ARTICLES
https://doi.org/10.1038/s41588-019-0432-9
nature genetics

Inferring protein 3D structure from deep mutation scans

Nathan J. Rollins[1,5], Kelly P. Brock[1,2,5], Frank J. Poelwijk[3,5], Michael A. Stiffler[3], Nicholas P. Gauthier[2,3], Chris Sander[2,3,4,6] and Debora S. Marks[1,4,6*]

We describe an experimental method of three-dimensional (3D) structure determination that exploits the increasing ease of high-throughput mutational scans. Inspired by the success of using natural, evolutionary sequence covariation to compute protein and RNA folds, we explored whether 'laboratory', synthetic sequence variation might also yield 3D structures. We analyzed five large-scale mutational scans and discovered that the pairs of residues with the largest positive epistasis in the experiments are sufficient to determine the 3D fold. We show that the strongest epistatic pairings from genetic screens of three proteins, a ribozyme and a protein interaction reveal 3D contacts within and between macromolecules. Using these experimental epistatic pairs, we compute ab initio folds for a GB1 domain (within 1.8 Å of the crystal structure) and a WW domain (2.1 Å). We propose strategies that reduce the number of mutants needed for contact prediction, suggesting that genomics-based techniques can efficiently predict 3D structure.

Amino acid pairs in a protein are considered epistatic when the combined effect of mutating both residues is different from what would be expected from the individual mutations if they had independent effects. Epistatic interactions have been observed between nearby residues in structure, suggesting that it may be possible to determine the three-dimensional (3D) fold of a protein from phenotype assays if direct contacts dominate the strongest epistasis. In this case, targeted genetics experiments that leverage the increasing ability to assay thousands of mutated sequences for functional effects might be sufficient to determine a protein's 3D fold (Fig. 1). Analogously, evolutionary coupling methods have used natural sequence variation to predict 3D structures, suggesting that 'laboratory', synthetic sequence variation might also yield accurate 3D structures. If genetic screens can provide enough structural information to predict the fold of a protein or RNA molecule, the increasing ease of mutant library generation and sequencing could be used to accelerate protein and RNA structure determination.

The success of computational approaches, such as evolutionary couplings, depends on large alignments of natural sequences to predict 3D structures ab initio by identifying pairs of residues likely to be in contact[1–8]. These computational methods, although powerful, are limited by the availability of large and diverse sequence families from the natural environment. Building the requisite alignments can be particularly challenging for mammalian-specific protein complexes and disordered regions. Even when considering individual protein domains, such as those in the Pfam database[9], roughly 70% of the domains of unknown structure have insufficient sequences for use in evolutionary covariation methods (unpublished data; models available at https://evcouplings.org). Extracting structural information from laboratory-created sequence variants could help solve the structure of some of these proteins.

In recent years, technological advances in sequencing have enabled high-throughput investigation of the effects of tens to hundreds of thousands of mutations in parallel (sometimes called deep mutational scanning studies)[10–38], opening the door to more systematic explorations. In these high-throughput genetic experiments, a large library of mutant sequences is synthesized, followed by selection for some phenotype of transformed cells or of the protein or RNA products (for example, ligand binding or structural stability[33]). By sequencing the library before and after selection, the fitness of each mutant can be defined according to the change in the corresponding sequence counts after selection. Therefore, high-throughput mutational scans can provide fitness measurements of thousands of sequence variants for a protein, where fitness is measured with respect to a particular phenotype.

However, being able to infer structure blindly from double mutation experiments relies critically on epistasis evidencing residues in direct contact. Studies have shown that epistasis can occur between residues that are spatially close in 3D structure[39–43], and experimentally determined epistatic pairs have even been used to discriminate incorrect decoys from correct structures generated from homology models[44,45]. Nevertheless, other studies have reported that strong epistasis between distant residues may reflect allostery or functional binding sites[38,46,47]. However, most studies have measured a low proportion of all mutant pairs, and therefore the relationship between epistasis and contacting residues has not been quantified systematically nor used to predict ab initio 3D folds.

Here, we test whether contacts can be predicted directly from epistasis data, by computing the epistasis at pairs of positions from high-throughput mutational scans on the GB1 domain of protein G in *Streptococcus* species group G[42], the WW domain of the human Yap1 (ref. [18]), the second RRM domain of *Saccharomyces cerevisiae* Pab1 (ref. [13]), the helical interaction in the Fos and Jun heterodimer[43], and the twister ribozyme of *Oryza sativa*[36]. For each study, we find that the strongest instances of positive epistasis reveal 3D contacts in the corresponding molecule. For the assays that measured pairs throughout most of the sequence—namely, the GB1 and WW proteins—we find the predicted contacts are sufficient to

[1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [2]Department of Cell Biology, Harvard Medical School, Boston, MA, USA. [3]cBio Center, Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. [4]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]These authors contributed equally: Nathan J. Rollins, Kelly P. Brock, Frank J. Poelwijk. [6]These authors jointly supervised this work: Chris Sander, Debora S. Marks. *e-mail: debbie@hms.harvard.edu

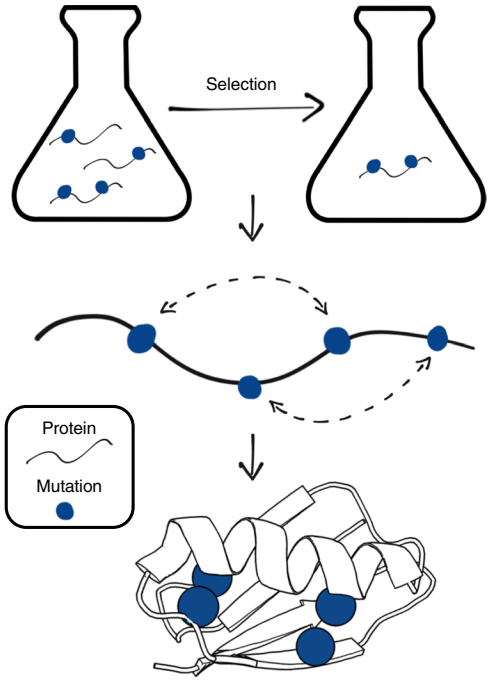

Fig. 1 | Genetic experiments can be used to discover epistatic interactions and solve the 3D fold. Mutant genes can be assayed (top) to reveal functional and structural interactions (middle). It is possible to create and test libraries sufficient enough to determine the 3D structure (bottom).

blindly determine the native 3D folds. Similarly, for Fos and Jun, the contacts predicted by epistasis are sufficient to determine the arrangement of the heterodimer complex. We also demonstrate that designed mutant libraries with fewer mutants can be used to determine 3D contacts and to fold structures to a similar accuracy as the full set of possible doubles. Together, our results indicate that high-throughput mutational scans coupled to functional assays can provide a method for determining protein and RNA structures.

Results

Epistasis reveals positions in 3D contact. To investigate whether epistasis can be used to blindly identify 3D contacts, we assembled five high-throughput mutational scan datasets that extensively measure double mutations. The scans of the GB1 domain[42], Fos–Jun dimer[43] and twister ribozyme[36] include nearly all double mutations, whereas those of the WW domain[18] and RRM domain[13] are much sparser (Supplementary Fig. 1 and Supplementary Table 1). For each dataset, we computed the epistasis of all measured double mutants where the single mutants are also measured, using the epistasis model best correlated with measured fitness (Supplementary Fig. 2 and Supplementary Table 1). A multiplicative model provided the best projection for every assay except that of Fos–Jun, which was better fit by a thermodynamic model[43].

Based on the idea that direct interactions in 3D might exhibit the strongest epistasis, we tested whether the most epistatic residue–residue pairs were proximal in known structures of each molecule. We identified the most epistatic pairs by sorting all of the pairs of positions by the corresponding double mutant with the strongest positive epistasis (Supplementary Fig. 3 and Supplementary Table 2). To evaluate 3D contact precision, we measure the fraction of the top $L/2$ and L pairs within 5 Å in the experimental structures (where L = sequence length), according to a convention in structure prediction that arose due to folded proteins having a number of contacts proportional to sequence length[2,6,8,48] (Methods and Supplementary Table 3). The precision of the top positive epistatic pairs compared with true 3D contacts is reported for all five macromolecules. Similarly, we found the pairs with the largest negative and largest-magnitude epistasis to often be proximal in 3D, but far less consistently than those with the largest positive epistasis (Supplementary Table 3).

GB1 domain. Olson et al.[42] assayed all single and almost all pairwise mutations of the 56-amino-acid GB1 domain of streptococcal protein G, including 535,917 out of 555,940 possible double amino acid mutations, for binding to human immunoglobulin G (IgG). While the experiment is very comprehensive, an experimental measurement floor interferes with the calculation of epistasis for at least 30% of double mutants. We then ranked all amino acid pairs by the maximum positive epistasis measured in corresponding double mutants, and found 68% of the top $L/2$ long-range pairs to be within 5 Å in any of the 3D structures of GB1 (refs. [49–57]). The probability of randomly drawing pairs with at least that many contacts is 1.26×10^{-13} by the hypergeometric test (Methods, Fig. 2a, Table 1 and Supplementary Table 3). As weaker epistatic pairs are included, the precision with respect to proximity drops dramatically (Supplementary Fig. 4 and Supplementary Table 3), suggesting why previous studies that are sparse or use a much lower threshold for epistasis would not have revealed a strong signal for structure[42].

The 'local' epistatic pairs (those separated by five or fewer residues in sequence) also provide useful information about secondary structure, as was seen in work on evolutionary couplings[5]. We scored residues according to the maximum positive epistasis measured at corresponding pairs expected to be close in an α helix or a β strand, and the resultant propensities largely overlap with the known secondary structure of GB1 (α helix, $P=6.84\times10^{-5}$; β strand, $P=1.03\times10^{-4}$; by t-test) (Fig. 2b, Supplementary Table 4 and Supplementary Fig. 5). Specifically, there are four peaks in β-strand propensity, roughly corresponding to the correct secondary structure, and one large peak in α propensity in the same location as the true helix (Fig. 2b). There are also two small α-helical signals (Supplementary Fig. 5) that are inconsistent with the second and third β strands, which could be noise or, more speculatively, could reflect known GB1 fold-switching[58–60].

Because the strongest positive epistatic pairs of GB1 were enriched in true residue–residue contacts, we were encouraged to infer a 3D model from the pairs (Results).

WW domain. Araya et al.[18] tested 47,000 variants of the 37-amino-acid human Yap1 WW domain for binding to a peptide ligand. Only 4% (8,797/202,521) of all possible double mutations can be tested for epistasis, and this level of sparsity may explain why the precision of the top $L/2$ long-range is much lower than for GB1 (39%; $P=1.60\times10^{-2}$). The sparsity of data also limited our ability to score secondary structure propensity (Supplementary Fig. 5). Nevertheless, many of the false positives (7/11) are still closer than 8 Å, and the predicted contacts reveal the correct overall fold topology[61–65] (Fig. 3a).

RRM domain. Melamed et al.[13] assayed 110,745 variants of the second RRM domain of Pab1 (75 amino acids). Mutations were confined to three 25-amino-acid fragments, such that double mutants occurred within an individual fragment, but not between fragments. Of the double mutations measured, 36,522 could be evaluated for epistasis (3.6% of the 1,001,775 possible double mutations across the length of RRM; 11.2% of the 324,900 possible double mutations within the three fragments mutated)[13]. Because the measurements are confined to fragments, we can only predict contacts between relatively local sequence positions (positions i and j, such that $|i-j|\leq25$) (Fig. 3b and Supplementary Fig. 1) and therefore include local pairs in the following reported precisions. The top

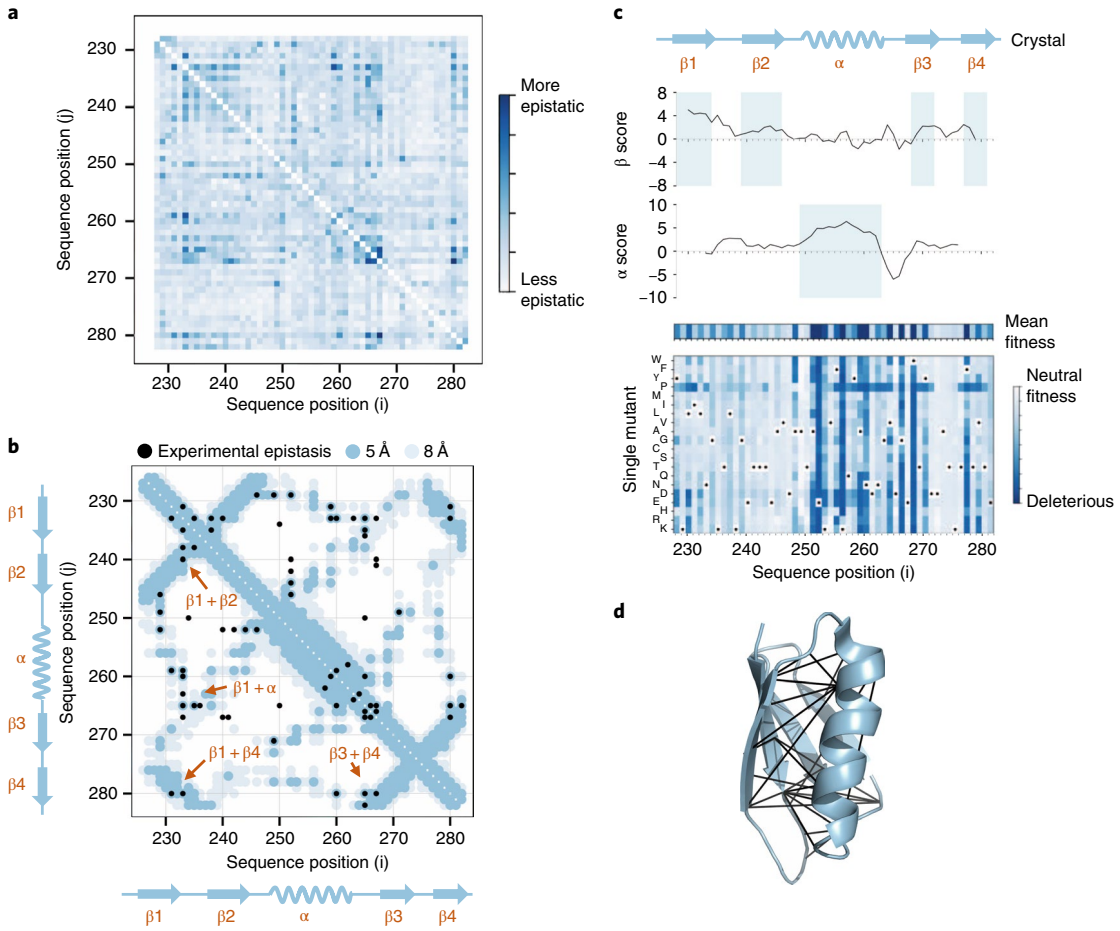

Fig. 2 | Experimental epistasis pairs reveal structural contacts in the GB1 protein. a, Maximum value of positive epistasis for each possible pair of residues in the GB1 domain, analyzed from Olson et al.[42] (Supplementary Tables 1 and 2). The most positive epistatic pairs (dark blue) suggest tertiary contacts. **b**, The 38 top positive epistatic pairs (black) include 28 ($L/2$) long-range ($|i-j|>5$) pairs and 10 local pairs (Supplementary Table 2). These pairs are used to fold the protein and to determine the topological arrangement of secondary structure elements (orange arrows). Epistatic pairs (black) are overlaid on the true contacting pairs in the NMR structure 2gb1 (ref. [48]) (minimum heavy atom distance between two residues: dark blue, 5 Å cut-off; light blue, 8 Å cut-off). **c**, Secondary structure. From top to bottom: observed secondary structure from 2gb1 (ref. [48]); β-strand scores from epistasis values; α-helical scores (Methods and Supplementary Table 4); and average-per-position and full single experimental mutation effect matrices showing concordance with local epistasis scores. **d**, Epistasis pairs (black lines) plotted on the 3D structure 2gb1.

Table 1 | Percentage of correctly predicted contacts (true positives) using various forms of epistasis

	Top 20 pairs	Top 30 pairs	Top 40 pairs	Top 50 pairs	Top 100 pairs
Positive epistasis (%)	90	80	78	68	61
Negative epistasis (%)	25	27	30	38	32
Absolute epistasis (%)	40	40	40	44	38

This table shows the percentage of predicted contacts according to residues with any heavy atom within 5 Å over multiple experimentally determined structures for GB1 (Protein Data Bank codes listed in Supplementary Table 3). Precisions are shown for the largest positive, negative or absolute measured epistasis with differing numbers of top-ranked pairs (including both long-range and local pairs).

$L/2$ (= 37) epistatic pairs have a precision of 54% <5 Å contacts ($P=7.72\times10^{-4}$)[66,67]. Although the mutation scan does not sample long-range pairs essential to determining the fold of the full protein, we do observe epistatic pairs consistent with the β hairpins in fragments 2 and 3 (Fig. 3b).

Fos–Jun heterodimer. Diss and Lehner[43] performed a high-throughput mutational scan of the 32-residue regions that heterodimerize between the bZip proteins Fos and Jun when binding DNA. These data allow us to test whether epistasis measurements can also reveal the interfaces and arrangement of protein complexes. The top $L/2$ (= 16) epistatic pairs between Fos and Jun have a contact precision of 50% <5 Å (distance <5 Å; $P=8.78\times10^{-8}$) (Supplementary Figs. 4 and 6). In general, far fewer than $L/2$ contacts are sufficient to determine the arrangement of a protein complex[3,68]. The top seven epistatic pairs are sufficient to reveal the parallel interface and helix–helix register, with five of these residue pairs within 5 Å in the experimental structure 1fos[69] (Supplementary Fig. 6).

Twister ribozyme. The twister ribozyme—a non-coding RNA molecule that self-cleaves—adopts a pseudoknot tertiary structure important for its catalytic activity[70,71]. Kobori and Yokobayashi[36] performed a high-throughput mutational scan of the *O. sativa* Osa-1-4 twister ribozyme, assaying all possible single and double mutants of the 48-nucleotide cleaved section[36]. Each variant was assayed for the fraction of copies cleaved, which we interpret as fitness, allowing

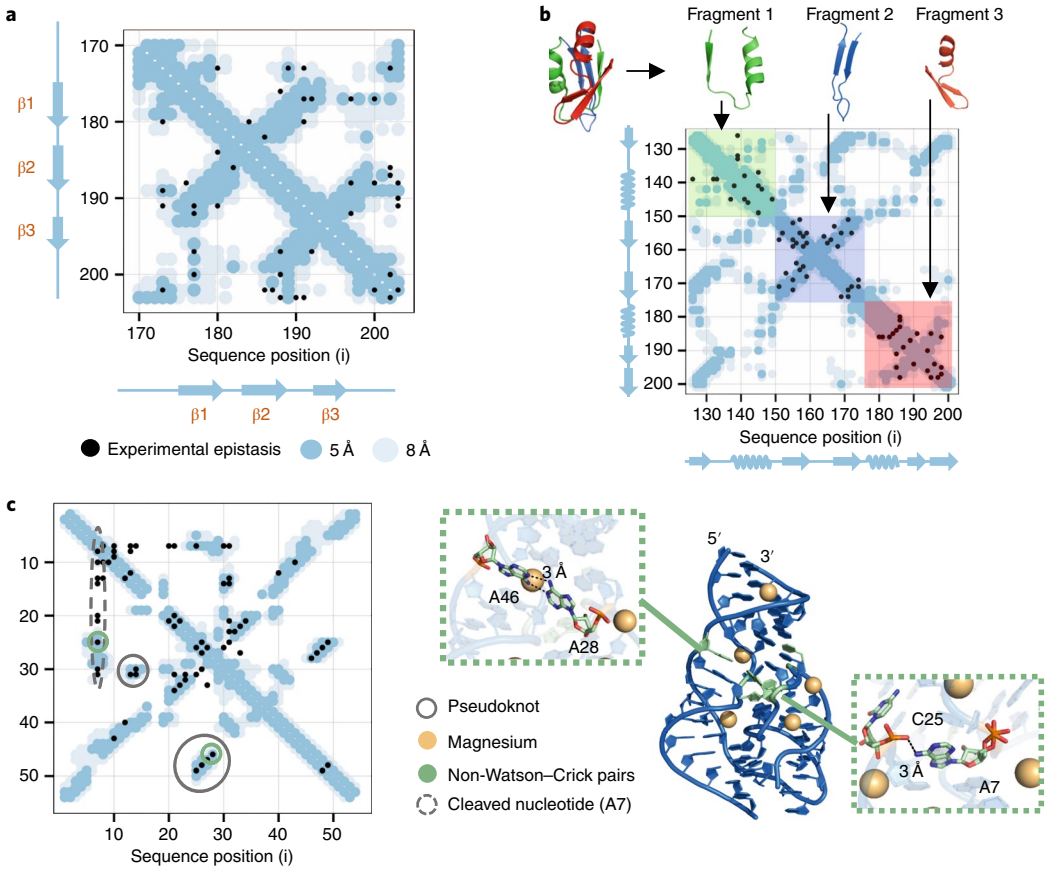

Fig. 3 | Experimental epistasis pairs reveal contacts in the WW domain, RRM domain and twister ribozyme. a, In the WW domain of human Yap1, the top 22 positive epistatic pairs (black) include 18 (*L*/2) long-range pairs, and are close in 3D, as analyzed from Araya et al.[18] (Supplementary Tables 1 and 2). **b**, Residue pairs that display strong positive epistasis in the second RRM domain in yeast Pab1, as analyzed from Melamed et al.[13] (Supplementary Tables 1 and 2). This experiment measured the effects of all pairs only within blocks of 25 residues in linear sequence (fragments 1, 2 and 3), and not between them. Therefore, experimental epistasis data exist only for the shaded square regions on the contact map. The top 38 (*L*/2) positive epistatic pairs (black) are close in the observed 3D structure 1cvj (ref. [65]). **c**, Contacts in the twister ribozyme. Left, contact map showing the 35 nucleotide pairs with the strongest positive epistasis (black), including 24 (*L*/2) long-range pairs ($|i-j|>5$), compared with true contacts from the crystal structure 4oji (ref. [70]). Strongly epistatic pairs are measured at the pseudoknot contacts (gray circles), and multiple nucleotides—both proximal and non-proximal in 4oji—share strong epistasis with the cleaved nucleotide A7 (dashed gray circle). Right, two non-standard pairs (green: A46 and A28, left insert; C25 and A7, right insert) are high-scoring epistatic pairs when compared with 4oji (RNA structure, blue; magnesium ions, yellow). In all panels, dark blue represents the 5 Å cut-off and light blue the 8 Å cut-off.

us to compute epistasis for all pairs of positions. Positive epistasis was again the most informative in identifying proximal nucleotides; 50% of the top long-range *L*/2 epistatic pairs of residues are within 5 Å ($P=2.01\times10^{-8}$), including multiple pseudoknot contacts[71,72] (Fig. 3c, left). Two of the top three most positive epistatic pairs (that is, C26–G48 and 14C–30G) correspond to the two long-range interactions that define the tertiary fold of this ribozyme forming a pseudoknot[71]. The top *L*/2 epistatic pairs also include interactions that are neither Watson–Crick nor wobble base pairings. For example, the *trans* non-Watson–Crick pairing A28–A46 is strongly epistatic and is thought to help position the active site nucleotide A7 in the structure, in addition to forming part of a pseudoknot (Fig. 3c, right)[71]. The A7–C25 pair (also in our top *L*/2 positive epistatic pairs) connects the active site nucleotide and the magnesium ion coordinating C25. Pseudoknot pairs, non-Watson–Crick pairs and metal-mediated interactions can be critical for 3D structure computation but are typically absent or poorly predicted by RNA secondary structure methods[73]. Since these high-throughput mutational scans can reveal these essential tertiary interactions, they could be an efficient method for 3D RNA structure determination.

Strong epistatic pairs not in contact are often part of functional sites. The non-contacting epistatic pairs in each molecule tended to involve residues at the binding or active sites (Supplementary Fig. 7). In GB1, all nine of the false positives in the top *L*/2 pairs are clustered at the binding surface with IgG, around residues A250 and G267. In WW, the majority (9 out of 11) of non-contacting epistatic pairs are clustered around Y188, N191 or T197, at the ligand interface. In RRM, 8 of 18 false positives are clustered around S155 and V198. Finally, in twister, 6 of the 12 false positives include the cleaved nucleotide A7. Although these epistatic pairs likely reflect functional relationships between distal residues, they can confound how we use epistasis measurements to predict folding. Assays for experimental phenotypes that more directly measure stability of the 3D fold may result in fewer false positives in predicted contacts by our method.

3D folds can be determined from epistasis. We tested whether the pairs of positions with high positive epistasis are sufficient to fold the protein ab initio (that is, from an extended polypeptide chain). By analogy to folding methods using evolutionary couplings[1,2,48], we applied constraints on up to *L* pairs of positions to generate

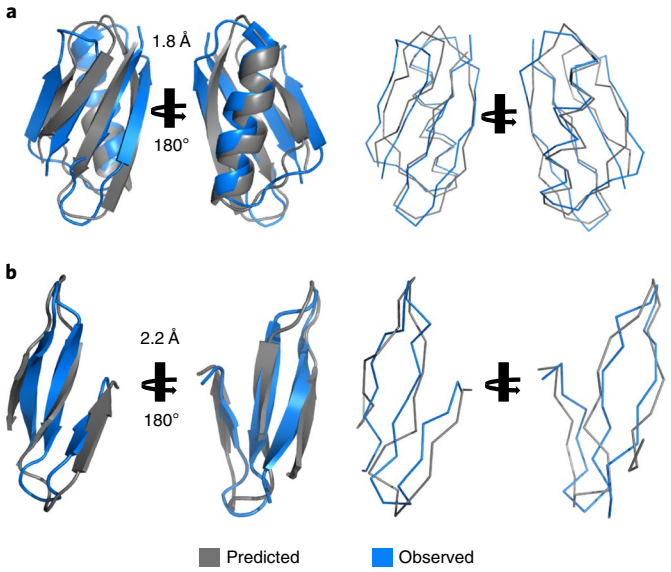

Fig. 4 | Predicted 3D structures from experimental epistasis scores alone. **a**, GB1 (gray), generated from positive epistatic pairs, compared with the NMR structure 2gb1 (ref. [48]; blue). The predicted structure is within 1.8 Å Cα RMSD of the known structure over 49/56 residues. **b**, WW domain, generated from positive epistatic pairs, compared with the NMR structure 1jmq (ref. [61]; blue). Models and structures are represented by secondary-structure cartoons (left) and backbone ribbons (right).

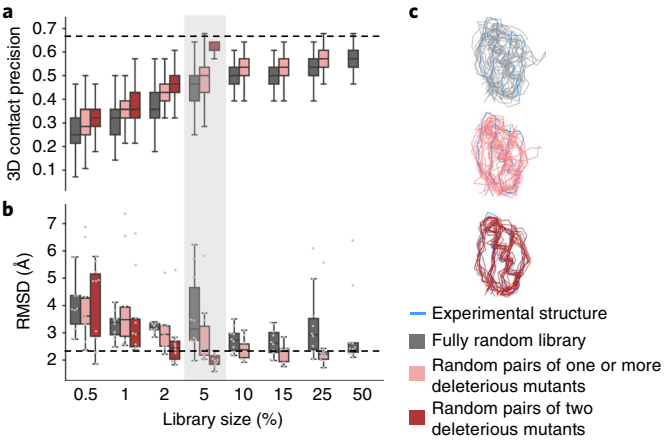

Fig. 5 | Only a small fraction of all double mutants is needed to determine the 3D fold. **a**, The precision of $L/2$ long-range epistatic pairs in contact (minimum heavy atom distance within 5 Å) is plotted for various fractional samples ($n=1{,}000$ each) of the full double mutant library, sampled according to three strategies: completely unguided mutations (gray), pairs of one or more deleterious single mutations (pink) and pairs of two deleterious mutations (red). Precision comparable to that of the full GB1 double mutant dataset (dashed line) is consistently achieved using just 50, 25 and 5% as many mutants, for each respective strategy. Central lines in all box-and-whisker plots correspond to the median, box boundaries represent the first and third quartiles, and whiskers show the range excluding suspected outliers (>quartile 3 + 1.5× interquartile range or <quartile 1-1.5× interquartile range). **b**, For each of these experimental strategies and library sizes, we folded from the epistatic pairs computed from ten different random samples and here plot the Cα RMSD of the final predictions. Notably, the third strategy consistently achieved folds more accurate than those of the full dataset (dashed line). Box-and-whisker plots are defined as above. **c**, 3D ensembles of the final folding results for each 5% subsample versus 2gb1 (ref. [48]; blue) illustrate how guided mutations can improve both the accuracy and consistency of models predicted from epistasis measured in small datasets ($n=10$).

several hundred models using the distance geometry and simulated annealing protocol in the Crystallography and NMR System package (CNS)[74]. Using a variable number of constraints allows us to test a wider variety of folds by applying different sets of distance restraints. Top models are then selected from all of the generated models by a blind ranking score (Methods).

GB1 folding. We folded GB1 from a fully extended polypeptide using the epistatic pairs as distance constraints, along with hydrogen bond constraints from predicted β-sheet topology and registrations. We ranked models blindly by how well they satisfied the input constraints (Methods and Supplementary Fig. 8). Of the 25 top-ranked candidates, the best structure is 1.8 Å Cα root-mean-square deviation (RMSD) over 49 residues to the nearest experimental structure (2.2 Å Cα to all 56 in 2gb1). Even folding without hydrogen bond constraints, the best model in the 25 top-ranked is 2.5 Cα RMSD over 49 residues (3.3 Cα to all 56 in 2gb1)[49] (Fig. 4a and Supplementary Table 5).

WW folding. We folded WW using the same procedure as GB1. Due to significant variation between experimental structures of WW (0.9–3.4 Å Cα RMSD), we restricted our comparison to the 22-residue region we found to be consistent across structures (177–198; 0.6–2.7 Å Cα RMSD) (Supplementary Table 6). The best model in the 25 top-ranked is 2.1 Å Cα RMSD over that full region in the closest structure 1jmq[62] (Fig. 4b, Supplementary Table 5 and Supplementary Fig. 8).

Fos–Jun docking. We docked idealized monomers using constraints on residues from the seven highest epistases resulting in 3D heterodimers with Cα RMSD of 0.99 Å over 58 residues (1.5 Å over 64 residues to 1fos)[69]. This result is much more accurate than a model docked without those constraints (5.4 Å over 58 residues; Supplementary Fig. 6).

In general, we found that folding with epistatic pair constraints results in more accurate structure prediction than by ab initio protocols alone; blind folding with Rosetta[75] achieves 4.0 Å Cα RMSD for GB1 and 3.8 Å for WW (Supplementary Fig. 9).

3D folds can be determined from much smaller mutant libraries. Generalizing this mutational scanning approach to large proteins would be infeasible if all possible double mutations needed to be assayed; for instance, testing a 300-residue-length protein would mean synthesizing and assaying 16 million sequences (scaling with L^2). We therefore considered whether partial libraries of fewer mutants could be used to solve 3D folds reliably. We tested three strategies of sampling just a fraction of all double mutants: (1) unguided sampling of any double mutations at random; (2) partially guided sampling of doubles including a detrimental single mutant; and (3) pairwise guided sampling of detrimental single-mutant pairs. Experimentally, these strategies can be implemented using error-prone PCR ((1) and (2)) or doped oligonucleotide synthesis ((1), (2) and (3)). We tested each strategy in silico at various library sizes by sampling subsets of the full GB1 dataset, evaluating the precision of predicted contacts and accuracy of 3D folds ($n=1{,}000$ and $n=10$ random draws, respectively) (Fig. 5 and Supplementary Table 7). For equivalent library sizes, the guided strategies had consistently higher 3D contact precision, raising both the lower bound and the median of sampling outcomes (Fig. 5a). Comparable folding accuracy to that of the full dataset (2.2 Å all-residue Cα RMSD) was achieved reliably for mutant libraries 50, 25 and 5% of the size of the full library for the three respective experimental strategies (Fig. 5b).

In summary, using guided filtering informed by single-mutation experiments reduces the search space of structurally meaningful epistatic pairs, suggesting that it may be possible to compute the structure of larger proteins with a fraction of the effort of all-pair scans.

Discussion

This work shows that the pairs of sequence positions with strongest positive epistasis are overwhelmingly close in 3D and can be systematically identified by mutation scans with sufficient coverage to determine protein folds. To generalize the use of genetic experiments for structure determination, several computational and experimental challenges must be addressed.

Computationally, we need better methods of inferring true contacts from phenotypic measurements, and of computing folds from those contacts. First, false positives could arise as the result of an insufficiently accurate model of epistasis and be reduced by models that account for nonlinear effects of independent mutations, correcting for systematic biases (Supplementary Fig. 2). Second, some true epistatic pairs may be distant in 3D structure (for example, through transitive interactions) and may be removed as predicted contacts using methods that have been applied to evolutionary couplings to deconvolve these types of indirect interactions[6,76,77]. Finally, folding biomolecules accurately from predicted contacts can be a challenge when there are false positives, and will benefit from recent advances in structure determination that iteratively discard non-satisfied constraints[78–80]. Meanwhile, folding RNA from base couplings is still a particular challenge, even with extra 3D information[81,82].

Regarding the genetic experiments, the challenges are the availability of assays and the ability to cover sufficient sequence diversity. First, mutational scans require a phenotypic assay that can be coupled one-to-one to sequences with appropriate dynamic range and functional mapping. The assays considered here make use of phenotypes specific to the studied molecule, and could be difficult to generalize to an arbitrary gene. Nevertheless, newer methods promise to address this problem (for example, by coupling green fluorescent protein to a target protein to assay for cellular abundance and thermostability[83]). Second, despite the falling costs of sequencing and synthesis, strategies for creating smaller libraries for measuring epistasis may be required to extend structure prediction to larger proteins, RNAs and complexes. We show here that simple experimental strategies can reduce the number of sequences necessary by at least an order of magnitude, and more sophisticated strategies could reduce the number even further.

In summary, these results highlight how small, laboratory-scale sequence diversity coupled to quantitative assays is sufficient to determine 3D structures of proteins and RNA, in contrast with the large amount of evolutionary sequence diversity previously used for structure prediction[2,68]. An independent effort by Schmiedel and Lehner[84] also yields high-quality 3D structures of the GB1 domain based on analysis of epistasis patterns in the Olson et al. mutation scan[42], suggesting that the results are robust to different approaches. Given that 3D structure could be determined with unguided libraries, we anticipate far broader applications with the use of designed libraries (for example, the 3D determination of large biomolecules and complexes).

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/s41588-019-0432-9.

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

References

1. Hopf, T. A. et al. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
2. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
3. Hopf, T. A. et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014).
4. Weinreb, C. et al. 3D RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963–975 (2016).
5. Toth-Petroczy, A. et al. Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170.e12 (2016).
6. 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).
7. Kosciolek, T. & Jones, D. T. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS ONE* **9**, e92197 (2014).
8. Ovchinnikov, S. et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248 (2015).
9. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
10. Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Natl Acad. Sci. USA* **112**, 7159–7164 (2015).
11. Roscoe, B. P. & Bolon, D. N. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *J. Mol. Biol.* **426**, 2854–2870 (2014).
12. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* **425**, 1363–1377 (2013).
13. 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).
14. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* **160**, 882–892 (2015).
15. 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).
16. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206 (2015).
17. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42**, e112 (2014).
18. 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).
19. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592 (2014).
20. Starita, L. M. et al. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).
21. Rockah-Shmuel, L., Toth-Petroczy, A. & Tawfik, D. S. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS Comput. Biol.* **11**, e1004421 (2015).
22. Jacquier, H. et al. Capturing the mutational landscape of the β-lactamase TEM-1. *Proc. Natl Acad. Sci. USA* **110**, 13067–13072 (2013).
23. Qi, H. et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. *PLoS Pathog.* **10**, e1004064 (2014).
24. Wu, N. C. et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS Genet.* **11**, e1005310 (2015).
25. Mishra, P., Flynn, J. M., Starr, T. N. & Bolon, D. N. Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. *Cell Rep.* **15**, 588–598 (2016).
26. Doud, M. B. & Bloom, J. D. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses* **8**, E155 (2016).
27. Deng, Z. et al. Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *J. Mol. Biol.* **424**, 150–167 (2012).
28. 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).
29. Aakre, C. D. et al. Evolving new protein–protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
30. Julien, P., Minana, B., Baeza-Centurion, P., Valcarcel, J. & Lehner, B. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* **7**, 11558 (2016).
31. Li, C., Qian, W., Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837–840 (2016).

32. Mavor, D. et al. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife* **5**, e15802 (2016).
33. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
34. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
35. Starita, L. M. et al. Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
36. Kobori, S. & Yokobayashi, Y. High-throughput mutational analysis of a twister ribozyme. *Angew. Chem. Int. Ed. Engl.* **55**, 10354–10357 (2016).
37. 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).
38. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
39. Chen, J. & Stites, W. E. Energetics of side chain packing in staphylococcal nuclease assessed by systematic double mutant cycles. *Biochemistry* **40**, 14004–14011 (2001).
40. 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).
41. Horovitz, A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Fold. Des.* **1**, R121–R126 (1996).
42. 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).
43. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, e32472 (2018).
44. Adkar, B. V. et al. Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* **20**, 371–381 (2012).
45. 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).
46. Melamed, D., Young, D. L., Miller, C. R. & Fields, S. Combining natural sequence variation with high throughput mutational data to reveal protein interaction sites. *PLoS Genet.* **11**, e1004918 (2015).
47. Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *eLife* **7**, e34300 (2018).
48. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc. Natl Acad. Sci. USA* **110**, 15674–15679 (2013).
49. Gronenborn, A. M. et al. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **253**, 657–661 (1991).
50. 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).
51. Tomlinson, J. H., Craven, C. J., Williamson, M. P. & Pandya, M. J. Dimerization of protein G B1 domain at low pH: a conformational switch caused by loss of a single hydrogen bond. *Proteins* **78**, 1652–1661 (2010).
52. Bouvignies, G., Meier, S., Grzesiek, S. & Blackledge, M. Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings from two alignment media. *Angew. Chem. Int. Ed. Engl.* **45**, 8166–8169 (2006).
53. Bouvignies, G., Markwick, P., Bruschweiler, R. & Blackledge, M. Simultaneous determination of protein backbone structure and dynamics from residual dipolar couplings. *J. Am. Chem. Soc.* **128**, 15100–15101 (2006).
54. Li, F., Grishaev, A., Ying, J. & Bax, A. Side chain conformational distributions of a small protein derived from model-free analysis of a large set of residual dipolar couplings. *J. Am. Chem. Soc.* **137**, 14798–14811 (2015).
55. Wylie, B. J. et al. Ultrahigh resolution protein structures using NMR chemical shift tensors. *Proc. Natl Acad. Sci. USA* **108**, 16974–16979 (2011).
56. Lian, L. Y., Derrick, J. P., Sutcliffe, M. J., Yang, J. C. & Roberts, G. C. Determination of the solution structures of domains II and III of protein G from *Streptococcus* by ^{1}H nuclear magnetic resonance. *J. Mol. Biol.* **228**, 1219–1234 (1992).
57. Derrick, J. P. & Wigley, D. B. The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.* **243**, 906–918 (1994).
58. Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. A minimal sequence code for switching protein structure and function. *Proc. Natl Acad. Sci. USA* **106**, 21149–21154 (2009).
59. He, Y., Chen, Y., Alexander, P., Bryan, P. N. & Orban, J. NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc. Natl Acad. Sci. USA* **105**, 14412–14417 (2008).
60. He, Y., Chen, Y., Alexander, P. A., Bryan, P. N. & Orban, J. Mutational tipping points for switching protein folds and functions. *Structure* **20**, 283–291 (2012).
61. Ferguson, N. et al. Using flexible loop mimetics to extend Φ-value analysis to secondary structure interactions. *Proc. Natl Acad. Sci. USA* **98**, 13008–13013 (2001).
62. 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).
63. Martinez-Rodriguez, S., Bacarizo, J., Luque, I. & Camara-Artigas, A. Crystal structure of the first WW domain of human YAP2 isoform. *J. Struct. Biol.* **191**, 381–387 (2015).
64. Aragon, E. et al. Structural basis for the versatile interactions of Smad7 with regulator WW domains in TGF-β pathways. *Structure* **20**, 1726–1736 (2012).
65. Aragon, E. et al. A Smad action turnover switch operated by WW domain readers of a phosphoserine code. *Genes Dev.* **25**, 1275–1288 (2011).
66. 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).
67. Safaee, N. et al. Interdomain allostery promotes assembly of the poly(A) mRNA complex with PABP and eIF4G. *Mol. Cell* **48**, 375–386 (2012).
68. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
69. 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).
70. Roth, A. et al. A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat. Chem. Biol.* **10**, 56–60 (2014).
71. Liu, Y., Wilson, T. J., McPhee, S. A. & Lilley, D. M. Crystal structure and mechanistic investigation of the twister ribozyme. *Nat. Chem. Biol.* **10**, 739–744 (2014).
72. Ren, A. et al. In-line alignment and Mg^{2+} coordination at the cleavage site of the *env*22 twister ribozyme. *Nat. Commun.* **5**, 5534 (2014).
73. Miao, Z. & Westhof, E. RNA structure: advances and assessment of 3D structure prediction. *Ann. Rev. Biophys.* **46**, 483–503 (2017).
74. Brunger, A. T. Version 1.2 of the Crystallography and NMR System. *Nat. Protoc.* **2**, 2728–2733 (2007).
75. Bradley, P., Misura, K. M. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).
76. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using psuedolikelihoods to infer Potts models. *Phys. Rev. E* **87**, 012707 (2013).
77. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnology* **30**, 1072–1080 (2012).
78. Tang, Y. et al. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat. Methods* **12**, 751–754 (2015).
79. Meiler, J. & Baker, D. Rapid protein fold determination using unassigned NMR data. *Proc. Natl Acad. Sci. USA* **100**, 15404–15409 (2003).
80. Sjodt, M. et al. Structure of the peptidoglycan polymerase RodA resolved by evolutionary coupling analysis. *Nature* **556**, 118–121 (2018).
81. Cheng, C. C. et al. Consistent global structures of complex RNA states through multidimensional chemical mapping. *eLife* **4**, e07600 (2015).
82. Das, R. et al. Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc. Natl Acad. Sci. USA* **105**, 4144–4149 (2008).
83. Matreyek, K. A. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
84. Schmiedel, J. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0431-x (2019).

Acknowledgements

The authors thank the Marks, Sander and Silver laboratories for discussion and support. The authors also thank S. Ovchinnikov for performing ab initio structure predictions with Rosetta for comparison. Partial financial support for C.S. was provided from the US NIH (RO1 GM106303). K.P.B. thanks the NIH (R01 R01GM120574) for financial support.

Author contributions

N.K.R., K.P.B. and D.S.M. performed the main analyses. N.J.R., K.P.B. and D.S.M. wrote the manuscript. F.J.P., M.A.S., N.P.G. and C.S. helped edit the manuscript. D.S.M. conceived the project. D.S.M. and C.S. 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/s41588-019-0432-9.

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

Correspondence and requests for materials should be addressed to D.S.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Calculation of epistasis from experimental data. We calculate epistasis (ϵ) using the multiplicative model, defined as the log ratio between double mutant fitness or activity values (W_{ab}) and the product of constituent single mutant fitnesses (W_a and W_b):

$$\varepsilon = \ln[W_{ab}] - (\ln[W_a] + \ln[W_b])_{\text{capped}}$$

Therefore, epistasis is defined as the signed deviation of observed fitness from fitness as projected by $(\ln[W_a] + \ln[W_b])_{\text{capped}}$. Where this projection exceeds the maximum or minimum fitness measured in an assay, we fix it to the maximum or minimum fitness value:

$$(\ln[W_a] + \ln[W_b])_{\text{capped}}$$

Additional information can be found in the Reporting Summary.

GB1 domain. Olson et al.[42] synthesized 99.97% of all double mutants (535,917/536,085) and all single mutants (1,045) in the first GB domain of protein G (GB1) by randomly combining variants of 11 five-residue cassettes, created by saturation mutagenesis. The fitness of each individual mutant was defined as the ratio of sequence reads before and after selection of mutant proteins by IgG binding, normalized by the ratio observed for the wild type. The preselection input counts of double mutants vary between 1 and 64,627. As lower input counts sensitize measurements to noise, we excluded all mutants with fewer than 20 preselection read counts from the analysis. This filtering step removes ~3% of the synthesized double mutants. Since non-specific adsorption onto IgG beads led to a fitness of approximately 0.01, all experimental or projected fitness values smaller than this were set to 0.01. This measurement floor makes negative epistasis particularly hard to measure, as >30% of double mutants may have been more deleterious than was measured in the assay.

WW domain. Araya et al. generated 47,000 mutants in the 34-residue WW domain of the hYAP65 protein by chemical assembly using a mixture of wild-type and mutant oligonucleotides[18,85]. A total of 4.4% of all possible double mutants (8,870/202,521) were synthesized that also had corresponding single mutants in the library. These proteins were presented by bacteriophage and selected by binding to a target peptide fixed to magnetic beads. Araya et al.[18] found fitness as the slope of the log ratio of counts before and after selection ('enrichment') over three rounds of selection, corrected for non-specific selection and normalized against the slope for the wild type.

RRM domain. Melamed et al.[13] created three separate mutation libraries for 25-residue regions of the second RRM domain of the essential yeast gene *Pab1*, expressed the mutants in BY4741 yeast, and selected under doxycycline until the log phase. Fitness was found as the ratio of initial sequence reads to those after selection. For this protein, we use the epistasis values calculated by the authors, who also used a multiplicative model and measured epistasis for 12.2% of possible double mutants within the pairwise sites mutated (39,608/334,900).

Fos–Jun heterodimer. Diss and Lehner[43] created all single mutations (608 each) in 32-amino-acid regions of both the Fos and Jun leucine zipper domains by overlap-extension PCR, then cloned random pairings of Fos and Jun mutants into plasmids by Gibson assembly to obtain 29% of all *trans* Fos–Jun double mutants (107,625/369,664). In their experiment, Fos and Jun were fused to separate fragments of dihydrofolate reductase, which confers yeast with resistance to methotrexate when the regions are complexed together. Yeast transformed with these mutant combinations was sequenced before and after competition under methotrexate selection. Diss and Lehner[43] computed a protein–protein interaction score as the $\log_2$ ratio of relative optical density (optical density × read fraction) after selection versus before selection, normalizing by that of the wild type. They removed background growth by subtracting the mean score of stop mutants. Diss and Lehner[43] computed epistasis by the multiplicative model, as well as a fitted thermodynamic model, which they showed to better describe the fitness of double mutants for this assay. We use the epistasis values computed for the thermodynamic fit, preferred by the authors.

Twister ribozyme. Kobori and Yokobashi[36] synthesized all double (10,296) and single (144) mutants of the Osa-1-4 ribozyme, excluding the 6-nucleotide region that is removed by self-cleavage. A pool of RNA mutants was synthesized from a mutant-doped oligonucleotide mixture in vitro, given time to self-cleave and then sequenced. The resultant sequence read records counts of cleaved and uncleaved mutants. The relative activity of a mutant was defined as the fraction of reads found cleaved for that variant, normalized by the fraction cleaved for the wild type.

Estimating secondary structure from patterns in epistasis. Epistasis between local residues was used to score the propensity of individual positions towards the α-helix or β-strand conformations. This score was developed by Toth-Petroczy et al.[5] to predict α and β propensities from local evolutionary couplings[5] corresponding to the spatial patterns in β strands ($i+1$ distant and $i+2$ proximal) and α helices ($i+1$, $i+2$ distant, $i+3$ and $i+4$ proximal):

$$\beta_{\text{score},i} = \frac{(A_{i+2} - A_{i+1})}{\text{Std}_{i+1}}$$

$$\alpha_{\text{score},i} = \frac{(A_{i+3} + A_{i+4} - A_{i+1} - A_{i+2})}{\text{Std}_{i+1}}$$

Here, A_{i+n} is the maximum positive epistasis averaged at $(i, i+n)$ and $(i, i-n)$, and normalized by the correlation between values at $i+1$ and $i+n$ determined by Toth-Petroczy et al.[5] across 3,800 protein families for evolutionary couplings (values in Supplementary Table 4; code available at https://github.com/debbiemarkslab/3D_from_DMS_Extended_Data). Scores are shown for individual positions, and when smoothed by averaging the β score across i to $i+1$ and the α score across i to $i+3$ (Supplementary Fig. 5).

Predicting β-sheet contacts from epistasis. We predicted which β-strand pairs were hydrogen bond partners according to which pairs of strands had the largest epistasis value for a residue–residue pair between the two strands. At maximum, each β strand was partnered with two other strands, and strands were only paired together if they were in each other's top two hits. To account for potential β hairpins, we assumed that strands with a linker of five or fewer residues were partners and had an antiparallel orientation. These simple rules were sufficient to predict the correct sheet topology for the GB1 and WW mutational scans.

The register between partner strands was selected as the strand alignment that places the largest epistatic pair in contact and that maximizes the number of strand–strand hydrogen bonds (or, in other words, maximizes the total overlap of the two strands). If the orientation (antiparallel versus parallel) was not identified by the length of the linker region connecting the two partnered strands, we used the highest and second-highest epistatic pair between the two strands to determine whether antiparallel or parallel strand bonding was more consistent with the two residue–residue pairs. Two possible patterns of hydrogen bonding based on this register were then separately applied to folding as distance restraints (3 ± 0.5 Å) between corresponding nitrogen and oxygen atoms in the protein backbone. Full code is provided at https://github.com/debbiemarkslab/3D_from_DMS_Extended_Data.

Folding from epistasis contacts. We generated 3D folds of the GB1 domain starting from an extended polypeptide by applying distance restraints between the top long-range (>5 amino acids apart) epistatic pairs. These constraints were input to the distance geometry and simulated annealing protocols in the CNS package as follows: (1) distance restraints (2–4 Å) between the most distal heavy atoms of side chains specified by the top epistatic pairs; (2) angle and distance restraints specified by secondary structure; and (3) when indicated, predicted β-sheet contacts as described above[1,2] (CNS input files at https://github.com/debbiemarkslab/3D_from_DMS_Extended_Data). Secondary structure specifications for GB1 were based on predictions from the PSIPRED 4.0 webserver[86,87]. The β-strand scores were ambiguous in some regions; therefore, we ran three corresponding secondary structure ranges (β1: 228–235; β2: 238/239/240–246; β3: 268–272; and β4: 276–282) and ranked all of the models together in one group. We computed ten models folded using the top-scoring 10, 11, …, ascending up to 56 (L) epistatic constraints using the previously described protocol[6], and blindly ranked these models (Methods). WW was folded according to the same method, using up to 36 (L) pairwise constraints, with the secondary structure ranges scored by PSIPRED (β1: 177–181; β2: 187–191; and β3: 196–199).

Ranking: blindly identifying the top ab initio model. We ranked each ab initio model by how well it satisfies the constraints used for folding. We calculate the equal-weighted sum of the extent to which: the top L epistatic pairs are contacting, as described in refs. [48,68]; the predicted hydrogen bond partners are contacting; and the backbone angles meet the constraints set by the predicted secondary structure according to a method described in ref. [2] (Supplementary Fig. 8; scoring code in https://github.com/debbiemarkslab/3D_from_DMS_Extended_Data):

$$\text{score} = \frac{\text{contact score} + \text{hbond score} + 2^{*}\text{score}}{3}$$

The contact score is computed according to the weighted sigmoid function described in Kamisetty et al.[46] to blindly score models based on the proximity of residue–residue pairs predicted to be in contact:

$$\text{contact score} = \sum_{n=1}^{L} w_n \times \text{sigmoid}(C\beta\ \text{dist}_n, \eta, \kappa)$$

where $C\beta$ dist_n is the C_β–C_β distance between residues in pair n. The parameters η and κ determine the activation distance and steepness of the sigmoid for a given amino acid pair, and are given by Kamisetty et al.[46]. We also used epistasis to infer β-sheet hydrogen bonds, which we scored analogously to the epistasis pairs,

but chose sigmoid parameters to describe the distance between partner residues in a β sheet:

$$\text{hbond score} = \sum_{n=1}^{L} \text{sigmoid}(C\alpha\ \text{dist}_n, 6\ \text{Å}, 2\ \text{Å})$$

Lastly, we measure how well dihedral angles within predicted α-helix and β-strand regions of each model agree with typical α-helix or β-strand angles by the method described by Marks et al.[2]. This code is part of the EVcouplings software package, available at https://github.com/debbiemarkslab/EVcouplings.

Docking Fos–Jun from epistatic contacts. We built two idealized helices (each 32 residues long) in PyMol and used these as the input monomer files to the Haddock2.2 webserver[88]. Monomer residues corresponding to Fos were numbered from 1–32, and residues for Jun were numbered from 33–64. Default settings for docking were used, besides specifying seven unambiguous restraints corresponding to the seven most positive epistatic pairs of residues. Each distance restraint was set to a distance of 2.0, with a possible range specified of 2.0 + 0 or −2.0. For the null model, we specified all residues to be active site residues but without any unambiguous distance restraints. All other settings were kept as default. For each run, we took the top-ranked model as supplied by Haddock and compared it with the crystal structure of the heterodimer 1fos (Protein Data Bank)[69].

Sampling, finding precision and folding of smaller mutation scans. To determine how precisely 3D contacts can be estimated from mutation scans of smaller libraries, we generated 1,000 independent random samples from the full GB1 dataset, measuring the precision of $L/2$ (=28) and L (=56) long-range epistatic pairs. We also tested how precisely the 3D fold can be solved from these predicted contacts, but were restricted to ten samples for each library size and strategy due to the computational cost of folding. The sampling code is provided at https://github.com/debbiemarkslab/3D_from_DMS_Extended_Data, and resulting precisions can be found in Supplementary Table 7. Folding was performed as described above in the Methods. For the guided library approaches, we define deleterious mutations as those in the lower fitness quartile (261/1,045) of single mutants. A total of 43% of the measured double mutants (229,421/536,085) include at least one deleterious mutant, and 13% (68,251/536,085) are pairs of deleterious mutants.

Ab initio folding with Rosetta. We benchmarked the precision of our folding results against that of folds determined without predicted contacts, via Rosetta ab initio folding. The Rosetta protocol works by assembling 10,000 models from short 3-mer and 9-mer fragments of experimental structures, and then scoring each according to approximate physical interactions and common bond angles observed in proteins[89]. We therefore generated 10,000 models for the GB1 and WW sequences, scored them with Rosetta and compared the structures to the native crystal structures.

Statistical tests. The hypergeometric distribution was used to compute the probability of obtaining, out of all pairs, at least the number of true contacts (<5 Å) observed in the epistatic pairs for each mutational scan, with the results reported in the text as *P* values. Enrichment of secondary structure elements was computed by one-tailed Student's *t*-test for two independent samples—positions outside versus within secondary structure regions (degrees of freedom = number of scored positions − 2; for α and β, respectively: GB1 = 45 and 49; WW = 24 and 28; and RRM = 49 and 61).

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

Data availability

The main data analyzed in this study are publicly available from the original publications (refs. [13,18,36,38,42,43]). All other data supporting the findings of this study are available within the article and its Supplementary Information files, and from the GitHub repository (https://github.com/debbiemarkslab/3D_from_DMS_Extended_Data).

Code availability

The code used in this study (along with folded models) is available at https://github.com/debbiemarkslab/3D_from_DMS_Extended_Data, and utilities for folding and ranking are available from the EVcouplings GitHub repository (https://github.com/debbiemarkslab/EVcouplings).

References

85. Fowler, D. M. et al. High-resolution mapping of protein sequence–function relationships. *Nat. Methods* **7**, 741–746 (2010).
86. Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED protein analysis workbench. *Nucleic Acids Res.* **41**, W349–W357 (2013).
87. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
88. Van Zundert, G. C. P. et al. The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
89. Bonneau, R. et al. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* **5**, 119–126 (2011).

nature research

Corresponding author(s): Debora Marks

Last updated by author(s): Apr 24, 2019

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.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒	A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No external data collection was performed in this study.
Data analysis	The code used in this study is available in the extended data, and utilities for folding and ranking are available at the EVcouplings github: <https://github.com/debbiemarkslab/EVcouplings>

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. 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

The main data analyzed in this study is publicly available from the original publications (13, 18, 36, 38, 42, 43). The authors declare that all other data supporting the findings of this study are available within the article and its supplementary information files.

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	All experimental data (and associated sample sizes) were used as published in their respective works. For our folding using subsampled sets, we used N=1000 subsamples for precision calculations (sufficient to converge to a stable distribution) and N=10 for folding samples (due to the computational cost).
Data exclusions	For epistasis calculations, we only considered mutants with pre-selection read counts >= 20 because lower input counts sensitize measurements to noise. No other data were excluded.
Replication	All experimental data come from published sources.
Randomization	All experimental data comes from published sources. For our computational subsampling results, mutations were chosen randomly and independently for the various sparsities applied to each library design strategy, as described in the text.
Blinding	The comparison of library design strategies was performed using identical code for precision and folding as for the full dataset, requiring no human interpretation and so no blinding was necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging