ARTICLES
https://doi.org/10.1038/s41592-018-0138-4
nature methods

Deep generative models of genetic variation capture the effects of mutations

Adam J. Riesselman[1,2,4], John B. Ingraham[1,3,4] and Debora S. Marks[1*]

The functions of proteins and RNAs are defined by the collective interactions of many residues, and yet most statistical models of biological sequences consider sites nearly independently. Recent approaches have demonstrated benefits of including interactions to capture pairwise covariation, but leave higher-order dependencies out of reach. Here we show how it is possible to capture higher-order, context-dependent constraints in biological sequences via latent variable models with nonlinear dependencies. We found that DeepSequence (https://github.com/debbiemarkslab/DeepSequence), a probabilistic model for sequence families, predicted the effects of mutations across a variety of deep mutational scanning experiments substantially better than existing methods based on the same evolutionary data. The model, learned in an unsupervised manner solely on the basis of sequence information, is grounded with biologically motivated priors, reveals the latent organization of sequence families, and can be used to explore new parts of sequence space.

A major unanswered question in biological research, clinical medicine, and biotechnology is how to decipher and exploit the effects of mutations on biomolecules. For efforts ranging from the identification of genetic variants underlying human disease, to the development of modified proteins with useful properties, to the synthesis of large molecular libraries that are enriched with functional sequences, there is a need to be able to rapidly assess whether a given mutation in a protein or RNA will disrupt function[1,2]. Although high-throughput technologies can now simultaneously assess the effects of thousands of mutations in parallel[1,3–27], sequence space is exponentially large and experiments are resource intensive. Accurate computational methods are thus an important component of high-throughput sequence annotation and design.

Most improvements to computational predictions of mutation effects have been driven by leveraging of the signal of evolutionary conservation among homologous sequences[28–33]. Historically, these tools have been used to analyze the conservation of single sites in proteins in a background-independent manner. Recent work has demonstrated that the incorporation of intersite dependencies in a pairwise interaction model of genetic variation can lead to more accurate prediction of the effects of mutations in high-throughput mutational scans[34–37]. However, numerous lines of evidence suggest that higher-order epistasis pervades the evolution of proteins and RNAs[38–41], and pairwise models are unable to capture this. Naive extension of pairwise models with third or higher terms is statistically unfeasible, as even third-order interaction models for a protein of 100 amino acids will have approximately 1 billion parameters. Even if such a model could be engineered or coarse-grained[42] to be computationally and statistically tractable, it would only marginally improve the fraction of higher-order terms considered, leaving fourth and higher-order interactions out of reach.

Direct parameterization of sequence-variation models with all possible interactions of order k leads to an intractable combinatorial explosion in the number of parameters to consider. An alternative to this fully observed approach for modeling data—in which the correlations between positions are explained directly in terms of position–position couplings—is to model variations in terms of 'hidden' variables to which the observed positions are coupled. Two widely used models for the analysis of genetic data, principal component analysis and admixture analysis[43–45], can be cast as latent-variable (i.e., hidden-variable) models in which the visible data (genotypes) depend on hidden variables (factors or populations) in a linear way. In principle, the replacement of those linear dependencies with flexible nonlinear transformations could facilitate the modeling of arbitrary-order correlations in an observed genotype, but the development of tractable inference algorithms for them is more complex. Recent advances in approximate inference[46,47] have made these kinds of nonlinear latent-variable models tractable for the modeling of complex distributions for many kinds of data, including text, audio, and even chemical structures[48], but their application to genetic data remains in its infancy.

Here we developed nonlinear latent-variable models for biological sequence families and leveraged approximate inference techniques to infer the families from large multiple-sequence alignments. We show how a Bayesian deep latent-variable model can be used to reveal latent structure in sequence families and predict the effects of mutations with accuracies exceeding those of site-independent or pairwise-interaction models.

Results

A deep generative model captures latent structure in sequence families. The genes observed across species today are the results of long-term evolutionary processes that select for functional molecules. We sought to model the constraints underlying the evolutionary processes of these sequences and to use those constraints to make reasonable inferences about what other mutations may be plausible. If we approximate the evolutionary process as a 'sequence generator' that generates a sequence $\mathbf{x}$ with probability $p(\mathbf{x}|\boldsymbol{\theta})$ and parameters θ that are fit to reproduce the statistics of evolutionary data, we can use the probabilities that the model assigns to any given

[1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [2]Program in Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Program in Systems Biology, Harvard University, Cambridge, MA, USA. [4]These authors contributed equally: Adam J. Riesselman, John B. Ingraham. *e-mail: debbie@hms.harvard.edu

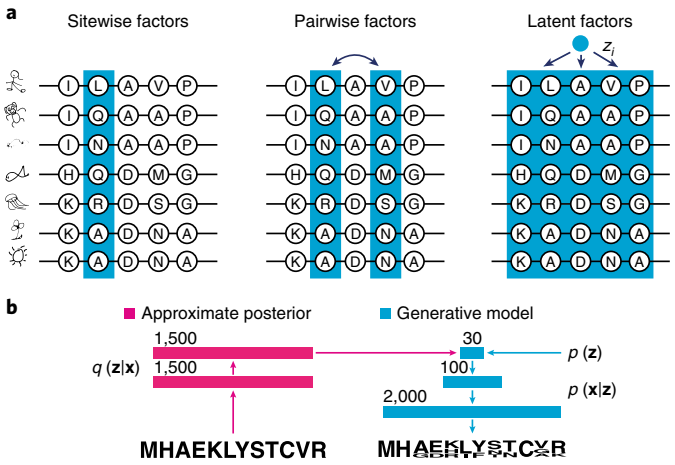

Fig. 1 | A nonlinear latent-variable model captures higher-order dependencies in proteins and RNAs. a, Comparison of a nonlinear latent-variable model with site-independent and pairwise models. **b**, The dependency $p(\mathbf{x}|\mathbf{z})$ (blue) of the sequence $\mathbf{x}$ on the latent variable $\mathbf{z}$ is modeled by a neural network, and inference and learning are made tractable by joint training with an approximate inference network $q(\mathbf{z}|\mathbf{x})$ (pink). This combination of model and inference is also known as a variational autoencoder. The size of the latent variables $\mathbf{z}$ and hidden dimensions of the neural network are shown.

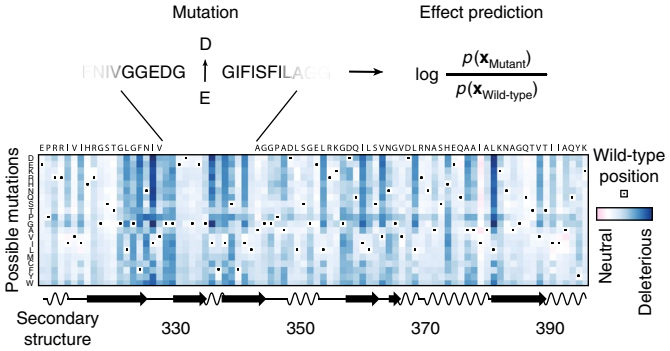

Fig. 2 | Mutation effects can be quantified by likelihood ratios. After fitting a probabilistic model to a family of homologous sequences, we heuristically quantified the effect of mutation as the log-ratio of mutant likelihood to wild-type likelihood (as approximated by the ELBO; Methods). Bottom: mutation-effect scores for positions 310–393 in the PDZ domain.

sequence as a proxy for the relative plausibility of a molecule satisfying functional constraints. We consider the log-ratio

$$\log \frac{p(\mathbf{x}^{(\text{Mutant})}|\boldsymbol{\theta})}{p(\mathbf{x}^{(\text{Wild-type})}|\boldsymbol{\theta})}$$

as a heuristic metric for the relative favorability of a mutated sequence, $\mathbf{x}^{(\text{Mutant})}$, compared with that of a wild-type sequence, $\mathbf{x}^{(\text{Wild-type})}$. This log-ratio heuristic has been shown to accurately predict the effects of mutations across multiple kinds of generative models $p(\mathbf{x}|\boldsymbol{\theta})$[34]. Our innovation here is to consider a nonlinear latent-variable model for $p(\mathbf{x}|\boldsymbol{\theta})$ that is capable of capturing higher-order constraints (Fig. 1a and Methods). This approach is fully unsupervised, as we do not train on observed mutation-effect data but rather use the statistical patterns in observed sequences as a signal of selective constraint.

We introduce a nonlinear latent-variable model $p(\mathbf{x}|\boldsymbol{\theta})$ to implicitly capture higher-order interactions between positions in a sequence. We imagine that when the data are generated, a hidden variable $\mathbf{z}$ is sampled from a prior distribution $p(\mathbf{z})$, in our case a standard multivariate normal, and a sequence $\mathbf{x}$ is in turn generated on the basis of a conditional distribution $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ that is parameterized by a neural network. If the system were fully observed, the probability of data would be simple to compute as $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z})$, but when $\mathbf{z}$ is hidden we must contend with the marginal likelihood,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z})d\mathbf{z}$$

which considers all possible explanations for the hidden variables $\mathbf{z}$ by integrating them out. Direct computation of this probability is intractable in the general case, but we can use variational inference[49] to form a lower bound on the (log) probability. This bound, known as the evidence lower bound (ELBO), $\mathcal{L}(\boldsymbol{\phi};\mathbf{x})$, takes the form

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\phi};\mathbf{x}) \triangleq \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})] - D_{\text{KL}}(q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})\|p(\mathbf{z}))$$

where $q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})$ is a variational approximation for the posterior distribution $p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})$ of hidden variables given the observed variables. We model both the conditional distribution $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ of the generative model and the approximate posterior $q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})$ with neural networks, which results in a flexible model-inference combination known as a variational autoencoder (VAE)[46,47] (Fig. 1b). After the model is fit to a given family through optimization of the variational parameters $\boldsymbol{\phi}$, it can be readily applied to predict the effects of arbitrary types and numbers of mutations. We quantify effects with an approximation to the log-ratio by replacing each log probability with the ELBO (Fig. 2).

We use a particular combination of priors, parameterizations, and learning algorithms to make the model more interpretable and more likely to generalize. First, we encourage sparse interactions[50] with a group sparsity prior on the last layer of the neural network for $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$. This prior encourages small subgroups of hidden units in the network to influence only a few positions at a time. Second, we encourage correlation between amino acid usage by transforming all local predictions of the amino acids at each position with a shared a linear map $\mathbf{C}$, which we refer to as a dictionary. Finally, and in deviation from standard practice for VAEs, we learn distributions over the weights of the neural network for $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ with a variational approximation over both the global model parameters and the per-datum hidden variables. This means that rather than learning a single neural network for $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$, we learn an infinite ensemble of networks.

We optimized the joint variational approximation over global and local parameters by stochastic gradient ascent on the ELBO to obtain a fully trained model (Methods). Because this is a nonconvex optimization problem with multiple solutions, we fit five replicas of the model from different initial conditions. Throughout the analysis, we considered both the average performance across these five fits and the performance of the ensemble prediction that averaged the predictions together.

Model probabilities correlate with experimental mutation effects. We compared predictions from DeepSequence to a collection of 42 high-throughput mutational scans (712,218 mutations across 108 sets of experiments on 34 proteins and a tRNA; Methods, Supplementary Tables 1 and 2, Supplementary Fig. 1). We found that the predictions of the DeepSequence ensemble correlated equally as well or better with experimental mutation effects across a majority of the datasets, compared with predictions from both a pairwise interaction model (EVmutation[34]; 33/42 datasets; median difference in rank correlation $\Delta\rho = 0.036$) and a site-independent model (34/42 datasets; median $\Delta\rho = 0.063$) trained

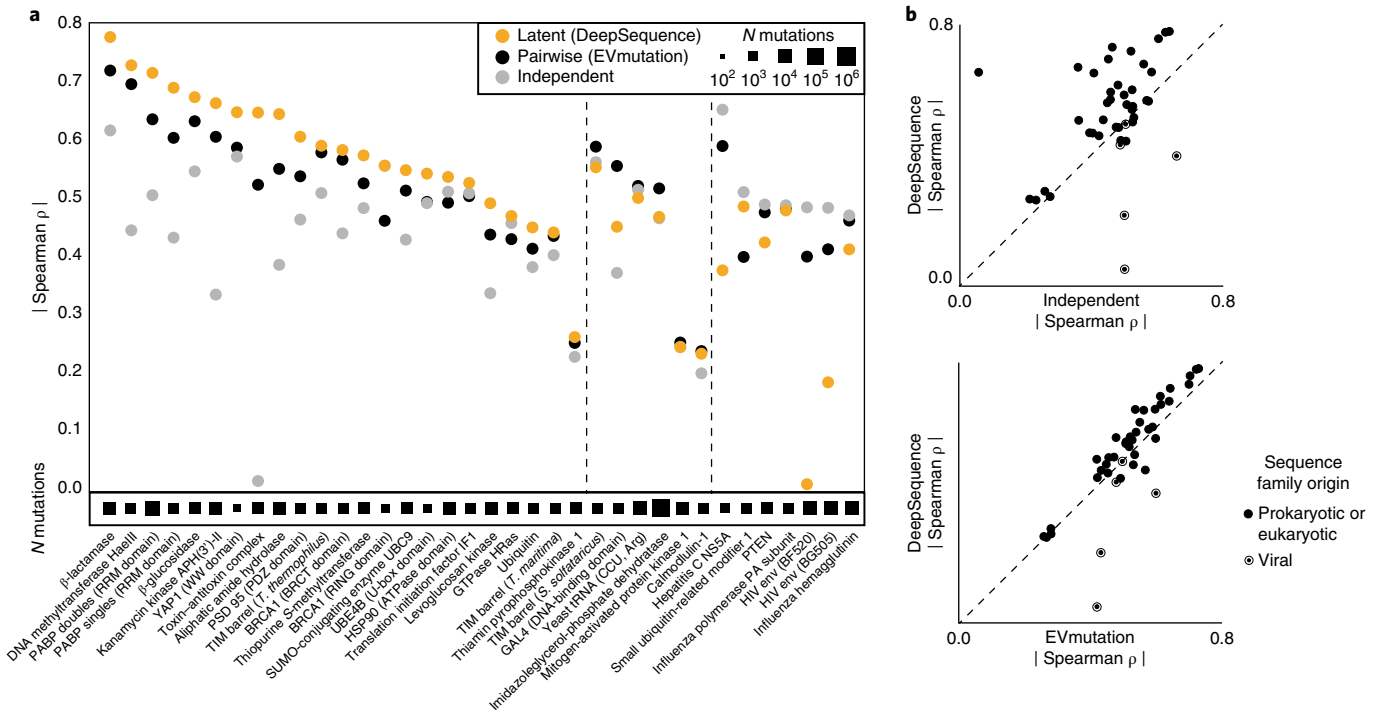

Fig. 3 | A deep latent-variable model predicts the effects of mutations better than site-independent or pairwise models. a, DeepSequence captures the effects of mutations across deep mutational scanning experiments as measured by rank correlation. **b**, Comparison of prediction accuracy of DeepSequence to that of a site-independent model (top) and EVmutation (bottom).

on the same data (Fig. 3, Supplementary Table 3). The average performance of DeepSequence without ensembling reproduced this overall advantage over EVmutation (32/42 datasets; median $\Delta\rho = 0.024$) and the site-independent model (31/42 datasets; median $\Delta\rho = 0.055$). The clear exceptions to the overall superiority of DeepSequence were the comparisons to the viral protein mutation experiment effects, and especially the two HIV env experiments, which suggests that the VAE approach is more dependent on a larger diversity of fit sequences on which to train the model. When compared against a subset of the data that was previously analyzed[34], the DeepSequence predictions were consistently more accurate than those of other commonly used methods, such as BLOSUM62 (20/20; median $\Delta\rho = 0.32$), SIFT[32] (20/20; median $\Delta\rho = 0.24$), and Polyphen2[31] (19/20; median $\Delta\rho = 0.20$) (Supplementary Table 4).

The deep mutational scans that we analyzed typically involved only one or a few mutational steps from assayed sequences ('test set') to the sequences that the model was trained on, which raised the question of how well the model can generalize when the number of steps is larger. To test this, we reran the experiments for TEM1 β-lactamase with artificially purged training sets in which sequences with 35%, 60%, 80%, 95%, or 100% identity to wild-type TEM1 were removed. We found that DeepSequence continued to outperform pairwise and site-independent models even when all sequences within 60% sequence identity were purged. The Spearman correlation decreased by only 0.07, despite the fact that all mutated test sequences were ~100 mutational steps away from the training set (Methods, Supplementary Fig. 2).

We observed a consistent amino acid bias in the prediction accuracy of all three evolutionary models (independent, EVmutation, and DeepSequence) when we compared the residuals of the rankings of the predictions with the experimental data for each amino acid transition (Supplementary Fig. 2), but we were unable to find consistent patterns for this discrepancy. For instance, we could not explain the observed bias by codon usage. Accounting for this bias by fitting a linear model on top of the predictions improved DeepSequence, but the improvements were small (Supplementary Fig. 3, Methods).

Sequence space in latent space. Examining the low-dimensional latent spaces learned by a latent-variable model can give insight into relationships between data points (sequences). To gain insight into the organization of latent space directly, we fit an otherwise identical copy of the model with 2-dimensional rather than 30-dimensional $\mathbf{z}$ to β-lactamase (Fig. 4). This visualization illustrated the comparative shallowness of deep mutational scans: all mutated sequences from the deep mutational scans of β-lactamase were tightly concentrated in a small region of latent space. We also observed an uneven distribution in latent space with phylogenetically coherent structure; however, we caution against overinterpreting this distribution, because it will depend strongly on the choice of prior and the variational approximations[51].

Bayesian learning prevents overfitting and facilitates interpretation. To test the importance of our specific choices for the architecture and learning algorithm, we carried out an ablation study of model design decisions across a subset of proteins. We found that all of the major design decisions contributed to improved performance, including the use of sparse priors on the last layer, a learned dictionary of amino acid correlations, a global inverse-temperature parameter, and a Bayesian variational approximation for the weights (Supplementary Table 5, Fig. 5, and Methods). The largest improvement seemed to result from the use of variational Bayes to learn distributions over the weights of the network rather than point estimates, even when point estimation was combined with group sparsity priors[52] or Dropout[53] regularization (Supplementary Table 5).

We considered two kinds of structured correlations in multiple-sequence alignments in the design of the final layer. The first was correlated bias in amino acid usage, where hydrophobic or polar amino acids tend to have correlated liability at a given position. We

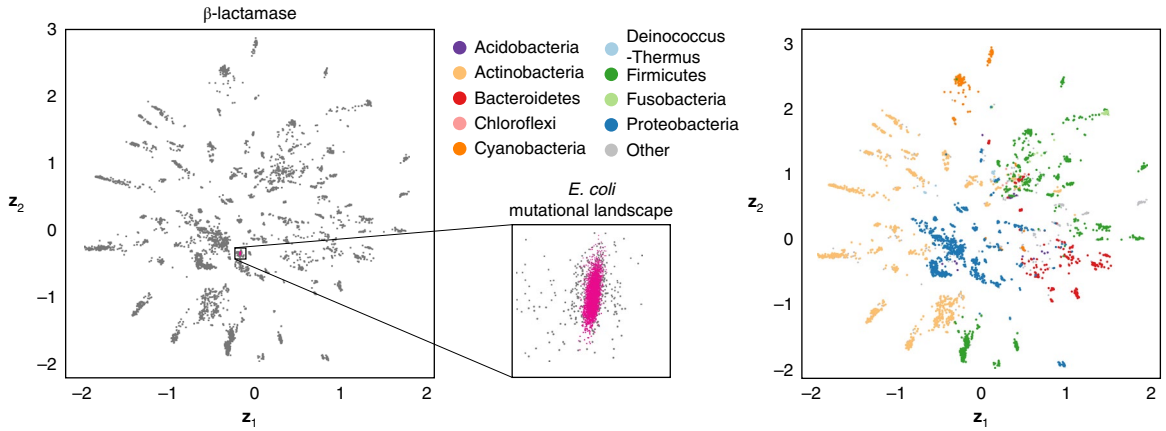

Fig. 4 | Latent variables capture the organization of sequence space. The β-lactamase family shown in two-dimensional latent space; a single deep mutational scanning experiment with variants (center: pink) is shown.

captured this with a shared linear transformation $\boldsymbol{C}$ that was tied across all positions and found that its implicit correlation structure ($\boldsymbol{C}^T\boldsymbol{C}$) reflected well-known amino acid similarities (Fig. 6a, Supplementary Table 6). The second kind of correlation in multiple-sequence alignments is between positions, and often coincides with structural proximity[54–57]. Our group sparsity prior captured these correlations by learning 500 soft subgroups of positions that were each connected to four hidden units. We computed the average pairwise distance between positions in each subgroup using representative structures from the Protein Data Bank (Methods) and found that most distances were less than a null expectation (Methods, Fig. 6a, Supplementary Tables 7 and 8), with subsets of residues close in 3D.

Interpretation of mutation-effect predictions. We then explored which sets of mutations were most differentially predicted by DeepSequence (using a subset of eight experiments with large overall differences from the independent methods; Supplementary Fig. 5, Supplementary Table 9, Methods). DeepSequence was most accurate for all proteins across all residue classifications that we explored, including both evolutionary ('conservation', 'frequency') and structural features ('interaction') (Fig. 6b). The overall increased accuracy of the latent model predictions was particularly strong for mutations that were deleterious in the experiment, and often where these deleterious sites were variable or proximal to interacting ligands. For example, in the RNA-recognition-motif domain of the poly(A) binding protein and the PDZ domain in PSD95 and kanamycin kinase, residues close to their ligands or cofactors are the most differentially accurate. These include a residue position involved in specificity switching G330 in the PDZ domain and RNA interaction sites in the RNA recognition motif (Fig. 6b). These results are consistent with the idea that the latent model makes better predictions for sites sensitive to change but still varied across evolution, and hence context dependent.

Discussion

We have presented a deep latent-variable model that can capture higher-order correlations in biological sequence families and shown how it can be applied to predict the effects of mutations across diverse classes of proteins and RNAs. We found that the predictions of the deep latent-variable model were more accurate than those of a previously published pairwise-interaction approach to model epistasis[34,36], which in turn was more accurate than commonly used supervised methods[58,59]. In addition, both latent variables and global variables of the model learned interpretable structure for both macrovariation and phylogeny, as well as the structural proximity of residues.

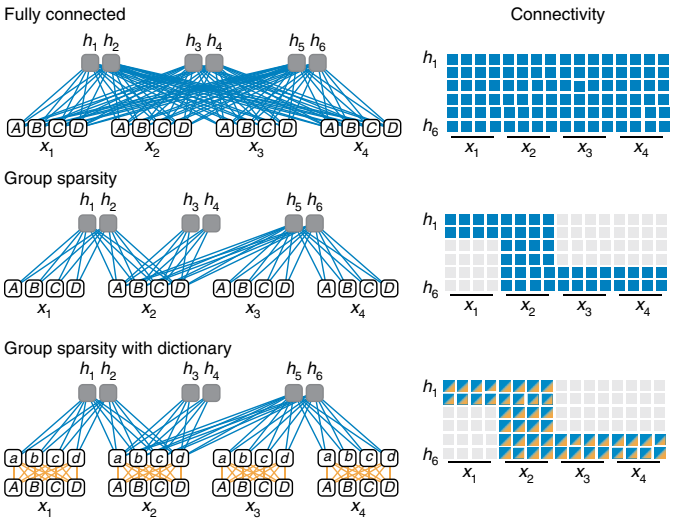

Fig. 5 | Structured priors over weights capture biological assumptions. Top: traditional fully connected layer that outputs per-position logits over different letters (*A*, *B*, *C*, *D*) at each position. Center: a group-sparsity prior over the weights encourages block sparsity such that groups of *k* hidden units tend to influence all the logits at a small number of positions. Bottom: an additional global transform at every position in the sequence by a shared weight matrix captures correlations between letter usage.

Deep latent-variable models introduce additional flexibility for modeling of higher-order constraints, but at the cost of reduced interpretability and increased potential for overfitting. Indeed, we found that even traditional approaches for regularization, such as Dropout[53] and sparsity priors, were often worse than the already-established pairwise models. While this work was in progress, other nonlinear latent-variable models were proposed for sequence families[60,61], evidencing the benefits of more parametrically powerful models for sequence variation. A key aspect of this work distinguishing it from both those and other models in our ablation study is its use of approximate Bayesian inference, whereby we estimated distributions over model parameters and propagated that uncertainty into model predictions. Although we found that mean-field approximate variational inference and group sparsity priors were sufficient to exceed the performance of a wide range of models, it is likely that future work would benefit from other biologically motivated priors, as well as more accurate approximations for variational inference[62,63]. Additionally, the incorporation

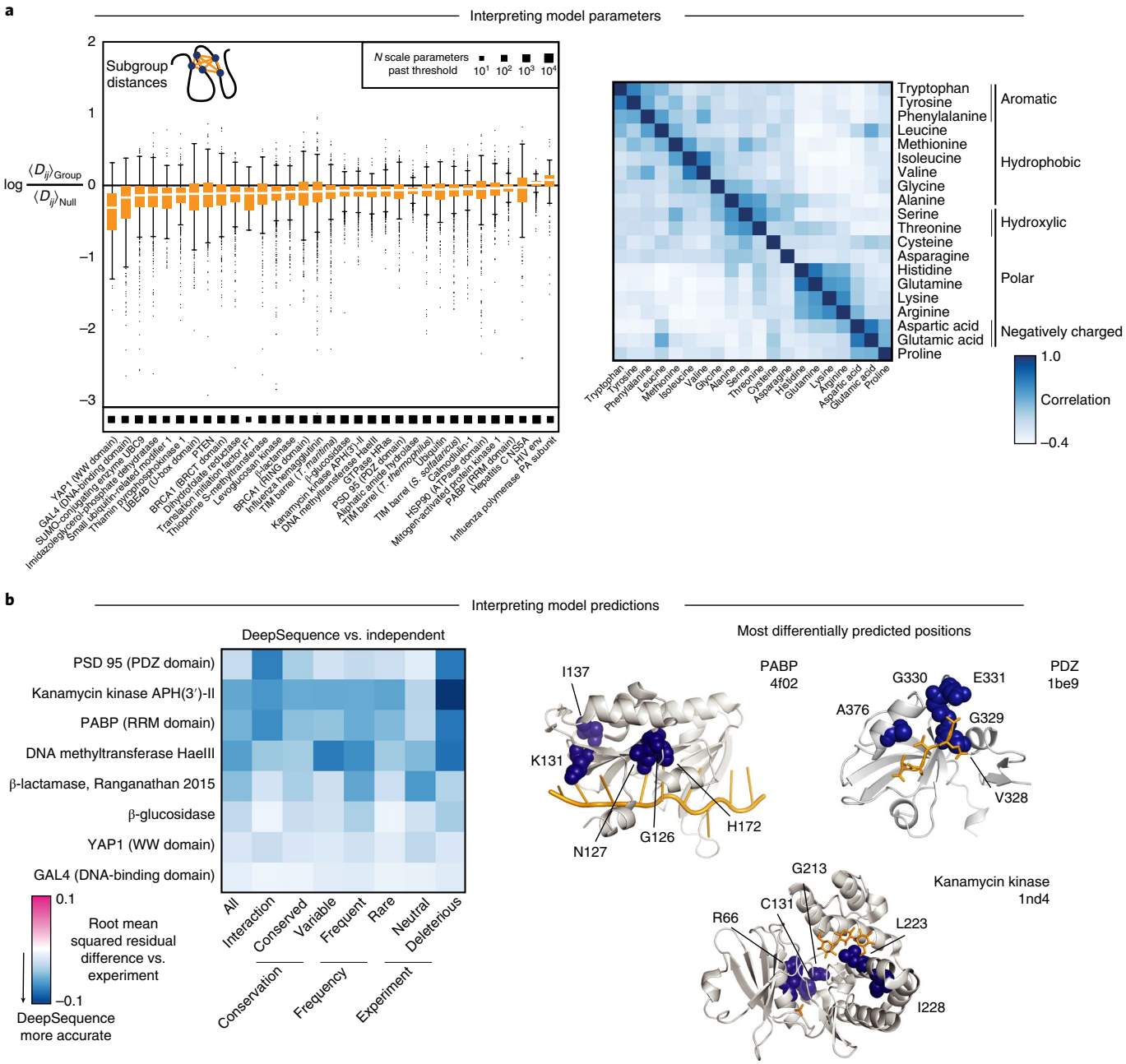

Fig. 6 | Interpretation of model and effect predictions. a, Left: log enrichment of 3D distances between residues that make up the sparse factors compared to a null distance (Methods). Box plots show the median (center line), interquartile range (hinges), and 1.5 times the interquartile range (whiskers); outliers are plotted as individual points. Right: correlations in the weights of the final width = 1 convolution taken across all models reflect known amino acid correlations and are correlated with a well-known substitution matrix BLOSUM62 (Spearman $\rho = 0.83$, $N = 210$). **b**, Left: comparison of relative rank error between DeepSequence and a position-independent conservation model across different positions and mutation attributes. Right: top five positions with the greatest reduction in rank error from the site-independent model to DeepSequence.

of more rigidly structured probabilistic graphical models in model dependencies between latent variables could improve generality and interpretability[64].

Lastly, the efficacy of DeepSequence for predicting experimental mutational scans will be contingent on both the quality and the relevance of the evolutionary sequence data in two ways. First, DeepSequence and other mutation-effect-prediction methods train on multiple-sequence alignments[34,55,65–67], for which the quality of the available data may vary across families and for which decisions on inclusion thresholds are somewhat ad hoc. Second, the premise that evolutionary data can be applied to predict outcomes of an experiment is highly contingent on the relevance of the experimental assay to long-term selective forces in the family. A mutation may be damaging with regard to some measurable protein feature—for example, enzyme efficiency—but harmless for stability or even organism fitness, as others and we have previously discussed[13,34,68]. Both of these issues could be partially addressed by the incorporation of DeepSequence predictions as features in a supervised learning framework.

Despite the challenges for deep models of sequence variation and data used to train them, they are likely to be of increasing importance for the high-throughput design and annotation of biological sequences. Evolution has conducted and continues to conduct an

unthinkably large number of protein experiments, and deep generative models can begin to identify the statistical patterns of constraint that characterize essential functions of molecules.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-018-0138-4.

Received: 1 May 2018; Accepted: 29 July 2018;
Published online: 24 September 2018

References

1. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
2. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
3. 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).
4. 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).
5. 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).
6. 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).
7. 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).
8. 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).
9. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206 (2015).
10. 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).
11. 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).
12. 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).
13. Starita, L. M. et al. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).
14. 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).
15. Jacquier, H. et al. Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl Acad. Sci. USA* **110**, 13067–13072 (2013).
16. 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).
17. Wu, N. C. et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS Genet.* **11**, e1005310 (2015).
18. Mishra, P., Flynn, J. M., Starr, T. N. & Bolon, D. N. A. Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. *Cell Rep.* **15**, 588–598 (2016).
19. Doud, M. B. & Bloom, J. D. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2016/04/07/047571 (2016).
20. 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).
21. 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).
22. Aakre, C. D. et al. Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
23. 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).
24. Li, C., Qian, W., Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837–840 (2016).
25. Mavor, D. et al. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife* **5**, e15802 (2016).
26. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
27. Starita, L. M. et al. Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
28. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
29. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16**, S1 (2015).
30. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
31. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
32. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
33. Finn, R. D. et al. HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
34. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
35. 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).
36. 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).
37. Lapedes, A., Giraud, B. & Jarzynski, C. Using sequence alignments to predict protein structure and stability with high accuracy. *arXiv* Preprint at https://arxiv.org/abs/1207.2484 (2012).
38. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).
39. Bendixsen, D. P., Østman, B. & Hayden, E. J. Negative epistasis in experimental RNA fitness landscapes. *J. Mol. Evol.* **85**, 159–168 (2017).
40. Rodrigues, J. V. et al. Biophysical principles predict fitness landscapes of drug resistance. *Proc. Natl Acad. Sci. USA* **113**, E1470–E1478 (2016).
41. Echave, J. & Wilke, C. O. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. *Annu. Rev. Biophys.* **46**, 85–103 (2017).
42. Schmidt, M. & Hamacher, K. Three-body interactions improve contact prediction within direct-coupling analysis. *Phys. Rev. E* **96**, 052405 (2017).
43. Roweis, S. & Ghahramani, Z. A unifying review of linear gaussian models. *Neural Comput.* **11**, 305–345 (1999).
44. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
45. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
46. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. *arXiv* Preprint at https://arxiv.org/abs/1312.6114 (2013).
47. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv* Preprint at https://arxiv.org/abs/1401.4082 (2014).
48. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *arXiv* Preprint at https://arxiv.org/abs/1610.02415 (2016).
49. Wainwright, M. J. & Jordan, M. I. *Graphical Models, Exponential Families, and Variational Inference* (Now Publishers, Hanover, MA, 2008).
50. Ingraham, J. & Marks, D. in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70 (eds Precup, D. & Teh, Y. W.) 1607–1616 (PMLR/Microtome Publishing, Brookline, MA, 2017).
51. Kingma, D. P. et al. in *Advances in Neural Information Processing Systems 29* (eds Lee, D. D. et al.) 4743–4751 (Curran Associates, Red Hook, NY, 2016).
52. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA, 2012).
53. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
54. Hopf, T. A. et al. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
55. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
56. 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).
57. Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999–1006 (2015).

58. Sim, N. L. et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
59. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20.1–7.20.41 (2013).
60. Tubiana, J., Cocco, S. & Monasson, R. Learning protein constitutive motifs from sequence data. *arXiv* Preprint at https://arxiv.org/abs/1803.08718 (2018).
61. Sinai, S., Kelsic, E., Church, G. M. & Nowak, M. A. Variational auto-encoding of protein sequences. *arXiv* Preprint at https://arxiv.org/abs/1712.03346 (2017).
62. Rezende, D. J. & Mohamed, S. Variational inference with normalizing flows. *arXiv* Preprint at https://arxiv.org/abs/1505.05770 (2015).
63. Burda, Y., Grosse, R. & Salakhutdinov, R. Importance weighted autoencoders. *arXiv* Preprint at https://arxiv.org/abs/1509.00519 (2015).
64. Johnson, M., Duvenaud, D. K., Wiltschko, A., Adams, R. P. & Datta, S. R. in *Advances in Neural Information Processing Systems 29* (eds Lee, D. D. et al.) 2946–2954 (Curran Associates, Red Hook, NY, 2016).
65. Ovchinnikov, S. et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248 (2015).
66. Weinreb, C. et al. 3D RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963–975 (2016).
67. Toth-Petroczy, A. et al. Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170 (2016).
68. 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).

Acknowledgements

We thank C. Sander, F. Poelwijk, D. Duvenaud, S. Sinai, E. Kelsic, the Cold Spring Harbor Laboratory Sequence-Function Relationship Journal Club and members of the Marks lab for helpful comments and discussions. A.J.R. is supported by DOE CSGF fellowship DE-FG02-97ER25308. D.S.M. and J.B.I. were funded by NIGMS (R01GM106303).

Author contributions

A.J.R., J.B.I., and D.S.M. designed the study. A.J.R. and J.B.I. performed the computations. All authors wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0138-4.

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

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

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

Methods

Datasets. We constructed a dataset of deep mutational scans that combined those analyzed in EVmutation[3–12,14–18,20–22,24,69] with additional studies published later[25,70–80]. Data underlying Fig. 3 include 41 measurements across 35 protein or RNA domains. The conditions, assay, references, maps to data for specific figures, and criteria for disambiguating multiple available measurements are listed in Supplementary Table 1.

Alignments. We repeated the same alignment-generation protocol used for EVmutation[34]. Briefly, for each protein (target sequence), we obtained multiple-sequence alignments of the corresponding protein family in five search iterations of the profile HMM homology search tool jackhmmer[81] against the UniRef100 database of nonredundant protein sequences[82] (release 11/2017 for the main analysis and 11/2015 for the ablation study). We used a bit score of 0.5 bits per residue 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 $\geq 10L$, where L is the sequence length). For $<10L$ sequences, we decreased the required average bit score until satisfied, and when the coverage was <80% we increased the bit score until satisfied. Proteins with $<2L$ sequences at <70% coverage were excluded from the analysis. See previous work for ParE–ParD toxin–antitoxin and tRNA alignment protocols.

Sequence weights. The distributions of protein and RNA sequences in genomic databases are biased by both (i) human sampling, as some parts of a phylogeny may be more studied and sequenced than others, and (ii) evolutionary sampling, as some groups of species have arisen from large radiations and will have proteins that are over-represented. We aim to reduce these biases by reweighting the data distribution. We use the previously established procedure[83] of computing each sequence weight π_s as the reciprocal of the number of sequences within a given Hamming distance cutoff. If $D_H(\mathbf{x}^{(s)}, \mathbf{x}^{(t)})$ is the normalized Hamming distance between the query sequence $\mathbf{x}^{(s)}$ and another sequence in the alignment $\mathbf{x}^{(t)}$ and θ_{ID} is a predefined neighborhood size (percent divergence), the sequence weight is

$$\pi_s = \left(\sum_t^N I\left[D_H\left(\mathbf{x}^{(s)}, \mathbf{x}^{(t)}\right) < \theta_{ID} \right] \right)^{-1}$$

The effective sample size of a multiple-sequence alignment can then be computed as the sum of these weights as

$$N_{\text{eff}} = \sum_t^N \pi_t$$

To fit a model to reweighted data, there are two common approaches. First, as was done previously[83], one can reweight every log-likelihood in the objective by its sequence weight π_s. Although this works well for batch optimization, we found that it led to high-variance gradient estimates with mini-batch optimization that made stochastic gradient descent unstable. We instead used the approach of sampling data points for each mini-batch with probability p_s proportional to their weight as $p_s = \pi_s / N_{\text{eff}}$.

Following prior work[34], we set $\theta_{ID} = 0.2$ for all multiple-sequence alignment sequences (80% sequence identity) except those for viral proteins, where we set $\theta_{ID} = 0.01$ (99% sequence identity) owing to limited sequence diversity and the expectation that small differences in viral sequences will have a higher probability of containing constraint information than the same diversity might from a sample of mammals, for instance.

Background: latent factor models. Probabilistic latent-variable models reveal structure in data by positing a partially unobserved generative process that created the data and then conducting inference to learn the parameters of the generative process. We focus on models in which an unobserved set of factors $\mathbf{z}$ are drawn from an independent distribution and each data point arises according to a conditional distribution $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ that is parameterized by θ. This process can be written as

$$\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_D)$$

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$$

Principal component analysis (PCA) has been a foundational model for the analysis of genetic variation and can be realized in this probabilistic framework as the zero-noise limit of probabilistic PCA[43,84]. With linear conditional dependencies $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$, PCA can model only additive interactions between the latent factors $\mathbf{z}$. This limitation could in principle be remedied through the use of a conditional model $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ with nonlinear dependencies on $\mathbf{z}$.

Here we consider a conditional model for sequences $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ that differs from PCA in two ways. First, the conditional distribution of the data $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ will be categorical rather than Gaussian to model discrete characters. Second, the conditional distribution $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ will be parameterized by a neural network rather than a linear map. In this sense, our latent-variable model may be thought of as a discrete, nonlinear analog of PCA.

Nonlinear categorical factor model. Our probabilistic model is specified by two components: a prior distribution $p(\mathbf{z})$ that specifies the marginal distribution of the hidden variables $\mathbf{z}$, and a conditional distribution $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ that specifies how a sequence $\mathbf{x}$ is generated given the hidden variables. In our model, the sequence $\mathbf{x}$ is a string of letters of length L drawn from an alphabet of size q, and the hidden variables $\mathbf{z}$ are a vector of real numbers with length D.

In this work, we structure our prior and conditional distribution similarly to original versions of the VAE[46]. That is, we model the prior distribution $p(\mathbf{z})$ as a multivariate normal of dimension D with mean 0 and identity covariance, and we model the conditional distribution $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ as a simple fully connected neural network with two hidden layers. Thus, the generative process for the joint distribution $p(\mathbf{z})p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ can be written as

$$\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_D)$$

$$\mathbf{h}^{(1)} = f_1\left(\mathbf{W}^{(1)}\mathbf{z} + \mathbf{b}^{(1)}\right)$$

$$\mathbf{h}^{(2)} = f_2\left(\mathbf{W}^{(2)}\mathbf{h}^{(1)} + \mathbf{b}^{(2)}\right)$$

$$\mathbf{h}^{(3,i)} = \mathbf{W}^{(3,i)}\mathbf{h}^{(2)} + \mathbf{b}^{(3,i)} \quad \text{for } i = 1 \ldots L$$

$$p(x_i = a|\mathbf{z}) = \frac{e^{h_a^{(3,i)}}}{\sum_b e^{h_b^{(3,i)}}} \quad \text{for } i = 1 \ldots L$$

where the nonlinearities are rectified linear units (ReLU) $f_1 = \max(0, u)$ on the first layer and sigmoidal $f_2 = 1/(1 + e^{-u})$ on the second. The sigmoidal nonlinearity was introduced to bound the magnitude of the preactivations that are multiplied by the structured matrix $\mathbf{W}^{(3,i)}$ (discussed in the next section). It is important to note that each letter x_i is conditionally independent of every other position given the hidden variables $\mathbf{z}$ and, as a result, all correlations between letters must be mediated by correlations with the hidden variables.

Structured parameterization. All biologically motivated aspects of our model are captured in a structured parameterization of the final weight matrix. Our parameterization is motived by three assumptions. (i) Sparsely interacting subsystems: hidden factors influence small subsystems of positions at a time, rather than jointly affecting the entire sequence. (ii) Correlated amino acid usage: certain amino acids are more likely to functionally substitute other amino acids in the same position on the basis of biochemistry. (iii) Selective uncertainty: differences in the strength of selection may lead to varying effective 'temperatures' of the constraint distribution. To capture these constraints, we parameterize at each position i as

$$\mathbf{W}^{(3,i)} = \lambda \mathbf{C} \tilde{\mathbf{W}}^{(3,i)} \text{diag}(\mathbf{S}_i)$$

where $\mathbf{W}^{(3,i)}$ is a $[q \times H]$ matrix that linearly combines the H activations in the final hidden $\mathbf{h}^{(2)}$ layer to q multinomial logits for the different characters (for example, amino acids) at position i. The parameterization consists of four terms: a matrix $\mathbf{C}$ that captures amino acid correlations, a matrix $\mathbf{S}$ with elements on (0,1) that gates which hidden units can affect which positions, a scalar constant λ capturing the overall selective constraint shared across all positions (inverse temperature), and underlying parameters $\tilde{\mathbf{W}}^{(3,i)}$ that combine with the other terms to capture site-specific constraints. We describe these elements in turn.

To capture correlations in amino acid usage, we split the weights themselves into a combination of local weights at each position $\tilde{\mathbf{W}}^{(3,i)}$ plus a global 'dictionary' matrix $\mathbf{C}$. The inner dimension of the product $\mathbf{C}\tilde{\mathbf{W}}^{(3,i)}$ is a hyperparameter E such that $\mathbf{C}$ is a global $[q \times E]$ matrix that left multiplies each of the $\tilde{\mathbf{W}}^{(3,i)}$ matrices, which are $[E \times H]$, to transform to the alphabet of the model (for example, the amino acids).

To capture sparsely depending subsystems, we learn an $[H \times L]$ matrix $\mathbf{S}$ that masks which neurons in the final hidden layer $\mathbf{h}^{(2)}$ can influence which positions in the sequence. Each column vector $\mathbf{S}_i$ captures the hidden units in $\mathbf{h}^{(2)}$ that affect position i in the sequence. We constrain the values of this matrix to be between 0 and 1, and also to tie some of the rows of $\mathbf{S}$ to be equal (Fig. 5; see parameterization below). In the above expression we write $\text{diag}(\mathbf{S}_i)$ to indicate the $[H \times H]$ matrix containing the column vector $\mathbf{s}_i$ along its diagonal, which allows us to frame the matrix $\mathbf{S}$ as a component of the parameterization of $\mathbf{W}^{(3,i)}$. In practice, we actually compute the effect of $\mathbf{S}$ with broadcasting as $h^{(3,i)} = \lambda \mathbf{C}\tilde{\mathbf{W}}^{(3,i)}(\mathbf{S}_i \odot \mathbf{h}^{(2)}) + \mathbf{b}^{(3,i)}$.

Lastly, a positive scalar λ captures the overall strength of the constraints. Typically, including a scalar prefactor in a weight matrix is fully redundant, but placing a prior over this parameter and then modeling uncertainty with a variational approximation can capture global, sequence-wide correlations in the selective strength.

We parameterize the constrained parameters $\mathbf{S}$ and λ with unconstrained forms $\widetilde{\mathbf{S}}$ and $\widetilde{\lambda}$ that can be optimized by gradient descent. The global 'inverse temperature' λ is parameterized by the softplus function as $\lambda = \log(1 + e^{\widetilde{\lambda}})$. The sparsity matrix $\mathbf{S}$ is constrained to both (i) have elements on (0,1) and (ii) have H/k blocks of k identical rows. To accomplish this, we parameterize it in terms of an $[H/k \times L]$ matrix $\widetilde{\mathbf{S}}$, transform it by a sigmoid, and tile the rows of $\widetilde{\mathbf{S}}$ k times with the transform $S_{ji} = (1 + \exp(\widetilde{\mathbf{S}}_{j \bmod (H/k),i}))^{-1}$. When paired with a Gaussian prior over $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{W}}^{(3,i)}$, the scale parameters $\mathbf{S}$ will be a priori logit-normally distributed, and the resulting product $\widetilde{\mathbf{W}}^{(3,i)}\mathrm{diag}(\mathbf{S}_i)$ can be seen as a continuous relaxation of a spike and slab prior (where it would be exact if the elements of $\mathbf{S}$ were Bernoulli).

Priors. We place Gaussian priors over all unconstrained parameters as $\widetilde{W}_{ij} \sim \mathcal{N}(0,1)$, $C_{ij} \sim \mathcal{N}(0,1)$, $\widetilde{S}_{ij} \sim \mathcal{N}(\mu_s, \sigma_s^2)$, and $\widetilde{\lambda} \sim \mathcal{N}(0,1)$. To set the hyperparameters μ_s, σ_s^2, we consider the effective logit-normal prior over $\mathbf{s}$, which can be thought of as a smooth relaxation of a Bernoulli that can be made sharper by an increase in the variance σ_s^2. We set $\mu_s = -12.36$ and $\sigma_s^2 = 16$.

Background: variational inference. Nonlinear latent factor models are difficult to infer. Because the latent variables $\mathbf{z}$ are not observed, computation of the marginal likelihood of the data requires that they be integrated out as

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log \int p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z}$$

We must apply this integral because we do not know a priori which $\mathbf{z}$ is responsible for each data point $\mathbf{x}$, and so we average over all possible explanations weighted by their relative probability. When this integral over $\mathbf{z}$ cannot be simplified, optimization of the marginal likelihood $\log p(\mathbf{x}|\boldsymbol{\theta})$ to fit a model to data will be intractable.

Variational inference forms a lower bound on $\log p(\mathbf{x}|\boldsymbol{\theta})$ that is easier to optimize. First, we introduce $q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})$, an approximate distribution for $\mathbf{z}$ given $\mathbf{x}$ that is flexibly parameterized by variational parameters $\boldsymbol{\phi}$. By Jensen's inequality, we can lower-bound the intractable marginal likelihood $\log p(\mathbf{x}|\boldsymbol{\theta})$ as

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log \int p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z}$$

$$= \log \int \frac{p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})} q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi}) d\mathbf{z}$$

$$\geq \int \log \frac{p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})} q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi}) d\mathbf{z}$$

We write this lower bound as

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\phi}) \triangleq \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})] - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi}) || p(\mathbf{z}))$$

which we refer to as the evidence lower bound (ELBO). Maximizing the ELBO has the side effect of minimizing the Kullback–Leibler (KL) divergence between the variational approximation $q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})$ and the true posterior distribution for $\mathbf{z}$ given $\mathbf{x}$,

$$p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}) = \frac{p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p(\mathbf{z})}{\int p(\mathbf{x}|\mathbf{z}',\boldsymbol{\theta}) p(\mathbf{z}') d\mathbf{z}'}$$

Variational approximation for local posteriors $p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})$. We structure the functional form of the variational approximation for $\mathbf{z}$ as

$$\mathbf{g}^{(1)} = f_1(\mathbf{W}_q^{(1)}\mathbf{x} + \mathbf{b}_q^{(1)})$$

$$\mathbf{g}^{(2)} = f_1(\mathbf{W}_q^{(2)}\mathbf{g}^{(1)} + \mathbf{b}_q^{(2)})$$

$$\boldsymbol{\mu} = \mathbf{W}_\mu^{(3)}\mathbf{g}^{(2)} + \mathbf{b}_\mu^{(3)}$$

$$\boldsymbol{\sigma} = \exp(\mathbf{W}_\sigma^{(3)}\mathbf{g}^{(2)} + \mathbf{b}_\sigma^{(3)})$$

$$q(\mathbf{z} \mid \mathbf{x},\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}\ (\boldsymbol{\sigma}^2))$$

We apply the 'reparameterization trick' of Kingma and Welling[46] and Rezende et al.[47] to write the latent variables $\mathbf{z}$ as deterministic transforms of a noise source $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_D)$ as $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, where the symbol $\odot$ is an element-wise product.

Variational approximation for global posteriors $p(\boldsymbol{\theta}|\mathbf{X})$. We briefly review[46,47] how to extend variational approximations to include both the latent variables $\mathbf{z}$ and the global parameters $\boldsymbol{\theta}$[46,47]. Because the posterior for the global parameters is conditioned on the entire dataset, we must consider the marginal likelihood of the full dataset $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, which integrates out all of the corresponding latent factors $\mathbf{Z} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}\}$. The likelihood of the entire dataset $\log p(\mathbf{X})$ can be lower-bounded by

$$\log\ p(\mathbf{X}) = \log \iint p(\mathbf{X} \mid \mathbf{Z},\boldsymbol{\theta}) p(\mathbf{Z}) p(\boldsymbol{\theta}) d\mathbf{Z} d\boldsymbol{\theta}$$

$$= \log \iint \frac{p(\mathbf{X} \mid \mathbf{Z},\boldsymbol{\theta}) p(\mathbf{Z}) p(\boldsymbol{\theta})}{q(\mathbf{Z},\boldsymbol{\theta} \mid \mathbf{X},\boldsymbol{\phi})} q(\mathbf{Z},\boldsymbol{\theta} \mid \mathbf{X},\boldsymbol{\phi}) d\mathbf{Z} d\boldsymbol{\theta}$$

$$\geq \iint \log \frac{p(\mathbf{X}|\mathbf{Z},\boldsymbol{\theta}) p(\mathbf{Z}) p(\boldsymbol{\theta})}{q(\mathbf{Z},\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\phi})} q(\mathbf{Z},\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\phi}) d\mathbf{Z} d\boldsymbol{\theta}$$

The ELBO can then be written as

$$\log p(\mathbf{X}) \geq \mathcal{L}(\boldsymbol{\phi}) \triangleq N \mathbb{E}_{\mathbf{x} \in \mathbf{X}}[\mathbb{E}_{q(\boldsymbol{\theta})q(\mathbf{z}|\mathbf{x})}(\log p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi}) || p(\mathbf{z}))] - \sum_{\boldsymbol{\theta}^{(i)}} D_{\mathrm{KL}}(q(\boldsymbol{\theta}^{(i)}) || p(\boldsymbol{\theta}^{(i)}))$$

We model all variational distributions over the parameters with fully factorized mean-field Gaussian distributions. We factorize the variational approximation across the global and local parameters as $q(\mathbf{Z},\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\phi}) = q(\mathbf{Z}|\mathbf{X},\boldsymbol{\phi}) q(\boldsymbol{\theta}|\boldsymbol{\phi})$, across $\mathbf{Z}$ as $q(\mathbf{Z}|\mathbf{X},\boldsymbol{\phi}) = \prod_i q(\mathbf{z}^{(i)}|\mathbf{x}^{(i)},\boldsymbol{\phi})$, and across the model parameters as $q(\boldsymbol{\theta}|\boldsymbol{\phi}) = \prod_i (\boldsymbol{\theta}^{(i)}|\boldsymbol{\phi})$. In accordance with our data-reweighting scheme, we set $N = N_{\mathrm{eff}}$, the effective number of sequences that is the sum of the sequence weights.

Model hyperparameters. We used a fixed architecture across all sequence families. The encoder has architecture 1500-1500-(30x2) with fully connected layers and ReLU nonlinearities. The decoder has two hidden layers: the first with size 100 and a ReLU nonlinearity, and the second with size 2,000 with a sigmoid nonlinearity. The dictionary $\mathbf{C}$ is a $q \times 40$ matrix where the alphabet size q is 20 for proteins and 4 for nucleic acids. We tied rows of the $\mathbf{S}$ matrix into groups of size $k = 4$. Dropout[53] was set to 0.5 when used in ablation studies. Models were optimized with Adam[85] with default parameters and a batch size of 100 until convergence, completing 300,000 updates.

Each model was fit five times to the same multiple-sequence alignment using a different random seed. We calculated the mutation effect prediction (ΔE) by taking the difference of the mean of 2,000 ELBO samples of the wild-type and a given mutated sequence.

Site-independent and pairwise model. We compared the VAE with two other kinds of probabilistic models, both of which can be characterized as undirected graphical models with probability

$$P(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z} \exp(E(\mathbf{x}))$$

For these distributions, $E(\mathbf{x})$ is the log-potential that describes the favorability of a sequence, and Z normalizes the distribution over sequence space. A site-independent model has site-additive terms for each amino acid in each position as

$$E_{\mathrm{site}}(\mathbf{x}) = \sum_i \mathbf{h}_i(x_i)$$

while the pairwise model includes additional parameters for every pairwise combination of amino acids as

$$E_{\mathrm{pair}}(\mathbf{x}) = \sum_i \mathbf{h}_i(x_i) + \sum_{i<j} \mathbf{J}_{ij}(x_i, x_j)$$

We estimated these models using the same methods previously described for EVmutation[34] (L_2-penalized maximum pseudolikelihood). To compute the effects of mutations, we again use the log-ratio

$$\log \frac{p(\mathbf{x}^{(\mathrm{Mutant})}|\boldsymbol{\theta})}{p(\mathbf{x}^{(\mathrm{Wild-type})}|\boldsymbol{\theta})}$$

which reduces to the difference of the (negative) energy values $E(\mathbf{x})$ for the mutated and wild-type sequence.

Predictors for non-epistatic mutation effects. We also compared a subset of protein datasets with three commonly used mutation-effect predictors—a BLOSUM62 matrix, SIFT[32], and Polyphen2[31]—as previously described for EVmutation[34].

Group sparsity analysis. We aimed to test whether positional activations driven by specific neurons in the last hidden layer corresponded to structural proximity. We gathered ground-truth distance data from the PDB[86] records of homologous sequences that were found via Jackhmmer (Supplementary Table 8). To aggregate distance information across multiple structures from the same family, we computed the median (across PDBs) of the minimum atom distances (across all atom pairs per position pair) for both intra- and homo-oligomer distances. The final aggregated distance matrix was the element-wise minimum of the intra- and homo-oligomeric distance matrices.

To identify the dominant connections between the last hidden layer of neurons and specific positions in the sequence, we consider the (approximate) sparsity structure of the matrix $\mathbf{S}$. The row vectors $\boldsymbol{S}_{a,:}$ represent positions in the sequence that are affected by the hidden unit $h_a^{(2)}$. We quantify the 'typical' distances in these groupings by means of a weighted average of the distances, where the weighting within each group is computed as $w_{ij}^{(a)} = S_{ai}S_{aj}$. The 'typical' distance within a group a is then

$$\hat{D}^{(a)} = \frac{\sum_{i<j} w_{ij}^{(a)} D_{ij}}{\sum_{i<j} w_{ij}^{(a)}}$$

The null expectation for each $\hat{D}^{(a)}$ is the expected average distance under permuted groups, which is simply the average pairwise distance across the whole structure. We discarded any 'disconnected' hidden units for which the entire row of $\mathbf{S}$ had a negligible value ($\forall j, S_{ij} < 0.001$).

Residual analysis. We calculated the Spearman ρ by transforming paired data to ranked quantiles and then computing the Pearson correlation between the ranks. To determine where the model over- or underpredicted the ΔE for each mutation, we transformed the experimental measurements and mutation-effect predictions to normalized ranks on the interval [0,1].

To quantify the error between predictions, we fit a least-squares linear fit from the normalized ranks of the predictions to the normalized ranks of the data for each method and then measured the residuals of this fit. Thus, we define the residual effects as the residuals of a least-squares linear fit between the normalized ranks. Given a least-squares fit with slope and bias m and b, respectively, the residuals are then

$$\varepsilon_{\Delta E} = d_{\text{Experiment}} - (m d_{\Delta E} + b)$$

Thus positive residuals $\varepsilon_{\Delta E} > 0$ represent underprediction of the rank of the experimental effect (and thus overprediction of deleteriousness), whereas negative $\varepsilon_{\Delta E}$ values represent overprediction of the experimental rank (and thus underprediction of deleteriousness). We analyzed deep mutational scans with only single mutations, using the most recent experimental data for each protein. Residuals were grouped by the identity of the amino acid either before mutation (wild-type) or after mutation (mutant).

Bias correction. To correct for biases between mutation-effect predictions and experimental measurements, we created a feature matrix for each mutation that included ΔE, amino acid identity before and after mutation, alignment column statistics (conservation and amino acid frequency), and residue hydrophobicity[87]. We used leave-one-out cross-validation (LOOCV) to correct the bias for each dataset. Using the most recent deep mutational scan (DMS) experiment as the representative of that protein family (28 DMS datasets), we used the mutants of 27 datasets to fit a regression model to predict the residuals of each known mutation, $\varepsilon_{\Delta E}$, given the feature matrix. After this model was fit, it was used to predict $\varepsilon_{\Delta E}$ for the mutants in the test dataset. This predicted residual bias $\hat{\varepsilon}_{\Delta E}$ was subtracted from the normalized predicted rank $\hat{d}_{\Delta E} = d_{\Delta E} - \hat{\varepsilon}_{\Delta E}$. These corrected predictions were then reranked and compared to the experimental results for calculation of the corrected Spearman ρ. To predict the effects of mutations solely from DMS data, we used the same LOOCV procedure but excluded all evolutionary information in the feature matrix for each mutation. In this case, the feature matrix was used to directly compute a rank $\hat{d}_{\text{DMS}}$. We subsequently reranked these values and compared them to the ranked experimental results to calculate a corrected Spearman ρ.

Generalizability analysis. We focused on the alignment and mutation effects of β-lactamase as a case study to test the generalizability of evolutionary models for the prediction of mutation effects (Supplementary Fig. 2). We first made four reduced alignments by removing the query sequence (BLAT_ECOLX) and all sequences with a normalized hamming distance greater than 0.95, 0.8, 0.6, and 0.35 to the query, respectively. Five latent variable models were fit to each alignment, as well as a pairwise and independent model using the same sequence weighting and model fitting techniques.

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

Data availability

The sequence data and code supporting this work are available at https://github.com/debbiemarkslab/DeepSequence. The mutation-effects data from all analyzed experiments, as well as all model predictions, are available in Supplementary Table 2.

References

69. Doud, M. B. & Bloom, J. D. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses* **8**, 155 (2016).
70. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8**, 15695 (2017).
71. Chan, Y. H., Venev, S. V., Zeldovich, K. B. & Matthews, C. R. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nat. Commun.* **8**, 14614 (2017).
72. Kelsic, E. D. et al. RNA structural determinants of optimal codons revealed by MAGE-Seq. *Cell Syst.* **3**, 563–571 (2016).
73. Brenan, L. et al. Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. *Cell Rep.* **17**, 1171–1183 (2016).
74. Bandaru, P. et al. Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife* **6**, e27810 (2017).
75. Findlay, G. M. et al. Accurate functional classification of thousands of *BRCA1* variants with saturation genome editing. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2018/04/05/294520 (2018).
76. Matreyek, K. A. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2018/01/16/211011 (2018).
77. Klesmith, J. R., Bacik, J.-P., Michalczyk, R. & Whitehead, T. A. Comprehensive sequence-flux mapping of a levoglucosan utilization pathway in *E. coli*. *ACS Synth. Biol.* **4**, 1235–1243 (2015).
78. Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J. & Bloom, J. D. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife* **7**, e34420 (2018).
79. Pokusaeva, V. et al. Experimental assay of a fitness landscape on a macroevolutionary scale. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2018/04/06/222778 (2018).
80. Weile, J. et al. A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).
81. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
82. 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).
83. 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).
84. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *J. R. Stat. Soc. Series B Stat. Methodol.* **61**, 611–622 (1999).
85. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *arXiv* Preprint at https://arxiv.org/abs/1412.6980 (2014).
86. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
87. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

nature research

Corresponding author(s): Debora S. Marks

Reporting Summary

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

Statistical parameters

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

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Deep Sequence code, INSTALL, README and examples are available on the Marks lab GitHub https://github.com/debbiemarkslab/DeepSequence DeepSequence is a generative, unsupervised latent variable model for biological sequences. Given a multiple sequence alignment as an input, it can be used to predict accessible mutations, extract quantitative features for supervised learning, and generate libraries of new sequences satisfying apparent constraints. It models higher-order dependencies in sequences as a nonlinear combination of constraints between subsets of residues. For more information, check out the paper on biorxiv and the examples below. For ease of analysis, we advise that alignments be generated with the EVcouplings package, though any sequence alignment can be used. Codebase is compatible with Python 2.7 and Theano 1.0.1. For GPU-enabled computation, CUDA will have to be installed separately. See INSTALL for more details.
Data analysis	DeepSequence is a generative, unsupervised latent variable model for biological sequences. Given a multiple sequence alignment as an input, it can be used to predict accessible mutations, extract quantitative features for supervised learning, and generate libraries of new sequences satisfying apparent constraints. It models higher-order dependencies in sequences as a nonlinear combination of constraints between subsets of residues. For more information, check out the paper on biorxiv and the examples below.

For ease of analysis, we advise that alignments be generated with the EVcouplings package, though any sequence alignment can be used. Codebase is compatible with Python 2.7 and Theano 1.0.1. For GPU-enabled computation, CUDA will have to be installed separately. See INSTALL for more details.

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

Data

Policy information about availability of data

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

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

Data analyzed in this study are publicly available and extracted from previous publications and in addition authors confirm that all relevant data are supplied in the Supplementary Tables
Data produced in this study are all supplied in the Supplementary Tables

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We used publicly available protein sequences extracted from UniProt database
Data exclusions	No data was excluded
Replication	We can the Deep Sequence model 10 times for each protein experiment to verify the results and produce confidence intervals
Randomization	This is not relevant to our study as we are blindy computing a statistical measure of the sequence from a neural net in an UN supervised fashion
Blinding	Blinding is not relevant to the study as the study is unsupervised learning

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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