bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Independent mutation effects enable design of combinatorial protein binding mutants

Authors: David Ding[1,*]

Affiliations:

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

*Corresponding authors: davidding@berkeley.edu

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Design and natural evolution of protein sequences can be profoundly impacted by the extent of epistasis between mutations. For most proteins and sets of residues, it's unclear how much epistasis there is. Here, we measure the effect of combinatorial variants at ten positions in the antitoxin ParD3 on its ability to neutralize its cognate toxin. Using this and two additional datasets, we show that a site-wise independent model without epistasis can explain virtually all of the combinatorial mutation effects. This model can be trained on few random observations and still predict combinatorial variant effects not observed during training. We then develop an unsupervised strategy to design functional and diverse protein sequences without experimental variant effect measurements by using a site-wise independent model trained on structural databases. Such independent approaches could enable the combinatorial design of therapeutically relevant binding proteins with desired binding properties with few or no observations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Understanding how substitutions in proteins manifest in functional changes is both a fundamental as well as applied protein engineering problem. The extent to which multiple mutations combine to result in functional changes can dramatically impact possible future mutational trajectories in natural evolution, as well as engineering of therapeutic proteins of interest, such as antibodies. For example, it's possible that the negative effect of a single mutation can only be tolerated in the presence of another enabling mutation, thereby constraining the possible subsequent mutational trajectories[1–3]. Conceptually, specific dependencies between mutations can cause 'rugged' fitness landscapes, in which Darwinian selection for fitness-increasing mutations does not necessarily result in globally optimal binders[4–6] (Fig. 1a). On the other hand, if multiple mutations combine without specific dependencies between each other, the sequence-fitness function will result in a convex function (Fig. 1a) on which selection can act more efficiently.

Epistasis – the non-additive effect of two or more mutations – can be defined in multiple ways depending on the null expectation of how mutations combine[7]. One null expectation is additivity in the observed functional dimension, such as the measured growth rate or fitness effects of the variants (Fig. 1). Using this expectation, there are many examples of proteins in which mutations pervasively act non-additively[8,9]. However, the non-additive effect of mutations along this observed dimension can often be explained by a 'global epistasis' models[7,9–11]. Here, the null expectation for the effect of combined mutations consists of independent, additive mutations effects along an unobserved dimension, which are then transformed through a nonlinear function to give rise to non-additive effects in the observed variable[7,9,11–18] (Fig. 2a). In this way, saturating effects, such as maximal growth rates of the cell, sigmoid binding curves and limitations to assay sensitivity can be captured. Using this null model, a second definition of epistasis is possible, where deviations from this new nonlinear, independent expectation form the set of specific epistatic interactions between mutations. Such models have been shown to 'explain away' many of the previously epistatic, but non-specific mutation effects without adding additional parameters, suggesting that such models are more parsimonious for explaining observed mutation effects[7].

To study the importance of epistasis in governing protein function, we study a bacterial ParE3-ParD3 toxin-antitoxin system from *Mesorhizobium opportunistum*. When the antitoxin ParD3 is

Fig 1

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

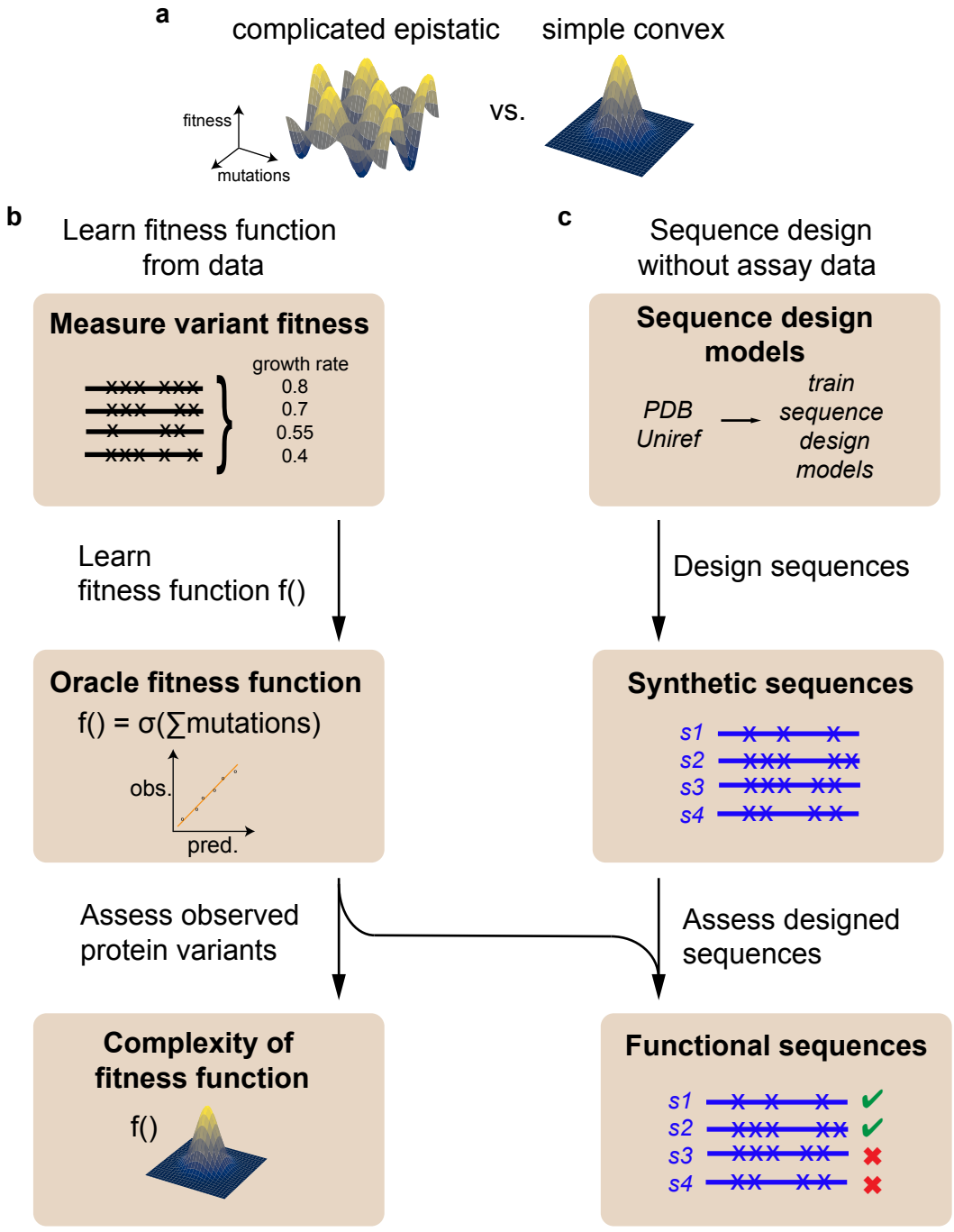


Fig. 1: Predictive fitness function reveals the local fitness landscape and enable assessment of sequence design models.

a, Local fitness landscapes can be conceptualized as simple convex functions arising from independent site-wise residue preferences (left) vs. complicated epistatic functions with dependence between residues (right).

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. This enables predicting unobserved mutation effects and revealing the shape of the fitness functions.

c, Unsupervised sequence design models learned from structural and sequence databases are used to design synthetic sequences. The designed sequences can be assessed for function based on the oracle function, f(), from **b**, revealing strategies for sequence design without requiring experimental observations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

co-expressed with the toxin ParE3 in *Escherichia coli*, only cells containing functional ParD3 are able to neutralize the toxin and proliferate. The *in vivo* function of thousands of antitoxin variants can be read out in bulk by following the variant frequencies over time using high-throughput sequencing[19–21]. Here, we measure random combinatorial variant effects at ten positions of the antitoxin ParD3 (Fig. 1b left). We discover that combinatorial variant effects in this and two other existing datasets can be explained by models which do not incorporate specific epistasis but only the residue site-wise amino acid preferences. These models consist of a linearly increasing number of parameters (20*n +1, where n is the number of mutated residues) to explain an exponentially growing number of combinatorial variant effects (20^n). This model can be trained accurately on few observed sequences, and be used as an 'oracle fitness function' to predict unobserved variant effects at these positions. Given the performance of such non-epistatic, but supervised models, we then devise a strategy to design functional and diverse protein variants that do not require any measurements by using an independent model trained on structural databases (Fig. 1c). Collectively, our findings illustrate the power of independent models in understanding protein evolution and suggest a concrete strategy to design protein binders with few or no variant effect measurements.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

High-throughput in vivo measurement of combinatorial variant effects at ten binding residues of the antitoxin ParD3 illuminates the functional set of antitoxin variants

To understand to which degree mutations depend on each other, we first set out the generate a dataset in which ten positions of the antitoxin ParD3 are randomly mutated and assayed for neutralization of the cognate toxin ParE3 (Fig. 2a). Cells containing different antitoxin variants are grown in bulk, and only those cells that contain functional antitoxin variants that neutralize the toxin can proliferate. The change in frequency of each antitoxin variant is followed via high-throughput sequencing over time (Fig. 2a). These variant effect measurements correlate with orthogonal growth rate measurements and have shown high reproducibility between biological replicates[19–21]. While previous studies have examined the effect of random mutations at three or four antitoxin positions only, we apply this assay to measure the effect of combinatorial variants at ten positions in the antitoxin, containing 7,923 unique amino acid variants and 2,615 truncated antitoxins. We calculate the growth rate score of each variant as the normalized log read ratio before and after selection, and find good separation between truncated and wild-type antitoxin variant effects (Fig. 2b). This assay shows high reproducibility between separate biological replicates (Pearson r: 0.93, Fig. 2c). As expected, the distribution of fitness effects shifts towards loss of function as more substitutions are introduced, with the majority of functional variants achieving at least half-maximal toxin neutralization when fewer than three substitutions are present (Fig. 2d-e, Fig. S1). As additional substitutions are introduced, the fraction of such functional variants decreases faster than exponentially (Fig. 2e, left). Using this maximum likelihood estimate of the fraction of functional variants and considering the total number of possible mutations at each mutation distance, we estimate that there are ~$3x10^{10}$ sequences – out of a possible 20^{10}~10^13 - that achieve half-maximal neutralization when these ten positions are mutated (Fig. 2e, right).

A nonlinear, independent model explains most combinatorial variant growth rate effects in the antitoxin ParD3.

We then set out to understand how well observed combinatorial variant effects can be explained by considering only site-wise amino acid preferences but no specific dependencies between

Fig 2

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

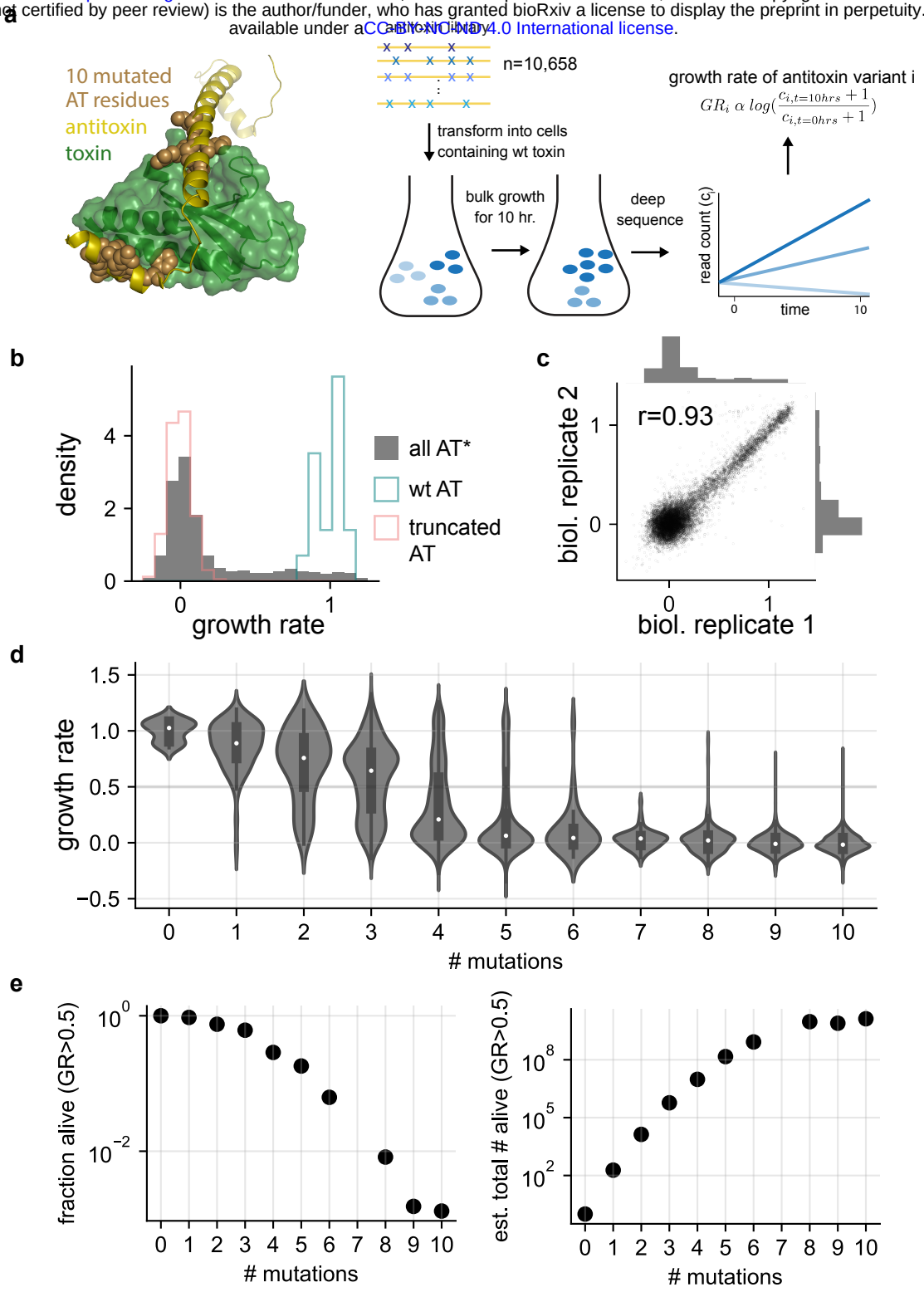


Fig. 2: High-throughput measurement of combinatorial variant effects at ten binding residues of the antitoxin ParD3 on neutralizing the cognate toxin ParD3.

a, 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), 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. Antitoxin variants that are able to bind and neutralize the toxin will show higher growth rates.

b, 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, The distribution of antitoxin variant growth rate effects split across the number of substitutions from the wild-type antitoxin sequence.

e, The fraction of antitoxin variants that reach half-maximal growth rate values as a function of substitution distance (left), and the resulting estimated total number of antitoxin variants that reach half-maximal neutralization at a given substitution distance (right).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutated residues. We first fit a nonlinear, independent model (Fig. 3a) to a dataset from Ding *et al.*[19]. Here, an exhaustive combinatorial set of antitoxin ParD3 variants at three residues (D61, K64 and E80) was assayed for neutralization of the cognate wild-type toxin ParE3 (Fig. 3b, left) as above. This dataset shows almost perfect reproducibility between separate biological replicates (Pearson r: 0.99). The independent model infers site-wise preference parameters (20*3+1 = 61 parameters), sums these inferred parameters and passes them through a nonlinear sigmoid function to predict two orders of magnitude more combinatorial variant effects (20*20*20=8,000 amino acid) (Fig. 3a). Strikingly, such a model predicts combinatorial variant effect almost perfectly (Pearson r: 0.99, explained variance R^2: 98%) with small deviations from the predicted growth rates. This suggests that mutations at these three non-contacting residues act independently without specific dependencies between each other, enabling the precise and rational tuning of variant neutralization strengths by considering only the site-wise amino acid preferences.

To test whether such independent models can perform well when more than three residues are mutated, we also examined a dataset from Aakre *et al.*[21] in which four partially contacting antitoxin positions (L59, W60, D61, K64) are combinatorially mutated (n=9,194 amino acid variants, Pearson r = 0.99 between biological replicates, Fig. 3b middle), as well as the dataset generated in this study, in which ten partially contacting antitoxin residues are randomly mutated (n=7,923 amino acid variants; Pearson r = 0.93 between biological replicates; Fig. 3b, right). The site-wise independent model predicts observed combinatorial variant effects well, both for the four position library (Pearson r: 0.98, explained variance R^2: 94%) as well as the ten position library (Pearson r: 0.91, explained variance R^2: 83%). This is achieved with a significantly smaller number of parameters than combinatorial observations (four position library: 4*20+1=81 parameters to explain 9,194 combinatorial variants; ten position library: 10*20+1 = 201 parameters to explain 7,923 combinatorial variants).

These results indicate that a site-wise independent model with much fewer parameters than observations is sufficient to explain almost all combinatorial variant effects among three non-contacting residues, and as well as predict combinatorial variant effects when four or ten partially contacting residues are mutated in the antitoxin ParD3.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig 3

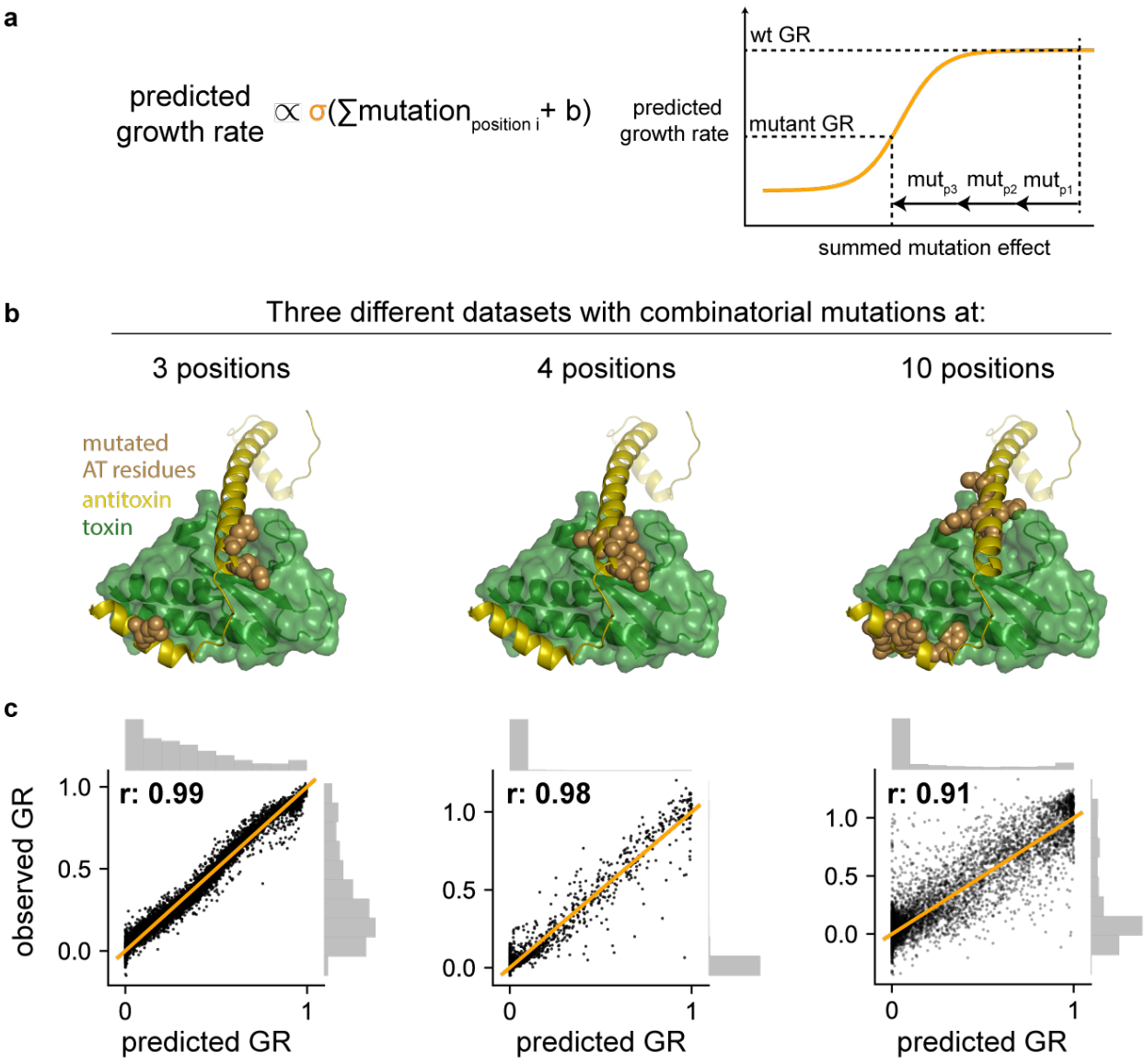


Fig. 3: Independent amino acid preferences can explain most combinatorial mutation effects in a protein binder across 3 datasets.

a, Schematic of the site-wise independent model for explaining combinatorial variant effects. The inferred mutation effects at each position are summed and passed through a sigmoid function to predict combinatorial mutation effects without dependence between mutations.

b, Randomly mutated antitoxin residues (spacefilled, brown) in the antitoxin at 3 positions (AT: D61, K64, E80, n=8,000; Data from Ding et al., 2022.), 4 positions (AT: L59, W60, D61, K64; n=9,194; Data from Aakre et al., 2014), and 10 positions (AT: L48, D52, I53, R55, L56, F74, R78, E80, A81, R82; n=10,658) (PDB ID: 5CEG).

c, Correlations of predicted vs. observed growth rates in 3 datasets with pearson correlation coefficient (r) indicated.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Few observations are sufficient to predict unobserved combinatorial mutation effects.

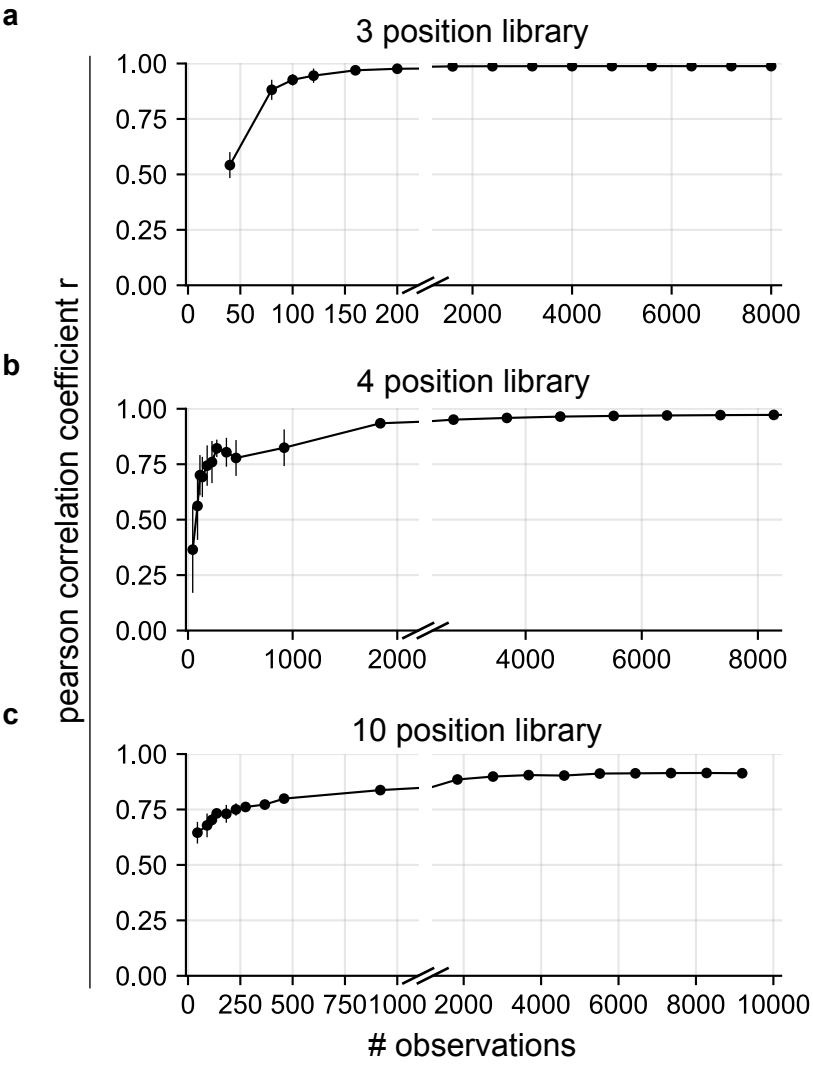
Since we were able to explain a large number of combinatorial variant effects using few independent parameters, we next asked whether a smaller subset of random observations is sufficient to infer these site-wise preferences. We repeated inference of the site-wise preferences of the nonlinear model using smaller random subsets of the total observed dataset (n=8), and evaluated these models on predicting the fitness effect of combinatorial variants in the complete dataset. Indeed, a much smaller number of random observations is sufficient to explain the majority of combinatorial mutation effects well (Fig. 4a-c). For the three position library, 100 or 200 random combinatorial variant effect measurements were sufficient to achieve high correlation (Pearson r:0.92 and 0.98, respectively) between the observed and predicted combinatorial variants (Fig. 4a).

We also tried to estimate how well the ten position, site-wise independent model trained on ~8,000 antitoxin variants can predict the possible 20^{10} possible combinatorial variants. To do so, we inferred parameters from a random 90% subset of the total observed combinatorial variants, and tested the predictive accuracy of the model on the remaining 10% of observed variants. The correlation between predicted and measured held-out test variants is high (train set Pearson r: 0.916 ± 0.001, test set Pearson r: 0.900 ± 0.011, Fig. 4d). This suggests that this model can be used as an 'oracle fitness function' to predict the effect of unobserved combinatorial variants at these ten positions.

Site-wise preferences can be affected by contacting residues.

We next asked whether site-wise preferences of these mutated binding residues can be changed by contacting residues. To do so, we examined the combinatorial variant effects in the three position (antitoxin residues D61, K64 and E80) library measured for neutralization of ten different single amino acid substitutions in the ParE3 toxin (data from Ding *et al.*[19]). We fit separate nonlinear, independent models to the antitoxin library in each toxin variant background. The inferred site-wise antitoxin preferences were almost perfectly correlated across 9 out of 10 toxin backgrounds, in which toxin substitutions do not contact any of the mutated antitoxin residues (Fig. S2b-c), but deviated in the background of toxin E87M (Fig. S2d). Inspection of the site-wise preferences reveals deviations particularly for antitoxin residue K63, which is found to be directly contacting

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



d

	10 position
90% train	*0.916 ± 0.001*
10% test	*0.900 ± 0.011*

Fig. 4: Few observed sequences are required to predict combinatorial mutation effects, and enable generalization to unobserved variants.

a-c,The nonlinear, additive model was trained on a subset of combinatorial variants, and then used to predict the combinatorial mutation effects in the entire dataset for the 3 position (**a**), 4 position (**b**) or 10 position library (**c**). 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 different subsampled numbers (n=8).

d, The nonlinear, additive model was trained on a random 90% subset of the combinatorial library data and then used to predict the remaining 10% of data. The mean pearson correlation coefficients and standard deviations between predicted and observed mutation effects is indicated (n=8).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the toxin E87M position on the crystal structure (Fig. S2a). This result suggests that the site-wise preferences can be altered by directly contacting residues surrounding a particular amino acid.

An independent, unsupervised model trained on structural microenvironments around a residue can generate diverse and functional sequences.

Is it possible to predict combinatorial variant effects without measurements? We hypothesized that the three-dimensional structural environment around a particular amino acid could be sufficient to learn site-wise amino acid preferences, and that these scores alone could be used to predict combinatorial variant effects. To test this, we trained a graph neural network that learns rotation equivariant transformations (RES model from Jing *et al.*[22]) to predict the identity of a masked amino acid from its structural, atom-level 3D environment across non-redundant structures from the Protein Database (PDB) (test accuracy = 52.8%, compared to ~5% for random guess, Fig. 5a, training and testing dataset from Townshend *et al.*[23]). We then predicted the amino acid preference at each of the mutated antitoxin residues given their respective structural environments (Fig. 5b), and summed these amino acid preferences to predict combinatorial variants effects. This strategy is able to predict observed mutation effects modestly well (three position library: Pearson r=0.72, four position library: Pearson r=0.45, ten position library: Pearson r=0.57; Fig. S3).

We then assessed whether this independent model trained on structural information alone is able to design diverse and functional sequences. Given these site-wise amino acid preference scores, we used a Boltzmann energy function to sample sequences (n=500) at various temperatures, t, to control the diversity of the generated sequences. We then assessed the predicted growth rate effect of these sequences using the above supervised ten residue oracle function, as well as the diversity of generated samples by calculating the number of mutations between sampled sequences and with respect to the wild-type sequence. At temperature t=1.5, this strategy generated 91 unique sequences of which ~70% are predicted by the oracle fitness function to achieve half-maximal fitness. These sequences showed an average of 6.7 substitutions between samples and 5.6 with respect to the wild-type antitoxin (Fig. 5c). Inspecting the generated sequences shows that multiple different substitutions can occur at each position (Fig. 5d).

Fig 5

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

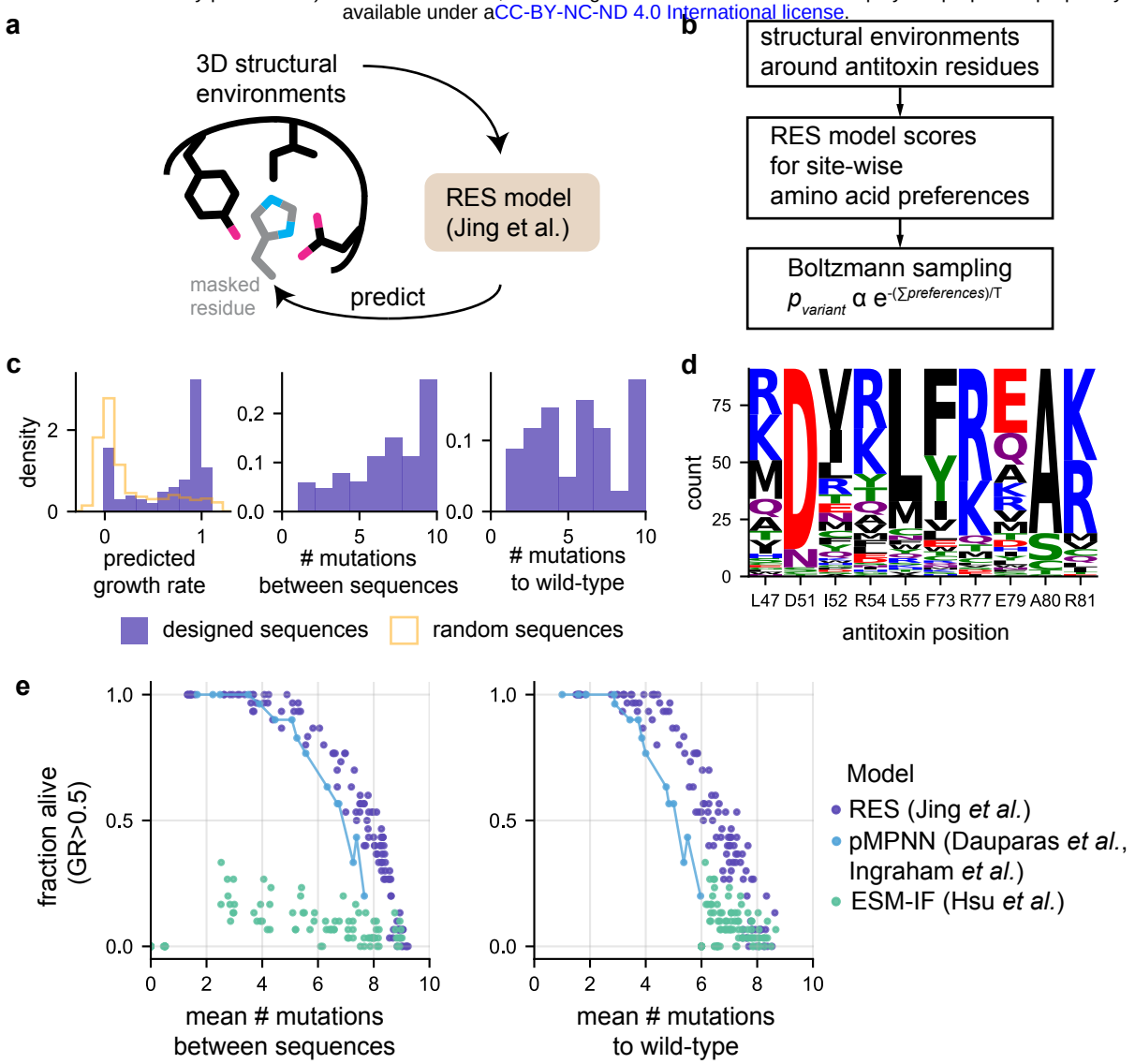


Fig. 5: An unsupervised, site-wise independent model trained on structural microenvironments can generate viable and diverse sequences.

a, Schematic of training the RES model (Jing *et al.,* Townshend *et al.*) to predict amino acid preferences from structural environments.

b, Workflow for designing protein sequences using the antitoxin structure. Microenvironments around residues of interest are extracted and fed into the RES model to generate variant preferences for each position. These preferences are then used to sample variants from a Boltzmann energy function.

c, Generated sequences (n=91 unique sequences, temperature = 1.5) from the RES model are evaluated for their growth rate effects, and their sequence diversity between generated sequences as well as to the wild-type sequence for the 10 position library.

d, Sequence logo of the unique generated sequences from the RES model (panel **c)**.

e, Comparison of generated sequences from the RES model vs. other state-of-the-art sequence design models in terms of the fraction of generated sequences predicted to be functional, their diversity and distance from wild-type antitoxin as the sampling temperature is varied.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

How well does this site-wise independent, unsupervised strategy compare to ones that consider dependencies between residues? To test this, we compared generated sequences from this independent model learned on structural information against sequences designed from two models that also consider the sequence context in an autoregressive manner (proteinMPNN from Dauparas *et al.*[24], Ingraham *et al.*[25]; ESM-IF from Hsu *et al.*[26]). For each model, we sampled sequences at a range of temperatures and assessed the sampled sequences for the fraction that is predicted to be functional by the 'oracle fitness function', as well as their diversity. We find that the simple Boltzmann sampler using the site-wise scores of the RES model outperforms both these models for generating functional sequences that diverge from each other and with respect to the wild-type antitoxin sequence (Fig. 5e).

These results indicate that sampling from a site-wise independent, unsupervised model is sufficient to generate functional and diverse sequences at these ten antitoxin residues, and can even outperform autoregressive sampling strategies.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

We were able to explain most combinatorial variant effects on the native *in vivo* function of the antitoxin ParD3 by considering only site-wise amino acid preferences with no specific dependence between residues. As a result, few observations are required to infer a linearly increasing number of site-wise preferences to predict an exponentially exploding number of variant effects in this antitoxin across three combinatorial variant datasets. The high correlation between predicted and observed variant effects is especially striking for the dataset in which three non-contacting positions are randomized. Here, an unprecedented 98% of the observed combinatorial variant effects can be explained. This is enabled by low noise of this dataset, which has almost perfect correlation between biological replicates.

Using a few thousand observations, we were able to fit a site-wise independent model to estimate the prediction error for the unobserved, remaining possible $\sim 20^{10} \approx 10^{13}$ combinatorial protein variant effects at ten positions of the antitoxin. By considering the noise in this predictor, we estimate that there are at least $\sim 1.7 \times 10^{10}$ combinatorial variants at these ten positions that achieve half-maximal neutralization of the toxin ParE3 (see Methods), similar to the empirical estimate of 3×10^{10} functional sequences above (Fig. 2). Similar to Ogden *et al.*[27], this supervised model itself could be used to design functional sequences for further experimental validation, but we leave this for future work.

The lack of specific dependencies between mutations also motivated successful strategies to generate functional and diverse sequences without assay-specific observations. In this case, simple sampling and modeling approaches that learn amino acid preferences from structural environments outperform approaches that do incorporate dependencies between residues. Future efforts to exploit such independent models could focus not just on improving the inference of site-wise preferences, but also the weighting of different residue contributions. For example, it is known that particular residues can be differentially important in contributing to binding, including the antitoxin ParD3 Tryptophan 60 residue[19]. In our sampling strategy, equal weighting of site-wise amino acid preferences performed well, but it's possible that unequal weighting and scaling of site-wise preferences could further improve prediction of combinatorial variant effects.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

It remains to be seen to what extent such independent residue approaches are successful for other proteins or selective functions, as well as the number and specific set of mutated residues considered. It is clear that the choice of residues is critical, since many high-order epistatic terms are required to explain combinatorial variant effects among particular chromophore residues of GFP[9], but not when random mutations throughout the protein are introduced[10]. The site-wise preferences from the three position library of the antitoxin can be altered by mutations in contacting positions, suggesting that the choice of non-contacting binding residues could be a defining factor leading to the lack of specific dependencies between residues. Since such supervised, site-wise independent models trained on measured variant effects can perform well in explaining combinatorial variant datasets[7,28], and generate unobserved functional sequences[10,27,29] across other proteins, we expect that our unsupervised strategy to design sequences without measurements could also enable the rational design of therapeutically relevant proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2**, e00631 (2013).
2. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science (New York, N.Y.)* **328**, 1272–5 (2010).
3. Weinreich, D. M. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **312**, 111–114 (2006).
4. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics* **31**, 24–33 (2015).
5. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet* **15**, 480–490 (2014).
6. Kauffman, S. A. & Weinberger, E. D. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology* **141**, 211–245 (1989).
7. Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E7550–E7558 (2018).
8. Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).
9. Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications* **10**, 1–11 (2019).
10. Gonzalez Somermeyer, L. *et al.* Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife* **11**, e75842 (2022).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11. Pokusaeva, V. O. *et al.* An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genetics* **15**, 1–30 (2019).
12. Poelwijk, F. J. Context-Dependent Mutation Effects in Proteins. *Methods in Molecular Biology* **1851**, 123–134 (2019).
13. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nature Genetics* **51**, 1177–1186 (2019).
14. 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).
15. Atwal, G. S. & Kinney, J. B. Learning Quantitative Sequence–Function Relationships from Massively Parallel Experiments. *Journal of Statistical Physics* **162**, 1203–1243 (2016).
16. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* 1–11 (2016) doi:10.1038/nature17995.
17. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, 1–31 (2018).
18. Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nature Genetics* **51**, 1170–1176 (2019).
19. Ding, D. *et al.* Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nat Ecol Evol* **6**, 590–603 (2022).
20. Lite, T.-L. V. *et al.* Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *eLife* **9**, e60924 (2020).
21. Aakre, C. D. *et al.* Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell* **163**, 594–606 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

22. Jing, B., Eismann, S., Soni, P. N. & Dror, R. O. Equivariant Graph Neural Networks for 3D Macromolecular Structure. Preprint at http://arxiv.org/abs/2106.03843 (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. 12.

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

27. Ogden, P. J., Kelsic, E. D., Sinai, S. & Church, G. M. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* **366**, 1139–1143 (2019).

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

34. 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).

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Material and Methods

Bacterial strains, vectors and media

E. coli TOP10 strains were grown at 37 °C in M9L medium (1x M9 salts, 100 μM $CaCl_2$, 0.4% glycerol, 0.1% casamino acids, 2 mM $MgSO_4$, 10% v/v LB). Antibiotics were used as follows: 50 μg/ml carbenicillin, 20 μg/ml chloramphenicol in liquid media, and 100 μg/ml carbenicillin, 30 μg/ml in agar plates. The toxins ParE3 was carried as before[8] on the pBAD33 vector ($chlor^R$ marker, ML3302 for wild-type ParE3) with expression repressed or induced with 1% glucose and L-arabinose at indicated concentrations, respectively, and the antitoxin ParD3 was carried on the pEXT20 vector ($carb^R$ marker, ML3296) with expression induced by IPTG[21].

High-throughput measurement of protein variant effects.

To measure the combinatorial variant effects in the antitoxin at ten residues, we constructed two sublibaries: one in which five residues are randomized, and one in which an additional five positions are randomized. In this way, we can guarantee the presence of sufficient number of variants with five or less substitutions, given the curse of dimensionality, in which random sampling at ten positions will generate a distribution of variants for which most will have a high number of substitutions.

Library construction

Both the ten position and five position libraries were constructed using a 2-step overlap-extension PCR protocol[30]. We first used primers DDP704+DDP142 and DDP705+DDP141 to introduce five randomized positions in the wild-type antitoxin pEXT20-*parD3* plasmid ML3296 (PCR cycling was: 30 sec. at 98°C; 20 cycles of: 10 sec. at 98°C, 20 sec. at 55°C, 1 min. at 72°C; 2 min. at 72°C,

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hold at 4°C; using KAPA). The PCR products were pooled, diluted 1:100 and amplified using the outer primers DDP141+DDP142 to generate full length, mutated antitoxin sequence. On this PCR product, we then used the primers DDP700+DDP142 and DDP705+DDP141 to introduce the next five randomized positions, and used the above strategy to generate full length antitoxin ParD3 gene with 10 positions randomized. We then cloned both the five and ten position randomized PCR product into the pEXT20 vector using restriction digests with SacI-HF and HindIII-HF (NEB) and ligation using T4 DNA ligase (NEB at 16 ºC for 16 hours with a 1:3 molar ratio of insert to vector. Ligations were dialyzed on Millipore VSWP 0.025 µm membrane filters for 90 minutes before electroporating into TOP10 cells, made using the protocol from Warren[31] (2mm cuvettes at 2.4kV). Cells were recovered in 1ml SOC for 1 hour. We propagated each library with at least 500,000 transformants, checked by spot plating 1:10 serial dilutions of recovered cells on LB/carb/chlor/1% glucose plates. We grew OD_{600} ~ 0.5 at 37 °C in M9L/carb/chlor/1% glucose, spun down (8000G, 5 minutes) and resuspended in 5ml M9L/carb/chlor/1%glucose/20% glycerol for storage at -80 °C. We then made these cells electrocompetent in replicate for each library using the above protocol[31], and transformed dialyzed wild-type toxin ParE3 into these cells. Cells were propagated with at least 500,000 transformants, and grown up to OD~0.6, before spinning down (8000G, 5 minutes) and resuspending in 5ml M9L/carb/chlor/1%glucose/20% glycerol. Cells were aliquoted into 1ml tubes and flash frozen in liquid nitrogen for storage.

High-throughput variant effect measurement

On the day of growth rate measurements, aliquots from two separate transformations were thawed and recovered in 50 ml M9L/carb/chlor/1% glucose at 30 °C for 3 hours. Subsequently, glucose was removed by washing 4 times with M9L, and cells were ready for growth rate measurement. Growth rate measurements were then performed as described previously[19]. Briefly, washed cells

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were resuspended in 250ml M9L/carb/chlor/IPTG to induce antitoxin expression, and toxin expression induced after 100 minutes by adding arabinose. Cells were diluted 1:10 with pre-warmed media when their OD_{600} reached ~0.3 to keep them in exponential growth throughout the duration of the experiment. 50ml of the cultures were sampled at the time of toxin induction, and 10 hours after. These cells were miniprepped, and a high-input (200 ng plasmid DNA), low cycle (14 rounds) PCR reaction performed to isolate amplicons of interest using primers DDP643-645/DDP648-651/DDP654-657 to introduce Illumina multiplexing indices and adapters. We then gel purified and sequenced these library amplicons as described previously[19]. We then performed sequencing using 250 basepair paired end reads using a Novaseq SP flowcell.

Analysis of high-throughput sequencing data

Paired-end reads were processed as described previously[19]. Briefly, paired-end sequencing reads were merged using FLASH 1.2.11[32]. Merged reads were quality filtered based on their phred-score using vsearch 2.13.0[33], with the following arguments: vsearch --fastq_filter {0} --fastq_truncqual 20 --fastq_maxns 3 --fastq_maxee 0.5 --fastq_ascii 33 --fastaout {1}.fasta. Reads were subsequently filtered for having defined mutations at the desired sites only, and the frequency of each variant at each timepoint was counted. We then calculated a log-read ratio for all variants with at least three reads pre- and post-selection, and normalized between 0 and 1 given the log read ratio of truncated and wild-type antitoxin variants.

Nonlinear, independent modeling of combinatorial variant effects.

We used a nonlinear site-wise model implemented in Tensorflow 2 to model combinatorial variant effects as before[19]. We used one-hot encoding of amino acid variants as a predictor and fit weights associated with each unobserved single amino acid mutant substitution as well as one additional bias parameter. The linear sum of these weights is passed through a sigmoid function to predict

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the normalized growth rate effect between zero and one for each combinatorial variant in both the five and ten position library. We used the Adam optimizer with a 0.1 learning rate to minimize the mean squared error of predicted to measured normalized log read ratios for each variant for several hundred epochs, until the training error stabilized.

Inference of site-wise preferences from subsampled variant effect measurements and estimation of generalization error.

To test how well models trained on few observations can predict the total observed variant datasets, we retrained the above models on a subset of data eight times, and tested their performance on the full dataset. In order to estimate the generalization error of prediction for unobserved variants, we split the training dataset into 90% training data and assessed the correlation between predicted and observed effects in the remaining 10% of test data. We repeated this procedure eight times. We do not consider a validation set since there are no hyperparameters to be optimized. We also do not filter our test or train set for less represented amino acids, as done in other studies[10] to enable prediction of all possible amino acid variants. We do not estimate the generalization error for either the three or four position library due to the nature of the observations. The three position library contains measurements for all combinatorial variant, leaving no additional variants to generalize to. For the four position library, only a subset of amino acid variants at each position were assayed. Hence, this model cannot be expected to generalize to variants that contain amino acids that have not been observed.

Estimating the total number of functional variants

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We estimate the total number of functional antitoxin sequences with at least half-maximal neutralization of the toxin using two ways orthogonal approaches, with both approaches predicting about 10^{10} functional sequences. The first estimate is based on the empirically measured distribution of fitness effects (Fig. 2). Here, we calculate the fraction of sampled sequences that are measured to be functional at each mutation distance from the wild-type sequence, and multiply this fraction by the total number of sequences that exist at this mutation distance. The total number of functional variants is the sum of all variants at each mutation distance. This approach assumes that the sampled sequences are a representative sample of all possible random mutations at each mutation distance, and does not consider the noise in the estimate of the fraction of functional sequences. For example, there are no functional sequences among the set of variants that show seven substitutions, likely due to the lower number of variants sampled at this mutation distance.

To address these issues, we also use the supervised, ten position oracle function to estimate the total number of functional sequences. Here, we generate one million random synthetic combinatorial variants, and ask the oracle function to predict what fraction of these variants reaches half-maximal fitness, p(predicted GR>0.5). We then estimate the fraction of true positive predictions, p(observed GR>0.5|predicted GR>0.5), by sampling with replacement the observed variants, for which observed and predicted growth rate data are available, to match the distribution of predicted growth rates for the synthetic sequences. Then we calculate the fraction of these subsampled observed variants with GR>0.5 among variants with predicted GR>0.5. Finally, we multiply these two fractions (p(predicted GR>0.5) and p(observed GR>0.5|predicted GR>0.5)) to estimate the fraction of synthetic variants that are both predicted and observed to be functional, p(observed GR<0.5, predicted GR>0.5). We then multiply this probability by the total number of possible variants (20^{10}) to estimate a lower bound on the total number of functional sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Unsupervised protein variant scoring and generation

We trained the RES classifier model as described previously[34]. Briefly, the RES model is trained on predicting the amino acid identity of a masked residue given a median of ~500 surrounding atoms, excluding hydrogens, from the protein database (PDB). Importantly, the training and test set are split according to domain-level structural CATH topology[23,25], and training examples are down-sampled to the least common amino acid to prevent biased learning of the classifier. The trained model achieves around 53% accuracy on the held-out test set, similar to what was achieved previously[34].

To score combinatorial variants, we fed the structural environments surrounding each mutated antitoxin residue to the RES model and obtained classifier scores for each amino acid at these positions. We then summed the classifier scores to result in a combinatorial variant effect prediction.

To sample from the RES model, we converted these site-wise amino acid scores into probabilities using the Boltzmann distribution ($p(amino\ acid)\ \alpha\ e^{\frac{-amino\ acid\ score}{t}}$). We then normalized these probabilities at each site. Using equal weighting between sites, we sampled 500 random sequences at varying temperatures (t $\in$ {0.1, 0.5,0.7,1,1.5,2,2.25,2.5,2.75,3,4,5}) and deduplicated the sampled sequences for evaluation.

Model weights and sampling strategies for proteinMPNN and ESM-IF followed previous studies[24,26]. For proteinMPNN, we generated 30 sequences at each temperature (t $\in$ {0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5}). For ESM-IF, we generated 100 sequences and deduplicated these variants at each temperature (t $\in$ {0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5}). We then evaluated the growth rate effects of generated variants using the supervised oracle function

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

trained on the ten position library data, which is estimated to have low generalization error when predicting unobserved variant effects (Fig. 3). In order to compare these sampled sequences fairly between models, we subsampled the number of unique generated variants for the RES and ESM-IF model to 30, which is the number of unique sequences sampled across most temperatures for the proteinMPNN model.

Code availability

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

Data availability

Raw sequencing read data will be made available upon publication.

Primers used in this study

See supplementary table ST1.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

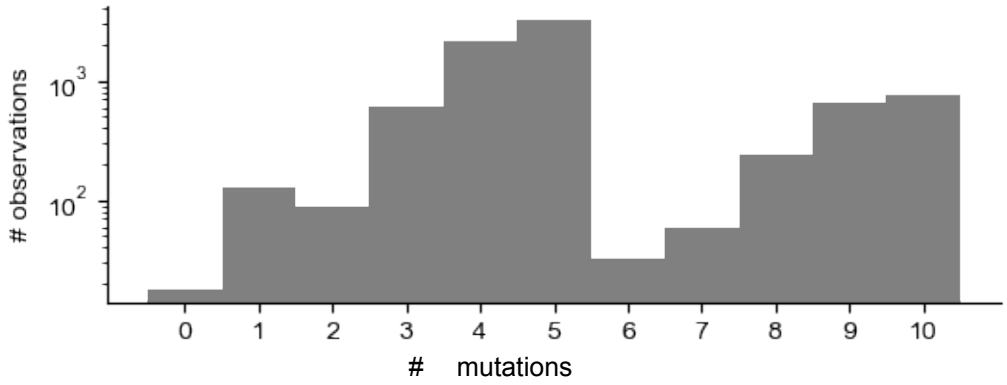
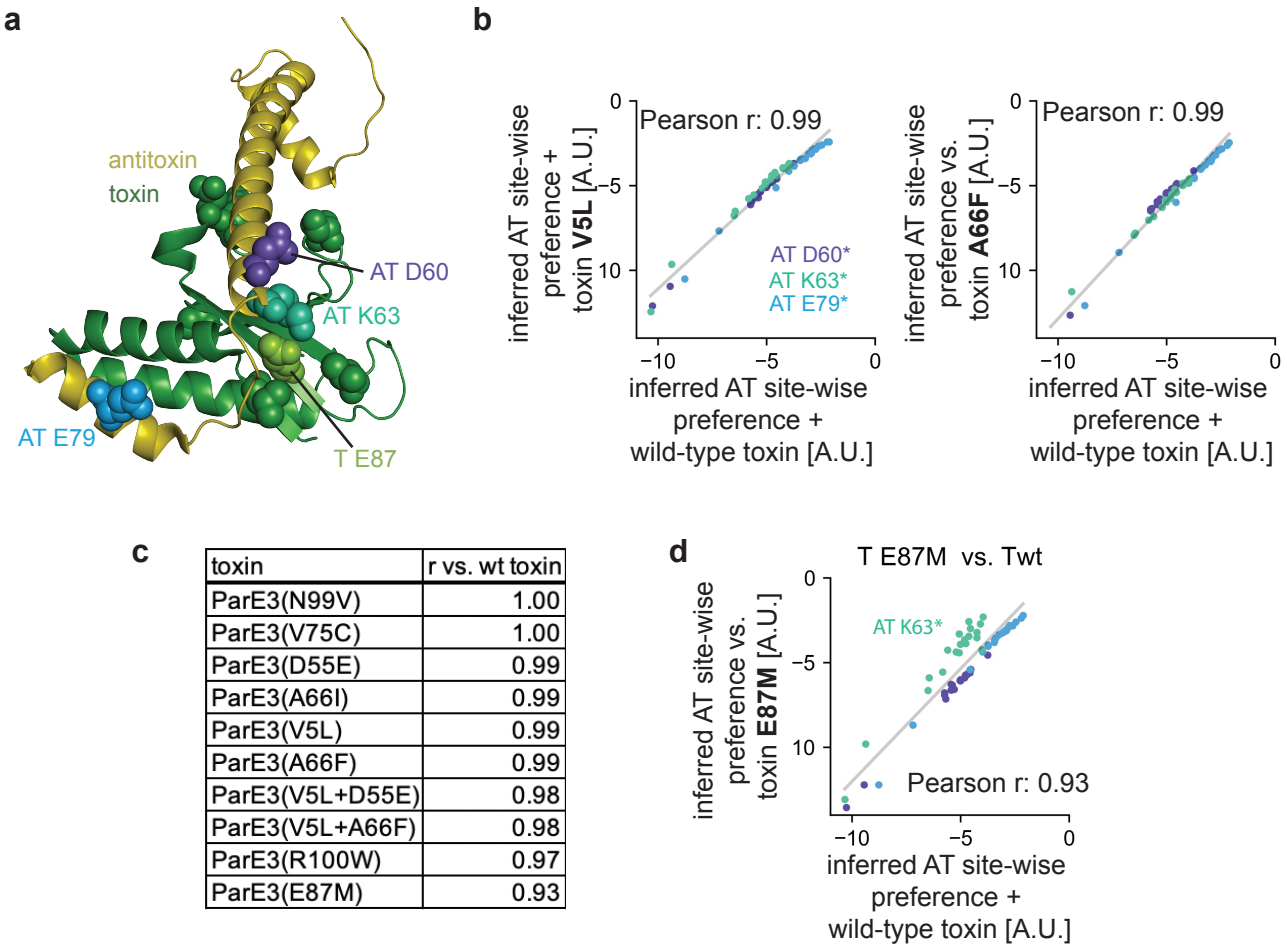


Fig. S1: Number of mutations measured for the ten position antitoxin library.
The number of observations is shown for each number of mutations of antitoxin variants with respect to the wild-type antitoxin.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



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

Figure 3: Site-wise preferences are altered by mutations in contacting residues.
a, Crystal structure (PDB ID: 5CEG) indicating that the combinatorial antitoxin ParD3 library from dataset 1 was screened for neutralization against 10 different ParE3 toxin single substitution variants. 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, Antitoxin ParD3 site-wise preferences inferred by the nonlinear, independent model correlate almost perfectly when inferred in wild-type toxin background or toxin mutant background ParE3(V5L) (left) or ParE3(A66F) (right).
c, Pearson correlation coefficients, r, shown for inferred sitewise antitoxin preferences in the background of 10 different toxin ParE3 single substitution variants vs. the wild-type toxin ParE3.
d, Antitoxin site-wise preferences differ only at antitoxin position K63 in the background of a toxin variants that contains a substitution, E87M, at a contacting position.

Fig S2

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.31.514613; this version posted November 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

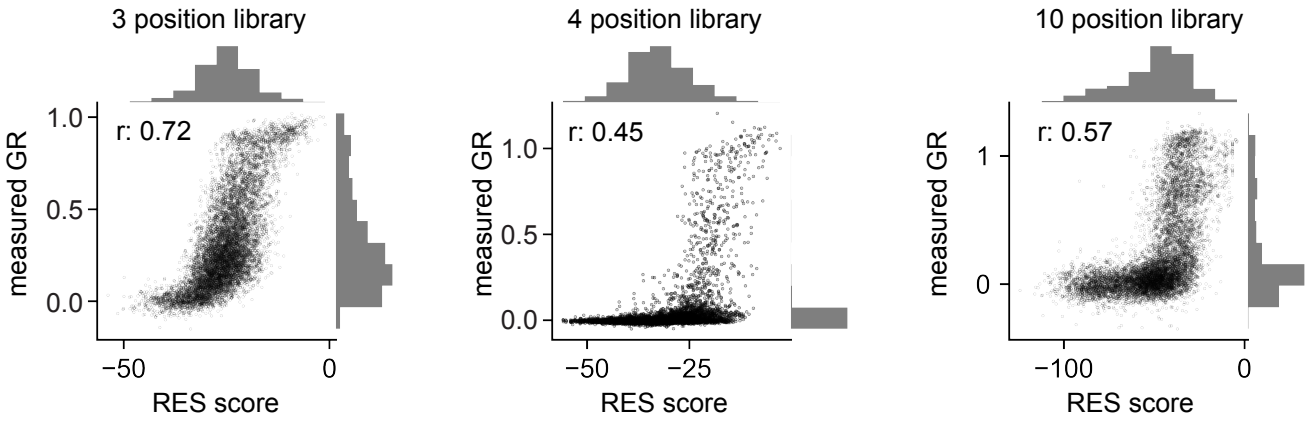


Fig. S2: RES model scores correlate with measured variant effects.
The summed scores for each position from the RES model are plotted versus the measured growth rate effects (GR) for each of the 3 position, 4 position and 10 position antitoxin libraries. The pearson correlation coefficients, r, are indicated.