# Protein design using structure-based residue preferences

**Authors:** David Ding[1*], Ada Shaw[2], Sam Sinai[3], Nathan Rollins[4], Noam Prywes[1], David F. Savage[1,5,6], Michael T. Laub[7,8*], Debora S. Marks[2*]

Affiliations:

[1] Innovative Genomics Institute, University of California; Berkeley, CA 94720, USA.

[2] Department of Systems Biology, Harvard Medical School; Boston, MA, 02115, USA.

[3] Dyno Therapeutics, Watertown, MA, 02472, USA.

[4] Seismic Therapeutics, Lab Central; Cambridge, MA, 02142, USA.

[5] Department of Molecular and Cell Biology, University of California; Berkeley, CA, 94720, USA.

[6] Howard Hughes Medical Institute, University of California; Berkeley, CA, 94720, USA.

[7] Department of Biology, Massachusetts Institute of Technology; Cambridge, MA 02139, USA.

[8] Howard Hughes Medical Medical Institute, Massachusetts Institute of Technology; Cambridge, MA 02139, USA.

*Corresponding authors: debbie@hms.harvard.edu, laub@mit.edu, davidding@berkeley.edu

## Abstract

Recent developments in protein design have adapted large neural networks with up to 100s of millions of parameters to learn complex sequence-function mappings. However, it is unclear which dependencies between residues are critical for determining protein function, and a better empirical understanding could enable high quality models that are also more data- and resource-efficient. Here, we observe that the per residue amino acid preferences - without considering interactions between mutations - are sufficient to explain much, and sometimes virtually all of the combinatorial mutation effects across 7 datasets ($R^2$ ~ 78-98%), including one generated here. These preference parameters (20*N, where N is the number of mutated residues) can be learned from as few as ~5*20*N observations to predict a much larger number (potentially up to $20^N$) of combinatorial variant effects with high accuracy (Pearson r > 0.8). We hypothesized that the local structural dependencies surrounding a residue could be sufficient to learn these required mutation preferences, and developed an unsupervised design approach, which we term CoVES for 'Combinatorial Variant Effects from Structure'. We show that CoVES outperforms not just model-free sampling approaches but also complicated, high-capacity autoregressive neural networks in generating functional and diverse sequence variants for two example proteins. This simple, biologically-rooted model can be an effective alternative to high-capacity, out of domain models for the design of functional proteins.

## Introduction

A key question in molecular evolution and protein engineering is how multiple mutations combine to impact both function and future mutational trajectories. The possible mutational trajectories of a given protein can be limited, if, for example, the negative effect of a single substitution can only be tolerated in the presence of another enabling mutation[1–3]. Conceptually, such specific dependencies between mutations cause 'rugged' fitness landscapes, in which natural or experimental selection for fitness-increasing mutations does not necessarily result in optimally functioning proteins[4]. On the other hand, if multiple mutations combine without specific dependencies between each other, the sequence-fitness function will result in a simple monotonic function on which selection can act more efficiently. Similarly, knowing such specific dependencies is critical in determining combinations of mutations for the design of protein therapeutics with desired functions.

While recent advances in modeling protein function in response to mutations have focused on increasing the capacity of models to fit more complex fitness landscapes[5–10], it remains unclear how complicated biological protein fitness landscapes are. The explicit number of required specific dependencies, *i.e.*, epistatic terms, grows combinatorially with the order of interactions considered ($20^k \left(\frac{L}{k}\right) \frac{1}{k!}$ terms of order k and protein length L). For example, for a protein of length 100, there are 2,000 first-order site-wise terms, but ~1 million second-order pair-wise interaction terms and ~200 million third-order interaction terms. Indeed, sequence models that are able to capture dependencies between residues show increased performance in predicting observed variant effects[5]. Recent efforts to predict combinatorial protein variant effects have adapted large neural networks from other domains, some with more than a billion parameters, to implicitly capture such dependencies when trained on protein sequences[6,10]. Such models can require not just vast amounts of data, but also costly compute resources, non-trivial tuning of hyperparameters and time for training. Overparameterized models are also prone to overfitting, notoriously hard to interpret, and hence can produce false positive predictions. For most proteins it is not clear how many and which dependencies are actually required for accurately predicting combinatorial protein variant effects. The complexity of the biological fitness landscape, *i.e.* how many dependencies need to be considered, will directly determine the required capacity of any model to approximate such fitness functions well.

To determine the importance of epistasis in protein fitness landscapes, we examined combinatorial variant effects across 6 proteins (using 7 separately collected datasets, including one generated here). We found that the measured combinatorial variant effects across these proteins can be well explained ($R^2$~0.78-0.95) by a function that only considers the 20*N residue mutation preferences, where N indicates the number of mutated positions, passed through a global nonlinearity[11], without considering specific dependencies between mutations. We observed that a small number of observations (5-fold oversampling the number of residue-wise mutation preference parameters, and, in one dataset, as few as 100-200 observations) is sufficient to enable high predictive accuracy on held-out combinatorial variant effects (Pearson $r > 0.8$), outperforming any unsupervised methods for predicting variant effects.

The high general applicability and performance of such per-residue mutation preference models led us to devise an unsupervised strategy that we term CoVES - for 'Combinatorial Variant Effects from Structure'. CoVES designs functional and diverse protein variants without experimental variant effect measurements by inferring the required residue-wise mutation preferences using an equivariant graph neural model which takes the structural context surrounding the residue as input (Fig. 1a). CoVES outperformed complicated, high-capacity neural models, which model long-range specific dependencies between primary sequence positions, in functional and diverse sequence design, when evaluated with near perfect oracle fitness functions for two proteins examined here (Fig. 1b-c). Collectively, our findings systematically illuminate the simplicity of local protein fitness landscapes in many cases, suggest minimum experimental measurement rules for enabling supervised protein variant design, and present a biologically-rooted unsupervised strategy, CoVES, to design combinatorial protein variants with negligible time and compute requirements in the absence of data. The performance of such microenvironment-only approaches suggests that structural context is sufficient for protein design.

Fig 1

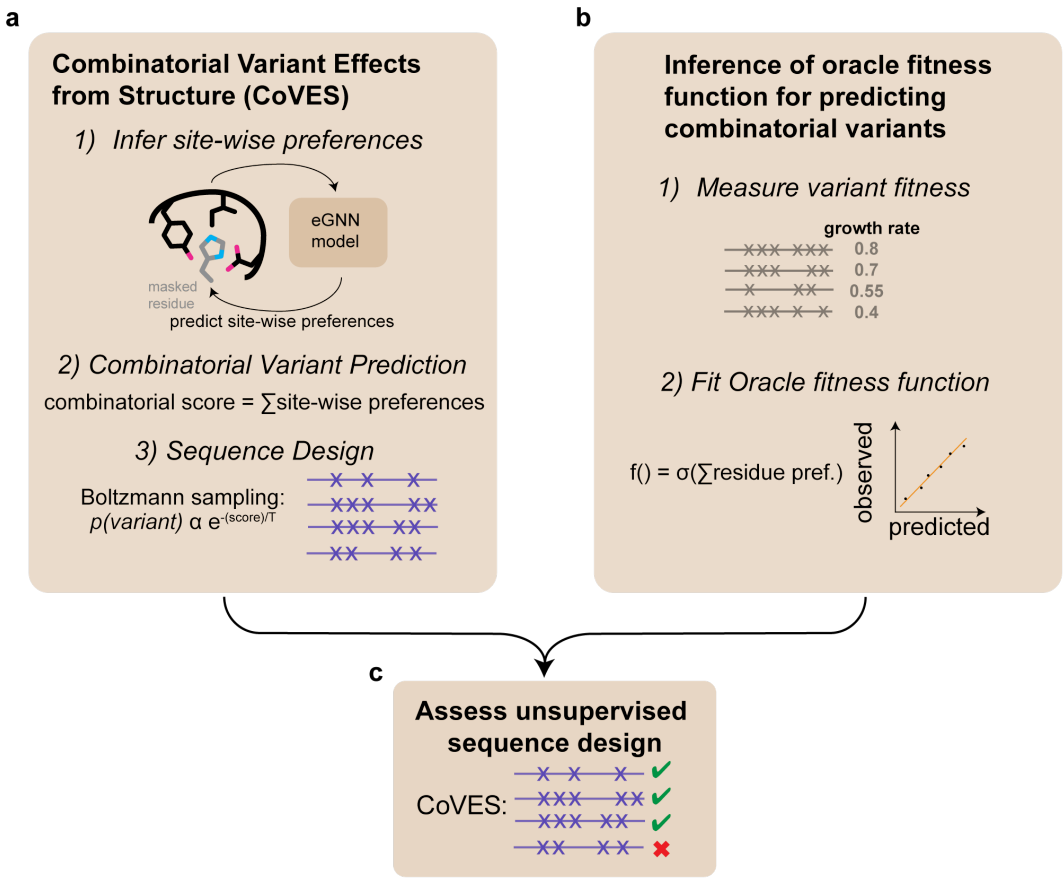



**Fig. 1: Design of protein sequences using structural information alone and assessment of designed sequences using oracle functions trained on experimental observations.**
**a,** Structural environment predicts residue mutation preferences and can be used for designing combinatorial protein variants.
**b,** Learning supervised fitness functions from experimental high-throughput variant measurements. The functional form of the fitness function, f(), can be learned by fitting to observed data, and enables predicting the function of unobserved sequence variants.
**c,** Assessment of designed sequences with oracle fitness function enables comparing different sequence design strategies.

97

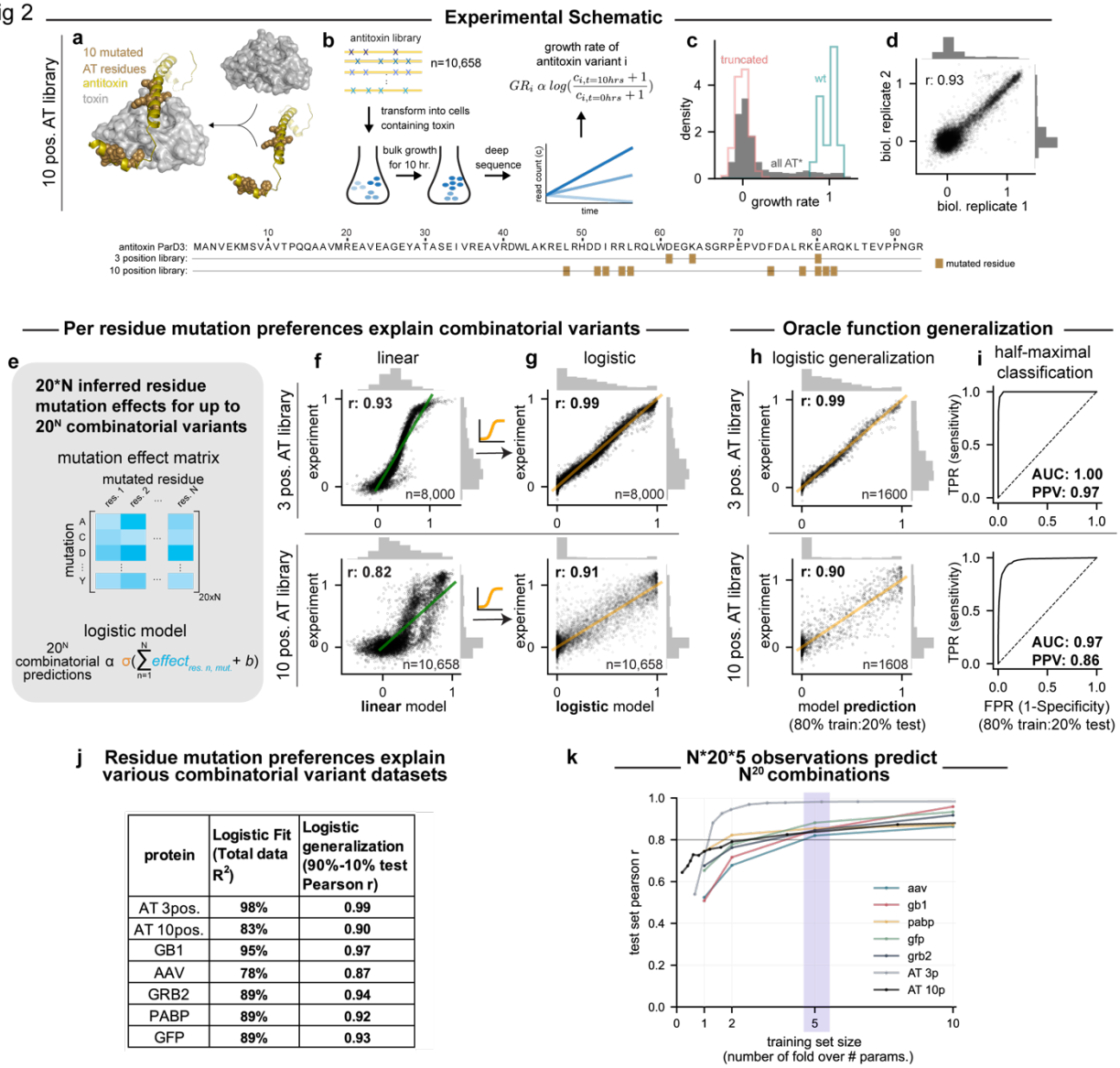
98 **Results**

99 ***High-throughput in vivo measurement of combinatorial variant effects in ParD3***

100 To understand how individual substitutions in proteins impact each other, we first focused on
101 measuring the effect of combinatorial variants in the *Mesorhizobium opportunistum* antitoxin
102 ParD3, which was randomly mutated at 10 positions and assayed for neutralization of its cognate
103 toxin ParE3 (Fig. 2a). This set of 10 residues was chosen as they directly contact the toxin, and are
104 among the top co-evolving residue pairs across the toxin-antitoxin interface[11]. A library of cells
105 containing different antitoxin variants was grown in bulk; only those cells containing antitoxin
106 variants that neutralize the toxin can proliferate. The change in frequency of each antitoxin variant
107 was followed via high-throughput sequencing over time (Fig. 2b). These variant effect
108 measurements correlate with orthogonal growth rate measurements and have high reproducibility
109 between biological replicates[11–13]. We used this assay to measure the growth rate effects of 7,923
110 combinatorial amino acid variants spanning the 10 binding residue positions and 2,615 truncated
111 antitoxins with stop codons (Fig. 2a, Figure S1, Table S1). We calculated the growth rate of each
112 variant as the normalized log read ratio before and after selection, and found good separation
113 between truncated and wild-type antitoxin variant effects (Fig. 2c). This assay exhibited high
114 reproducibility between separate biological replicates (Pearson r: 0.93, Fig. 2d), but we note that
115 nonfunctional antitoxin variants with growth rate values below ~0.2 could not be resolved. As
116 expected, the distribution of fitness effects shifted towards loss of function as more substitutions
117 were introduced, with the majority of variants achieving at least half-maximal toxin neutralization
118 when fewer than three substitutions were present (Fig. S1a, Supplementary Table 1). With more
119 than three substitutions, the fraction of functional variants decreases faster than exponentially (Fig.
120 S1b). Using this maximum likelihood estimate of the fraction of functional variants and
121 considering the total number of possible mutations at each mutation distance, we estimated that
122 there are ~$3x10^{10}$ combinatorial variants – out of a possible $20^{10}$~$10^{13}$ combinatorial variants
123 among these 10 residues - that achieve half-maximal neutralization when these ten positions are
124 mutated (Fig. S1c).

125

Fig 2




| protein | Logistic Fit (Total data $R^2$) | Logistic generalization (90%-10% test Pearson r) |
|---|---|---|
| AT 3pos. | 98% | 0.99 |
| AT 10pos. | 83% | 0.90 |
| GB1 | 95% | 0.97 |
| AAV | 78% | 0.87 |
| GRB2 | 89% | 0.94 |
| PABP | 89% | 0.92 |
| GFP | 89% | 0.93 |

**Fig. 2: Residue amino acid preferences can explain combinatorial mutation effects for multiple proteins and enable predicting the function of unobserved combinatorial variants.**

**a-d**, Ten binding residues (AT: L48, D52, I53, R55, L56, F74, R78, E80, A81, R82; n=10,658) of the antitoxin ParD3 were randomized (shown space-filled on PDB ID:5CEG, panel **a**) and highlighted on the antitoxin sequence (bottom), transformed into cells containing wild-type toxin ParE3, and the growth of individual antitoxin variants followed by high-throughput sequencing over time to calculate the normalized log read ratio (growth rate, GR) for each variant (**b**). Antitoxin variants that are able to bind and neutralize the toxin will show higher growth rates.The distribution of measured growth rate values for all antitoxin variants, wild-type antitoxin, and truncated antitoxins is shown (**c**). The reproducibility of growth rate values between two biological replicates (**d**).

**e-g**, The logistic regression model learns the 20*N per residue mutation effects (**e**) before passing through a sigmoid function (orange) to predict $20^N$ combinatorial variants. Logistic regression (**g**) fits the observed combinatorial variants better than linear regression does (**f**). Top row shows fit to 3 position randomized antitoxin library, bottom row shows 10 position randomized antitoxin library.

**h-i**, The logistic regression model enables predicting held-out 20% of the random combinatorial variants (**h**), and enables classification of half-maximal neutralization (**i**).

**j**, Total variance explained and held-out correlation of site-wise logistic regression model across 7 combinatorial variant datasets.

**k**, A subset of the observed combinatorial variants is sufficient to infer the site-wise preference parameters to explain the remaining held-out combinatorial variant effects.

126

127 ***A simple, logistic regression model explains combinatorial variant effects in the***
128 ***antitoxin ParD3***

129 We sought to understand how well observed combinatorial variant effects could be explained by
130 considering only per position amino acid preferences, without specific interaction terms between
131 mutated residues. To do so, we examined two datasets that measured combinatorial variant effects
132 in the ParD3 antitoxin for cognate toxin neutralization (Fig. 2a): The 10 position library dataset
133 generated in this study, as well as a 3-position, combinatorially exhaustive library from Ding *et*
134 *al.*[11] (8,000 combinatorial variants at positions D61, K64 and E80, with Pearson r = 0.98 between
135 biological replicates). We first fit a linear regression model to each of these datasets, with per
136 residue mutation preferences as predictors, but without specific interaction terms between residues.
137 This model estimates a latent residue preference effect for each individual amino acid variant at
138 each position, and sums these effects to predict combinatorial variants. This approach performed
139 well (3 position dataset Pearson r: 0.93, 10-position dataset: 0.82), but shows markedly biased
140 residuals between fitted and observed growth rate effects (Fig. 2f, Fig. S2a). We also examined
141 whether simply adding the observed single variant effects are predictive of combinatorial variant
142 effects (Fig. S2b). The predicted and observed growth rate effects correlate poorly (3-position
143 dataset Pearson r: 0.48, 10-position dataset: 0.57).

144 We then fit a nonlinear, site-wise model (*i.e.*, a logistic regression model with only per residue
145 amino acid preferences as predictors, Fig. 2e,g; Fig S2c) to each dataset. This model infers
146 mutational preferences for each residue using the observed combinatorial variants, sums the
147 relevant inferred preference parameter for each mutation in a particular combination of variants,
148 and then passes this sum through a nonlinear sigmoid function to predict the effect of a
149 combinatorial variant (Fig. 2e, Fig. S2c). This sigmoid function could account for nonlinearities
150 arising from growth rate saturation or binding curves[11]. This model enabled significantly better,
151 and in the case of the 3 position library, almost perfect fit to the observed combinatorial variant
152 effects (Fig. 2g; 3-position dataset: 0.98 Pearson r or 98% explained variance $R^2$; 10-position
153 dataset: 0.91 Pearson r or 83% explained variance $R^2$), without biased residuals between fitted and
154 observed combinatorial variant effects.

155 Importantly, this non-linear, mutation preference model used a linearly increasing number of
156 parameters (20*N preference parameters + 1 bias parameter, where N indicates the number of

157 mutated positions) to explain an exploding number of combinatorial variants ($20^N$) (Fig. 2e), *i.e.*
158 the 8,000 (=$20^3$) observed combinatorial variant effects in the antitoxin 3 position library were
159 explained by a logistic regression model with 61 (=20*3 residue mutation preference +1 bias)
160 parameters.

161 ***Per residue mutation effect models can explain combinatorial variant effects across***
162 ***proteins***

163 We wondered whether the performance of such non-epistatic, per residue mutation effect models
164 for fitting observed combinatorial variants generalizes to other protein functions and folds. Indeed,
165 the explanatory power of such models have been described in previous cases[11,14–17]. To
166 systematically examine the applicability of such models, we identified 5 existing variant datasets
167 that contain combinatorial variant effect measurements, including variants in the IgG-binding
168 domain of protein G (GB1[18]), green fluorescent protein (GFP[16]), poly(A)-binding protein
169 (PABP[19]), the SH3-domain of the human growth factor receptor-bound protein 2 (GRB2[20]), and
170 model-designed combinatorial variant effects in adeno-associated virus (AAV[21]). We then fit a per
171 residue, nonlinear regression model to these variant datasets (Fig. S3a-b, Table S2). Even though
172 these assays have varying experimental noise, we observe that similar to the toxin-antitoxin case,
173 a nonlinear, non-epistatic mutation preference model could explain 78-95% of the observed mutant
174 effects across these datasets (Fig. 2j). A linear regression model with the same number of mutation
175 preference parameters, but lacking the non-linear transformation function performed less well in
176 all cases (Fig. S3a). These results suggest that summed per residue mutation preferences passed
177 through a nonlinear transformation, without considering specific dependencies between mutations,
178 are a powerful tool for explaining combinatorial variant effects across multiple proteins.

179 ***Few observations are sufficient to predict unobserved combinatorial mutation effects***

180 Because we were able to explain a large number of combinatorial variant effects using few
181 mutation preference parameters, we next asked whether a smaller subset of random observations
182 would be sufficient to infer these preferences. We repeated inference of the mutation preferences
183 of the nonlinear, residue mutation effect regression model using smaller, random subsets of the
184 total observed dataset in each case, and evaluated these models for predicting the fitness effect of
185 combinatorial variants in the remaining dataset. Indeed, models trained on a much smaller number
186 of random observations were sufficient to explain the unobserved combinatorial mutation effects

well (Fig. S4). Strikingly, for the ParD3 antitoxin 3 position library, 100 or 200 random combinatorial variant effect measurements were sufficient to achieve almost perfect correlation (Pearson r = 0.98) between the observed and predicted held-out combinatorial variants (Fig. S4a top).

We hypothesized that the required training set size is dependent on the number of randomized positions, and the resulting number of mutation preference parameters, which in turn differ as a function of the number of mutated positions in each dataset. We therefore visualized the performance of models trained on a subset of datapoints relative to the number of mutation preference parameters in each model. This demonstrated that a mere oversampling of 5-fold in terms of training set size relative to the number of parameters in all 7 datasets was sufficient to enable prediction of unobserved combinatorial variants with a Pearson correlation coefficient r > 0.8, outperforming state-of-the-art unsupervised models[9] (Fig. 2k).

These results indicate that measuring only 5*N*20 combinatorial variant effects are sufficient to fit the N*20 mutation preference parameters of this non-epistatic, nonlinear model at N residues in order to predict observed combinatorial variants effects across 7 protein variant effect datasets.

### *Mutation preference non-linear regression can serve as an 'oracle fitness function' to predict the effect of unobserved combinatorial variants*

We also tried to estimate how well such per residue, nonlinear regression models can predict unobserved variant effects. For example, how well can this model trained on ~8,000 antitoxin variants at ten randomized positions be used to predict the possible $20^{10}$ possible combinatorial variants? To estimate the generalization error beyond the observed variants, we inferred parameters from a random 80% subset of the total observed combinatorial variants, and tested the predictive accuracy of the model on the held-out remaining 20% of observed variants. The correlation between predicted and measured held-out test variants was high (Fig. 2h, Pearson r: 0.99 for the 3 position library and Pearson r: 0.90 for the 10 position library). This model had high performance for classifying held-out combinatorial variants that achieve at least half-maximal fitness (3 position antitoxin library AUROC: 1.00, positive predictive value (PPV): 0.97; 10 position antitoxin library AUROC: 0.97, PPV: 0.86). These findings suggest that the non-epistatic logistic model can be used as an 'oracle fitness function' to predict the effect of unobserved combinatorial variants among these ten antitoxin positions.

***A per residue, unsupervised method, CoVES, trained on the structural microenvironments around each residue, can recover mutation preferences and score observed combinatorial variant effects across datasets.***

Is it possible to recover the required per residue mutation preferences and predict combinatorial variant effects without measurements? We observed that the inferred per residue mutation preferences in the 3 position antitoxin library can be altered by mutations in contacting residues in the toxin (Fig. S5). Given this and the performance of non-epistatic models to capture combinatorial variants, we hypothesized that the three-dimensional structural environment around a particular amino acid could be sufficient to learn the per residue mutation preferences of interest and thereby predict combinatorial variant effects.

To test whether the structural microenvironment is sufficient to learn per residue mutation preferences and predict combinatorial variants, we developed a strategy - which we call CoVES for 'Combinatorial Variant Effects from Structure' (Fig. 3a) - by training a graph neural network[22] that learns rotation equivariant transformations to predict the identity of a masked amino acid from its atomic level 3D environment across non-redundant structures from the Protein Database (PDB) (test accuracy = 52.8%, compared to ~5% for random guess, training and testing dataset from Townshend *et al.*[23]). We then predicted the amino acid preference at each of the mutated positions across the combinatorial variant datasets given their respective structural environments (Fig. 3a). Indeed, this graph neural network was able to recover the per residue mutation preferences which explained the observed combinatorial variant effects in the antitoxin (Fig. S6, average pearson r across 3-position antitoxin library residues: 0.80, average pearson r across 10-position antitoxin library residues: 0.57). To predict combinatorial variant effects, we summed the log probability of amino acid preferences at each mutated position to predict the effects of combinatorial variants. This strategy was able to predict observed combinatorial mutation effects similarly well to other state-of-the-art unsupervised variant effect predictors across 5 variant effect datasets with proteins containing structural information (Fig. 3b, Fig. S7).

Fig 3

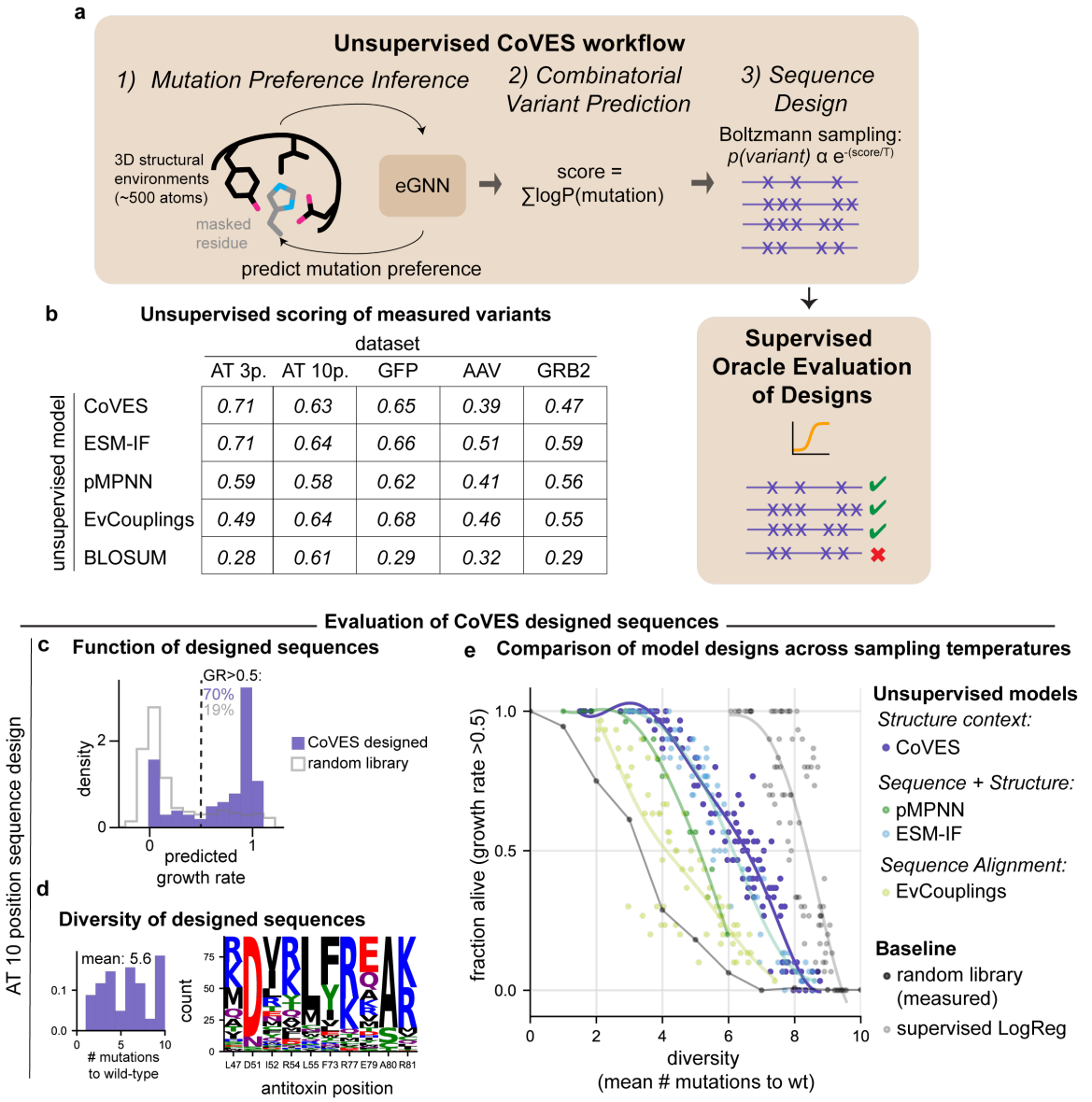



**Fig. 3: CoVES, an unsupervised approach to learn residue mutation preferences from structural microenvironments, can predict variant effects and generatively design viable and diverse sequences.**

**a,** Schematic of the CoVES workflow: First, an equivariant graph neural network[19,20] is used to predict amino acid preferences from the structural environment around a particular residue. The mutation preference for each residue are converted into log probabilities, and summed to predict combinatorial variant effects. Finally, these scores can be used to design combinatorial variants at the desired residue positions by sampling with the Boltzmann energy function. Designed sequences are evaluated using the supervised oracle function which is trained on observed variant effects.

**b**, Spearman correlation coefficients between unsupervised model scores and observed combinatorial variant effects.

**c-d**, CoVES designed antitoxin sequences evaluated for their predicted functionality and diversity. Generated sequences (n=91 unique sequences, temperature = 1.5) from CoVES are evaluated for their growth rate effects (**c**), and their mutations with respect to the wild-type antitoxin (**d**).

**e**, Comparison of generated sequences using CoVES vs. other state-of-the-art sequence design models in terms of the fraction of generated antitoxin sequences predicted to be functional and their average number of mutations as the sampling temperature is varied. Each dot represents a collection of sampled sequences summarized by their average number of mutations and their predicted fraction of functional sequences. Random library measurements indicated in dark grey, and energy-based Boltzmann sampling from the supervised, site-wise only logistic regression model trained on the observed variant data indicated in light grey. Lines represent polynomial fits.

245 ***CoVES trained on structural microenvironments can generate diverse and functional***
246 ***antitoxin sequences.***

247 The high predictive performance of the supervised 'oracle function' on held-out observed
248 combinatorial variants in the 10 position antitoxin library (Fig. 2h-j) gives us an opportunity to
249 assess the function of designed sequences that are not observed in our high-throughput assays. We
250 assessed whether the unsupervised, structural context-only CoVES approach was able to design
251 diverse and functional sequences at these 10 antitoxin positions (Fig. 3a). This task differs from
252 combinatorial variant score prediction as emphasis is placed on sampling truly high functioning
253 mutations, rather than predicting the entire range of mutation effects. Given the amino acid
254 preference scores learned from each mutated residue's structural environment, we used a
255 Boltzmann energy function to sample sequences (n=500) at various temperatures, t, to control the
256 diversity of the generated sequences. We then assessed the predicted growth rate effect of these
257 generated sequences using the above supervised oracle function, as well as the diversity of
258 generated samples by calculating the number of mutations between sampled sequences and with
259 respect to the wild-type sequence. At temperature t=1.5, CoVES generated 91 unique sequences
260 of which ~70% are predicted by the oracle fitness function to achieve half-maximal fitness (Fig.
261 3c). These sequences showed an average of 6.7 substitutions between samples and 5.6 with respect
262 to the wild-type antitoxin (Fig. 3d). Inspecting the generated sequences indicated that multiple
263 different substitutions can occur at each position (Fig. 3d).

264 How well does this per residue, unsupervised strategy compare to ones that can consider arbitrary
265 specific dependencies between residues? To address this question, we compared generated
266 sequences from CoVES against sequences designed from two models that can consider the
267 sequence context in an autoregressive manner (proteinMPNN from Dauparas *et al.*[24], Ingraham *et*
268 *al.*[25]; ESM-IF from Hsu *et al.*[8]). Because sampling temperatures do not have a 1:1 correspondence
269 between models and sampling algorithms, we sampled sequences for each model at a range of
270 temperatures and assessed each set of sampled sequences for the fraction that is predicted to be
271 functional (at least half-maximally neutralizing) by the 'oracle' fitness function, as well as their
272 diversity. We found that CoVES outperformed both of these models for generating functional
273 sequences that diverge from each other and with respect to the wild-type antitoxin sequence (Fig.
274 3e, Fig. S8). All of these unsupervised models, which consider structural features, also

275 outperformed simulated annealing sampling from EvCouplings (Fig. 3e yellow), which learns site-
276 wise and pair-wise mutation preferences from the natural sequence alignment, and outperformed
277 model-free random sampling of sequences observed in the combinatorial variant library (Fig. 3e
278 dark grey). We note that simple Boltzmann sampling using the scores from the supervised 'oracle'
279 model, *i.e.*, the logistic regressor trained on observed combinatorial variant effects to learn amino
280 acid preferences, outperformed all unsupervised strategies outlined above (Fig. 3e, light grey).

281 ***CoVES can generate diverse and functional GFP sequences.***

282 We next asked whether CoVES sampling can generate functional and diverse sequences in a
283 different protein fold and function, as well as across more than fixed 10 residues. We chose to
284 focus on designing GFP sequences, for which an accurate oracle function could be trained using
285 random mutation data from ref[16], to predict 10% held-out combinatorial variant datasets with up
286 to 15 mutations spread over 233 GFP residues (Pearson r: 0.93, Fig. 4a). Almost all combinatorial
287 held-out variants that are predicted by the oracle function to achieve at least half-maximal
288 fluorescence do achieve this threshold in the experiment (Fig. 4b-c, positive predictive value=0.99,
289 AUROC for half-maximal fluorescence classification: 0.99). We note that this oracle function is
290 expected to perform well only for in-distribution generalization, *i.e.* for predicting combinatorial
291 variants consisting of individual mutations that are observed during training (1810 out of 4427
292 residue mutations at 233 residues due to error-prone PCR library construction limitations), and for
293 the limited range of observed variant mutation numbers (range of mutations: 1-15).

294 We used CoVES, as well as ESM-IF and proteinMPNN for sampling combinatorial GFP variants
295 at various temperatures and randomly chosen sets of positions to mutate. We then filtered the
296 generated sequences for the ones that the oracle can be expected to predict well, and evaluated
297 them for the fraction predicted to be at least half functional, as well as their average mutation
298 distances. Sequences designed with CoVES had a higher fraction of functional sequences and
299 higher average number of mutations than sequences designed by both ESM-IF and proteinMPNN,
300 as well as the measured random error-prone PCR library sequence samples (Fig. 4d-e, Fig. S9).

301 These results indicate that, in the absence of observed combinatorial variant measurements, simple
302 and fast sampling from a per residue, unsupervised model that learns amino acid preferences from
303 the respective amino acid microenvironments is sufficient to generate functional and diverse
304 combinatorial variants at ten ParD3 antitoxin residues as well as across 233 residues with up to 15

305 mutations in GFP, and outperforms not just random sampling from experimental libraries but also
306 autoregressive models and their sampling strategies.

Fig. 4

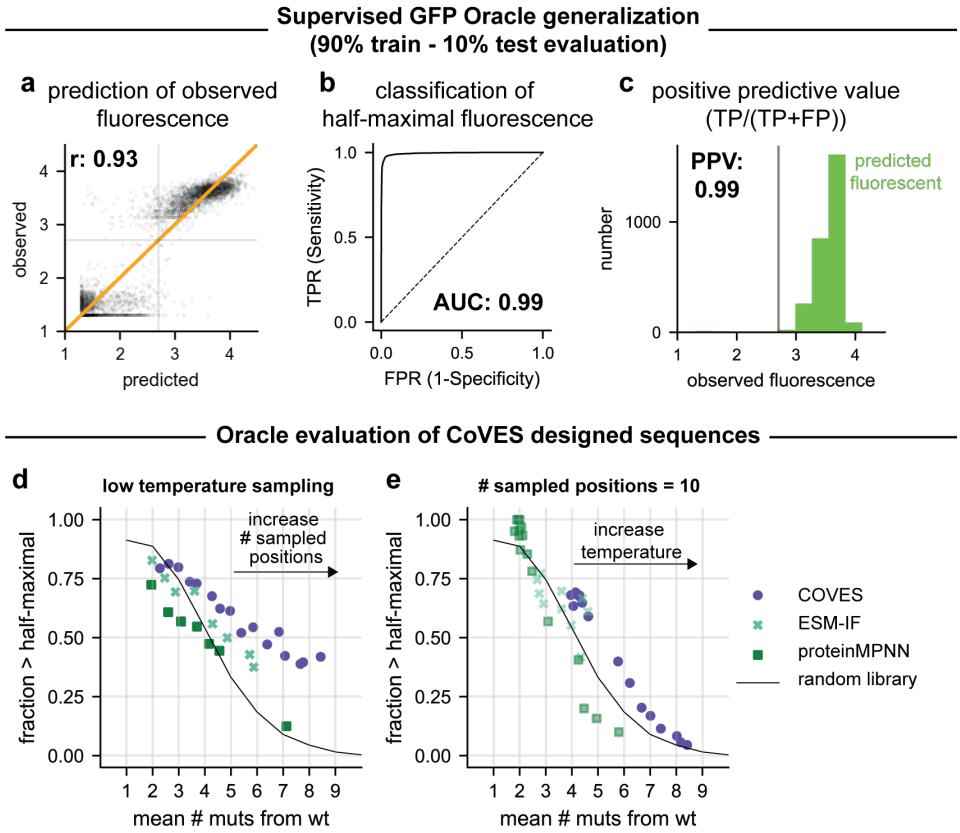



**Fig. 4: Evaluation of function and diversity of CoVES-designed GFP variants.**
**a-c,** GFP nonlinear regression oracle performance evaluated on held-out 10% test data, when trained on 90% random training data. **a**, The correlation between oracle predicted and observed fluorescence values is shown. The pearson correlation coefficient r is indicated. **b**, The receiver-operating curve of half-maximal fluorescence classification is shown. **c**, The positive predictive value, ie. fraction of truly above half-maximal fluorescing variants over the total number of predicted above half-maximal fluorescing GFP variants is shown for the held-out 10% training data.
**d-e**, The diversity and function of sampled sequences from unsupervised models is shown. Each dot represents a set of sampled sequences at a given temperature and a set number of positions that are randomized, summarized by the fraction of sequences predicted to be above half-maximal fluorescence and the average number of mutations. **d**, A low sampling temperature (CoVES T=0.5, ESM-IF T=1.0, proteinMPNN T=1.5) is used while the number of random mutated positions (out of 233 GFP residues) is increased. **e**, The maximum number of mutated positions is set to 10, and the sampling temperature is increased. Only combinatorial variants consisting of individual variants, which the oracle logistic regression has been trained on and has been tested on in panels **a-c,** are shown.

307

## Discussion

We found that a large fraction of observed combinatorial variant effects across 6 proteins, including therapeutically relevant ones, can be explained with an interpretable, simple model that only considers per residue mutational preferences, *i.e.*, without explicit terms for interactions between mutations. While previous studies have described such supervised additive models with a nonlinear transformation (termed 'global epistasis' or 'nonlinear correction' models) for prediction[14–17,26,27] and supervised protein design[21], here we show that this insight can be exploited for generalizable and unsupervised protein design. Specifically, we show that one can effectively design functional and diverse variants just with independent accounting of residue micro-environments using CoVES, and that this approach can outperform state-of-the-art high-capacity neural methods in protein design. While our assay conditions do not permit measuring gain-of-function antitoxin variants (due to saturation in measured growth rate effects), it is likely that such effects can be captured by such residue preference models. Indeed, the structural microenvironment has been used successfully to identify gain-of-function mutations before[28].

The performance of such mutation preference models, which do not explicitly capture dependencies between mutated residues, does not exclude the existence of higher order epistasis. First, the per residue mutation preferences inherently capture an implicit notion of dependence on neighboring residues, and we indeed observe that mutations at contacting residues can alter the mutation preferences at a given residue. Second, while 78-98% of the observed combinatorial variant effects can be explained by mutation preferences alone, it is likely that some of the remaining variation in some datasets will be explained by true biological specific dependencies between residues. To identify such dependencies requires a careful treatment of the experimental noise and considerations of statistical power in each selection system, such as done in refs[11,17], and experimental validation of significant epistatic dependencies, which we leave for future work.

A number of observations suggest that the close structural context is largely sufficient for variant effect prediction and design: 1) The performance of CoVES, which samples and scores residue preferences solely on the local structural context, over autoregressive methods that can learn arbitrary mutation dependencies, suggests the structural context is sufficient. 2) We observed that site-wise preferences in the supervised global epistasis models can be altered by mutations in contacting residues. 3) We observed that the explanatory power of the nonlinear, per residue

338 mutation preference model is the least high for combinatorial variant datasets in which spatially
339 close residue are mutated (AAV and 10 position antitoxin dataset). In contrast, the performance is
340 strikingly high for datasets in which strictly or largely non-contacting residues are mutated (*i.e.*
341 when mutating a small number of random positions spread across an entire domain). 4) Positive
342 deviations from the nonlinear, per residue mutation preference model predictions of combinatorial
343 variants (positive epistasis) have been shown to be in close contact[11], and can indeed be used to
344 fold proteins[29,30]. Together, these observations suggest that future efforts in modeling protein
345 variant effects should focus on considering the structural context of a given protein.

346 While we demonstrated performance of such residue mutation preference models in datasets that
347 contain variants with up to 10, 15, or 21 mutations (for the antitoxin, GFP and AAV, respectively),
348 it will be interesting to examine for which number of mutations and choice of residues such simple
349 residue mutation models can be used. Choosing special residues will affect performance: Previous
350 efforts examining mutations at 13 active site residues of GFP[17], have demonstrated that dominant
351 higher order dependencies are required to explain combinatorial variant effects well. Our
352 observations suggest that choosing residues with non-contacting, 'independent' structural
353 microenvironments will enable simple, per residue mutation effect models to predict exploding
354 numbers of combinatorial variant effects well. Nonetheless, because the supervised mutation effect
355 predictors show almost perfect classification accuracy for half-maximal function of GFP and
356 antitoxin variants, we believe that such models can be used as useful additional benchmarks to
357 overcome the limitation of evaluating sequence design methods only by recovery of the wild-type
358 sequence for a given structure[8,25].

359 For the supervised residue mutation effect models, we were able to learn 'rules' for choosing
360 minimum numbers of experimental variant measurements necessary for accurate combinatorial
361 protein variant effect prediction. Specifically, we found that a number of random combinatorial
362 variant observations that five-fold oversample the number of mutational preference parameters
363 (i.e., random observations = N*20*5, where N is the number of mutated residues) was sufficient
364 to give high accuracy in predicting held-out variant effects (>0.8 pearson correlation coefficient)
365 across 7 combinatorial variant effect datasets, outperforming even the best unsupervised sequence-
366 based variant effect predictors[9]. Choosing a low number of mutated, non-contacting sites could
367 enable predictive tuning of combinatorial variant effects for proteins where high-throughput
368 measurements are not possible, and instead rely on medium throughput methods, such as plate-

369 based screening ones. Even in the case of adeno-associated virus, where neighboring residues are
370 mutated, the supervised non-linear, residue mutation preference model was shown to outperform
371 high-capacity neural models for design[21].

372 It will be interesting to examine CoVES performance at higher numbers of mutations, or even in
373 full sequence design. In the case of full sequence design, the exact atom-neighborhood around a
374 position is unknown because neighboring residue's side-chains are unknown. Here, it could be
375 possible to use an equivariant graph neural network trained on predicting residue mutation
376 preferences given only the surrounding backbone atoms instead[31]. Alternatively, iteratively
377 choosing residues using CoVES for each site, followed by structural modeling of neighbouring
378 side-chain positions could enable generating functional sampled sequences. Additionally, future
379 work could examine using predicted structures[32] instead of experimentally measured structures to
380 infer such per residue mutation preferences.
381 We note that the assessment of generated sequences depends both on model quality as well as
382 sampling strategies. Improvements in sampling alone could boost the fraction of generated
383 functional sequences. For example, the current sampling algorithm for the ESM-IF model[8] works
384 in an autoregressive manner from the N-terminus without consideration of amino acid identities
385 towards the C-terminus of the currently sampled amino acid.

386 In sum, our results highlight the power of simple, interpretable residue mutation preference models
387 for explaining observed combinatorial variant effects, their use as an 'oracle' fitness function to
388 predict unobserved combinatorial variants, even when trained on a small number of observations,
389 and present a structure-based design method, CoVES, for unsupervised protein design applications
390 that can outperform high-capacity approaches.

391

392

393

394

395

396

397 **References**


398 1. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the
399 evolution of an influenza protein. *eLife* **2**, e00631 (2013).

400 2. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution
401 of influenza oseltamivir resistance. *Science (New York, N.Y.)* **328**, 1272–5 (2010).

402 3. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in
403 sequence space. *Trends in Genetics* **31**, 24–33 (2015).

404 4. Kauffman, S. A. & Weinberger, E. D. The NK model of rugged fitness landscapes and its
405 application to maturation of the immune response. *Journal of Theoretical Biology* **141**, 211–
406 245 (1989).

407 5. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation
408 capture the effects of mutations. *Nat Methods* **15**, 816–822 (2018).

409 6. Madani, A. *et al.* Large language models generate functional protein sequences across diverse
410 families. *Nat Biotechnol* (2023) doi:10.1038/s41587-022-01618-2.

411 7. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning
412 to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).

413 8. Hsu, C. *et al. Learning inverse folding from millions of predicted structures*.
414 http://biorxiv.org/lookup/doi/10.1101/2022.04.10.487779 (2022)
415 doi:10.1101/2022.04.10.487779.

416 9. Notin, P. *et al.* Tranception: protein fitness prediction with autoregressive transformers and
417 inference-time retrieval. Preprint at http://arxiv.org/abs/2205.13760 (2022).

418 10. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for
419 protein design. *Nat Commun* **13**, 4348 (2022).

420 11. Ding, D. *et al.* Co-evolution of interacting proteins through non-contacting and non-specific
421 mutations. *Nat Ecol Evol* **6**, 590–603 (2022).

422 12. Aakre, C. D. *et al.* Evolving New Protein-Protein Interaction Specificity through Promiscuous
423 Intermediates. *Cell* **163**, 594–606 (2015).

424 13. Lite, T. L. V. *et al.* Uncovering the basis of protein-protein interaction specificity with a
425 combinatorially complete library. *eLife* **9**, 1–57 (2020).

426 14. Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis.
427 *Proceedings of the National Academy of Sciences of the United States of America* **115**, E7550–
428 E7558 (2018).

429 15. Pokusaeva, V. O. *et al.* An experimental assay of the interactions of amino acids from
430 orthologous sequences shaping a complex fitness landscape. *PLoS Genetics* **15**, 1–30 (2019).

431 16. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**,
432 397–401 (2016).

433 17. Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking
434 genotype and phenotype in a protein. *Nature Communications* **10**, 1–11 (2019).

435 18. Olson, C. A., Wu, N. C. & Sun, R. A Comprehensive Biophysical Description of Pairwise
436 Epistasis throughout an Entire Protein Domain. *Current Biology* **24**, 2643–2651 (2014).

437 19. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational
438 scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA*
439 **19**, 1537–1551 (2013).

20. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).

21. Bryant, D. H. *et al.* Deep diversification of an AAV capsid protein by machine learning. *Nat Biotechnol* **39**, 691–696 (2021).

22. Jing, B., Eismann, S., Soni, P. N. & Dror, R. O. Equivariant Graph Neural Networks for 3D Macromolecular Structure. *arXiv:2106.03843 [cs, q-bio]* (2021).

23. Townshend, R. J. L. *et al.* ATOM3D: Tasks On Molecules in Three Dimensions. *arXiv:2012.04035 [physics, q-bio]* (2021).

24. Dauparas, J. *et al. Robust deep learning based protein sequence design using ProteinMPNN*. http://biorxiv.org/lookup/doi/10.1101/2022.06.03.494563 (2022) doi:10.1101/2022.06.03.494563.

25. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. *33rd Conferece on Neural Information Procssing Systems (NeurIPS 2019)* 12.

26. Tareen, A., Posfai, A., Ireland, W. T., Mccandlish, D. M. & Kinney, J. B. MAVE-NN : learning genotype-phenotype maps from multiplex assays of variant effect. *bioRxiv* 1–19 (2020).

27. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, e32472 (2018).

28. Lu, H. *et al. Deep learning redesign of PETase for practical PET degrading applications*. http://biorxiv.org/lookup/doi/10.1101/2021.10.10.463845 (2021) doi:10.1101/2021.10.10.463845.

29. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nature Genetics* **51**, 1177–1186 (2019).

462 30. Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nature Genetics*
463 **51**, 1170–1176 (2019).

464 31. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. & Dror, R. Learning from Protein
465 Structure with Geometric Vector Perceptrons. *arXiv:2009.01411 [cs, q-bio, stat]* (2021).

466 32. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–
467 589 (2021).

468 33. Stiffler, M. A., Subramanian, S. K., Salinas, V. H. & Ranganathan, R. A protocol for functional
469 assessment of whole-protein saturation mutagenesis libraries utilizing high-throughput
470 sequencing. *Journal of Visualized Experiments* **2016**, 1–11 (2016).

471 34. Warren, D. J. Preparation of highly efficient electrocompetent Escherichia coli using
472 glycerol/mannitol density step centrifugation. *Analytical Biochemistry* **413**, 206–207 (2011).

473 35. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome
474 assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

475 36. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source
476 tool for metagenomics. *PeerJ* **4**, e2584 (2016).

477 37. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. in *12th USENIX*
478 *Symposium on Operating Systems Design and Implementation* 265–283 (2016).

479

480

481 ## Material and Methods

482 ### *Bacterial strains, vectors and media*

483 *E. coli* TOP10 strains were grown at 37 °C in M9L medium (1x M9 salts, 100 μM $CaCl_2$, 0.4%
484 glycerol, 0.1% casamino acids, 2 mM $MgSO_4$, 10% v/v LB). Antibiotics were used as follows: 50
485 μg/ml carbenicillin, 20 μg/ml chloramphenicol in liquid media, and 100 μg/ml carbenicillin, 30
486 μg/ml in agar plates. The toxins ParE3 was carried as before[11] on the pBAD33 vector ($chlor^R$
487 marker, ML3302 for wild-type ParE3) with expression repressed or induced with 1% glucose and
488 L-arabinose at indicated concentrations, respectively, and the antitoxin ParD3 was carried on the
489 pEXT20 vector ($carb^R$ marker, ML3296) with expression induced by IPTG.

490 ### *Antitoxin ten position library construction*

491 To measure the combinatorial variant effects in the antitoxin at ten residues, we constructed two
492 sublibaries: one in which five residues are randomized, and one in which an additional five
493 positions are randomized. In this way, we can guarantee the presence of sufficient number of
494 variants with five or less substitutions, given the curse of dimensionality, in which random
495 sampling at ten positions will generate a distribution of variants for which most will have a high
496 number of substitutions. Both the ten position and five position libraries were constructed using a
497 2-step overlap-extension PCR protocol[33]. We first used primers DDP704+DDP142 (see
498 Supplementary Table 1) and DDP705+DDP141 to introduce five randomized positions in the wild-
499 type antitoxin pEXT20-*parD3* plasmid ML3296 (PCR cycling was: 30 sec. at 98°C; 20 cycles of:
500 10 sec. at 98°C, 20 sec. at 55°C, 1 min. at 72°C; 2 min. at 72°C, hold at 4°C; using KAPA). The
501 PCR products were pooled, diluted 1:100 and amplified using the outer primers DDP141+DDP142
502 to generate full length, mutated antitoxin sequence. On this PCR product, we then used the primers
503 DDP700+DDP142 and DDP705+DDP141 to introduce the next five randomized positions, and
504 used the above strategy to generate full length antitoxin ParD3 gene with 10 positions randomized.
505 We then cloned both the five and ten position randomized PCR product into the pEXT20 vector
506 using restriction digests with SacI-HF and HindIII-HF (NEB) and ligation using T4 DNA ligase
507 (NEB at 16 ºC for 16 hours with a 1:3 molar ratio of insert to vector. Ligations were dialyzed on
508 Millipore VSWP 0.025 μm membrane filters for 90 minutes before electroporating (2mm cuvettes
509 at 2.4kV) into TOP10 cells, made using the protocol from ref[34]. Cells were recovered in 1ml SOC
510 for 1 hour. We propagated each library with at least 500,000 transformants, checked by spot plating

511 1:10 serial dilutions of recovered cells on LB/carb/chlor/1% glucose plates. We grew $OD_{600}$ ~ 0.5
512 at 37 °C in M9L/carb/chlor/1% glucose, spun down (8000G, 5 minutes) and resuspended in 5ml
513 M9L/carb/chlor/1%glucose/20% glycerol for storage at -80 °C. We then made these cells
514 electrocompetent in replicate for each library, and transformed dialyzed wild-type toxin ParE3 into
515 these cells. Cells were propagated with at least 500,000 transformants, and grown up to OD~0.6,
516 before spinning down (8000G, 5 minutes) and resuspending in 5 ml
517 M9L/carb/chlor/1%glucose/20% glycerol. Cells were aliquoted into 1 ml tubes and flash frozen in
518 liquid nitrogen for storage.

519 ***High-throughput variant effect measurement***

520 On the day of growth rate measurements, aliquots from two separate transformations were thawed
521 and recovered in 50 ml M9L/carb/chlor/1% glucose at 30 °C for 3 hours. Subsequently, glucose
522 was removed by washing 4 times with M9L, and cells were ready for growth rate measurement.
523 Growth rate measurements were then performed as described previously[11]. Briefly, washed cells
524 were resuspended in 250ml M9L/carb/chlor/IPTG to induce antitoxin expression, and toxin
525 expression induced after 100 minutes by adding arabinose. Cells were diluted 1:10 with pre-
526 warmed media when their $OD_{600}$ reached ~0.3 to keep them in exponential growth throughout the
527 duration of the experiment. 50ml of the cultures were sampled at the time of toxin induction, and
528 10 hours after. These cells were miniprepped, and a high-input (200 ng plasmid DNA), low cycle
529 (14 rounds) PCR reaction performed to isolate amplicons of interest using primers DDP643-
530 645/DDP648-651/DDP654-657 to introduce Illumina multiplexing indices and adapters. We then
531 gel purified and sequenced these library amplicons as described previously[11]. We then performed
532 sequencing using 250 base-pair paired end reads using a Novaseq SP flowcell for each timepoint
533 and replicate.

534 ***Analysis of high-throughput sequencing data***

535 Paired-end reads were processed as described previously[11]. Briefly, paired-end sequencing reads
536 were merged using FLASH 1.2.11[35]. Merged reads were quality filtered based on their phred-score
537 using vsearch 2.13.0[36], with the following arguments: vsearch --fastq_filter {file_name} --
538 fastq_truncqual 20 --fastq_maxns 3 --fastq_maxee 0.5 --fastq_ascii 33 --fastaout
539 {output_file_name}.fasta. Reads were subsequently filtered for having defined mutations at the
540 desired sites only, and the frequency of each variant at each timepoint was counted. We then

541 calculated a log-read ratio for all variants with at least three reads pre- and post-selection with one
542 pseudocount, and normalized these log-read ratios to fall between 0 and 1 given the log read ratio
543 of truncated and wild-type antitoxin variants. We note that non-functional antitoxin variants that
544 have a growth rate around 0 cannot be resolved. This noise for the non-functional antitoxin variants
545 arises due to sequencing depth limitations in pre- and post-selection read counts: Variants might
546 drop from different pre-selection read counts to unobserved in the post-selection library and hence
547 show variance in their calculated log read ratios.

548 ***Nonlinear, residue mutation effect modeling of combinatorial variant effects***

549 We used a nonlinear per residue effect model implemented in Tensorflow 2[37] to model
550 combinatorial variant effects in the antitoxin as before[11]. We used one-hot encoding of amino acid
551 variants as a predictor and one additional bias feature, and fit weights associated with each single
552 amino acid mutant substitution as well as the bias parameter. The linear sum of these weights is
553 passed through a sigmoid function to predict the normalized growth rate effect between zero and
554 one for each combinatorial variant in the three and ten position library. We used the Adam
555 optimizer to minimize the mean squared error of predicted to measured normalized log read ratios
556 for each variant, until the training error stabilized. For each protein, a number of learning rates,
557 number of training epochs and batch sizes was tried to minimize the training error. For models fit
558 to GFP, AAV, GRB2, PABP, and GB1, we used an additional modification: The sigmoid output
559 is scaled and shifted with a learnable parameter each to enable modeling experimentally measured
560 values that do not fall within a range between 0 and 1. This also enabled higher flexibility in
561 choosing nonlinear effect. For example, the GFP fitted nonlinearity utilizes only the 'lower' part
562 of the sigmoid output function (Fig. S9a). It is possible to find better nonlinear transformations for
563 some of these proteins, but we observed that the scaled and shifted sigmoid function can suffice
564 for high explanatory power (Fig. 2j, Fig. S3). We note that this model cannot generalize to
565 combinatorial variants that contain individual substitutions that have not been observed during
566 training, ie. for the GFP dataset, only combinatorial variants that consist of the 1810 observed
567 individual residue mutations (out of a possible 4427 residue mutations at 233 residues) due to
568 error-prone PCR library construction limitations, can be predicted.

569 ***Inference of site-wise preferences from subsampled variant effect measurements and***
570 ***estimation of supervised oracle generalization error***

571 To test how well models trained on few observations can predict the remaining held-out observed
572 variants across 7 variant datasets, we retrained the above models on a subset of data and tested
573 their performance on the full dataset. In order to estimate the in-distribution generalization error
574 of predicting unobserved variants, we split the training dataset into 90% training data and assessed
575 the correlation between predicted and observed effects in the remaining 10%, respectively, of held-
576 out test data. We did not consider a validation set since there are no hyperparameters to be
577 optimized. To estimate the generalization error more robustly for the antitoxin 10 position (and 3
578 position) library, where the majority of combinatorial variants are non-functional, we trained on
579 an 80%-20% train test split, to increase the number of test examples that are able to neutralize the
580 toxin. Similarly, to estimate the error in classifying variants that achieve at least half-maximal
581 function in the antitoxin or GFP, we calculated area under the receiver operator curves and positive
582 predictive values (true positives/(true positive + false positives)) on the random held-out test sets.

583 ***Inferring changes in antitoxin ParD3 site-wise mutation preferences by contacting***
584 ***residues***

585 A previous study has examined the effect of the combinatorial three position antitoxin binding
586 residue library (at antitoxin residues D61, K64 and E80) in the presence of various toxin variants
587 (single residue toxin variants N99V, V75C, D55E, A66I, V5L, A66F, R100W, E87M)[11]. To
588 examine the site-wise preferences at each antitoxin position in each toxin variant background, we
589 fitted separate site-wise logistic regression models to the antitoxin library in each toxin variant
590 background. The inferred antitoxin residue mutation preferences were almost perfectly correlated
591 across 9 out of 10 toxin backgrounds, in which toxin substitutions do not contact any of the mutated
592 antitoxin residues, but deviated in the background of toxin E87M (Fig. S5). Inspection of the site-
593 wise preferences revealed deviations specifically for antitoxin residue K63, which is found to be
594 directly contacting the toxin E87M position (Fig. S5a).

595 ***Estimating the total number of functional variants***

596 We estimated the total number of functional antitoxin sequences with at least half-maximal
597 neutralization of the toxin using two orthogonal approaches, with both approaches predicting ~$10^{10}$

598 functional sequences. The first estimate is based on the empirically measured distribution of fitness
599 effects (Fig. S1). Here, we calculated the fraction of sampled sequences that are measured to be
600 functional at each mutation distance from the wild-type sequence, and multiplied this fraction by
601 the total number of sequences that exist at this mutation distance. The total number of functional
602 variants is the sum of all variants at each mutation distance. This approach assumes that the
603 sampled sequences are a representative sample of all possible random mutations at each mutation
604 distance, and does not consider the noise in the estimate of the fraction of functional sequences.
605 For example, there are no functional sequences among the set of variants that show seven
606 substitutions, likely due to the lower number of variants sampled at this mutation distance.

607 To address these issues, we also used the supervised, ten position oracle function to estimate the
608 total number of functional sequences. Here, we generated one million random synthetic
609 combinatorial variants, and asked the oracle function to predict what fraction of these variants
610 reaches half-maximal growth rate, p(predicted growth rate (GR)>0.5). We then estimated the
611 fraction of true positive predictions, p(observed GR>0.5|predicted GR>0.5), by sampling with
612 replacement the observed variants, for which observed and predicted growth rate data are
613 available, to match the distribution of predicted growth rates for the synthetic sequences. Then we
614 calculated the fraction of these subsampled observed variants with GR>0.5 among variants with
615 predicted GR>0.5. Finally, we multiplied these two fractions (p(predicted GR>0.5) and p(observed
616 GR>0.5|predicted GR>0.5)) to estimate the fraction of synthetic variants that are both predicted
617 and observed to be functional, p(observed GR<0.5, predicted GR>0.5). We then multiplied this
618 probability by the total number of possible variants ($20^{10}$) to estimate a lower bound on the total
619 number of functional sequences.

620 ***Unsupervised protein variant scoring***

621 To implement CoVES, we trained the RES classifier model as described previously[22]. Briefly, this
622 equivariant graph neural network is trained on predicting the amino acid identity of a masked
623 residue given ~500 surrounding atoms, excluding hydrogens, from the protein database (PDB).
624 Importantly, the training and test set were split according to domain-level structural CATH
625 topology[23,25], and training examples were down-sampled to the least common amino acid to
626 prevent biased learning of the classifier. The trained model achieved ~ 53% accuracy on the held-
627 out test set.

628 To score combinatorial variants for the antitoxin, we fed the wild-type structural environments,
629 including both toxin and antitoxin atoms in their biological octameric assembly, surrounding each
630 mutated antitoxin residue to the graph neural network and obtained classifier scores for each amino
631 acid at each positions. We then summed the log probability normalized classifier scores at each
632 site (log $p(amino\ acid)\ \alpha\ \frac{-\ amino\ acid\ score}{t}$) using t=0.1) to result in a combinatorial variant
633 effect score. We followed an analogous approach for GFP (PDB ID: 2WUR), GRB2 (PDB ID:
634 1GCQ) and AAV (PDB ID: 1LP3). To score mutants in GFP, GRB2 and AAV using
635 proteinMPNN and ESM-IF, we calculated the conditional autoregressive log-probabilities for each
636 of the sequences further conditioned on the PDB structure.

637 ***Unsupervised protein variant generation***

638 To sample antitoxin variants using CoVES, we converted the structure-learned amino acid
639 preference scores into probabilities at each site using the Boltzmann distribution
640 ($p(amino\ acid)\ \alpha\ e^{\frac{-amino\ acid\ score}{t}}$). Using equal weighting between sites, we sampled 500
641 random sequences at varying temperatures (t ∈ {0.1, 0.5,0.7,1,1.5,2,2.25,2.5,2.75,3,4,5}) at each
642 mutated site and deduplicated the sampled sequences for evaluation. Model weights and sampling
643 strategies for proteinMPNN and ESM-IF followed previous studies[8,24]. For proteinMPNN, we
644 generated 30 sequences at each temperature (t ∈ {0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4,
645 1.5}). For ESM-IF, we generated 100 sequences and deduplicated these variants at each
646 temperature (t ∈ {0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5}). For all models, we fixed
647 amino acids that were not mutated, and also tied any mutations in corresponding positions between
648 the four antitoxin chains in the octameric toxin-antitoxin sequence from PDB ID: 5CEG. We then
649 evaluated the growth rate effects of generated variants using the supervised oracle function trained
650 on the ten position library data, which is estimated to have low generalization error when predicting
651 unobserved variant effects (Fig. 2). We used EvCouplings to score antitoxin variants as done
652 previously[11]. We then sampled sequences, while fixing all non-mutated positions, by performing
653 simulated annealing with a temperature cycle of 1.0*scaling_factor, 0.5*scaling_factor, and
654 0.2*scaling factor, each for 160 steps. We note that EvCouplings is not able to generate samples
655 for the antitoxin position L48, as this column does not have sufficient coverage in the alignment
656 to infer robust site-wise and coupling parameters. In order to compare these sampled sequences
657 fairly between models, we subsampled the number of unique generated variants from all models

658 to 30, which is the number of unique sequences sampled across most temperatures for the
659 proteinMPNN model. Finally, to help in visualization of Fig. 3, we fit polynomial fits of various
660 polynomial order to each model's generated sequence summary statistics.

661 ***Unsupervised GFP variant generation and evaluation***

662 GFP variants were generated using CoVES, proteinMPNN and ESM-IF by both varying the
663 sampling temperature, as well as the number of randomly chosen residues (among the following
664 mutation ranges: {5,6,7,8,9,10,11,12,13,14,15, 16, 17, 18, 19, 20}) that are designed, with the
665 remaining positions fixed to their wild-type amino acid. The examined CoVES sampling
666 temperatures were among {0.1, 0.3,0.4, 0.5, 0.6, 0.7,1,2, 3, 4, 5, 8, 10, 20}. The examined
667 ProteinMPNN sampling temperatures were among {0.0001, 0.1, 0.15, 0.2, 0.25, 0.3, 0.5, 0.7,
668 1.0,1.5,1.75, 2,2.25, 2.5,5,10}, and ESM-IF sampling temperatures among {1e-2,1e-1,0.5,0.75,1,
669 1.25, 1.5, 1.75, 2, 3, 4, 5, 10}. In order to ensure that only designed variants are assessed which
670 are not out of distribution and fall within the set of sequences that the oracle function has predictive
671 power on and has been evaluated on during random 90%-10% testing, designed variants were then
672 post-filtered to a) contain fewer than 15 residue mutations and b) contain individual variants that
673 have been observed by the oracle when trained on the observed combinatorial variant dataset (1810
674 per residue variant effects among 233 residues).

675 ***Supervised protein variant generation***

676 To report on the number of functional variants at various mutation distances from the random
677 library (Fig. 3e), we fetched all the observed library observations, and calculated the function of
678 observed variant effects at a given hamming distance from the wild-type sequence, without
679 sampling.

680 To sample from the supervised, logistic regression 'oracle' function, we converted site-wise
681 preferences fit to the total combinatorial dataset into energies by min-max normalization and using
682 the Boltzmann distribution to obtain probabilities for each amino acid variant at each position. We
683 then used various temperatures ($t \in \{0.5, 0.7, 0.8, 0.9, 1, 1.5, 2\}$) to generate sets of sampled
684 sequences.

## Author contributions

Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft and preparation: DD

Data Curation, Resources: DD, AYS

Software: DD, AYS, NR

Writing – review and editing: DD, SS, DSM, MTL, DFS, NP, AYS

Supervision, Project Administration: DSM, MTL

Funding Acquisition: DSM, MTL, DFS

## Acknowledgements

We thank members of the Laub and Marks labs for helpful feedback, and Chloe Hsu for helpful discussions on sampling from ESM-IF. This work was supported by the Howard Hughes Medical Institute (MTL, DFS), National Institutes of Health grant R01CA260415 (DSM) and Chan Zuckerberg Initiative CZI2018-191853 (DSM).

## Competing interests

SS is employed by Dyno Therapeutics. DSM is an advisor for Dyno Therapeutics, Octant, Jura Bio, Tectonic Therapeutics, and Genentech, and a co-founder of Seismic. NR is employed by Seismic. DFS is a co-founder and scientific advisory board member of Scribe Therapeutics. The remaining authors declare no competing interests.

## Code availability

Custom scripts will be available at: https://github.com/ddingding/CoVES

## Data availability

Raw sequencing read data will be made available upon publication.

## Supplementary tables

Supplementary table ST1: Primers used in this study.

709 Supplementary table ST2: Data statistics across combinatorial variant effect datasets.

710

711 **Supplementary Figures 1-9**

Fig S1

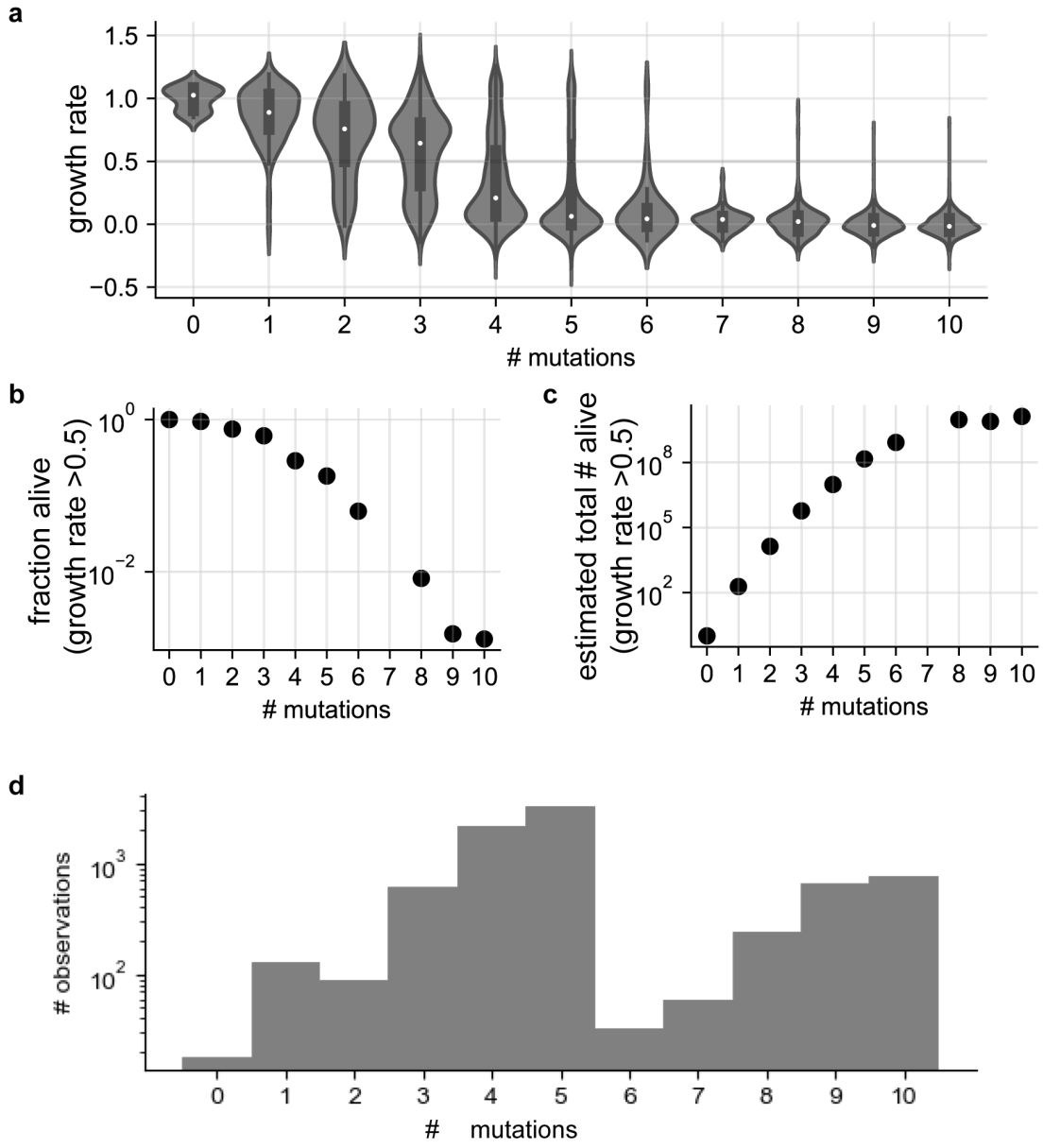



**Fig. S1: Observed statistics of 10 position antitoxin ParD3 library enable predicting the total number of functional antitoxin variants.**
**a**, The distribution of antitoxin variant growth rate effects split across the number of substitutions.
**b-c**, The fraction of antitoxin variants that reach half-maximal growth rate values as a function of substitution distance (**b**), and the resulting estimated total number of antitoxin variants that reach half-maximal neutralization at a given substitution distance (**c**).
**d**, The number of observations for each number of substitutions of antitoxin variants is shown.

712

713

Fig. S2

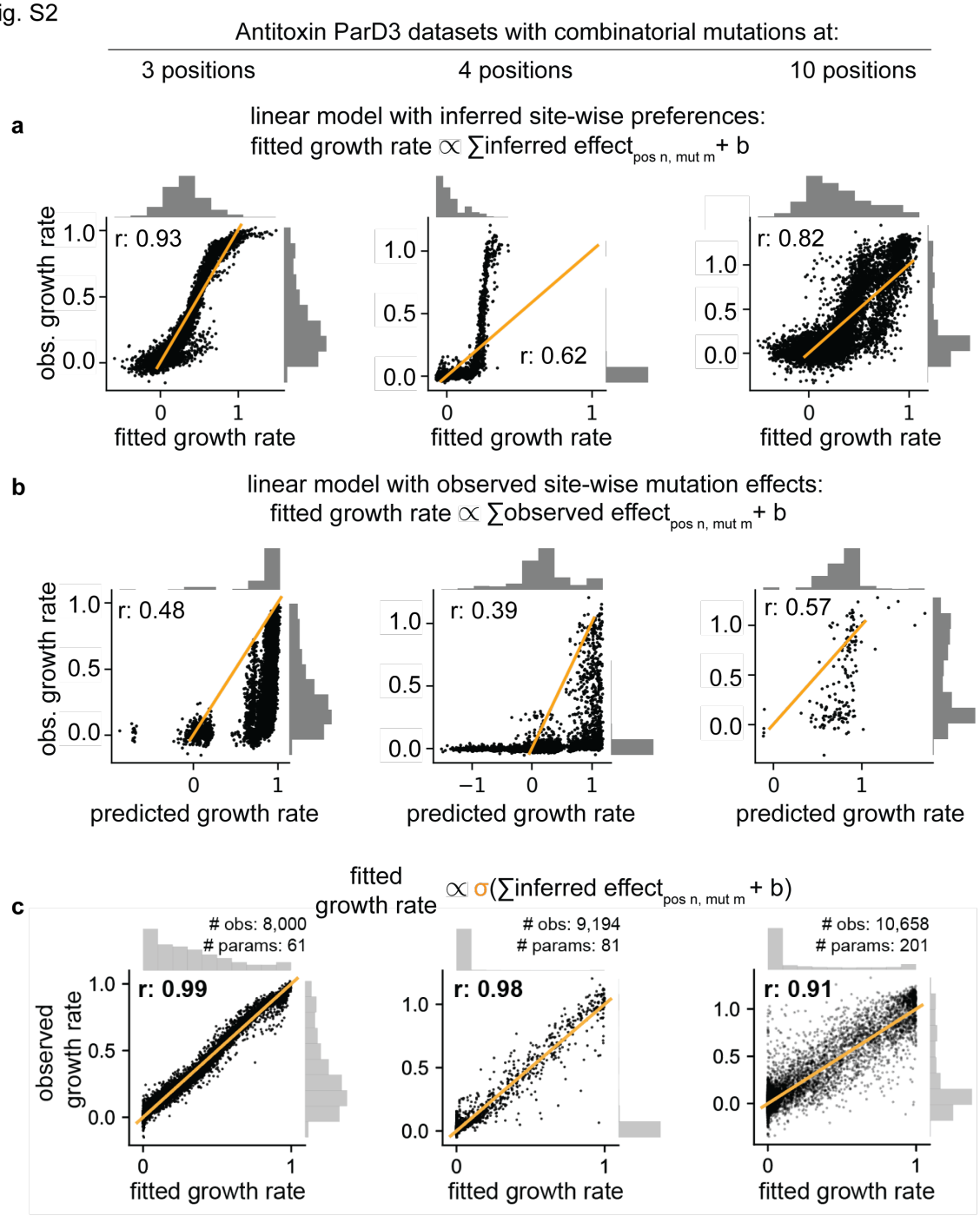



**Figure S2: Fit of linear site-wise preference model and observed site-wise mutation effects.**
**a**, Fit of a linear model using inferred site-wise preferences as predictors, without a sigmoid nonlinearity, for observed combinatorial variant effects.
**b,** Fit of a linear model using observed single variant effects as predictors for observed combinatorial variant effects.
**b,** Fit of a logistic model using inferred single variant effects as predictors for observed combinatorial variant effects. This corresponds to Fig. 2g.
Pearson correlation coefficients (r) are indicated.

714

Fig. S3

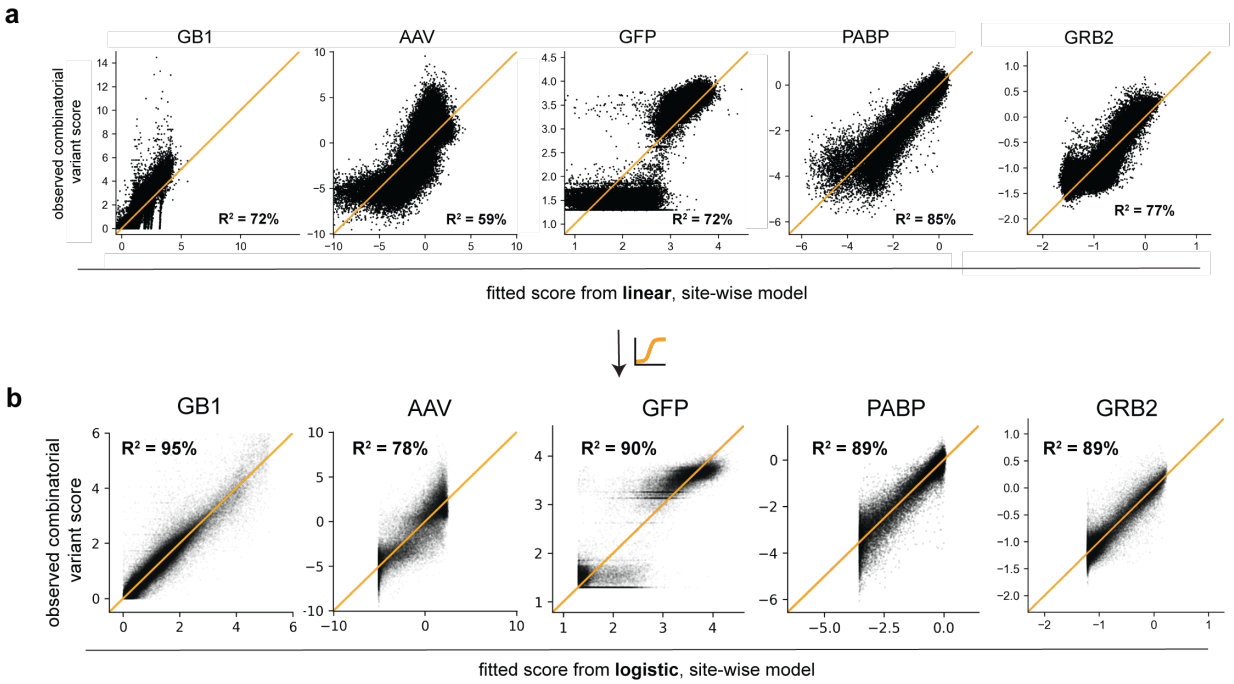



**Fig. S3: Logistic regression with site-wise mutation preferences enables better fit to combinatorial variant effect datasets across proteins.**
**a-b**, Site-wise linear regression (**a**) and logistic regression (**b**) fits to the total combinatorial variant effects for various proteins, with the total explained variance $R^2$ indicated.

715

Fig S4

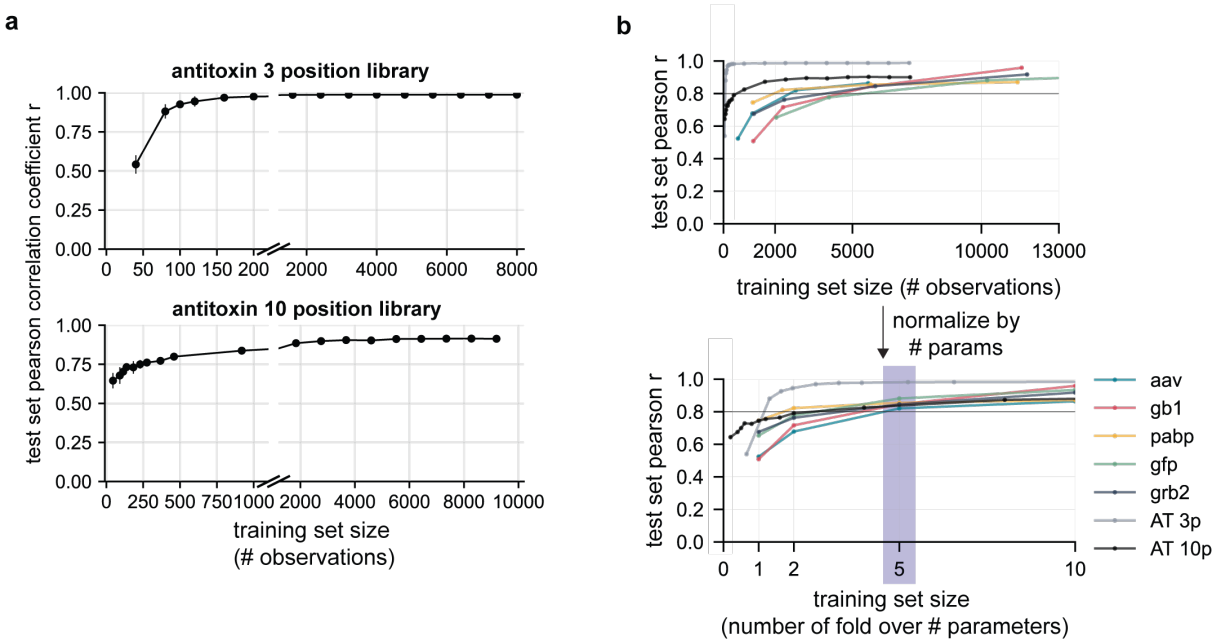
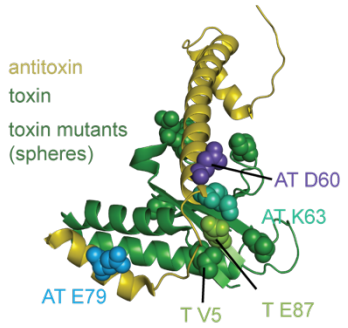



**Fig. S4: Few observed sequences are required to predict combinatorial mutation effects with a logistic model using residue mutation preferences.**
**a**,The nonlinear, additive model was trained on a subset of combinatorial variants, and then used to predict the remaining held-out combinatorial mutation effects for the 3 position (top) or 10 position antitioxin library (bottom). The pearson correlation coefficient, r, between predicted and observed variant effects is indicated vs. the number of subsampled observations used to infer the site-wise amino acid preferences. Errorbars represent standard deviations from repeated random subsampling (n=8).
**b**, A subset of the observed combinatorial variant effects were used to fit the site-wise preference parameters of a logistic model for 6 proteins and their performance on the remaining held-out combinatorial variant effects in each dataset is shown. The top graph shows the absolute number of training examples used, while the bottom graph shows the number of training examples used to fit the model in terms of fold over the number of site-wise preference parameters used by the model depending on the number of mutated residues in each dataset.

716

Fig. S5

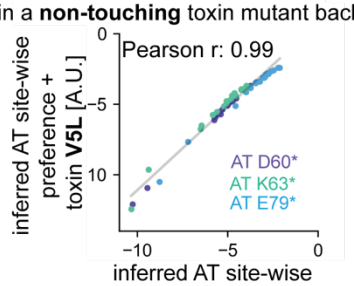
**a** 3 position AT library vs. 10 toxin mutant backgrounds



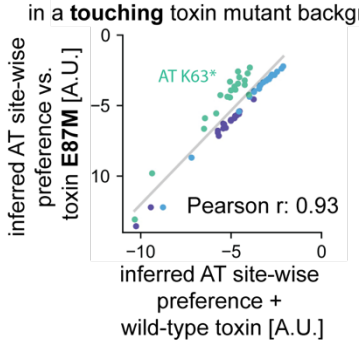

**b** Correlation of inferred antitoxin residue preferences between different toxin mutant backgrounds

| toxin | r vs. wt toxin |
|---|---|
| ParE3(N99V) | 1.00 |
| ParE3(V75C) | 1.00 |
| ParE3(D55E) | 0.99 |
| ParE3(A66I) | 0.99 |
| ParE3(V5L) | 0.99 |
| ParE3(A66F) | 0.99 |
| ParE3(V5L+D55E) | 0.98 |
| ParE3(V5L+A66F) | 0.98 |
| ParE3(R100W) | 0.97 |
| ParE3(E87M) | 0.93 |

**c** Antitoxin site-wise preferences are unaltered in a **non-touching** toxin mutant background




**d** Antitoxin site-wise preferences change in a **touching** toxin mutant background




**e**

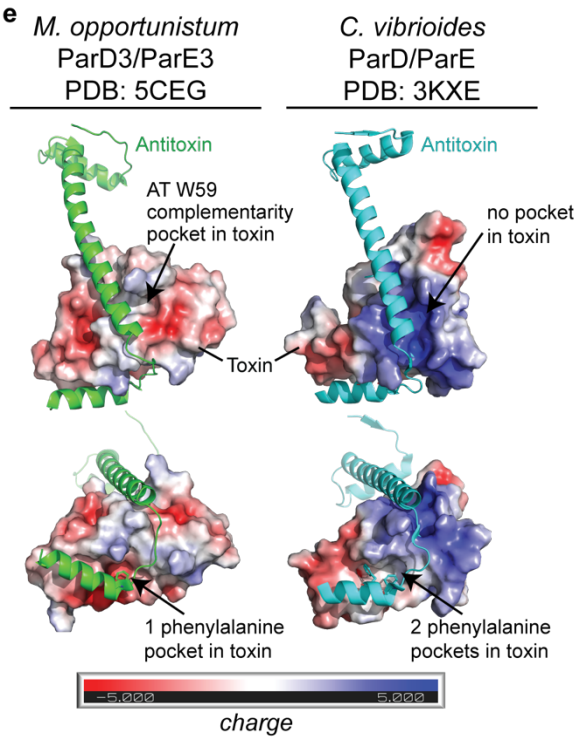



**Figure S5: Antitoxin site-wise preferences are altered by mutations in a contacting residue, and homologous structure comparison suggest differences in site-wise preferences between homologous toxin-antitoxin pairs.**

**a**, Crystal structure (PDB ID: 5CEG) indicating the 10 different ParE3 toxin single substitution variants that the 3 position combinatorial antitoxin ParD3 library was screened against. The toxin is shown in green, with the mutated positions spacefilled. The antitoxin is shown in yellow with purple, cyan and blue indicating the combinatorially mutated residues. Only one mutated position in the toxin, E87, contacts the antitoxin at position K63 (cyan).

**b-d**,Antitoxin ParD3 site-wise preferences inferred by the nonlinear, independent model correlate almost perfectly unless a contacting residue is mutated. **b,** Pearson correlation coefficients, r, for inferred sitewise antitoxin preferences at each of the 3 mutated antitoxin positions in the background of 10 different toxin ParE3 single substitution variants vs. the wild-type toxin ParE3. Scatterplot of antitoxin residue mutation preferences when screened against toxin with a non-contacting mutation (V5L) vs. wild-type toxin (**c**). Differences in antitoxin position K63 mutation preferences in the background of a toxin variant that contains a substitution, E87M, at a contacting position (**d**).

**e**, Differences in local interface microenvironments of binding between ParD3/E3 from *Mesorhizobium opportunistum* (PDB ID: 5CEG) vs. homologous structure ParD/E (PDB ID: 3KXE) from *Caulobacter vibrioides*, suggesting that alignment-averaged amino acid preferences do not reflect the site-wise preference of each individual structural context well.

717

Fig. S6

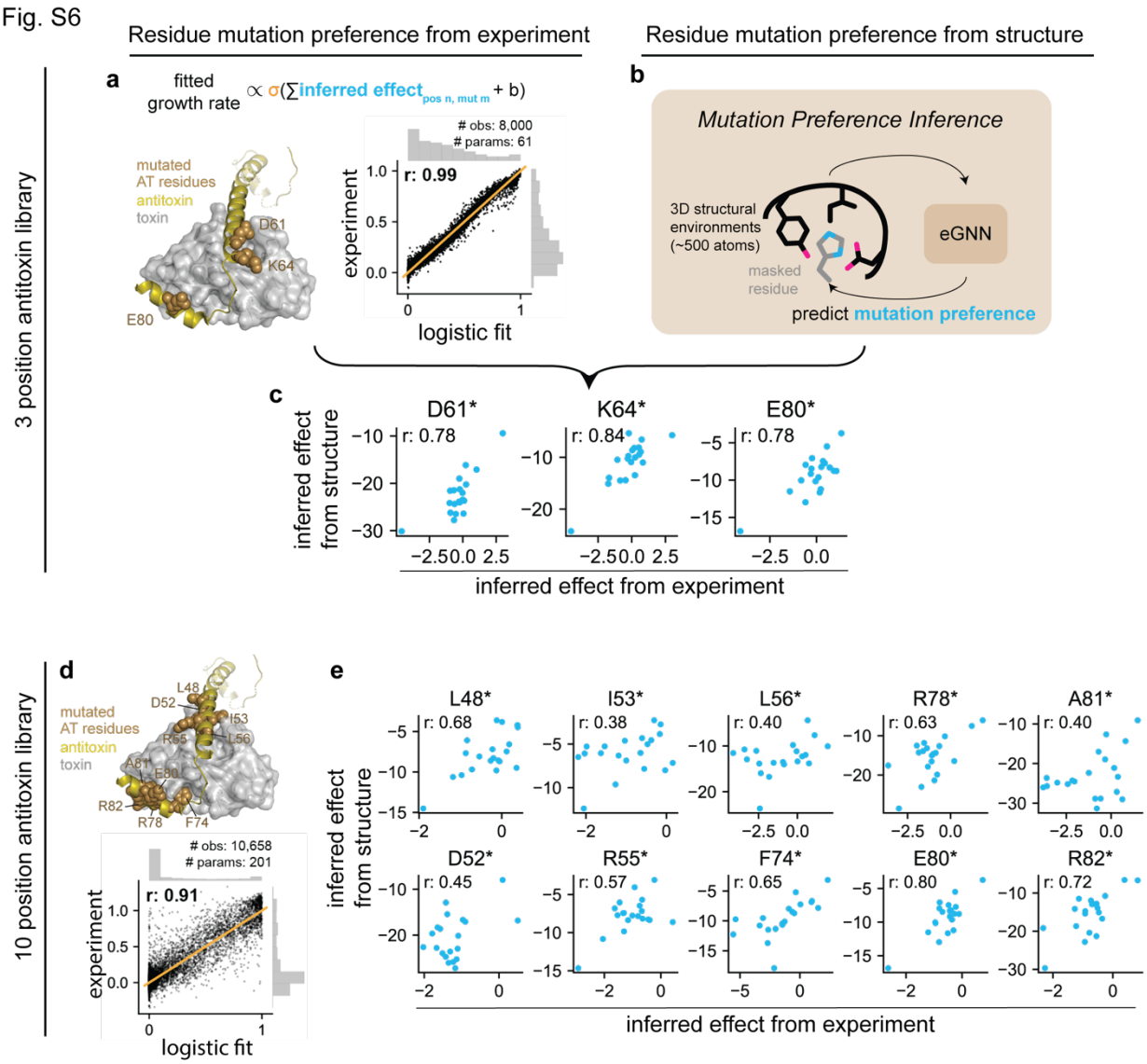



**Figure S6: Antitoxin residue mutation preferences can be learned from structural context.**
**a**, 60 residue mutation preferences at three antitoxin positions learnd from the experiment can explain almost all of the combinatorial variant effects.
**b**, Mutation preferences can be learned from structural context without variant effect measurements.
**c**, Correlation between experimental and unsupervised, structure-based inference of mutation preferences.
**d-e**, Per residue mutation preferences can be learned from structural context for the 10 position antitoxin library, with the explanatory power of the inferred residue preferences shown (**d**).
Pearson correlation coefficients r indicated.

Fig. S7

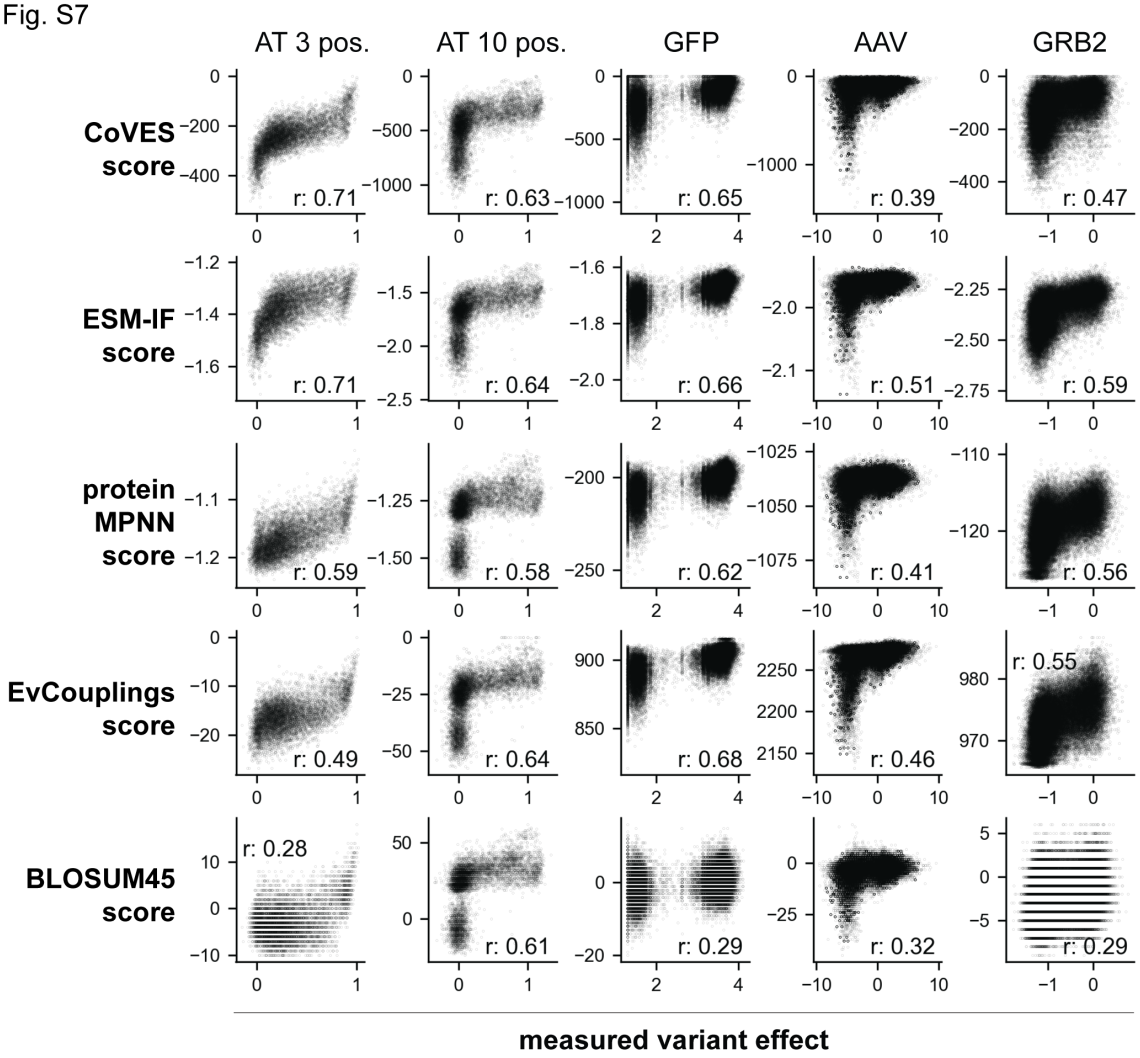



**Fig. S7: Unsupervised model scores correlate with measured combinatorial variant effects.**
The scores for various models (CoVES, ESM-IF, proteinMPNN, EvCouplings and BLOSUM45) are displayed versus the measured variant effects for the antitoxin ParD3 3 position library, antitoxin 10 position library, GFP, AAV and GRB2 datasets. The spearman correlation coefficients (r) are indicated.

Fig. S8

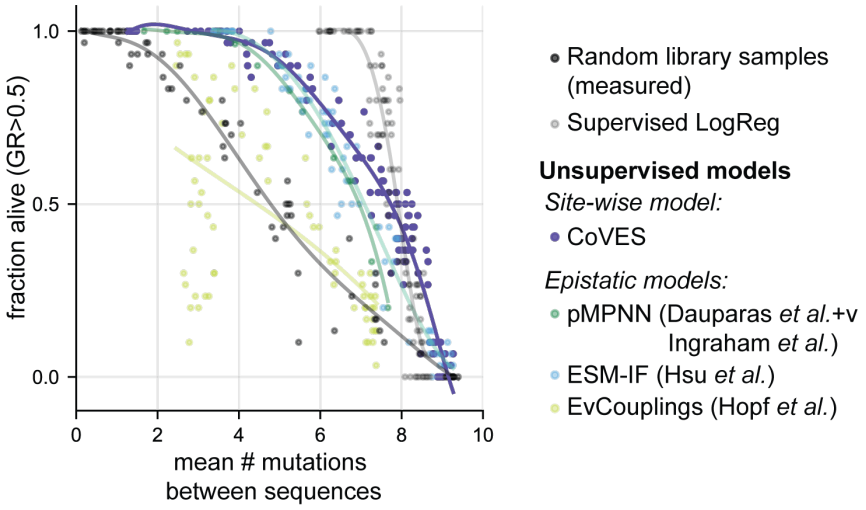



**Fig. S8: Designed antitoxin sequences at 10 positions from unsupervised models evaluated for function and between sequence diversity.**
Scatterplot showing the fraction of designed sequences that are predicted to be functional and the average number of mutations between designed sequences. Each dot represents 30 unique sequences sampled by each supervised model at a given sampling temperature.

Fig. S9

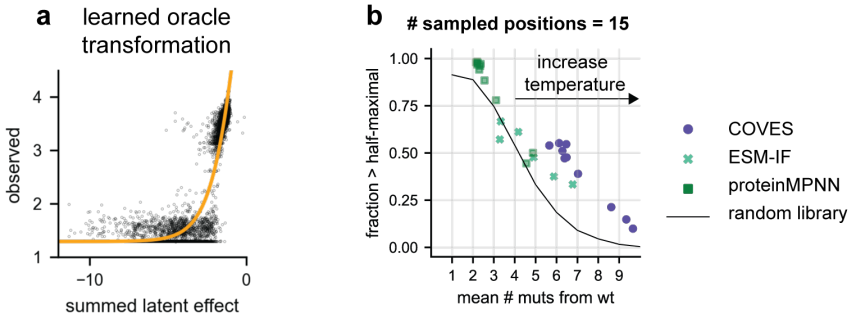



**Fig. S9: Oracle evaluation of designed GFP variants using unsupervised sequence models.**
**a**, The learned nonlinear transformation of the supervised oracle using the summed inferred residue-wise mutation preferences to observed combinatorial variant effects is shown.
**b**, The diversity and function of sampled sequences from unsupervised models is shown. Each dot represents a set of sampled sequences at a given temperature and number of positions that are randomized, summarized by the fraction of sequences predicted to be above half-maximal fluorescence and the average number of mutations. The maximum number of mutated positions is set to 15, and the sampling temperature is increased. Only combinatorial variants consisting of individual variants, which the oracle logistic regression has been trained on and tested are shown.