein sequence–function relationships. *Nat. Methods* **7**, 741–746 (2010).
59. McLaughlin, R. N. Jr, Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).
60. Bolognesi, B. et al. The mutational landscape of a prion-like domain. Preprint at *bioRxiv* https://doi.org/10.1101/592121 (2019).
61. Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* **33**, 4721–4729 (1994).
62. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).

Acknowledgements

We thank Y. Liu and J. Peng for making their DeepContact code available and for their advice; members of the Lehner laboratory, T. Gross, G. Mönke, M. Bolognesi and C. Camilloni for discussions and feedback. This work was supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Economy, Industry and Competitiveness (MEIC; BFU2017-89488-P), the AXA Research Fund, the Bettencourt Schueller Foundation, Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322), the EMBL-CRG Systems Biology Program and the CERCA Program/Generalitat de Catalunya. J.M.S. was supported by an EMBO Long-Term Fellowship (ALTF 857-2016). This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 752809 (J.M.S.). We acknowledge support from the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership and the Centro de Excelencia Severo Ochoa.

Author contributions

J.M.S. and B.L. conceptualized the study; J.M.S. developed the methods and carried out the study; J.M.S. and B.L. wrote the paper; B.L. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0431-x.

Reprints and permissions information is available at www.nature.com/reprints.

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

Methods

Datasets and preprocessing. *The GB1 domain.* The DMS data for GB1 were obtained from the supplementary information of a previous study[13]. The data consist of summed read counts of three replicate experiments assaying the binding affinity of GB1 variants to IgG.

Read frequencies of each single- or double-mutant variant in the input library and output library (after the binding affinity assay) were calculated as variant read counts relative to wild-type variant read counts. The fitness of a variant was calculated as the natural logarithm of the ratio of output to input read frequency, that is, $f_i = \log\left(\frac{n_i^{\text{out}} / n_{\text{WT}}^{\text{out}}}{n_i^{\text{in}} / n_{\text{WT}}^{\text{in}}}\right)$ where n is the read counts, superscripts denote the input (in) or output (out) sequencing library and subscripts denoting variant i or wild-type (WT) variant.

The standard error of fitness estimates was calculated from read counts under Poissonian assumptions[63], that is, $\sigma_i = \sqrt{\frac{1}{n_i^{\text{in}}} + \frac{1}{n_i^{\text{out}}} + \frac{1}{n_{\text{WT}}^{\text{in}}} + \frac{1}{n_{\text{WT}}^{\text{out}}}}$. We note that this is a lower bound estimate of the actual error, owing to the lack of replicate information.

Each measurement assay has a lower measurement limit due to unspecific background effects (Supplementary Fig. 1a). In the case of the IgG-binding assay for GB1, this is presumably mainly due to unspecific carryover on beads[13]. The fitness values derived from the measurement are therefore a convolution of the actual binding affinities to IgG and nonspecific carryover, that is, $\exp(f_i^{\text{measured}}) = \exp(f_i^{\text{binding}}) + \exp(f^{\text{carryover}})$ and fitness values of variants close to the lower measurement limit of the assay are dominated by unspecific carryover effects. The lower measurement limit of the assay was estimated by two approaches that yielded similar estimates. The first used a kernel density estimate of the single-mutant fitness distribution (R function density with parameter bw set to 0.15), where the position of the lower mode of the data corresponded to $f^{\text{carryover}} = -5.85$. The second examined the fitness distribution of double mutants with expected fitness lower than −8 log-transformed units, that is, double mutants resulting from two lethal or nearly lethal single-mutant variants, for which the fitness values are thus expected to be dominated by background effects. The median of this background fitness distribution yielded an estimate of $f^{\text{carryover}} = -6.14$. The mean of the two estimates, that is, $f^{\text{carryover}} = -6$ (approximately 0.25% on a linear scale) was used for downstream analyses.

In addition, 7% of double-mutant variants were discarded because of a too low sequencing coverage in input or output libraries (Supplementary Fig. 1b). That is, variants with 10 or less input read counts were discarded because of too high errors in fitness estimates. Moreover, variants with less than 200 input reads and no output reads were discarded, because it is not possible to determine their fitness. Above 200 input reads, variants without output reads are certain to be dominated by nonspecific carryover effects. These variants were retained and their fitness was calculated by setting their output read count to 0.5.

Downsampling of the GB1 data. Downsampling of the full GB1 dataset was performed in three different ways. First, to downsample the sequencing read coverage, the read count of each variant was drawn from a binomial distribution with the number of sequencing reads in the full datasets as trials and the target downsampling rate (25%, 10% or 2.5%) as chance of success. Second, in the doped datasets, only amino acid changes created by one nucleotide mutation from the wild-type sequence (ENA entry M12825) were retained. For the read downsampled and doped datasets (and combinations of both), the analysis workflow for the full dataset was repeated.

For the downsampled datasets in which only positive or negative epistatic information was taken into account, enrichment and correlation scores were calculated from epistatic enrichment matrices and partial correlation matrices of only positive or negative epistasis information. Instead of merging positive and negative matrices and then calculating z-scores, z-scores were calculated with the individual errors from only positive or negative epistasis information. The combined scores (for which results are reported) for each set were then calculated as for the full dataset by summing standardized enrichment and correlation scores.

The human YAP65 WW domain. Data for the human YAP65 WW domain were obtained from Sequence Read Archive (SRA) entry SRP015751[43]. Paired-end reads were merged with USearch[64] and merged reads for which any base had a Phred base quality score below 20 were discarded. Read counts from the two technical sequencing replicates were merged and read counts for the same amino acid variants with at most one synonymous mutation in one other codon were summed. The dataset consists of an input library and three output libraries after consecutive rounds of selection in a phage display assay. Fitness was estimated as the slope of log-transformed frequency (variant counts divided by wild-type counts) changes over the rounds of selection experiment[43]. For each variant at each selection step a Poissonian error of $\sigma_{i,x} = \sqrt{\frac{1}{n_i^x} + \frac{1}{n_{\text{WT}}^x}}$ was calculated, where x denotes the selection step. Slopes were calculated as weighted straight-line least-square fits[65]. Comparison of library-wide changes in variant frequencies between selection rounds suggested differential selection pressures across the rounds. We thus applied a non-equidistant spacing of 0.6, 1.17 and 1.22 between selection rounds when calculating slopes. Only variants that have more than 10 reads in the input library and at least one read after the first selection were retained for further analysis (45% of constructed double mutants). The lower fitness limit was calculated as the weighted mean fitness of all variants that contained STOP codons (−0.78 in log units).

PAB1 RRM2 domain. PAB1 RRM2 domain data were obtained from Supplementary Table 5 of a previous study[12]. Reported variant read enrichment scores were log-transformed to obtain fitness values. Output reads per variant were deduced from the number of input reads times the read enrichment score and used to calculate a Poissonian error of the fitness estimate. Single-mutant count data are not provided and we thus estimated the error of single-mutant fitness estimates to be 0.01. The lower bound of the fitness assay was estimated as the weighted mean fitness of all double-mutant variants that contained STOP codons (−3.1 log units). In the dataset, three 25-amino acid segments were mutated independently, and we restricted analysis to the middle segment (position 26–50) containing a significant number of non-local contacts.

FOS–JUN interaction. Raw count tables were provided by G. Diss[11]. The dataset consists of input and output sequencing libraries after selection for physical interactions between the two proteins in a protein complementation assay in three biological replicates. Per sequencing library, read counts from all synonymous variants were summed up. Only variants that had more than 10 reads in each of the three input libraries were used for further analysis (43% of double mutants). Per input–output replicate, fitness of each variant was calculated as the log-transformed change in frequency compared to the wild-type variant (as for GB1). A Poissonian error for the fitness estimate of each variant was derived. The lower measurement bound of the fitness assay was estimated as the weighted mean fitness of all double STOP mutant variants (−8.6 log units). A Bayesian estimator of fitness values was implemented to overcome variant dropout due to a large dynamic range in the fitness assay (see Supplementary Note).

Epistasis classification. Epistasis was calculated from a non-parametric null model—running quantile surfaces—to account for nonlinearities close to the lower limit of the measurement range of the fitness assay, nonspecific epistatic behavior resulting from, for example, thermodynamic stability thresholds as well as differential uncertainty of fitness measurements across the fitness landscape, due to lower read counts in the output for low fitness variants (Fig. 1b).

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

The evaluation of positive or negative epistasis was, however, restricted to specific subsets of the data for which measurement errors do not impede epistasis classification (see Supplementary Note and Supplementary Fig. 1c). As a result of these restrictions as well as differences in initial coverage, the number of double-mutant variants that can be used to assess positive and negative epistasis varies substantially across position pairs and datasets (see Table 1 and Supplementary Figs. 1d–f, 4c, 5a).

Interaction scores. Several interaction scores were derived to estimate which position pairs are in close contact in the tertiary structure (Fig. 2a,c; see Supplementary Fig. 8 for an overview of the workflow). These scores are based on summarizing epistasis information on the position pair level and accounting for the varying uncertainty of the summarized estimates due to differential error of fitness estimates across the measurement range as well as varying numbers of double mutants amenable to epistasis classification (see Table 1 and Supplementary Figs. 1d–f, 4c, 5a). To summarize epistasis information on the position pair level, the fraction of positive or negative epistatic variants per position pair was calculated (number of epistatic variants divided by the number of variants amenable for epistasis classification; Supplementary Fig. 8, step 5b). Because enrichments with positive and negative epistatic variants per position are anti-correlated (Supplementary Fig. 2a), positive and negative enrichments were treated separately and only aggregated to derive the final interaction scores. The uncertainty of interaction scores was calculated from a resampling procedure in which fitness values for the variants as well as the resulting epistatic fractions were drawn from appropriate probability distributions (see Supplementary Note for details and Supplementary Fig. 8, step 5).

Enrichment scores, which quantify how often positions interact epistatically, were derived by merging positive and negative epistatic fractions by weighted averaging, that is, $e_{xy} = \frac{e_{xy}^{+} \times \sigma_{e_{xy}^{+}}^{-2} + e_{xy}^{-} \times \sigma_{e_{xy}^{-}}^{-2}}{\sigma_{e_{xy}^{+}}^{-2} + \sigma_{e_{xy}^{-}}^{-2}}$ where $e_{xy}^{\pm}$ is the mean positive/negative epistatic fraction and $\sigma_{e_{xy}^{\pm}}$ is the standard deviation of positive/negative epistatic

fractions across resampling runs. These merged epistatic fractions were further normalized by their uncertainty, that is, $E_{xy} = e_{xy}/\sigma_{xy}$ with $\sigma_{xy} = (\sigma_{e_{xy}^{+}}^{-2} + \sigma_{e_{xy}^{-}}^{-2})^{-1/2}$ to arrive at the final enrichment score (Supplementary Fig. 8, step 6).

Correlation scores are derived from the similarity of epistasis interaction profiles 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 profiles of epistatic interactions with all other positions. First, symmetric matrices (of size mutated amino acid positions × mutated amino acid positions) of positive and negative epistatic enrichments were constructed (Supplementary Fig. 8, step 5c). Missing values (position pairs without observed variants) were imputed by drawing a random value from the overall distribution of epistatic fractions. A pseudo-count equal to the first quartile of the epistatic fraction distribution was added to all matrix entries. Diagonal elements (epistatic fractions of a position with itself) were set to 1. The matrix values were transformed by the natural logarithm and for each pair of columns the Pearson correlation coefficient was calculated to arrive at the correlation matrix (Supplementary Fig. 8, step 5d). The correlation matrix was regularized using a shrinkage approach[66] to minimize the mean-squared error between the estimated and true correlation matrix and to obtain a positive definite and well-conditioned correlation matrix suitable for inversion (R package corpcor). Next, partial correlations of epistatic interaction profiles between each position pair were calculated by inverting the regularized correlation matrix and normalizing each off-diagonal entry of the inverted matrix by the geometric mean of the two respective diagonal entries, that is $a_{xy}^{+} = \frac{r_{xy}^{-1}}{\sqrt{r_{xx}^{-1} r_{yy}^{-1}}}$, with r_{xy}^{-1} as the (x,y) entry of the inverted correlation matrix (Supplementary Fig. 8, step 5d). We note that this approach is similar to how mean-field approaches can help to discriminate direct from indirect evolutionary couplings in multiple-sequence alignments[24,30,67]. Equivalent to the enrichment score, positive and negative partial correlation estimates were merged by calculating weighted averages of their mean estimates across resampling runs, with weights as the inverse variances across resampling runs, that is $a_{xy} = \frac{a_{xy}^{+} \times \sigma_{a_{xy}^{+}}^{-2} + a_{xy}^{-} \times \sigma_{a_{xy}^{-}}^{-2}}{\sigma_{a_{xy}^{+}}^{-2} + \sigma_{a_{xy}^{-}}^{-2}}$, and the final correlation score was obtained by normalizing by the combined uncertainty $A_{xy} = a_{xy}/\sigma_{xy}$ with $\sigma_{xy} = (\sigma_{a_{xy}^{+}}^{-2} + \sigma_{a_{xy}^{-}}^{-2})^{-1/2}$ (Supplementary Fig. 8, step 6).

Finally, a combined score was derived by summing the standardized enrichment and correlation scores, that is $C_{xy} = \frac{E_{xy} - \bar{E}}{\sigma_E} + \frac{A_{xy} - \bar{E}}{\sigma_A}$, to prioritize position pairs that are enriched for epistatic interactions and have similar epistasis profiles. We note that this is a naive approach to combining the information from these two complementary sources, and surely more sophisticated approaches that further improve proximity estimates can be developed in the future.

Protein distance metrics. The minimal distance between side-chain heavy atoms of two residues (in the case of glycine, C_α) was used as the distance metric. A direct contact was defined as a minimal side-chain heavy atom distance <8 Å. Only position pairs with linear sequence separation greater than 5 amino acid were considered when evaluating tertiary contact predictions. Evaluating contact predictions only on side-chain heavy atom distances instead of all heavy atoms increases true-positive rates over random expectation, thus suggesting that epistatic interactions are mostly mediated by structural interactions of amino acid side chains (Supplementary Fig. 7).

We used the following reference structures for comparison:

- GB1 domain: PDB entry 1PGA, X-ray diffraction structure[61]
- WW domain: PDB entry 1K9Q, solution NMR structure[68]
- RRM2 domain: PDB entry 1CVJ (chain A), X-ray diffraction[69] structure of human PAB1; note that the central section of the analyzed yeast RRM2 domain is one nucleotide longer than the corresponding homologous region in the human RRM domain. We therefore arbitrarily removed position 14 (in the loop region, as previously described[12]) when comparing the DMS-derived predictions to the human PAB1 structure.
- FOS–JUN interaction: PDB entry 1FOS (chains E and F), X-ray diffraction structure[70]

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

Secondary structure prediction. Secondary structure elements were predicted using a two-dimensional kernel smoothing approach on the interaction score matrices (Fig. 3a–c). For a given amino acid position in the linear chain (on the diagonal of the interaction score matrix), the perpendicular dimension of the kernels define how interactions with adjacent positions (off-diagonal entries close to the diagonal) should be integrated given the interaction patterns expected from the stereotypical periodicities of secondary structures, that is, 3.6 amino acids in α-helices and 2 amino acids in β-strands. Moreover, the diagonal dimension of the kernels averages the stereotypical interaction patterns of secondary structures across several adjacent positions. Similarly, modified β-strand kernels were used to detect β-sheet interactions for all pairs of positions. Significance of secondary structure element predictions was assessed from a permutation test, for which kernel smoothing was performed on 10^4 randomly permutated interaction score maps. More details on secondary structure predictions are provided in the Supplementary Note.

Protein structure prediction. Protein structures were modeled ab initio with structural constraints derived from the DMS data using simulated annealing molecular dynamics (XPLOR-NIH modeling suite[42], see Supplementary Note for details).

DeepContact learning. DeepContact software was obtained from GitHub (https://github.com/largelymfs/deepcontact)[44]. We are grateful to Y. Liu and J. Peng for also making their basic DeepContact network architecture available on their GitHub repository and helping us with the implementation. The DeepContact architecture used here only takes one two-dimensional input of predicted contact scores and returns a two-dimensional map of transformed scores (denoted as 'DeepContact CCMPred only' in the previous study[44] and described in the first paragraph of the results section therein). The DeepContact architecture that was used came with a pre-trained network model that had been trained by comparing tertiary contact predictions from correlated evolution (using CCMpred[71]) to experimentally determined structures of proteins in the 40% homology-filtered ASTRAL SCOPe 2.06 dataset (see GitHub repository and the previous study[44]). Because CCMpred scores[71] are distributed in the range of 0 to 1, DMS-derived interaction scores were pre-normalized to a range between 0 and 1 before providing them as an input to DeepContact. As negative control, we created, for each dataset, three random permutations of combined score matrices (while preserving matrix symmetry; in the case of the FOS–JUN dataset, non-symmetric enrichment score matrices were permutated), which were transformed by the DeepContact algorithm. These control datasets show no increased precision over random expectation (Fig. 6c).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

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

Code availability

Paired-end sequencing reads were merged with USearch v.10.0.240. Data were analyzed with custom scripts written and executed in the R programming language, v.3.4.3. Structural simulations were performed with Xplor-NIH modeling suite v.2.46. TM-Score[72] (update 23 March 2016) was used to evaluate accuracy of structural models. PSIPRED v.3.3 was used to predict secondary structure elements from amino acid sequences. PyMOL v.1.8.6.0[73] was used to visualize protein structures. All custom scripts needed to repeat the analyses are available at https://github.com/lehner-lab/DMS2structure.

References

63. Rubin, A. F. et al. A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 741 (2017).
64. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
65. Barlow, R. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences* (Wiley, 1989).
66. Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4**, 32 (2005).
67. Stein, R. R., Marks, D. S. & Sander, C. Inferring pairwise interactions from biological data using maximum-entropy probability models. *PLoS Comput. Biol.* **11**, e1004182 (2015).
68. Pires, J. R. et al. Solution structures of the YAP65 WW domain and the variant L30 K in complex with the peptides GTPPPPYTVG, *N*-(*n*-octyl)-GPPPY and PLPPY and the application of peptide libraries reveal a minimal binding epitope. *J. Mol. Biol.* **314**, 1147–1156 (2001).
69. Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98**, 835–845 (1999).
70. Glover, J. N. & Harrison, S. C. Crystal structure of the heterodimeric bZIP transcription factor c-Fos–c-Jun bound to DNA. *Nature* **373**, 257–261 (1995).
71. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30**, 3128–3130 (2014).
72. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
73. The PyMOL Molecular Graphics System v.1.8 (Schrödinger LLC).

nature research

Corresponding author(s): Ben Lehner

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	no software was used to collect data
Data analysis	Paired-end sequencing reads were merged with USearch v10.0.240. 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. TM-Score (update 2016/03/23) was used to evaluate accuracy of structural models. PSIPRED v3.3 was used to predict secondary structure elements from amino acid sequence. PyMOL v1.8.6.0 was used to visualize protein structures. All custom scripts needed to repeat the analyses are available at https://github.com/lehner-lab/DMS2structure.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All published deep mutational scanning datasets of proteins in which at least 25amino acid positions were assayed with more than 2% of amino acid double mutant coverage were analysed (n = 4).
Data exclusions	Sequencing read mapping (only applies to WW domain data): Reads were discarded if they had any base call with lower than Phred quality score 20 after merging paired-end reads. Double mutants were excluded from further analyses if the expected errors of their fitness estimates were too high or if fitness was not determinable. GB1 - variants equal or less than 10 input read counts or those with less than 200 input read counts and 0 output read counts. WW - variants with equal or less than 10 input read counts or zero read counts in after the first selection round. RRM - variants with equal or less than 10 read counts. FOS-JUN - variants with equal or less than 10 input read counts. The following data exclusion rules were applied to epistasis classification in order to prevent classification being dominated by noise in fitness measurement: Positive epistasis classification: Double mutants were not considered for positive epistasis classification if the 95th percentile fitness surface at their location is equal to or above wild-type fitness; if not at least one single mutant fitness value is significantly smaller than wild-type fitness at two standard errors of fitness estimate; if the expected fitness (sum of single mutant fitness values) is not significantly lower than wild-type fitness at two standard errors. Negative epistasis classification: Double mutants were not considered for negative epistasis classification if the 5th percentile fitness surface at their location is equal to or below the 95th percentile of the background effect distribution; if at least one single mutant fitness value is not significantly higher than the lower measurement limit of the fitness assay at two standard errors; if the expected fitness is significantly higher than wild-type fitness at two standard errors.
Replication	Workflow was developed on protein G B1 domain and replicated on WW and RRM domain datasets as well as FOS-JUN interaction dataset and downsampled versions of protein G B1 domain.
Randomization	No samples were allocated into experimental groups.
Blinding	No blinding was performed as there was no group allocation.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging