Supplementary Note

Bayesian estimator of variant fitness in FOS-JUN dataset

The FOS-JUN dataset has a large dynamic range (8.6 log-units), thus many low-fitness variants with low input read coverage have very low or no output read counts (per replicate ~1/3 of variants have below 3 output counts, ~15% of variants have zero output counts), effectively reducing the dynamic range of the assay for low input variants and distorting the estimate of the overall fitness distribution (see Supplementary Figure 4d). To overcome this, a Bayesian estimator of fitness was implemented. First, for each double mutant variant, the 1000 nearest neighbors in single mutant fitness space (i.e. those double mutants whose respective single mutant fitness values are similar to the single mutant fitness values of the variant under consideration) with sufficient input coverage (more than 100 reads in the input library) were identified. From this set of 1000 nearest neighbors the expected distribution of double mutant fitness values was calculated, which served as a prior distribution. For the variant under consideration the likelihood distribution of fitness values given its input and output read counts was calculated under Poissonian assumptions. Fitness was then estimated as the mean of the distribution resulting from the multiplication of prior and likelihood distributions. Error of fitness estimate was estimated as the standard deviation of the resulting distribution. Estimated fitness from the three replicate experiments were merged by weighted averaging.

Restricted epistasis classification close to the measurement range of fitness assays

The subset of variants deemed suitable for positive epistasis classification is limited to regions where

- the 95th percentile fitness surface is below wild-type fitness
- at least one single mutant fitness value is significantly smaller than wild-type fitness at two standard errors
- the expected fitness (sum of both single mutant fitness values) is not significantly higher than wild-type at two standard errors

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

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

- the 5th percentile fitness surface is above the 95th percentile of the background effect distribution; this value is derived from the 95th percentile of double mutant fitness distribution of lethal variants (GB1 - doubles with expected log-fitness

below -8; RRM and WW domains and FOS-JUN interaction - variants with STOP codons)
- both single mutant fitness values are significantly higher than the lower limit of the fitness assay measurement range at two standard errors
- the expected fitness (sum of both single mutant fitness values) is not significantly higher than wild-type at two standard errors

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

Uncertainty estimates from re-sampling procedure

A re-sampling approach was used to estimate the uncertainty in interaction score estimates (Supplementary Figure 8, step 5, described here for positive epistatic variants, but equivalent for negative epistatic variants). In each of 10.000 re-sampling runs:

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

Secondary structure prediction via two-dimensional kernels

The alpha kernel takes on a sinusoidal profile perpendicular to the diagonal to weight interactions according to whether the position pair considered should have congruent side-chain orientations (see diagonal and perpendicular profiles in Figure 3b). The kernel was defined as $k_\alpha(d,p) = \left(\cos\left(p * \frac{2\pi}{3.6}\right) + 1/3\right) * e^{-\frac{d^2}{c^2}}$, with $d = |2x - i - j|$ as the diagonal distance of the interaction *ij* (off the diagonal) to the reference position *x* (on the diagonal) and $p = |i - j|$ as the perpendicular distance of the interaction off the diagonal. The kernel weight for positions with p > 5 was set to 0, thus only including interactions across little more than the first helical turn. Finally, $c = 4$ is the integration scale for the Gaussian kernel along the diagonal. While smaller integration scales do yield nosier results and longer integration scales can lead to non-detection of shorter secondary

structure stretches, we found that in practice our whole approach (including the actual detection algorithm described below) is robust to alterations of the integration length.

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

The beta kernel takes an alternating profile perpendicular to the diagonal to weight interactions according to alternating side-chain orientations in a beta strand and was defined as $k_\beta(d,p) = \left((p+1)\ mod\ 2 - {}^1/_3\right) * e^{-\frac{d^2}{c^2}}$, with $c = 4$. Only interactions with perpendicular distances equal or smaller than two (i.e. $k_\beta(d, p > 3) = 0$) were included.

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

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

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

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

3. select the most significant seed
4. test whether extension to any side yields lower combined p-value
   4.1. if yes: extend seed in this direction and repeat step 4
   4.2. else: repeat step 4 once to see whether further extension in same direction yields lower combined p-value
      4.2.1. if yes: extend and repeat step 4
      4.2.2. else: proceed to step 5

5. fix as secondary structure element and delete all ‘used’ p-values (and combined seed p-values), such that other elements cannot incorporate them
6. check whether other already fixed elements are adjacent or at most one position away
   6.1. if yes: merge both elements and update p-value
7. repeat steps 3-6 until no more seeds with combined p-value < 0.05 are left

This yields a list of predicted locations of secondary structure elements. We note that the secondary structure elements predicted from deep mutational scanning data could be compared to and combined with predictions derived from other tools, such as PSIPRED[1], to further improve reliability.

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

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

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

The steps were modified as follows:

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

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

8. compute association of seeds with known beta strands (e.g. seed positions overlap strand 1 on one side and coincide with strand 3 on the other side)
9. while there are seeds with p < 0.05: pick most significant seed from either the parallel or anti-parallel sheet subset
10. check consistency with secondary structure elements
    10.1. discard the seed and jump back to step 9 if

10.1.1. it is overlapping or too close to an alpha helix or the linker region between two beta strands that interact (minimal distance smaller one)
10.1.2. at least one of the two strands it is associated with already has two other beta sheet interactions or the total number of beta sheet interactions exceeds 2*(#beta strands – 1)
10.2. modify secondary structure elements and start anew from step 3 if
10.2.1. one side of the seed is not associated to a known beta strand: create this beta strand
10.2.2. if both sides of the seed are associated with the same known strand: split the strand and create a linker region in-between the strands
10.3. else fix the beta sheet interaction and delete all other interactions that are associated with the same strands and haven't been fixed yet, jump back to step 9
11. if no more seeds with p<0.001, finish
12. update beta strands: keep only those positions that are part of a beta sheet interaction

For beta sheet pairing detection in the GB1 domain (as reported in Figure 3d and Supplementary Figures 3b-d and 9) secondary structure element predictions derived from the deep mutational scanning data were used as input (as shown in Figure 3a-c and Supplementary Figure 3a). For the RRM and WW domains, PSIPRED[1] (v3.3) predicted secondary structure elements were used as input, due to the insufficient signal from secondary structure element predictions from deep mutational scanning data.

Protein structure prediction

Protein structures were modeled *ab initio* with structural restraints derived from the deep mutational scanning data using simulated annealing molecular dynamics (XPLOR-NIH modeling suite[2]).

Distance restraints from top L predicted contacts (position pairs with highest interaction scores and linear chain separation greater than 5 positions, L - mutated length of protein) were implemented as NOE (nuclear Overhausser effect) potential by setting Cβ-Cβ atom distances (Cα in case of Glycine) between positions to range between 0 and 8Å and weighting the restraints according to their relative interaction score (interaction score divided by mean interaction score of all predicted contacts used). NOE potential type was set to "soft" for stages 1 and 2 and "hard" for the final simulation stage. Using fewer or more predicted tertiary contacts to derive restraints yielded similar, albeit mostly slightly worse structural models (Supplementary Figure 3f).

Restrains from secondary structures elements were implemented as dihedral angle restraints (CDIH potential). Dihedral angles of both beta strands and alpha helices were set to range between values commonly observed in crystal structures[3], for alpha helices $\Phi_\alpha = -63.5° \pm 4.5°$ and $\Psi_\alpha = -41.5° \pm 5°$ and for beta strands $\Phi_\beta = -118° \pm 10.7°$ and $\Psi_\beta = 134° \pm 8.6°$.

Restraints for beta sheet interactions were implemented by setting H-N:O=C hydrogen bond distances between interacting positions to range between 1.8 and 2.1Å (Ref. 3), with weight one. Predictions of beta sheet interactions derived from deep mutational scanning data yield a string of interacting positions, but hydrogen bonding in beta sheets

occurs in specific non-continuous patterns between position pairs (between alternating positions off the interaction diagonal in parallel beta sheets and between every second set of position pairs in anti-parallel beta sheets). Specifically, for each set of interacting positions there are two alternative patterns of hydrogen bonding possible. These alternative possibilities of pairing were implemented as mutually exclusive selection pairs with the "assign … or" syntax in Xplor-NIH.

Simulations were performed in three stages, in each of which 500 structural models were generated. Stages 1 and 2 served to identify inconsistencies among defined structural restraints, and simulations in both stages were started from an extended chain configuration. Stage 3 served to refine a final set of best models, here simulations were started from the average structure of the best 10% of models obtained at the end of stage 2.

After simulation stages 1 and 2, restraints were checked for their consistency with predicted structural models and were down-weighted if they were violated in too many of the best structural models. First, structural models were clustered based on whether they fulfill or violate similar sets of the given distance and dihedral angle restraints (k-means clustering, k = 4). Then the top cluster was identified by the mean total XPLOR energy (from all energy potentials used) of its 50 models with lowest total energy. These 50 models with the lowest total energy from the top-ranked cluster was used to identify which restraints are commonly violated. For the subsequent simulation stage, distance restraints were down-weighted according to the fraction of top structural models that violated them, $w_{x,i} = w_{x,i-1} * (1 - f_x)^2$, and distance restraints with a weight below 0.1 were discarded. There is no option to weight dihedral angle restraints, thus instead dihedral angle restraints that accumulated down-weighting to below 1/3 were discarded for the subsequent simulation stages.

The top 5% structural models, as judged by total XPLOR energy, from simulation stage 3 were evaluated against the reference structure; results are robust to the cut-off used (Supplementary Figure 3g). The TM-score program (update 2016/03/23) was used to calculate the Cα root mean squared deviation and the template modeling score[4].

Several types of control simulations were performed to judge the predictive power of restraints derived from deep mutational scanning data. As a negative control we performed simulations without restraints from predicted contacts and beta sheet interactions, but with restraints from secondary structure elements predicted by PSIPRED[1] (version 3.3). As a positive control we performed simulations with restraints derived from the reference structure. Here, L true contacts of position pairs with linear chain distance greater than 5 amino acids were randomly sampled and beta sheet hydrogen bonding were determined using PyMOL[5]. These simulations serve as a positive control and give the maximally achievable accuracy of our Xplor-NIH workflow.

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

For structural simulations of down-sampled GB1 datasets (and DeepContact transformed versions thereof) we used distance restraints derived from top predicted contacts and secondary structure restraints derived from PSIPRED predictions, but no restraints for beta sheet pairing. This was done to avoid skewed results due to the fact that especially beta sheet pairing predictions are often false in low quality datasets (Supplementary Figure 9). For structural simulations from DeepContact-transformed predictions, we found that using more tertiary contacts resulted in better models. We conclude that this is because the deep learning algorithm focuses many strong predictions in few structural features (such as interactions of secondary structure elements), which are therefore the top contacts. Restraints in other regions of the protein are therefore only included if more predicted contacts are used for restraint calculations, therefore improving structural predictions. Because of this, when comparing structural simulations from scores derived before and after deep learning, we compare the top 5% of structural models derived with the top L predicted contacts from original scores with those derived with the top 1.5*L predicted contacts from DeepContact transformed scores.

References

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

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

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

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

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