nature biotechnology

Mutation effects predicted from sequence co-variation

Thomas A Hopf[1–3,6], John B Ingraham[1,6], Frank J Poelwijk[4], Charlotta P I Schärfe[1,5], Michael Springer[1], Chris Sander[2,4] & Debora S Marks[1]

Many high-throughput experimental technologies have been developed to assess the effects of large numbers of mutations (variation) on phenotypes. However, designing functional assays for these methods is challenging, and systematic testing of all combinations is impossible, so robust methods to predict the effects of genetic variation are needed. Most prediction methods exploit evolutionary sequence conservation but do not consider the interdependencies of residues or bases. We present EVmutation, an unsupervised statistical method for predicting the effects of mutations that explicitly captures residue dependencies between positions. We validate EVmutation by comparing its predictions with outcomes of high-throughput mutagenesis experiments and measurements of human disease mutations and show that it outperforms methods that do not account for epistasis. EVmutation can be used to assess the quantitative effects of mutations in genes of any organism. We provide pre-computed predictions for ~7,000 human proteins at http://evmutation.org/.

Understanding the phenotypic effects of genetic variation is a central challenge for bioengineering and basic biology. Molecular technologists strive to engineer biologics that are safe and effective[1], design new genomes[2,3], and develop 'smart' libraries of synthesized molecules[4,5]. Biologists seek to identify the genetic changes that underlie organism phenotypes or complex diseases[6] and identify them in the ever-expanding catalog of variation in humans and model organisms. New technologies have emerged that respond to these needs and simultaneously assess the effects of thousands of mutations in parallel[4,7–27] (sometimes referred to as 'deep mutational scans'[28] or 'MAVEs'[29]). In these assays, the measured attributes or processes vary, ranging from ligand binding, splicing, and catalysis[7,11,14,16,24,26,30] to cellular or organismal fitness under selection pressure[8–10,12,15,17,20,22,23].

However, although high-throughput mutational scans have improved our understanding of the consequences of genetic variation, their relevance to organism fitness and physiology critically depends on the choice of the measured phenotype[29,31]. It is reasonable to assume that the relationship between a biochemical phenotype and organism fitness may be nonlinear[32] and even non-monotonic[31]. For the foreseeable future, scans are also limited in their scalability, which requires researchers to focus on a modest number of either positions or alleles. Nature has not been subject to these limitations, however, and proteins diverging by hundreds of positions can often functionally replace one another[33].

Statistical models of natural sequence variation can complement experimental approaches. These models take advantage of the fact that evolution has been performing its own massively parallel mutagenesis and selection experiments over time. Most computational methods, including SIFT, PolyPhen-2, and CADD, exploit evolutionary conservation to predict the effects of mutations[34–36]. However, these methods do not explicitly consider genetic interactions between mutations and the sequence background, despite widespread evidence for epistasis, which is the non-independence of the effects of mutations[37,38].

Recent work has applied unsupervised statistical models based on the co-evolution of natural sequences to accurately predict three-dimensional (3D) contacts in protein and RNA structures[39–45], suggesting that dependencies between residues can be systematically captured from natural sequences. Although related models have been applied to predict the effect of mutations on a few proteins[46–48], these methods have not been tested systematically across many proteins or against human disease mutations.

We present a method, EVmutation, which accounts for epistasis by explicitly modeling interactions between all the pairs of residues in proteins (and bases in RNAs), and which then quantifies the effects of mutations, including multiple mutations, simultaneously. We compare predictions generated by our models with measurements from 34 mutagenesis experiments and in classification of human disease mutations showing that the modeled mutational landscapes are in better agreement with experimental data than existing non-epistatic methods.

RESULTS

Probabilistic model captures residue dependencies

Extant proteins and RNAs show strong signatures of selective pressures that have acted throughout their evolution. It is not uncommon for genes separated by hundreds of millions of years, having negligible sequence identity, to nevertheless exhibit remarkable conservation of their structures and functions. Here, we aimed to build statistical models that can recover some of the dominant constraints that define families of homologous sequences.

[1]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Informatics, Technische Universität München, Garching, Germany. [4]cBio Center, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [5]Applied Bioinformatics, Department of Computer Science, University of Tübingen, Tübingen, Germany. [6]These authors contributed equally to this work. Correspondence should be addressed to D.S.M. (debbie@hms.harvard.edu).

Received 10 June 2016; accepted 9 December 2016; published online 16 January 2017; corrected online 30 January 2017 (details online); doi:10.1038/nbt.3769

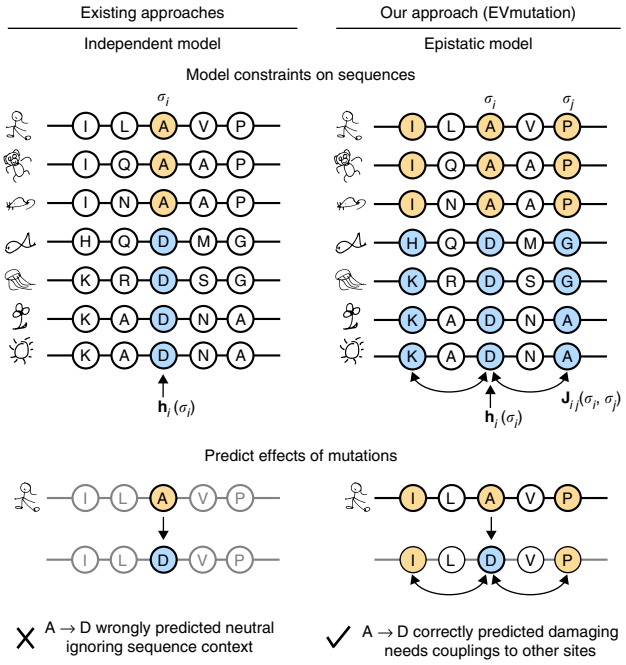

Figure 1 Inferring context-dependent effects of mutations from sequences. Evolution has generated diverse families of proteins and RNAs with varied sequences that perform a common function. An unsupervised probabilistic model trained to generate the natural diversity in a multiple sequence alignment of a family can be used to predict the relative favorability of unseen mutations. Existing models describe functional constraints on each position i in a sequence σ independently, averaging over the effect of background positions j. This can lead to incorrect predictions of neutrality. Our approach infers a global probability model with pairwise interactions between positions i and j ($\mathbf{J}_{ij}$) as well as background biases at single positions ($\mathbf{h}_i$). For a more detailed graphical schematic of the calculation, see **Supplementary Figure 1**.

We model the evolutionary process that has produced each family as a sequence generator at equilibrium that produces a sequence σ with probability $P(\sigma)$ as

$$P(\sigma) = \frac{1}{Z}\exp\{E(\sigma)\} \quad (1)$$

Different parametric forms of the 'energy' function $E(\sigma)$ enable the model to capture different types of constraints on the sequences (epistatic or not) and Z normalizes the distribution to sum to one over all possible sequences of a fixed length. $E(\sigma)$ may be thought of as a (negative) energy of a model from statistical physics or as proportional to the scaled fitness N_eF in toy equilibrium models of population genetics[49]. We use an energy function $E(\sigma)$ with two types of constraints: pairwise constraints that describe co-dependencies in combinations of amino acids or nucleotides for each pair of sites, and site-specific constraints reflecting bias toward or away from specific amino acids or nucleotides at each position. The total energy for a specific sequence $E(\sigma)$ is the sum of coupling terms $\mathbf{J}_{ij}$ between every pair of residues and a sum of site-wise bias terms $\mathbf{h}_i$ (fields),

$$E(\sigma) = \sum_i \mathbf{h}_i(\sigma_i) + \sum_{i<j} \mathbf{J}_{ij}(\sigma_i,\sigma_j) \quad (2)$$

Combining the sequence generator model (equation (1)) with our energy function (equation (2)) produces a model that is known as a pairwise undirected graphical model in computer science and a Potts model in statistical physics. When fit to data, these models explain the global correlations observed between variables in a system in terms of direct pairwise interactions $\mathbf{J}_{ij}$ that are typically simpler and more localized. Determining these interactions involves simultaneous accounting for all possible couplings between all pairs of positions, which is not possible with local measures of correlation such as 'mutual information'[50]. When models of this form are fit to natural sequence families, the magnitudes of the $\mathbf{J}_{ij}$ terms have consistently predicted contacts in the 3D structures of proteins, with sufficient accuracy to predict the folds of proteins[41,43,51,52].

We applied these models to make sequence-specific predictions about the relative favorability of mutations. Starting from a multiple alignment of a sequence family, we estimated the site and coupling parameters $\mathbf{h}$ and $\mathbf{J}$ using regularized maximum pseudo-likelihood[40,53–56]. After the parameters were inferred, we quantified the effects of single or higher-order substitutions on a particular sequence background with the log-odds ratio of sequence probabilities between the wild-type and mutant sequences (**Fig. 1** and **Supplementary Fig. 1**):

$$\Delta E(\sigma^{\text{mut}},\sigma^{\text{wt}}) = \log\frac{P(\sigma^{\text{mut}})}{P(\sigma^{\text{wt}})} = E(\sigma^{\text{mut}}) - E(\sigma^{\text{wt}}) \quad (3)$$

The summation over coupling terms $\mathbf{J}_{ij}$ between all pairs of positions in the evolutionary statistical energy E directly incorporates sequence context, that is, the effects of pairwise epistasis, into the calculation of mutation effects. We refer to this model, which is the basis of EVmutation, as the 'epistatic model'.

Model captures experimental fitness landscapes

We assessed the extent to which the statistical energy landscapes computed using EVmutation corresponded with experimentally measured changes of phenotypes. We collected data from both saturation mutagenesis experiments of genes encoding proteins and RNA and focused low-throughput studies[15,46,57–60], resulting in 34 data sets from 29 non-redundant experiments (21 proteins, and a tRNA gene[27]; **Supplementary Table 1**) that include all currently available mutation experiments where the protein or RNA had a sufficiently large and diverse alignment (Online Methods). For each protein or RNA molecule tested, we generated a multiple sequence alignment of the sequence family, inferred the parameters $\mathbf{h}$ and $\mathbf{J}$ of the epistatic model, and compared the change in statistical energy (ΔE) to the corresponding experimental measurements of the effects of the mutations (**Fig. 2** and **Supplementary Table 2**). Since relationships between protein function and organismal fitness are not expected to be linear[31], we focused on reporting rank correlations as the primary metric for evaluating predictive performance, but our results are robust to a variety of measures (**Supplementary Table 3**).

We found significant correlations between the computed ΔE of the epistatic model and the experimental measurements of phenotype and fitness for all high-throughput experimental data sets (Spearman's ρ 0.4–0.7; P-values from $<10^{-300}$ to $<10^{-27}$ and $\rho = 0.2$, $P < 10^{-12}$ for one of the BRCA1 experiments not expected to correlate well (see below)). We found that the agreement between ΔE and experimental data is higher if the assayed phenotype is closely linked to an essential process, and that agreement depends on the strength of purifying selection applied in the experiment (**Fig. 3**, **Supplementary Table 3**, and **Supplementary Fig. 2**). For instance, some of the strongest correlations between ΔE and experimental data were for the enzymatic activity of a methyltransferase that protects DNA from degradation[17],

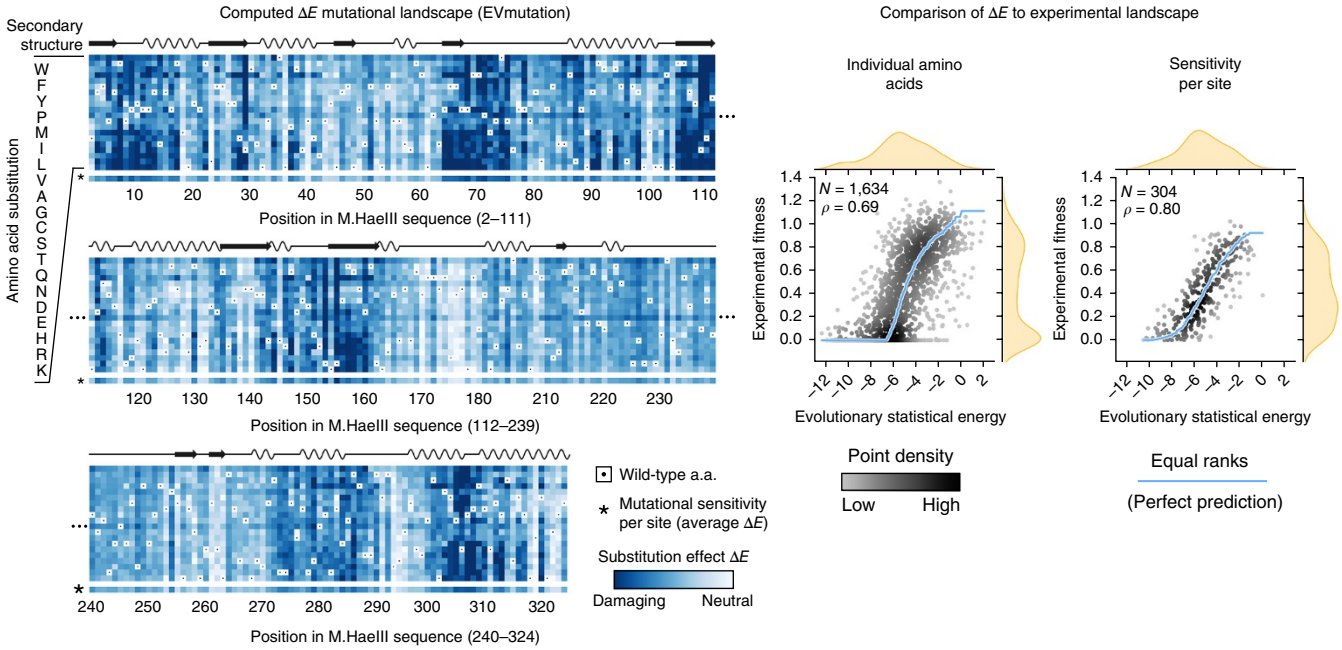

Figure 2 Saturation mutagenesis experiments provide a quantitative test of context-dependent predictions. The computed ΔE mutational landscape of the DNA methyltransferase M.HaeIII (left, color range from 5th percentile to 0) agrees quantitatively with experimental measurements of M.HaeIII fitness under selection by restriction enzyme cleavage (right, $\rho = 0.69$, $N = 1{,}634$; marginal distributions in orange). The average mutational sensitivity per position shows improved correlation beyond individual effects ($\rho = 0.80$, $N = 304$).

and of a β-glucosidase that hydrolyzes biomass[4]. Likewise, ΔE was more weakly correlated with data from experiments where the laboratory selection pressures may not match those in the natural environment, and where the distribution of experimental mutation effects tends to be skewed toward mostly neutral effects (e.g., BRCA1 binding to BARD1) or toward mostly very deleterious effects (a bacterial kinase subject to high doses of kanamycin; **Supplementary Table 4** and **Supplementary Fig. 3**). The dependence on the strength of selection is nicely exemplified by the increasing correlation of our predictions with the experiments on β-lactamase and kanamycin kinase as the antibiotic doses are titrated to reveal the full dynamic range of effects (**Supplementary Figs. 4** and **5**).

The epistatic model also successfully captured the effects of mutations in low-throughput experiments on thermostability of trypsin[58] ($\rho = 0.77$, $N = 23$) and SH3 (ref. 57) ($\rho = 0.69$, $N = 48$) and the catalytic efficiency of β-lactamase[15] ($\rho = 0.87$, $N = 30$), including double and triple substitutions. Although these experiments consisted of a biased choice of mutants that may benefit our correlation, in principle, the results suggest EVmutation could be used to design stabilizing or catalytically stronger proteins.

To assess whether ΔE distinguishes human variants associated with diseases from putatively neutral variants, we compared predictions for 9,008 variants (1,553 proteins) annotated as pathogenic in ClinVar[61], to variants in the 60,000 human exome sequencing collection (ExAC[6]) at increasing allele frequencies (AF; 2,190 proteins). The ΔE scores for the disease variants were significantly different from those common alleles, presumed mostly to be neutral, from the ExAC collection (area under the ROC curve, AUC = 0.92–0.96, P-values < 10^{-300}, two-sided sample Kolmogorov–Smirnov tests; **Fig. 3b**).

Comparison of epistatic model with published methods

We sought to evaluate the utility of adding epistatic interactions to mutation prediction by comparing EVmutation with commonly applied tools on both the assembled experimental data as well as human genetic variants. For the experimental data, we compared predictions from EVmutation with those of SIFT[34] and PolyPhen-2 (ref. 62), along with the substitution matrix BLOSUM62. The rank correlations of ΔE (epistatic model) were higher than those of the other methods for the majority of the analyzed mutation experiments (**Fig. 3c** and **Supplementary Table 5**). The average increase in ρ was 0.30 over the BLOSUM62 matrix (EVmutation better on 30/31 data sets), 0.18 over SIFT (better on 29/31 data sets), and 0.18 over PolyPhen-2 (better on 26/29 data sets). While this work was in review, a few studies used an epistatic model to predict effects of mutations in β-lactamase and three other proteins[48], but used different statistical inference and alignment generation protocols. EVmutation predictions were mildly more accurate when compared directly across the four proteins tested (average increase in correlation: 0.04) and on a par with a computationally expensive approach for inference for a small number of HIV protein mutations[46] (EVmutation $\rho = 0.81$ vs. $\rho = 0.80$; **Supplementary Table 6**).

We compared predictions from the unsupervised EVmutation method with PolyPhen-2, a supervised mutation predictor, on human disease variants. Despite no training on clinical variants, the performance of ΔE predictions was comparable with those of PolyPhen-2 and SIFT on the HumVar benchmark set[62–64] (ΔE: AUC = 0.89; PolyPhen-2: AUC = 0.88; SIFT: AUC = 0.85) and better than PolyPhen-2 with the subset of variants that are predicted differently by PolyPhen-2 and SIFT (AUC = 0.71 versus AUC = 0.61; **Supplementary Fig. 6**). It has been reported that prediction methods, such as PolyPhen-2, that use supervised machine learning, can have inflated measures of accuracy due to multiple routes of circularity in the training and test sets[16,64]. Therefore, it is surprising that our unsupervised model performed as well, or better than, supervised methods that have been directly trained to segregate deleterious human variants.

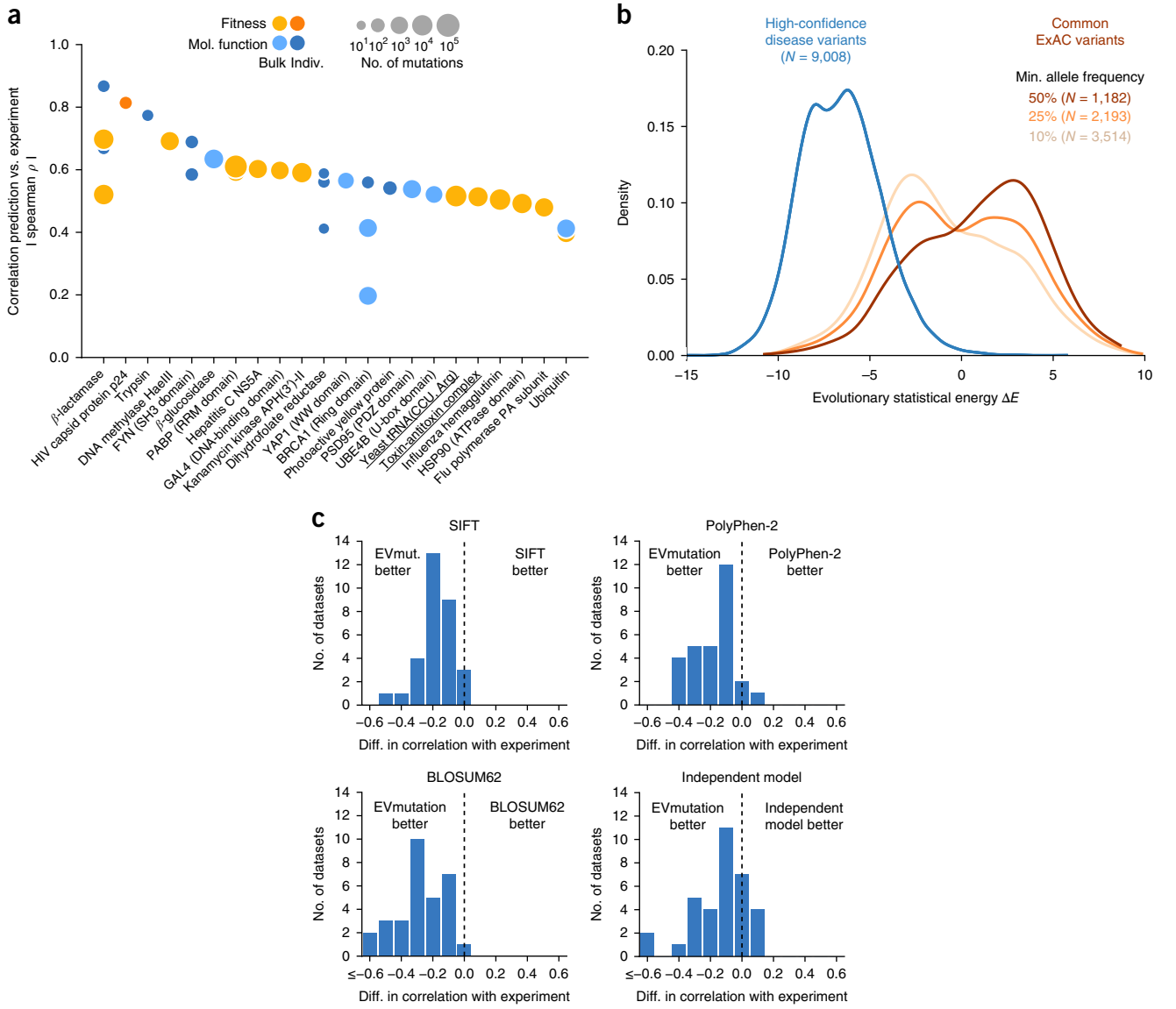

Figure 3 ΔE captures experimental fitness landscapes and identifies deleterious human variants. (**a**) Computed effects of specific mutations (difference in evolutionary statistical energy ΔE) based on the epistatic model agree with diverse experimental measurements of fitness and molecular function for 34 experiments for 20 proteins, a protein complex, and an RNA molecule (underlined), as measured by Spearman's rank correlation coefficient ρ (for equivalent site average plot, see **Supplementary Fig. 2**; for correlations across all different assays tested in the experiments, see **Supplementary Fig. 4**). (**b**) Evolutionary statistical energies ΔE distinguish human disease-associated variants from common alleles in the population. This separation increases with the minimum allele frequency (AF) of the variants assumed to be neutral (area under the ROC curve (AUC) = 0.92 for AF ≥ 0.1, AUC = 0.94 for AF ≥ 0.25, AUC = 0.96 for AF ≥ 0.5). (**c**) The epistatic model shows stronger agreement with experiments than do the established methods SIFT and PolyPhen-2, a baseline model based on the BLOSUM62 substitution matrix, and a corresponding independent model without pairwise interactions (differences of ρ of more than 0.6 were included in the bin at 0.6).

Interaction terms underlie improved predictions

In order to understand whether the improved predictions of our epistatic model resulted from the added pairwise interactions **J**, we used a control, non-epistatic, 'independent' model that is identical to ΔE but lacks interactions between sites. We fit this model to the same alignments as the epistatic model. The independent model is closely related to the log-frequency-based position-weight matrices, and other scores of conservation[16,34,36,65] that are commonly used by existing methods. The ability of ΔE (epistatic model) to capture mutational effects was better (difference in $\rho \geq 0.05$) than the independent model across 23 of 34 of the analyzed experiments (**Fig. 3c** and **Supplementary Table 5**; average increase of ρ = 0.14, including high-throughput fitness experiments with well-defined selective pressures such as β-lactamase[10,15,18,48], kanamycin kinase[13], and DNA methyltransferase M.HaeIII)[17], and was also more accurate for human disease mutations (AUC = 0.92 versus 0.88 for AF ≥ 0.1), despite the ascertainment bias of these mutations[64].

One particularly striking contrast between the epistatic and independent models is for the toxin-antitoxin complex ParED[25], especially when the two proteins are modeled jointly as a protein interaction[40]. In this case the epistatic model captured the experimental effects with much higher correlation (ρ = 0.51 for singles, doubles, and triples, N = 3335 and ρ = 0.43, N = 9194 for all mutations) than the independent model, which hardly correlated at all (ρ = −0.05 for all up to triplets).

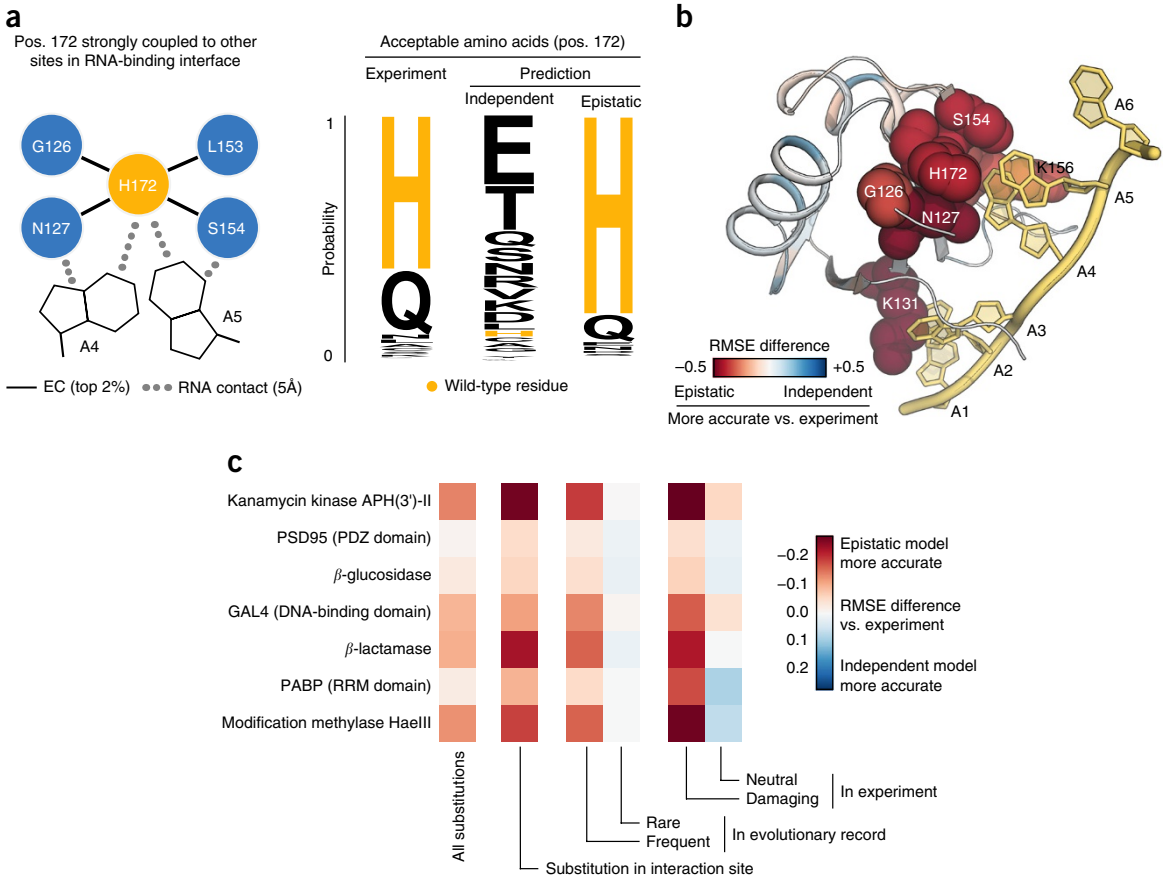

Figure 4 Improvements of the epistatic model for functional sites. (**a**) Left. The RNA-binding residue H172 of PolyA-binding protein (PABP) is strongly coupled to other residues in the binding interface that are close in 3D. Right. The epistatic coupling leads to strong constraints on acceptable amino acids in position 172, as observed in an experimental mutation scan of PABP. Only the epistatic model correctly identifies these co-constraints, while a model without sequence context (independent model) suggests many more substitutions would be acceptable (range of experimental preferences scaled to range of predicted preferences based on full set of mutants for entire domain). (**b**) Positions in PABP for which prediction accuracy improves the most by considering epistasis (≥2σ difference in root mean squared prediction error, spheres) cluster around the RNA ligand (yellow sticks, PDB: 4f02). (**c**) For seven high-throughput data sets where the correlation ρ of the epistatic and independent models differs more than 0.05, the epistatic model is more accurate overall (1st column), specifically for the effects of mutations of residues in interaction and ligand-binding sites (2nd column), where the residue mutation is frequent rather than rare in the evolutionary sequence alignment (3rd and 4th columns), and where the residue change is damaging rather than neutral in the experiment (5th and 6th columns) (**Supplementary Table 8**).

Interestingly, if ΔE was computed on couplings from ParD alone, the match to experiment was lower ($\rho = 0.33$), and if computed only on couplings between (but not within) the complex subunits, the predictions were comparable to the full model ($\rho = 0.53$ vs. $\rho = 0.51$). Hence, the interactions learned in the epistatic model support accurate prediction of the interaction specificity of the complex[25].

Experiments where the epistatic model performed comparably to the independent model (11/34 sets) tended to be where the correlations were below average for either method (**Supplementary Table 5**). This included three of the four viral proteins and might have been a consequence of the limited diversity of the sequence alignments or, alternatively, a discrepancy between the proxy for viral fitness in the laboratory and the *in vivo* fitness of the virus[22].

We next asked how the results for the epistatic and independent models depend on the evolutionary depth of the alignments used for inference. As sequence alignments became narrower around the mutated protein, that is, evolutionarily distant sequences were progressively excluded, the epistatic model and independent model made similar predictions that both performed less well against experimental data. For specific alignment depths of some proteins, the independent model can capture the experimental data as accurately as the epistatic model (**Supplementary Table 7**). However, since we do not know which alignment depth to choose a priori, the ability of the epistatic model to capture constraints without prior knowledge of the optimal alignment depth may be an advantage. The choice of alignment in our method is blind to the mutation experiments and is based on the algorithm used for alignment choice when computing 3D structure contacts using EVfold[41].

Where is the epistatic model better?

To investigate where inclusion of epistatic interactions leads to improved predictions, we compared the epistatic and independent models on a mutation-by-mutation basis. Direct comparison of residuals can be misleading when the functional relationship between predictions and experimental measurements is unknown and/or nonlinear, so we computed deviations between the epistatic and independent models after mapping through a quantile-quantile transformation. This allowed us to identify specific mutations that

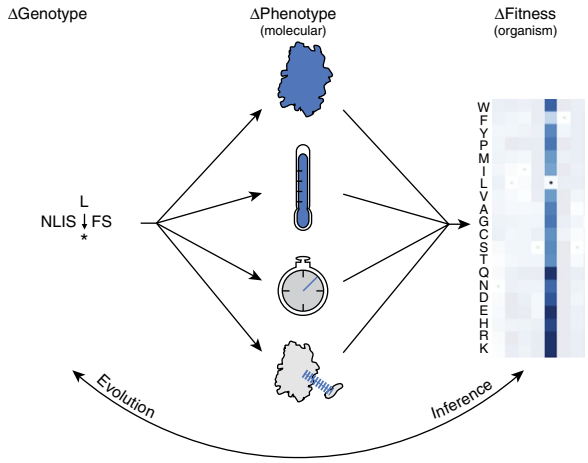

Figure 5 Computational predictions complement experimental measurements. Various molecular phenotypes (center) such as structure, thermostability, activity, and ligand-binding affinity are determined by genotype and contribute to fitness in a complicated manner that is not known a priori. However, the distribution of contemporary genotypes (left) provides a record of historical fitness values (right), which can roughly be inferred by computational methods. For instance, upon substitution of L with any other residue (L→*), what is the effect on molecular phenotype or organism fitness? Identifying those phenotypes that connect to inferred fitness may shed light on which molecular phenotypes have historically been the most relevant to the organism.

had the largest reduction of prediction error when using the epistatic model. For instance, in an RNA-recognition motif of the poly(A)-binding protein, the three mutations with the largest reduction in error using the epistatic model were H172T, N127R, and S154D. The epistatic model agreed with experimental data showing the deleterious effects of these mutations, but the independent model suggested that the same mutations were neutral or even beneficial (**Fig. 4a** and **Supplementary Table 8**). Although these substitutions were frequently observed in other sequences in the protein family alignment, only the epistatic model captured that they were not acceptable in the background of the target sequence. These three residues form a network of spatially proximal residues that contact RNA (**Fig. 4a**), and are highly coupled when their total coupling is summarized across all amino acid combinations (**Supplementary Table 9**). The predictions that differed most between the epistatic and independent models (defined as >2σ of error change distribution) are in eight positions that are within 6 Å of the bound RNA, which includes the positions identified above (**Fig. 4b**).

We then explored whether these observations generalize to other proteins by applying the above analysis for all proteins with single substitution high-throughput scans in which the epistatic and independent models differed ($\rho(\Delta E(epistatic)) - \rho(\Delta E(independent))) > 0.05$) but still showed reasonable correlation to the experimental data ($\rho > 0.50$; **Fig. 4c** and **Supplementary Fig. 7**). Comparison of predictions made by the two models suggests that the advantages of the epistatic model stem at least in part from more accurate modeling of positions that facilitate interactions with ligands or other proteins, and show strong deleterious effects in experiments (**Fig. 4c** and **Supplementary Fig. 7**).

DISCUSSION

We report here that a prediction method built on natural sequence variation can partially capture the experimental effects of mutations in a variety of biological contexts. This can be applied to predict or interpret the effects of genetic variation in any species of interest and for higher-order mutations that change multiple positions at the same time.

Limitations of the model include biases that arise from evolutionarily younger families of limited diversity, non-uniform selective constraints across a sequence family, and higher-order epistasis. Although incorporating epistatic terms into the model results in a practical improvement over other methods, there are remaining challenges regarding the interpretation and inference of the model parameters. For interpretation, it is difficult to distinguish those couplings in the model that may be due to subfamily-specific selection from those that may be due to more universal epistatic constraints across the whole family[66], and future methods may begin to address these issues by consideration of phylogeny. For inference of the parameters, the statistical challenge remains that the typical number of free parameters ($\sim 10^6$–10^8) vastly exceeds the number of available sequences ($\sim 10^3$–10^5), and since evolutionary sampling is highly correlated and consequently highly redundant, this may lead to considerable challenges in extrapolating a reasonable space of possible sequences.

The success of this and other models based on sequence variation at recapitulating high-throughput mutation experiments depends in part on the extent to which experimental assays capture phenotypes that are under direct, long-term selection (**Fig. 5**). For example, thermostability, activity, or binding energetics of a protein will generally not all contribute to fitness in the same way, so even a perfect model and perfect measurements might have imperfect correlations. The excellent correlation between model and experimental data for β-lactamase plausibly reflects how survival 'in the wild' in the presence of β-lactam antibiotics depends directly on that specific protein. For other assays, such as nonessential peripheral enzymes or signaling proteins, the property being tested in the laboratory may have only an indirect, context-dependent impact on the organism. So the evaluation of agreement between model and experiment is two-way, reliant on whether the model is effective as a method and, when multiple kinds of measurements are available for the same protein, to what extent the lab assay captures evolutionarily conserved properties.

Nevertheless, our probabilistic approach is readily applicable for analysis of specific variants and combinations thereof for numerous families of proteins and RNAs and the interactions between them. We anticipate that analyses of genetic variation and mechanisms of evolution will benefit from global probability models of sequence families that explicitly incorporate interactions between positions, such as the model presented here. The consistency of our predictions with prior biological knowledge on functional sites highlights how inclusion of epistatic interactions will facilitate more accurate assessment of mutations and aid the design of libraries of protein sequences. To enable the community to analyze ΔE for their proteins of interest, we have made software (**Supplementary Code**) and pre-computed mutation effects for ~7,000 human proteins available at http://evmutation.org/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors would like to thank A. Lapedes, B. Rost, and members of the Marks laboratory for scientific discussion, and J. Reeb for help with existing

mutation prediction software. C.S. was funded by NIGMS (R01GM106303). D.S.M. and T.A.H. were funded by NIGMS (R01GM106303) and the Raymond and Beverley Sackler Foundation. J.B.I. was funded by an NSF Graduate Research Fellowship (DGE1144152).

AUTHOR CONTRIBUTIONS

D.S.M., T.A.H. and C.S. initiated the project. T.A.H. and J.B.I. developed algorithms and wrote software. T.A.H., J.B.I. and D.S.M. analyzed the data with contributions from M.S. F.J.P. advised on the interpretation of experiments. C.P.I.S. supplied processed human genetic variation data. T.A.H., J.B.I., C.S. and D.S.M. wrote the paper. D.S.M. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Miersch, S. & Sidhu, S.S. Intracellular targeting with engineered proteins. *F1000Res.* **5** http://dx.doi.org/10.12688/f1000research.8915.1 (2016).
2. Boeke, J.D., *et al.* GENOME ENGINEERING. The Genome Project-Write. *Science* **353**, 126–127 (2016).
3. Ostrov, N. *et al.* Design, synthesis, and testing toward a 57-codon genome. *Science* **353**, 819–822 (2016).
4. 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).
5. Currin, A., Swainston, N., Day, P.J. & Kell, D.B. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.* **44**, 1172–1239 (2015).
6. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
7. 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).
8. 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).
9. 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).
10. 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).
11. 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).
12. Kitzman, J.O., Starita, L.M., Lo, R.S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206, 4, 206 (2015).
13. 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).
14. 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).
15. 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).
16. Starita, L.M. *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).
17. Rockah-Shmuel, L., Tóth-Petróczy, Á. & 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).
18. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl. Acad. Sci. USA* **110**, 13067–13072 (2013).
19. 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).
20. Wu, N.C. *et al.* Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS Genet.* **11**, e1005310 (2015).
21. 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).
22. Doud, M.B. & Bloom, J.D. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *bioRxiv* **8**, E155 (2016).
23. 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).
24. 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).
25. Aakre, C.D. *et al.* Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
26. Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J. & Lehner, B. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* **7**, 11558 (2016).
27. Li, C., Qian, W., Maclean, C.J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837–840 (2016).
28. Fowler, D.M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
29. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
30. Sarkisyan, K.S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
31. Boucher, J.I., Bolon, D.N. & Tawfik, D.S. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Sci.* **25**, 1219–1226 (2016).
32. Gong, L.I., Suchard, M.A. & Bloom, J.D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2**, e00631 (2013).
33. Kachroo, A.H. *et al.* Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348**, 921–925 (2015).
34. Sim, N.L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452 (2012).
35. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
36. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20 (2013).
37. Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C. & Kondrashov, F.A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).
38. McCandlish, D.M., Shah, P. & Plotkin, J.B. Epistasis and the dynamics of reversion in molecular evolution. *Genetics* **203**, 1335–1351 (2016).
39. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
40. Hopf, T.A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3** http://dx.doi.org/10.7554/eLife.03430 (2014).
41. Hopf, T.A., *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
42. Marks, D.S., Hopf, T.A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
43. Marks, D.S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
44. 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).
45. Jones, D.T., Buchan, D.W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
46. Mann, J.K. *et al.* The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput. Biol.* **10**, e1003776 (2014).
47. Lapedes, A., Giraud, B. & Jarzynski, C. Using sequence alignments to predict protein structure and stability with high accuracy. Preprint at https://arxiv.org/pdf/1207.2484v1.pdf (2012).
48. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* **33**, 268–280 (2016).
49. Sella, G. & Hirsh, A.E. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. USA* **102**, 9541–9546 (2005).
50. Giraud, B.G., Heumann, J.M. & Lapedes, A.S. Superadditive correlation. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **59**, 4983–4991 (1999).
51. Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248 (2015).
52. Kosciolek, T. & Jones, D.T. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS One* **9**, e92197 (2014).
53. Besag, J. Statistical analysis of non-lattice data. *Statistician* **24**, 179–195 (1975).
54. Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.I. & Langmead, C.J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
55. 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).
56. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87**, 012707 (2013).
57. Di Nardo, A.A., Larson, S.M. & Davidson, A.R. The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J. Mol. Biol.* **333**, 641–655 (2003).
58. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).

59. Philip, A.F., Kumauchi, M. & Hoff, W.D. Robustness and evolvability in the functional anatomy of a PER-ARNT-SIM (PAS) domain. *Proc. Natl. Acad. Sci. USA* **107**, 17986–17991 (2010).
60. Bershtein, S., Mu, W. & Shakhnovich, E.I. Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *Proc. Natl. Acad. Sci. USA* **109**, 4857–4862 (2012).
61. Landrum, M.J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D1, D862–D868 (2016).
62. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
63. Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729–2734 (2006).
64. Grimm, D.G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
65. Bromberg, Y., Yachdav, G. & Rost, B. SNAP predicts effect of mutations on protein function. *Bioinformatics* **24**, 2397–2398 (2008).
66. van Nimwegen, E. Inferring contacting residues within and between proteins: what do the probabilities mean? *PLoS Comput. Biol.* **12**, e1004726 (2016).

ONLINE METHODS

Generation of multiple sequence alignments. For each analyzed protein (target sequence), multiple sequence alignments of the corresponding protein family were obtained by the default five search iterations of the profile HMM homology search tool jackhmmer[67] against the UniRef100 database of non-redundant protein sequences[68] (release 11/2015). To control for comparable evolutionary depth across different families, we used length-normalized bit scores to threshold sequence similarity rather than E-values[40]. A default bit score of 0.5 bits/residue was used as a threshold for inclusion unless the alignment yielded <80% coverage of the length of the target domain or if there were not enough sequences (redundancy-reduced number of sequences ≥10L); in the first case, the threshold was increased in steps of 0.05 bits/residue until sufficient coverage was obtained; in the second case, the threshold was decreased until there were sufficient sequences (≥10L). If these two objectives were conflicting, precedence was given to maintaining more than 10L sequences. Since the sequence diversity of viral protein families is typically much lower than that of bacterial and eukaryotic families, the alignment depth for viral proteins was chosen as a default of 0.5 bits/residue even if the redundancy-reduced number of sequences was lower than 10L. The alignments were post-processed to exclude positions with more than 30% gaps and to exclude sequence fragments that align to less than 50% of the length of the target sequence. For the ParE-ParD toxin-antitoxin interaction, a joint sequence alignment with matched homologs of both interaction partners was generated using our previously described approach EVcomplex[40]. Alignments for RNA sequence families were obtained from the Rfam database[69] and redundancy-reduced at the same 80% identity cutoff as proteins. The tRNA alignment was filtered to contain only sequences with a CCU anticodon.

Inference of epistatic models of biological sequences. *Model.* Each family is modeled as a distribution over the space of all possible sequences that is parameterized by two types of constraints: site-specific constraints on the biases for specific amino acids or nucleotides at each position, and pairwise constraints for combinations of amino acids or nucleotides for each pair of sites. The use of such a global probabilistic approach is motivated by the idea that the induced correlations observed in a multiple sequence alignment can be explained by a simpler set of underlying couplings between positions. Local measures of coupling, such as the log ratios that define 'mutual information', cannot deconvolve these transitive correlations between positions[50].

The form of the chosen distribution can be thought of as the least-structured (i.e., maximum entropy) distribution over sequence space that is consistent with the single-site and pairwise marginal distributions of amino acids or nucleotides observed in the alignment[43,50,55,70]. Under the model, the probability of a sequence σ of length N is defined as

$$P(\sigma) = \frac{1}{Z}\exp\{E(\sigma)\}$$

The partition function Z normalizes the distribution by summing over the relative weights (Boltzmann factors) of all possible sequences σ'. The strength of the Boltzmann factors for each sequence σ is defined by the evolutionary statistical energy $E(\sigma)$ as the sum of all its pairwise coupling constraints $\mathbf{J}_{ij}$ and single-site constraints $\mathbf{h}_i$ (fields), where i and j are positions along the sequence:

$$E(\sigma) = \sum_i \mathbf{h}_i(\sigma_i) + \sum_{i<j} \mathbf{J}_{ij}(\sigma_i,\sigma_j)$$

In a maximum likelihood fit without regularization, the site-specific parameters $\mathbf{h}_i(\sigma_i)$ and pair-specific parameters $\mathbf{J}_{ij}(\sigma_i,\sigma_j)$ implement the constraints that the marginal probability distributions of the model agree with the empirical marginal frequencies $f_i(\sigma_i)$ and $f_{ij}(\sigma_i, \sigma_j)$ in the sequence data. This model is also as a Markov Random field or a Potts model in statistical physics.

Insertions and deletions. While the model readily describes fixed-length sequences such as strings of 20 amino acids or 4 nucleotides, it is unclear how to meaningfully represent the insertions and deletions in a way that reasonably reflects the generative process occurring in evolution. Traditionally, statistical approaches have tended to model indels either as (i) an extra character or (ii) missing data. Both of these approaches are conceptually problematic, since the former partitions a single deletion event into many separate 'gap characters' while the latter forces the model to impute a hidden variable in gapped positions that can affect other variables in the system despite knowledge that no such coding variable exists. We used an alternative approach for modeling indels that does not invoke a gap character and does not give missing data the ability to influence the system. We regard the indel process as a separate, observed process, and instead model the conditional distribution of the amino acids given an indel pattern. If z_i are Bernoulli indicators that are 1 when a site is coding and 0 when gapped, the conditional energy function for coding regions is:

$$E(\sigma;\mathbf{z}) = \sum_i z_i\mathbf{h}_i(\sigma_i) + \sum_{i<j} z_i z_j \mathbf{J}_{ij}(\sigma_i,\sigma_j)$$

Given an observed set of gaps, the marginal distribution of the amino acids at the ungapped positions defines a conditional model for a sequence. In principle, this approach could be made fully generative by combining it with a model for gaps, but for the purposes of this study we only fit the conditional distribution.

Sample reweighting. Natural sequences descend from a common ancestry and consequently may be highly correlated. Moreover, uneven sampling of the family by both sequencing projects and natural processes may result in overabundance of sequences from some parts of phylogeny (e.g., model organisms) and an underabundance of others. To partially account for this redundancy, we use the established approach of reweighting sequences by a measure of their uniqueness[43,44]. Briefly, we define the sequence weight π_s of a given sequence S as

$$\pi_s = \left(\sum_t I[D_H(\sigma^s,\sigma^t) < \theta]\right)^{-1}$$

where $D_H(\sigma^s, \sigma^t)$ is the normalized Hamming distance (i.e., 1.0 – percentage identity) between sequence s and all sequences t in the alignment and θ is a threshold for the percent of maximum divergence of sequences classified as similar. We used an 80% identity cutoff ($\theta = 0.2$) for all proteins except viral proteins, for which we used a 99% identity cutoff ($\theta = 0.01$).

Inference. Given observed sequence data (with potential reweighting for redundancy), the model parameters $\mathbf{h}_i$ and $\mathbf{J}_{ij}$ could, in principle, be estimated by maximum likelihood, i.e., by finding the parameters that maximize the probability of observing the sequence data. This approach, however, is deterministically intractable due to the 4^N (RNA) or 20^N (protein) sequences in sequence space that must be computed for the partition function Z. As a replacement for the full likelihood, we used a site-factored pseudolikelihood approximation[53].

Regularization. The number of parameters for the pairwise model, which for typical protein families of 50–500 amino acids will range from 10^5 to 10^7, outnumbers the typical number of sequences available (10^2 to 10^5) for even the largest families by several orders of magnitude. In this undersampled regime, standard maximum likelihood estimation is highly prone to overfit the sample data. We penalized model complexity with l_2-regularization[55,56], which may also be interpreted as maximum a posteriori (MAP) inference under zero-mean Gaussian priors on the parameters. The strength of l_2-regularization was set as $\lambda_h = 0.01$ for the single-site constraints $\mathbf{h}_i$ and $\lambda_J = 0.01{\cdot}q(N-1)$ for the pairwise coupling constraints $\mathbf{J}_{ij}$, where N is the number of sites in the model and q corresponds to the number of possible states ($q = 20$ for proteins, $q = 4$ for RNA).

Combining the pseudolikelihood approximation, sequence reweighting, treatment of indels, and regularization, we estimate the parameters by maximizing the objective

$$\hat{\mathbf{h}}\ \left\{\hat{\mathbf{h}},\hat{\mathbf{J}}\right\} = \arg\max_{\mathbf{h},\mathbf{J}} \sum_{s,i} \pi_s \log P_i^s(\sigma) - \lambda_{\mathbf{h}} \sum_{i,a} \mathbf{h}_i(a)^2 - \frac{\lambda_{\mathbf{J}}}{2} \sum_{i,j,a,b} \mathbf{J}_{ij}(a,b)^2$$

We use σ to denote the collection of all sequence data, σ_i^s to denote the letter of sequence s at position i, and z_i^s to denote the gap state at this location. The conditional likelihood $P_i^s(\sigma)$ of sequence s at position i is defined whenever that location is ungapped ($z_i^s = 1$) as

$$P_i^s(\sigma) = \exp\left(\mathbf{h}_i(\sigma_i^s) + \sum_{j\neq i} z_j^s \mathbf{J}_{ij}(\sigma_i^s,\sigma_j^s)\right) / \sum_a \exp\left(\mathbf{h}_i(a) + \sum_{j\neq i} z_j^s \mathbf{J}_{ij}(a,\sigma_j^s)\right)$$

doi:10.1038/nbt.3769

and is 1 otherwise. We solve this optimization problem with a quasi-Newton method (L-BFGS), and make our C software implementing this available at github.com/debbiemarkslab/plmc.

Calculation of context-dependent mutation effects (evolutionary statistical energy). Using the inferred probability models, one can then quantify the effect of single or higher-order substitution on a particular sequence background by computing the log-odds ratio of probabilities between the mutant sequence σ^{mut} and the wild-type sequence σ^{wt}. This ratio of probabilities is simply the difference of the evolutionary statistical energies of the two sequences, which is

$$\Delta E(\sigma^{\text{mut}},\sigma^{\text{wt}}) = \sum_i \left(\mathbf{h}_i(\sigma_i^{\text{mut}}) - \mathbf{h}_i(\sigma_i^{\text{wt}})\right) + \sum_{i<j}\left(\mathbf{J}_{ij}(\sigma_i^{\text{mut}},\sigma_j^{\text{mut}}) - \mathbf{J}_{ij}(\sigma_i^{\text{wt}},\sigma_j^{\text{wt}})\right)$$

For a given mutant, the evolutionary statistical energy difference ΔE will be the sum of differences of the single-site constraints $\mathbf{h}_i$ for all substituted sites, plus the sum of differences of the coupling parameters for all pairs of positions involving at least one mutated site. By evaluating the change of couplings to other sites, the sequence context and therefore epistatic effects are explicitly incorporated into the computed mutation effects. Values of ΔE above 0 correspond to more probable mutant sequences (putatively beneficial), values below 0 to less probable mutant sequences (putatively deleterious) and values equal to 0 to equally probable sequences (putatively neutral).

Throughout the manuscript, mutation effects computed using this model are referred to as statistical energy differences from the epistatic model.

Calculation of evolutionary couplings. Between any two positions i and j in the protein family, the coupling matrix $\mathbf{J}_{ij}$ describes the co-constraint on all possible 20^2 amino acid or 4^2 nucleotide combinations. To quantify the total epistatic constraint between pairs of sites across all sequences in the alignment, the Frobenius norm was used to summarize each matrix $\mathbf{J}_{ij}$ into a single number that is proportional to the s.d. of the (zero-mean) matrix[56]. After computing the norm scores for every pair of sites, we remove background coupling that is presumed to be caused by limited sampling and phylogenetic relationships between sequences by applying the average product correction (APC)[71]. This is equivalent to setting the dominant singular value or eigenvalue of the coupling matrix to zero. Significantly constrained pairs (evolutionary couplings, ECs) were then selected by a mixture-model-based strategy quantifying the probability of any pair to belong to the high-scoring tail of the score distribution rather than the background noise distribution[40,72].

Inference of independent statistical models. To assess the contribution of epistatic interactions to ΔE, additional maximum entropy models were inferred that describe protein sequences using only site-specific amino acid constraints $\mathbf{h}_i$, without considering explicit interdependencies between sites (i.e., predict mutation effects independently of the sequence context). Since the parameters of each independent model were inferred from the same sequence alignment as the epistatic model, it serves as a direct control for the contribution of epistatic interactions. The probability of any amino acid sequence σ under this "independent" model is given by

$$P(\sigma) = \frac{1}{Z}\exp\left\{\sum_i \mathbf{h}_i(\sigma_i)\right\}$$

Consistent with the regularization applied to the epistatic model, the strength of the l_2 penalty was set to $\lambda_h = 0.01$ when estimating the model parameters $\mathbf{h}_i$. The statistical energy difference between two sequences can be analogously inferred by calculating the statistical energy difference between the mutant and the wild-type sequences, which again corresponds to the logs-odds ratio of their probabilities. This formalism is closely related to the log-odds conservation scores used in many methods to predict mutation effects from sequence[36,73].

Mutational landscape data sets. Mutation effect data sets were identified by a comprehensive literature search for quantitative high-throughput mutagenesis experiments of entire proteins, protein domains, or RNA molecules. All experiments that targeted proteins or RNAs with insufficient sequence diversity (redundancy-reduced number of sequences $<10L$, where L = length of protein or domain), covered only small subregions, or tested synthetic wild-type proteins were excluded from the final compilation of data sets (**Supplementary Table 1**). For further comparisons, the data set was extended with low-throughput measurements of molecular phenotypes (stability, catalytic activity, binding), with a focus on sequence co-evolution studies. If a data set reported more than one measurement for the wild-type sequence (as was the case for 11 of the high-throughput scans with one wild-type value per position in the sequence), we averaged these redundant experimental measurements into a single value (**Supplementary Table 2**).

Classification of experimental mutation effects. 18 of the experimental mutation scans had visible bimodality in their effect distribution.

Since this can lead to biased Pearson and Spearman correlations, we tested these with a binary classification measure, the Matthews correlation coefficient, in addition to the other analyses. We classified mutations as damaging or neutral by (i) fitting a two-component Gaussian mixture models to each data set, and (ii) by assigning individual mutations to the mixture model component returning the higher posterior probability. (Mutation effects (enrichment ratios of sequencing reads before and after functional selection) were transformed into log-space where the experiment was reported in linear space.)

Correspondence between computed and experimental mutational landscapes. The agreement between computational and experimental mutation effects was evaluated using standard metrics of bivariate dependence. Due to strongly skewed experimental effect distributions and the expected nonlinear relationships between protein function and organism fitness, we focused on Spearman's rank correlation coefficient with dense ranking as our main evaluation metric. To test the robustness of our results, we additionally evaluated all relationships using distance correlations[74] and, where applicable, Pearson correlation coefficients and the corresponding linear regression R^2 values. For data sets showing bimodal effect distributions, we also tested prediction quality in a binary setting using the Matthews correlation coefficient[75]. Quantitative ΔE values from the epistatic and independent models were assigned as damaging or neutral if they were below or above the median of the respective effect distribution, respectively.

Analysis of properties of distributions. Systematic biases in experimental or computed effect distributions can influence which correlation measures are applicable, and what bivariate relationship between the two can be expected. To describe the overall shape of distributions, in particular deviations from normality and strong biases toward damaging or neutral effects, for each distribution we calculated its skewness (scipy.stats.skew) and the R^2 of the sample data against a normal distribution in a probability plot (scipy.stats.probplot). The latter quantity corresponds to the test statistic of the Shapiro–Francia test for normality.

Comparison to existing mutation effect prediction methods. For the comparison of ΔE to existing approaches for the prediction of mutation effects, we computed mutational landscapes using local installations of SIFT and PolyPhen-2 applied to the same input sequences as for our sequence alignments. Quantitative effect scores and binary classifications (neutral/damaging) were obtained from the respective columns in the output files. Mutation effects computed using the BLOSUM62 matrix correspond to the respective matrix entry of wild-type and substituted residue. Correlations were calculated on the joint set of variants that could be predicted by all methods.

Human gene and disease variant analysis. Besides the evaluation on quantitative mutation effect data, we assessed the ability of ΔE to discriminate between known disease mutations and neutral amino acid variants in a binary classification setting. For this evaluation, we computed quantitative effect scores and tested how well both types of variants are separated by ΔE. Unlike the existing method PolyPhen-2, which trains on known disease variants, our mutation effects are purely based on sequences without learning on the outcome variable.

Alignments for protein sequences with disease and/or neutral variants were generated by identifying Pfam domains in the respective sequence using *hmmscan* from HMMER and running our own alignment protocol for each domain region. If possible, regions were extended by 10 residues on either side to correct for the lack of N- and C-terminal coverage often observed in Pfam. Following the same protocol as for mutational scans, alignments were inferred at a bitscore threshold of 0.5 bits/residue. Families where the number of sequences in the alignment exceeded our computational resources were run at a bitscore of 0.8 bits/residue. Alignments were filtered at a 95% sequence identity cutoff to reduce computation time. Epistatic and independent models were then inferred for all alignments and used to compute the effect of amino acid substitutions. In the evaluation, we used all variants in domains that had alignments with $M_{eff}/L \geq 1$ and could be unambiguously mapped to UniProt sequences.

To assess if ΔE separates disease and neutral variants, we derived a data set based on high-confidence disease and neutral variants. We obtained 15,405 amino acid variants in 2,387 proteins from ClinVar[61] that were unambiguously annotated with clinical significance "pathogenic." Similarly, we derived sets of amino acid variants assumed to be neutral because of their high allele frequency in the ExAC exome sequence data set[6], at increasing levels of stringency (allele frequency (AF) ≥ 0.1: 13,643 variants/6,993 proteins; AF ≥ 0.25: 8,595 variants/5,193 proteins; AF ≥ 0.5: 4,700 variants/3,282 proteins). Of the pathogenic ClinVar variants, 10,556 were covered by an alignment (in 1,848 proteins), and 9,008 (1,553 proteins) had sufficient sequences ($M_{eff}/L \geq 1$) and were used for evaluation. Of the ExAC variants, 3,514 variants (2,190 proteins) remained at AF ≥ 0.1; 2,193 variants (1,524 proteins) at AF ≥ 0.25; and 1,182 variants (937 proteins) at AF ≥ 0.5.

For comparison to the existing mutation effect classifiers SIFT and PolyPhen-2, we chose the HumVar data set (in post-processed form with classifier predictions added by Grimm *et al.*)[62,64] as this is commonly used as a benchmark. However, using this benchmark set in the comparison to our unsupervised method is conservative, since most current supervised methods, including PolyPhen-2, train in a supervised way on these known variants and may have a tendency to overestimate their own accuracy[64]. Ideally, one would construct a truly unbiased test set by excluding these and related variants, but since this leads to a strong reduction in the number of variants that can be used for evaluation, we left the benchmark set as used by others[64]. We started from 40,389 variants (9,231 proteins) in the data set, and after using the above domain identification and alignment protocol, arrived at a set of 21,915 variants in 5,067 proteins jointly predicted by all methods. Of these, 18,001 variants in 3,912 proteins had $M_{eff}/L \geq 1$ and were used for evaluation. In addition to the full set of variants with sufficient alignment depth, we also evaluated performance on 'difficult' examples where the predictions between SIFT and PolyPhen-2 disagreed (3,126 variants in 1,459 proteins).

The statistical energy distributions between neutral and disease variants were compared using two-sample Kolmogorov-Smirnov tests (two-sided; function scipy.stats.ks_2samp). To assess the discrimination between both types of variants, the area under the receiver operating characteristic curve (AUC) was calculated using scikit-learn (function sklearn.metrics.roc_auc_score) and compared between the different methods. We also evaluated the discovery of damaging variants with high specificity by computing the partial AUC up to a false-positive rate of 20% using the pROC R package[76].

Pre-computed mutation landscapes for human proteins. Since we calculated epistatic models for all human proteins that have variants annotated as pathogenic, ExAC variants with allele frequency ≥ 0.1, and others in the course of this analysis, this results in a resource of models for 6,762 unique human proteins (9,935 alignment regions, $M_{eff}/L \geq 1$). For these proteins, we provide single substitution landscapes, sequence alignments, and evolutionary couplings (summarized epistatic constraint for pairs of positions) at http://evmutation.org/. Mutational landscapes can be explored using interactive visualization. Together with the provided software (**Supplementary Code**), the downloadable files allow us to reproduce our analyses and compute higher-order mutation landscapes for any of these proteins.

Error analysis of evolutionary statistical energy predictions. Besides the overall correlation analysis between prediction and experiment, we wanted to understand how large the prediction error is for individual mutations, and if particular mutations were predicted with lower error by either the epistatic or the independent model. To calculate individual error terms, we need to compare the predicted value to the respective experimental value. In our case, this is, however, complicated by the fact that ΔE and the experimental data are on different scales and the relationships between both are often nonlinear, with no particular expectation on the shape of the relationship. This problem prevents the use of standard regression approaches to calculate error terms.

Instead, we chose to transform predicted values into the space of the experiment by a quantile mapping strategy (a perfect prediction would mean that the normalized ranks of each variant in the predicted and experimental distribution agree). For any substitution S, let $\Delta E(S)$ correspond to the p-quantile of the predicted distribution. Then we transform a prediction $\Delta E(S)$ into experimental space by mapping it onto the respective p-quantile of the experimental distribution. We denote this mapping by $Q(\Delta E(S))$. For robustness against a small number of experimental outliers, Q assigns the respective experimental 0.01/0.99 experimental quantile for any value below/above the predicted 0.01/0.99 quantile. By defining a quantile-quantile mapping for each of the models, we can now calculate an error term $\varepsilon(\Delta E(S))$ for each variant S by comparing the mapped prediction $Q(\Delta E(S))$ to the actual experimental value $y(S)$ of the variant S: $\varepsilon(\Delta E(S)) = Q(\Delta E(S)) - y(S)$. The error term $\varepsilon(\Delta E(S))$ is normalized by the range of the full experimental distribution between the 0.01 and 0.99 quantiles to make it comparable between different experiments.

To evaluate if the use of epistatic interactions improves or decreases the error of individual variant predictions (i.e., variants that are predicted with different log-odds scores $\Delta E(S)$ compared to the independent model rather than similar log-odds scores), we had to modify the above approach because differently predicted, but wrong variants might bias the quantile–quantile mappings in favor of each method. Instead, we defined quantile–quantile mappings Q' not on the full distributions but only on those variants that are predicted similarly between both models ($|\Delta E(\text{epistatic}) - \Delta E(\text{independent})| \leq 2$, threshold chosen so that there are enough variants to define a smooth curve). Q' allows to map any arbitrary score based on the nearest quantile that was used during calculation of the mapping. Comparison of the respective absolute error terms $\varepsilon(\Delta E(S)) = Q'(\Delta E(S)) - y(S)$ between the epistatic and the independent model allows us to assess for any substitution S if the use of epistatic interactions increases or decreases the error compared to the experiment.

The computation of error terms for individual variants makes it possible to assess if certain groups of variants are systematically predicted more or less accurately using epistatic interactions. For this analysis, we compared the root mean square errors (RMSE, summing the individual error terms $\varepsilon(\Delta E(S))$) across all individual substitutions within a particular group between the epistatic and the independent model. Subgroups were defined as follows: (1) high/low experimental effect: Gaussian mixture model of effect distribution as described above; (2) frequent/rare substitution: frequency of substitution in respective column of alignment ≥ 0.01 or < 0.001; (3) ligand binding/protein interaction: all substitutions to residues within 4Å minimum atom distance of ligand or interaction partner, but excluding evolutionarily conserved cofactors; (4) buried/exposed: relative solvent accessibility calculated using DSSP < 0.1 or > 0.25; (5) conserved/variable: column conservation of position in alignment > 0.4 or < 0.2; (6) strongly/weakly coupled: at least 4/no coupled pairs in list of L top-ranking long-range ($|i - j| > 5$) evolutionary couplings.

Analysis of structural features. Evolutionary couplings calculated from multiple sequence alignments were compared to experimental protein 3D structures from the PDB[77] to assess if the identified epistatic constraints correspond to structural contacts. Structures and mappings to the target sequence were obtained using jackhmmer-based searches against the PDB (one search iteration), and residue pair distances calculated for up to ten of the most significant hits with a normalized bit-score of at least than 0.5 bits/residue to the target sequence. Two residues were considered to be in contact if any of their atoms are closer than 5 Å in any of the identified structures; a distance threshold of 4 Å was applied to interactions between amino acid residues and ligands.

Data analysis and method availability. All data analysis was conducted using Jupyter notebooks[78] and the scientific Python stack[79,80]. Supplemental Web

doi:10.1038/nbt.3769

Data, human protein predictions and code (**Supplementary Code**) to calculate statistical models and mutation effects from sequence alignments are available at http://evmutation.org/.

67. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
68. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B. & Wu, C.H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
69. Nawrocki, E.P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2015).
70. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620 (1957).
71. Dunn, S.D., Wahl, L.M. & Gloor, G.B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 (2008).
72. Toth-Petroczy, A. *et al.* Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170.e12 (2016).
73. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* **8**, R232 (2007).
74. Kosorok, M.R. Brownian distance covariance and high dimensional data. *Ann. Appl. Stat.* **3**, 1266–1269 (2009).
75. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
76. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
77. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
78. Pérez, F. & Granger, B.E. IPython: a system for interactive scientific computing. *Comput. Sci. Eng.* **9**, 21–29 (2007).
79. Van der Walt, S., Colbert, S.C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
80. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).