PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

In vivo functional phenotypes from a computational epistatic model of evolution

Sophia Alvarez[a,1], Charisse M. Nartey[a,1], Nicholas Mercado[a], Jose Alberto de la Paz[a], Tea Huseinbegovic[a], and Faruck Morcos[a,b,c,2]

Edited by Terence Hwa, Department of Physics, University of California San Diego, La Jolla, CA; received May 26, 2023; accepted December 19, 2023

Computational models of evolution are valuable for understanding the dynamics of sequence variation, to infer phylogenetic relationships or potential evolutionary pathways and for biomedical and industrial applications. Despite these benefits, few have validated their propensities to generate outputs with in vivo functionality, which would enhance their value as accurate and interpretable evolutionary algorithms. We demonstrate the power of epistasis inferred from natural protein families to evolve sequence variants in an algorithm we developed called sequence evolution with epistatic contributions (SEEC). Utilizing the Hamiltonian of the joint probability of sequences in the family as fitness metric, we sampled and experimentally tested for in vivo β-lactamase activity in *Escherichia coli* TEM-1 variants. These evolved proteins can have dozens of mutations dispersed across the structure while preserving sites essential for both catalysis and interactions. Remarkably, these variants retain family-like functionality while being more active than their wild-type predecessor. We found that depending on the inference method used to generate the epistatic constraints, different parameters simulate diverse selection strengths. Under weaker selection, local Hamiltonian fluctuations reliably predict relative changes to variant fitness, recapitulating neutral evolution. SEEC has the potential to explore the dynamics of neofunctionalization, characterize viral fitness landscapes, and facilitate vaccine development.

evolutionary dynamics | epistasis | sequence evolution | direct coupling analysis | sequence–fitness landscape

Significance

The ability to observe viable step-wise changes to functional sequences evolved computationally from extant sequence data is a powerful tool. We developed a model of neutral evolution capable of preserving the statistics of observed proteins while generating sequences with extensive changes that, nevertheless, preserve the functional characteristics of their ancestors. We validated experimentally, in an antibiotic resistance protein of the beta-lactamase family, how this model produces evolved enzymes that maintain or improve their ability to inactivate ampicillin. Sequence-based computational models of evolution such as those presented herein provide us better insight into the process of neutral evolution and increase our understanding of the dynamics needed to preserve functional fitness. Potential applications include protein design and critical predictive power regarding pathogen landscapes during unrelenting epidemics.

Important features of protein structure, their functional capabilities, and the constraints imposed during the course of evolution can, in principle, be elucidated from sequence data and used to develop models of sequence evolution. The value of such models rests in the fact that these tools help us to understand not only past events, but the driving forces of protein-sequence change. Traditionally, models characterize subsets of statistical features found in natural sequence data, often requiring the application of multiple theories and practices to paint a comprehensive picture of evolution. We developed a model that unifies such features and uses them to guide unexplored evolutionary trajectories for sequences in specific protein families (1). This model called sequence evolution with epistatic contributions (SEEC) utilizes a global inference model to recapitulate family sequence statistics determined from evolutionarily related epistatic interactions. The algorithm proceeds to sequentially evolve novel protein sequences with the potential to retain wild-type (WT) functionality based on a conditional probability that takes into account epistasis and the sequence context at each evolutionary step.

The SEEC model incorporates epistatic information provided by direct coupling analysis (DCA) (2), a joint probability covariance model that utilizes both pairwise and single-site statistics to infer the family couplings (e_{ij}) and local fields (h_i) parameters of the Potts model of the protein family sequence space (2, 3). SEEC models neutral evolution by exploring new sequence space while preserving family-like function. In these simulations, SEEC unifies various evolutionary models with epistasis and the emergent properties of overdispersion, gamma distribution of substitution rates across sites, heterotachous sites, and evolutionary Stokes shifts or entrenchment (1). To demonstrate the significance of epistatic constraints, other groups have found that the application of coevolutionary information within molecular binding affinity highlighted the incorporation of epistasis and a changing mutational fitness which better modeled the dynamics of antibody evolution (4), while the development of an epistatic inference model that utilized time-series genetic data better simulated complex selection that can be applied to the analysis of virus, bacteria, and cancer allele evolution (5). Entrenchment is another epistatic sequence evolution phenomenon that has been recently observed experimentally (6–9). Another key aspect of the SEEC evolutionary simulations involves the predictive power

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]S.A. and C.M.N. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: faruckm@utdallas.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2308895121/-/DCSupplemental.

Published January 29, 2024.

of the sequence Hamiltonian, a statistical energy calculated for each evolved sequence based on the probability of retained family-likeness. Many studies have explored the interconnection between the Hamiltonian, free energy landscapes, and protein fitness showcasing how family alignments encode biological information such as folding (10, 11), melting (12), mutational space (13), and molecular interaction potential (14). Subsequently, the Hamiltonian energies are often correlated with distinguishing fitness such that sequences with lower statistical energies are more probable in regards to the landscape of the family (15). Experimental data also show strong correlations between empirically measured fitness differences and changes within the Hamiltonian energy of a given family (16–18). The retention of these aforementioned evolutionary statistical features affirms the power of the Hamiltonian fitness landscape and its use in understanding the dynamics of sequence–function adaptation (19, 20). Altogether, the potential for SEEC to produce functional proteins motivates our rationale to further assess this evolutionary model with a direct experimental counterpart.

Our primary goal in this work is to validate the capabilities of this model of epistatic evolution to produce sequences that are viable in vivo, so we focused on the *Escherichia coli* (E. coli) antibiotic resistance protein TEM-1 (UniProt ID P62593). The β-lactamase family is a convenient biological system for testing evolutionary models due to the ease of assaying protein activity with survival in the presence of antibiotics. Additionally, the plethora of sequence information and previous research available including the successful analysis of coevolutionary data for the β-lactamase family (17, 21, 22) makes this an ideal system to assess the performance of SEEC experimentally. We utilized SEEC to computationally evolve TEM-1 using parameters inferred from both mean-field and Boltzmann machine learning models (2, 23). We identified and synthesized key variants that, when expressed from a plasmid, protected *E. coli* from ampicillin on par with or even better than the WT enzyme. Remarkably, some of these successful variants undergo about 448 substitutions and reversions leading to 47 point mutations when compared to the WT TEM-1 β-lactamase. The number of potential sequences represented by this number of changes is enormous and requires a trustworthy model to navigate this unexplored mutational space. Observing viable step-wise changes to functional sequences evolved dynamically from extant sequence data is a powerful tool; as such, SEEC exemplifies a model of neutral evolution capable of preserving the statistics of observed evolution as well as specifying generated sequences with the functional characteristics of its ancestors.

Results

In order to assess SEEC's evolutionary modeling experimentally, we first simulated computational protein evolution (Fig. 1*A*). To do this, we compiled a multiple sequence alignment (MSA) for the antibiotic resistance family of β-lactamase enzymes (PF13354) which is used as the input for DCA to infer the family couplings (e_{ij}) and local fields (h_i) parameters of our Potts model (Fig. 1 *A*, *Top*). We generated models using both mean-field (2) and Boltzmann machine learning (23) implementations of DCA (mfDCA and bmDCA) to compare the performance of each method. There are a variety of epistatic model inference methods that could yield similar statistical properties (24) and features of sequence evolution (1), but here, we focus on the ones we extensively analyzed previously. Further on, we will discuss differences found in the variant phenotypes from each of these inference methods. We expect other generative models such as mi3-GPU (25), autoregressive (ar)DCA (26), and adaptive cluster expansion (27) and nongenerative models like pseudolikelihood method (plm) DCA (28) and GREMLIN (29, 30) to produce related but distinct results and may be intriguing alternative inference methods for future studies.

Starting with the *E. coli* WT TEM-1 (UniProt ID P62593), variants are computationally evolved through the SEEC model. Specifically, sites to be mutated are chosen at random over the entire protein. For a particular chosen site, a substitution is made based on a conditional probability distribution for mutations that is calculated based on the family model parameters (1) (*Materials and Methods*). Once the simulation of N number of steps is completed, a Hamiltonian, or statistical energy, is calculated for each sequence along the evolutionary trajectory. Note that a step in the simulation represents an evolutionary event related to mutation but is not intended to represent evolutionary time. Using these scores and various other selection parameters discussed further on, key variants are selected for the next phase of in vivo experimental testing. In the experimental phase (Fig. 1 *A*, *Bottom*), we obtained synthesized versions of these variants, cloned into an expression vector under the control of an inducible promoter (*Materials and Methods*). *E. coli* clones were grown under the challenge of ampicillin and the optical density was monitored over time. The results from the variant growth assays were used to feedback into the SEEC model for the optimization of computational evolution.

We employed three strategies for choosing variants from our evolutionary trajectories. First, in order to sample the evolutionary trajectory, we established three areas for variant selection: the beginning, middle, and late portions of the trajectory (Fig. 1*B*). In doing this, we set to explore whether there existed a threshold for the number of changes one single protein could undergo while still retaining the original function. Second, we targeted sets of variants that fell sequentially across evolutionary steps (Fig. 1*C*). These sets would include triplets of protein sequences that differed only by a single point mutation between each step. Our hypothesis is that SEEC recapitulates neutral evolution, in that fitness might be compromised in some steps, but as long as the gene remains viable, later changes can improve fitness in vivo; this strategy for variant exploration aims to assess the presence of this feature. Last, we analyzed variants across a range of Hamiltonian scores: many with favorable scores that were increasingly negative, often found in wells within the Hamiltonian trajectory, some with more positive scores, found in peaks along the trajectory, and other sequences with average scores picked from areas between wells and peaks (Fig. 1*D*). There is evidence that with optimized simulations and generative models that utilize a Markov search process, you can produce biologically active sequences (31, 32). It follows then that with such models, local changes in Hamiltonian energy that result from the simulation parameters could also be predictive of functionality (33). In using these strategies, we aimed to further understand the multifaceted predictive power of simulated evolution via SEEC.

Inclusion of Simulation Restrictions Optimizes SEEC Functional Phenotypes. In Phase I of testing, we used models inferred using bm and mfDCA with an input alignment based on the Pfam domain model (N = 202). Predicted contact maps revealed the high quality of these models (*SI Appendix*, Fig. S1), and yet of the 18 variants chosen across both trajectories, bacterial cultures expressing the 3 bmDCA variants (with 2, 3, and 4 amino acid changes) immediately died and exhibited function and growth only after a delayed period, a phenotype we have termed reanimation (*SI Appendix*, Fig. S2). The earliest mfDCA variant,

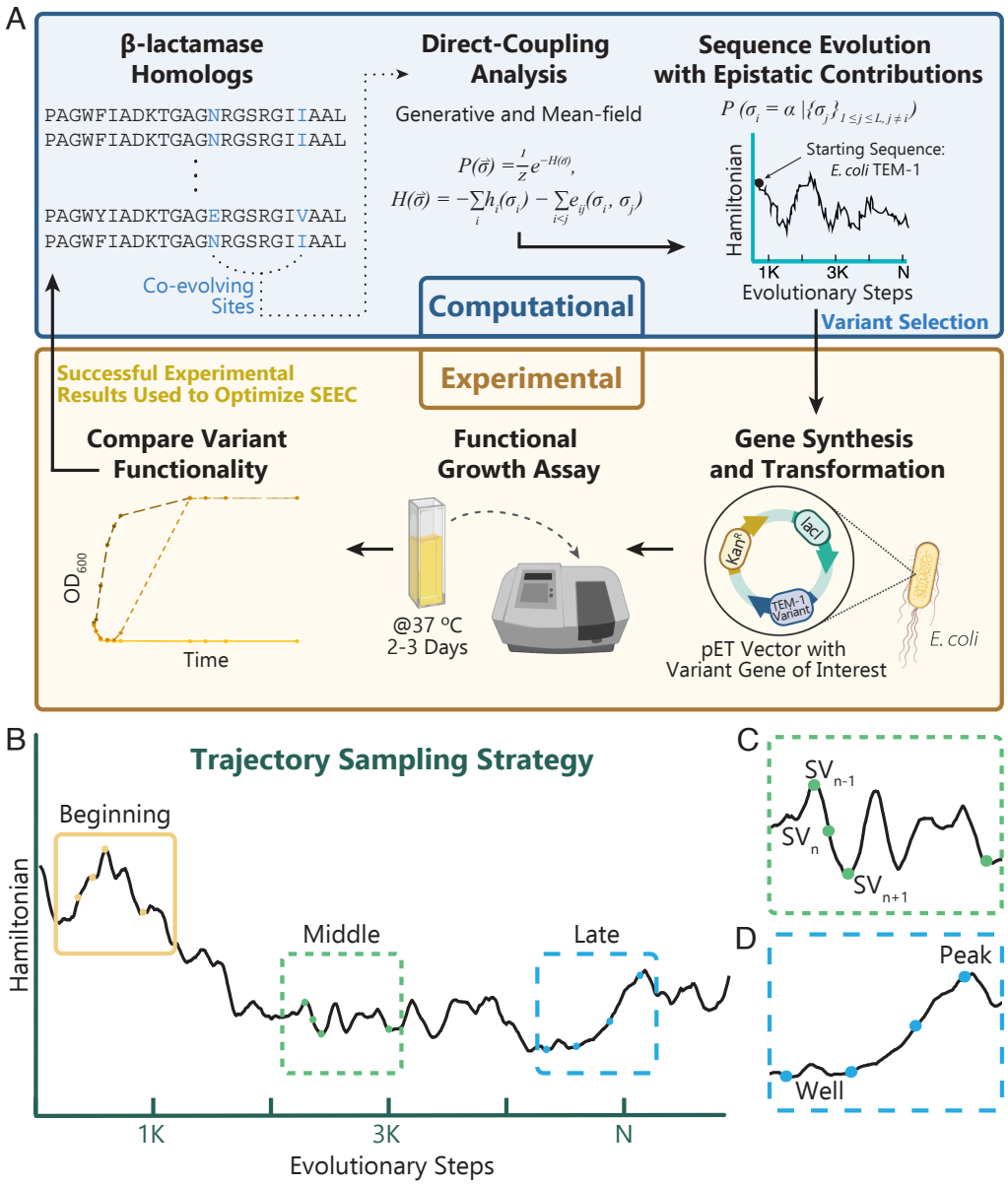

Fig. 1. Schematic representation of the computational evolution and experimental validation cycle. (*A*) The MSA informs the statistical parameters for both the mean-field and generative (Boltzmann machine) DCA models. Starting with the *E. coli* wild-type TEM-1, variants are computationally evolved through a process of iterative site mutation over N evolutionary steps (SEEC) and key variants are selected for testing. Each TEM-1 variant is cloned into a pET vector with essential expression and selection genes. Once transformed, the cultures are tested in the presence of antibiotics. The results of variant functionality are used as feedback for the optimization of the computational evolution. (*B*) Strategy for sampling variants across evolutionary trajectories. To completely sample the evolution, three areas are established: the beginning, middle, and late portions of the trajectory. Within a given area, additional selection attributes include (*C*) sampling sequential evolutionary steps and (*D*) selecting variants with increasingly negative Hamiltonian scores, described as wells of sequence space, and variants with more positive scores, sampled from peaks along the evolutionary trajectory.

with three substitutions, did survive on par with the WT at both minimum inhibitory concentration (MIC) and standard levels of ampicillin (50 μg/mL) (*SI Appendix*, Fig. S5). The remaining 14 variants, however, were completely inactive. All SEEC-amino acid (SEEC-aa) variant sequences can be found in Dataset S1, and (*SI Appendix*, Fig. S4) shows the Hamiltonians of each variant relative to the input family distribution. A more detailed summary of these results can be found in *SI Appendix*, *Text*.

With the majority of Phase I variants being nonfunctional, we adjusted relevant aspects of the computational process for further optimization. One concern we identified was that the evolutionary model parameters were inferred from the Pfam domain family, which only contained 202 out of 263 total amino acids (*SI Appendix*, Fig. S1*C* and *Materials and Methods*). Since mutations acquired during the simulation only occurred in this area, there was a potential that the evolved sequence might have lost compatibility with the retained WT portions outside of the domain. To address this issue, we chose to use an MSA queried on the entire *E. coli* TEM-1 sequence without the signal peptide (*SI Appendix*, Fig. S6*C*). In doing so, we increased the specificity of our MSA for TEM-1-like proteins but also decreased overall diversity as the effective number of sequences changed from 3,834 to 1,152. The impact of this greater specificity can be seen in the comparison between the Pfam contact maps (*SI Appendix*, Fig. S1) where the number of true positives is greater than the number of

hits for the TEM-1 whole protein contact maps (*SI Appendix*, Fig. S6). Although our current hypothesis is that our model can produce functional proteins, they might not be optimized for a specific cellular context. Increasing sequence diversity is known to improve structure prediction, but granting the simulation access to broader sequence space could be pushing simulated proteins toward different realms of taxonomic evolution that are discordant with the organismal context of the input sequence (17, 34). While all the β-lactamase family members are catalyzing essentially the same chemical reactions, the proteins themselves are acting in different biological contexts. The physiology of this process includes protein–protein interactions, optimal growth temperatures, expression/translation regulation, and so on—all which require specific context provided by the host organism. Therefore, considering the taxonomic context of an evolved gene influences the potential of not just its biochemical functionality but its physiological implications as well (35, 36). We reasoned that funneling the explored sequence space to that which corresponded to the specific physiological context of *E. coli* would enhance the goal of pursing evolutionary trajectories with functional, novel proteins. In addition to changes made to the input-aligned sequences, we also made adjustments to the SEEC algorithm to address obstacles experienced in the first iteration of experimental trials. For instance, we modified the mutation criterion so that the algorithm could no longer introduce gaps or select gaps for mutation during the simulation which prevented shortened or lengthened output proteins. Additionally, we again selected amino acid substitutions based on a conditional probability distribution, but this time, only residues that were accessible via a single nucleotide change could be accepted. Bisardi et al. found that in making these changes, simulations from a similar model better correlated with in vivo functional data (37). These changes further established the biological relevance of the SEEC model (hereafter referred to as SEEC-nucleotide or SEEC-nt) with the goal of improving variant functionality.

The final improvement came from optimizing the simulations themselves by modulating the model selection temperature (14, 31). During the initialization of the evolutionary simulations, there exists the opportunity to regulate the family couplings (e_{ij}) and local fields (h_i) statistics with selection temperature (T) from Eq. **3** (*Materials and Methods*). Using the Hamiltonian from Eq. **2** as a representation of statistical energy, adjusting the temperature of these parameters learned from the protein family serves as a method to adjust the energetic exploration of the mutational space. During Phase I, we did not select specific temperatures for each simulation as both mfDCA and bmDCA inferred models were ran at $T = 1$. From these experiments, we saw improved results with the lower sequence energy trends from the mfDCA SEEC-aa variants. Moving into Phase II, we applied the same concept to our bmDCA model to sample at temperatures less than one for resulting sequence trajectories trending toward favorable Hamiltonian values.

SEEC-Nucleotide Produces Variants with Improved Functionality. Using SEEC-nt, we again ran simulations using parameters inferred from both mfDCA and bmDCA and generated 5000-step evolutionary trajectories starting from the WT *E. coli* TEM-1 sequence. In contrast to error-prone polymerase chain reaction (PCR) mutagenesis or saturation mutagenesis, our global model informs the propensity for a change at the current step by all of the mutations that have come before in the evolutionary simulation. From these simulations, following the same strategies described earlier, we chose thirteen variants using both mfDCA and bmDCA to infer parameters (Dataset S2). For these trajectories, the late variants that had evolved the most had acquired 34 point mutations for a variant called Late_mf_3_NT and 58 point mutations for Late_bm_3_NT compared to the WT TEM-1. When these late variants were queried on BLASTp (38), the top hits were to the *Pseudomonas aeruginosa* TEM-136 that has 33 different point mutations from Late_mf_3_NT and a *Citrobacter freundii* class A β-lactamase that has 48 different point mutations from Late_bm_3_NT. This further supports that these variants are exploring novel sequence space rather than becoming another extant protein. Fig. 2*A* details the accumulation of mutations for the variants chosen from the bmDCA evolutionary trajectory. While there are some mutation steps that are reversions back to the WT sequence (blue dots), most mutations carry the protein into novel sequence space with the late variants having close to 60 different substitutions. Note that variants Beg_bm_4 through 6 were sampled from the earliest part of the trajectory, and as such have the least accumulated mutations. We selected these variants in order to make direct comparisons to the mfDCA sequences with equal numbers of mutations.

Panels *B* and *D* on Fig. 2 contain the trajectories with points and areas marked for the variants chosen for testing. For the mfDCA model, due to the changes made in Phase II, we increased the simulation temperature to 1.5 to see noticeable changes in the protein evolution (Fig. 2*B*, *SI Appendix*, Fig. S7, and *Materials and Methods*). Here, we can see that the trajectory favorably becomes more negative, and the variants' Hamiltonian scores relative to the simulation temperature remain similar to WT and remain within the distribution of scores for the family (*SI Appendix*, Fig. S8).

Initially, we tested in vivo functionality of these variants at the MIC for ampicillin (5 μg/mL) and found that growth was not challenged at all, so we raised the ampicillin concentration to 50 μg/mL (Fig. 2*C*). Final samples were collected and Sanger sequenced; quality chromatograms covering the entire gene were obtained for the majority of the samples, enabling us to conclude that our sequences of interest remained intact during the experiment. Also, no significant population with additional compensatory mutations acquired during the assay can account for the observed phenotype (see *SI Appendix* for Chromatograms). For the mfDCA SEEC-nt variants, even this level of 50 μg/mL ampicillin was not a challenge as they all grew at rates on par and sometimes better than WT (*SI Appendix*, Fig. S9).

Likewise, the bmDCA trajectory also becomes favorably negative; however, most of the middle and late variants' scores are outside of the family range (Fig. 2*D* and *SI Appendix*, Fig. S8). As expected, the simulation temperature for the bmDCA model had to be modulated to achieve a favorable Hamiltonian trend, so this simulation was run at $T = 0.75$ (*SI Appendix*, Fig. S10). Interestingly, despite requiring a low selection temperature, the bmDCA simulations still explored vaster sequence space than the mfDCA simulation ran at twice the temperature (compare Percent ID between *SI Appendix*, Figs. S7 and S10 and compare the pairwise percent ID histograms in *SI Appendix*, Fig. S11, which show that the diversity of the simulation outputs increases with temperature for both models, but is always higher for bmDCA). When tested at 50 μg/mL ampicillin, select beginning and middle variants (Beg_bm_5_NT, Beg_bm_6_NT, and Mid_bm_4_NT) grew on par or faster than WT, while all other variants facilitated only slow growth or reanimation (Fig. 2*E* and *SI Appendix*, Fig. S12). When considering the difference in results between the SEEC-aa and the SEEC-nt variants, the beginning variants for bmDCA SEEC-aa only had 2, 3, and

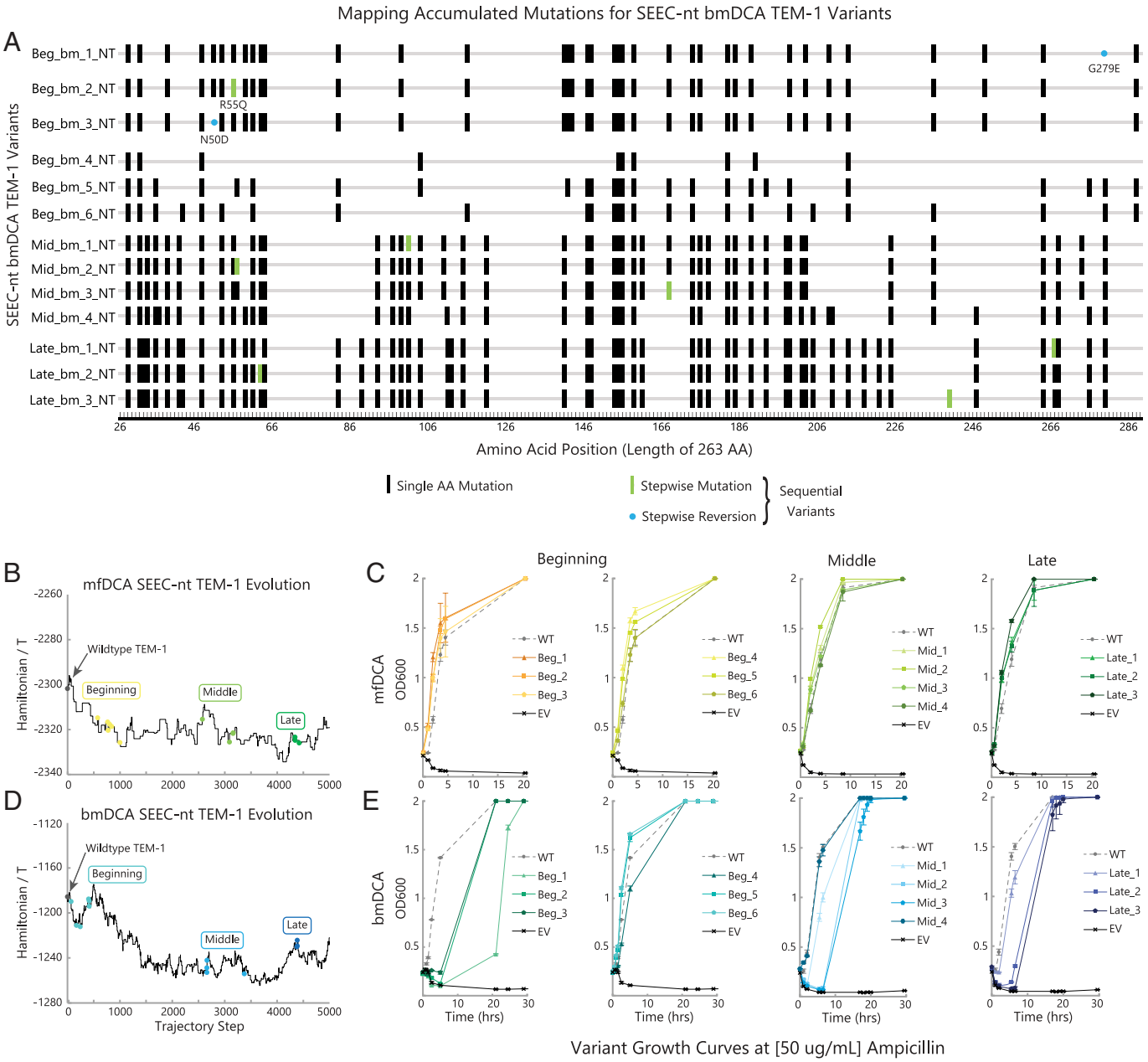

Fig. 2. Nucleotide sequence evolution with epistatic contributions (SEEC-nt) Phase II variant selection and function results. (*A*) Sequence diagram of variant mutations for the SEEC-nt bmDCA TEM-1 variants. Singular change between sequential variants displayed as green bars for mutations or blue circles for reversions. (*B*) Mean-field (*mf*) DCA model SEEC-nt trajectories with selected variants at $T = 1.5$ and (*C*) *E. coli* growth curves for mfDCA variants. (*D*) Boltzmann machine learning (bm) DCA model SEEC-nt trajectories with selected variants at $T = 0.75$ and (*E*) *E. coli* growth curves for selected bmDCA variants. Cultures were grown in 50 μg/mL ampicillin. Data points are the mean of 3 experimental replicates, and error bars represent SDs. EV = Empty Vector.

4 mutations that nevertheless resulted in an impairment of function and delayed growth potential leading to reanimation. Correspondingly, while there exists the threshold model that random mutations would impact stability and not function directly, once many of said mutations accumulate, their effect will exponentially decrease the protein's overall fitness (39). Specifically within TEM-1, Bershtein et al. (40) found that the synergistic accumulation of eight or more random mutations more than exponentially diminished fitness. Thus, the fact that bmDCA SEEC-nt generated a variant that retained WT-like activity, such as Mid_bm_4_NT with 47 substitutions, is a remarkable result. The potential number of sequences restricted by mutation in nucleotide space is approximately 9^{47} positional changes; hence, this variant exists in a possibility space that exceeds the total number of water molecules on the earth. Therefore, finding this sequence by chance is not plausible.

SEEC-nt Informed by mfDCA Evolves Variants with Enhanced Enzymatic Activity. To glean further insights into the difference in functionality among Phase II variants, we further analyzed the sequence trends and their behavior at even more challenging levels of antibiotics. In Fig. 3, we can see the expanded SEEC-nt trajectories with individual variants pinpointed along the simulation (Fig. 3*A*). To push the functional boundaries of these variants, we tested them all in cultures containing a challenging level of 100 μg/mL ampicillin (Fig. 3*B*). In comparing the

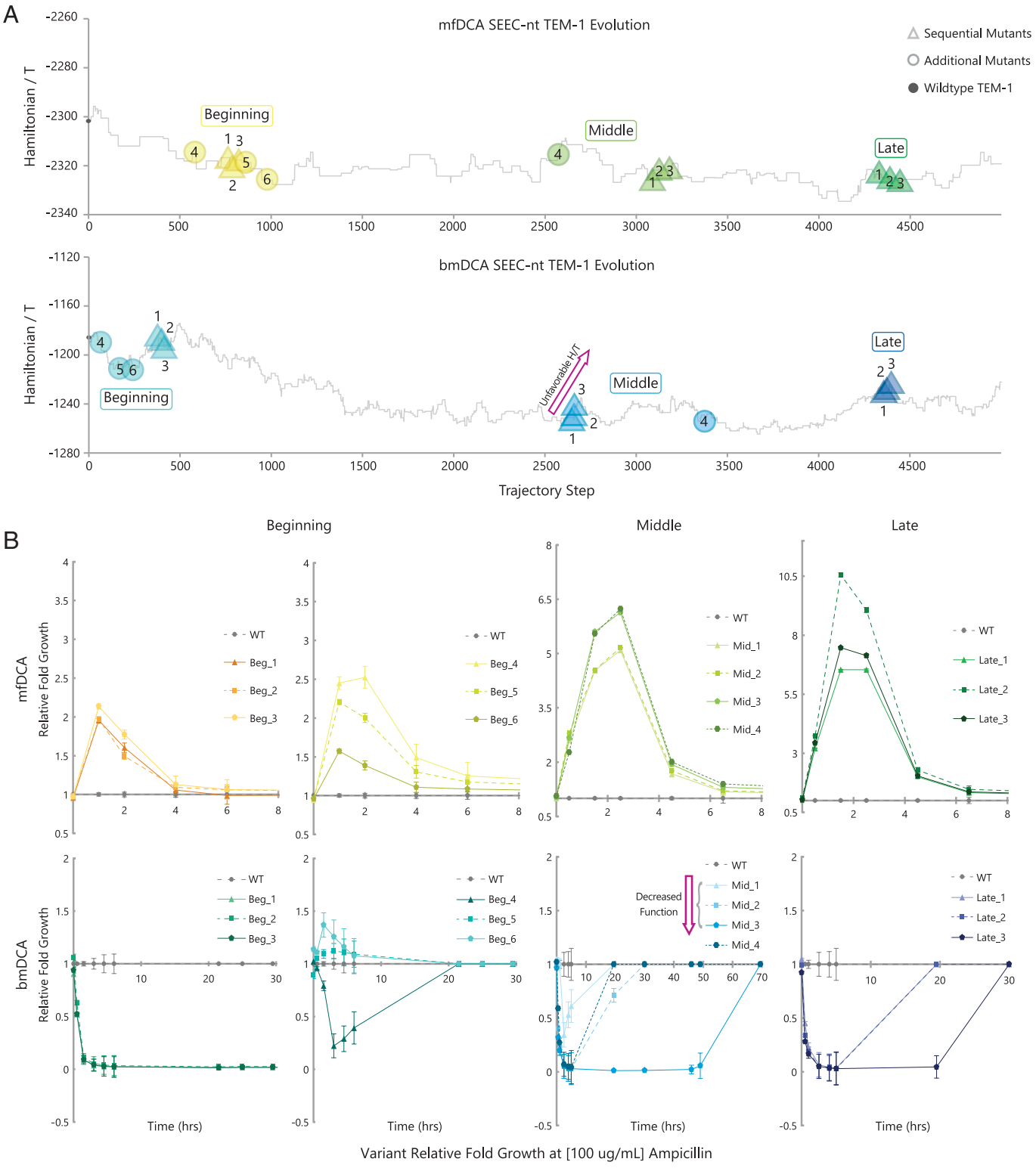

Fig. 3. SEEC-nt Phase II variant relative fold change in function compared to WT *E. coli* TEM-1. (*A*) Mean-field (T = 1.5) and Boltzmann machine learning (T = 0.75) expanded evolutionary trajectories and areas of variant selection. Points of variant selection from the trajectories are indicated with corresponding colors and shapes based on area and mutant type. The purple arrow highlights the pattern of the Hamiltonian divided by temperature score becoming increasingly unfavorable for the middle bmDCA sequential variants. (*B*) Relative fold growth graphs for variants picked from the mfDCA and bmDCA model SEEC-nt trajectories. Cultures were grown in 100 μg/mL ampicillin. Within the relative fold graph for the middle bmDCA variants, the purple arrow emphasizes the pattern from panel *A* where the Hamiltonian score becomes unfavorable, a decrease in variant functionality follows. Data points are the mean of three experimental replicates, then normalized to the mean of three positive controls detailed in Eq. **5**, and then normalized to the WT rationalized growth from Eq. **6** (*Materials and Methods*). The error bars represent the addition of propagated errors from the SDs of the measured samples.

difference in activity for the variants and WT protein, we calculated the ratio between the mean optical density of the experimental samples and their positive controls and proceeded to compare the rationalized growth between each variant and the WT (detailed in Eqs. **5** and **6**; see *Materials and Methods* for details). By performing this analysis, we found that many variants exhibited multiple relative folds of change in activity. Here, cultures that recovered from this challenging level of antibiotic at the same time or quicker than WT exhibit positive peaks of relative fold growth or a positive difference in growth that was 2 to 10-fold greater than that of WT. For variants with sub-optimal function, relative fold growth was less than 1 for the periods of the assay where bacterial growth was inhibited. Surprisingly, every mfDCA variant outperformed the antibiotic resistance activity of the WT protein by several folds (Fig. 3*B*, mfDCA Relative Fold Growth, and *SI Appendix*, Fig. S13). On the whole, for this mfDCA simulation, almost every variant achieved an increase in relative fold growth when compared to the activity of variants from the previous section of the trajectory. Remarkably, the best relative fold growth was seen from Late_mf_2_NT (Fig. 3*B*) with over a 10-fold increase in the rate of culture growth during the exponential phase of the growth assay. Therefore, even with an increasing number of accumulated mutations, the global downward trend of the Hamiltonian/T mfDCA SEEC-nt trajectory is highly predictive of improved variant functionality.

SEEC-nt Informed by bmDCA Highlights Predictive Power of the Local Hamiltonian Context. Beyond the global trend of the trajectory becoming more negative, we find that crucial, predictive patterns also occur in the local context; while using the SEEC model as a predictive tool to engineer proteins is not our main goal, a meaningful connection between the Hamiltonian and variant fitness adds to SEEC's robustness as a model of evolution through adaptive landscapes. For example, in the bmDCA simulation, the sequential middle variants increase in Hamiltonian/T scores in a step-wise fashion, gradually becoming less favorable (Fig. 3 *A*, *Bottom*). In the case of these variants, as the Hamiltonian becomes increasingly unfavorable, we see a decrease in ampicillin resistance as it takes each sequential mutant a longer period of time to reanimate (magenta arrows, Fig. 3). A similar pattern can be said of the bmDCA late variants as well. The same phenomenon occurs, but in the opposite direction, for the bmDCA beginning sequential variants. These sequences, albeit selected from an early, higher energy portion of the trajectory, are sequentially moving toward a local well of Hamiltonian energy. While these variants could not survive at this challenging level of ampicillin (100 μg/mL) (Fig. 3*B*, bmDCA Relative Fold Growth, and *SI Appendix*, Fig. S14), in Fig. 2 where the concentration was only 50 μg/mL, we again see a step-wise pattern in function. This time, however, the final sequential variant with the lowest Hamiltonian, Beg_bm_3_NT, retains the best functionality between the three proteins and can grow to saturation the quickest, followed by the second and then the first beginning variant.

While the global trend of the trajectory is meaningful, it is not enough to predict individual variant functionality. If the global movement were sufficient, then each variant from a negative trending trajectory would be functional. However, our data show this is not the case, as demonstrated with the difference in function between the bmDCA middle and late variants when compared to Beg_bm_5_NT and Beg_bm_6_NT. Despite having more positive Hamiltonians, these two variants outperform their later variant counterparts, even in the most challenging levels of ampicillin. This effect is not easily explained by the early positioning of these variants in the trajectory, as each has already accumulated 26 mutations. At the same time, in Fig. 3 we have two examples in which sequential variants that head toward a local peak coincide with decreased functionality (middle and late bmDCA variants), and one example in Fig. 2 in which sequential variants that head toward a local well coincide with increased functionality (beginning bmDCA variants), indicating that it is in fact these local Hamiltonian trends that are a more reliable factor for optimizing functionality.

Boltzmann Machine Learning Model Allows SEEC-nt to Explore Greater Sequence Space. Our results support the idea that there is a clear difference between the functionality of our variants chosen from models inferred using mean-field versus Boltzmann machine learning implementations of DCA. While both models produced functional variants, mfDCA consistently produced variants that outperformed the WT controls. Part of the reason for this could be that the variants chosen from the bmDCA model explore more sequence space, even with the constraints provided by making mutations in nucleotide space. The logos in Fig. 4 *A* and *B* depict the frequencies of amino acids found at each site in the two SEEC-nt variant pools that we experimentally synthesized and assayed. At several positions, highlighted by orange dots, the bmDCA model samples more amino acids (Fig. 4). This is especially meaningful because the mfDCA group has more variants with a substitution at these sites (*SI Appendix*, Fig. S15 *A* and *B*).

We also looked at sites that are mutated at least once in both trajectories, compared the substitutions made for the mfDCA and bmDCA trajectories (*SI Appendix*, Fig. S16) and noted the amino acids that were shared or unique to either of the models. As with the logos, this analysis similarly reveals that sequence space exploration is more extensive in the bmDCA model, as the number of unique amino acids are overwhelmingly greater in the bmDCA-based trajectory (compare orange versus green). Our chosen variants are representative of the simulation as a whole, as we see the same phenomenon occurring there as well (*SI Appendix*, Fig. S16, *Bottom*). To further demonstrate the greater freedom of bmDCA sequence space exploration, we analyzed the amino acid frequencies for each site across the 5K variants generated by each simulation (*SI Appendix*, Fig. S17 *A* and *B*). In the heat maps, columns with a dark red rectangle (frequency = 1) in a dark background (frequency = 0) indicate conserved sites that never changed throughout the simulation. Comparison of the two heat maps reveals that the bmDCA model produces far more sites in between these two extremes, signifying more sequence space exploration. Importantly, a scatter plot of the amino acid frequencies from both trajectories highlights that often, a position is conserved during the mfDCA trajectory ($X = 1$) but mutated during the bmDCA ($Y < 1$, 158 points, purple oval); it is rare, however, to see this in the other direction (four points, light blue oval, *SI Appendix*, Fig. S17*D*).

Do all amino acids get treated the same by both models, or are some amino acids more substituted in one model versus the other? To address this question, we compared amino acid frequencies for mfDCA versus bmDCA models by calculating the correlation coefficients for each set of amino acid frequencies across all sites (*SI Appendix*, Fig. S17*E*). In panel *F*, this calculation is shown for the 5K trajectory sequences as well as the subset that were chosen for experimental testing. For the trajectory data (x-axis), all of the Pearson correlations of these amino acid frequency vectors for the mfDCA and bmDCA models are over 0.85 for all amino acids

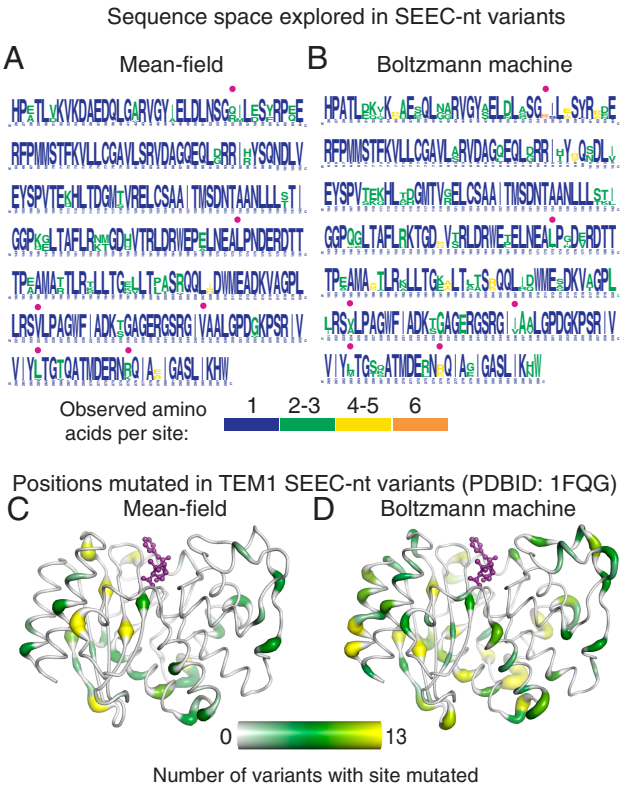

Fig. 4. Comparison of sequence space exploration within mean-field (mf) and Boltzmann machine learning (bm) Phase II variants. Logo of amino acid frequencies across the 26 variants tested, 13 from mf (*A*) and 13 from bm (*B*) models. Orange dots highlight examples of positions which showed more amino acid diversity in the bm group of variants, despite being mutated more often in the mf group of variants. Logo generated as a frequency plot with small sample correction using Weblogo version 2.8.2. (41, 42). Colors indicate groups of similar amino acids. (*C* and *D*) The frequencies of each position being mutated across the variants derived from mf (*C*) and bm (*D*) models are mapped onto the TEM-1 structure (PDBID: 1FQG) using a color gradient and putty thickness. Penicillin D, covalently linked to the nucleophilic serine residue, is depicted in purple ball and sticks and highlights the active site region.

except Asn, which is ∼ 0.70. The correlations are even higher for the chosen variants (y-axis). The amino acids that are most similar between the trajectories for the two inferred models (that is, close to the diagonal) are cysteine, tyrosine, and tryptophan, while the residues that change the most are asparagine and glutamine. Overall, the scatter plot of these correlations reveals that i) positional amino acid frequencies for each residue type are different across the mfDCA and bmDCA models, ii) the correspondence is high ($R = 0.88$) between the amino acid frequencies of the whole trajectory and the mutants chosen for experimental testing, and iii) points fall on either side of the line of unity indicating minimal systematic bias, confirming that our experimentally tested cohort are representative of the trajectory.

Substitutions for Both mfDCA and bmDCA Models Are Spread across the Structure While Avoiding Sites with Vital Functional Roles. In addition to looking at the differences in sequence space exploration between the two models, we also asked whether the locations within the structure of the substituted positions differed significantly. Localization of mutation sites on the three-dimensional structure of TEM-1 β-lactamase reveals that mutations from both models are distributed across the entire structure except for two regions: one buried helix (E64-L81),

Table 1. Mutations to residues nearby active site of TEM-1 for the two different inference methods: mfDCA (MF) and bmDCA (BM)

Position	WT residue	Mutated residue	MF	BM
168	Glu	Ala	2	-
173	Ile	Leu	13	-
235	Ser	Thr	8	8
103	Val	Ile	-	8
167	Pro	Thr and Ala	-	8
239	Glu	Asp	-	1

Columns MF and BM show the number of variants that contain the specified mutation.

which houses essential catalytic residue Ser70 and is nestled between a second conserved region, a helix turn helix (E121-L139)(orange regions in Fig. 4 *C* and *D* and *SI Appendix*, Fig. S15 *A* and *B*). Despite the conclusion from various computational mutation models predicting solvent-exposed sites to be more robust to mutation than buried sites (43–45), in our variant pool, there is no connection between the propensity for a site to be mutated and the degree of solvent exposure of that position in the structure (*SI Appendix*, Fig. S15 *C* and *D*). There is also conserved information about the active sites and correspondingly, the variants retain most of the active site areas. Specifically, the active site residues, Ser70, Gln166, and Asn170 are conserved in both variant cohorts (*SI Appendix*, Fig. S15 *A* and *B*). Even though these critical residues are never mutated, some mutated residues are within 8 Å of these active site residues (red dashes, *SI Appendix*, Fig. S18 *A* and *B*). Interestingly, the majority of the substitutions change to similar amino acids. In the cases where this is not true, the bmDCA pool has more variants with nonconservative substitutions (see bold rows, Table 1). Thus, compensatory changes can occur even within contact distance from sensitive active site residues.

Clearly, there is general conservation in the active site region, and while this is most likely constrained by catalytic requirements, we also investigated the role of other biophysical constraints such as protein–protein interactions. β-lactamase binding protein (BLIP) and BLIPII are proteins that abrogate the catalytic activity of β-lactamases by binding the conserved helix–loop–helix region (red region, *SI Appendix*, Fig. S18 *C* and *D*) of class A β-lactamases and inserting loops into the active site (46, 47). Biochemical studies have identified the key residues mediating the inhibitory binding activity of these two proteins (48). Mutations within our experimentally tested cohort are almost nonexistent among the BLIP and BLIPII binding hot spot positions in TEM-1: E104,L102, Y105, P107, K111, and M129 *SI Appendix*, Fig. S18 *C* and *D*). Taken together, these observations highlight the fact that SEEC-nt does not merely target positions that are "low hanging fruit," or changes that would be obviously nondisruptive, like surface residues. At the same time, the preserved regions are likely to be essential across the family for either structure, catalytic efficiency or even protein–protein interactions; this all demonstrates the benefit of a model informed by family statistics.

Discussion

In this work, we provide evidence on how SEEC, an epistatic evolutionary model, leads to sequence changes that preserve function. We observe this result not only for a few mutations, but many across the entire protein. As we saw with the initial SEEC-aa trials, the potential for a handful of mutations disrupting

function is evident; hence, the number of mutations acquired with SEEC-nt that continue to retain functionality is significant. In the case of the mean-field DCA model, we created variants with sequential changes that still preserved function, and in every case, resulted in improved antibiotic resistance capacity when compared to WT. These results further demonstrate that the SEEC model recapitulates neutral evolution, and that with mfDCA-inferred parameters, each sampled evolutionary step retained fitness and viability in vivo. To summarize these results, we have included an overview table of the influence of the varied parameters on the results presented here (*SI Appendix*, Table S1). This model presents a relatively faithful reflection of neutral evolution. Through these simulations, we observe how the proteins traverse sequence space, how their functions are impacted by various mutations, and the overarching, cyclical nature of evolution. Over time, sequences fluctuate for better or worse functional activity. These resulting proteins represent the culmination of thousands of evolutionary events, and in the end, the number of functional mutations after this magnitude of potential changes is remarkable. Although we have noticed in previous work that SEEC had convincing statistical features for evolutionary models, we now provide evidence of plausible evolutionary trajectories that lead to functional phenotypes in vivo.

In comparison, the enduring pursuit of protein engineering has resulted in the exploration of a vast multitude of modern methods. From directed evolution, to simulated biophysics, and now with recent machine learning technologies, more and more avenues of de novo protein design are being discovered. These methods have resulted in innovations such as greatly diversified capsid protein variants developed by machine learning (49), functional variants of a bacterial luciferase generated from variational autoencoders (50), and even the ability for deep learning methods such as ProteinMPNN to recover function for previously failed sequences created by physics-based designs such as Rosetta or Alphafold (51, 52). One of the latest advancements includes the deep language model ProGen trained on millions of sequences over thousands of families that can generate variants with specific properties that can function similarly to natural proteins (53). Although attaining functional sequences was a significant outcome of these simulations, the main goal prevailed in developing SEEC as an interpretable model of sequence-based evolution with comprehensive parameters and a connected series of changes over time. Not only do these results produce diverse proteins that biochemically function on par or even better than native proteins they also reveal the series of thousands of evolutionary steps of potential single mutations that the simulation explored to reach that point. Besides the functional results, the SEEC model presents explicit pathways that an individual protein experiences over the course of in silico evolution all while using native protein datasets to access previously unexplored sequence space.

When comparing the two statistical inference methods, we aimed to fairly evaluate simulated sequences by selecting a handful of variants from both models that had accumulated the same number of mutations during their respective simulations. Comparatively, SEEC-nt informed by bmDCA headed into novel sequence space faster than the mfDCA simulation, so variants had to be selected from the beginning portion of the trajectory to match the percent identity of the mfDCA selected variants (Dataset S3). In every comparison between variants with equivalent numbers of changes, mfDCA led to more fit, functional variants than their bmDCA counterpart. Concurrently, we found that mfDCA informed SEEC-nt consistently resulted in functional sequences that all had better antibiotic resistance activity than WT. While we have not ruled out the possibility that our inferred bmDCA Hamiltonians are optimized for natural rather than the semisynthetic antibiotic, ampicillin, we believe this phenomenon might not be a result of model selection but parameter selection. Indeed, from additional investigation, if the selection temperature is further lowered, bmDCA is capable of outputting sequences that are similar to those found in mfDCA trajectories (*SI Appendix*, Figs. S17*C* and S19). This includes variants that we assayed, potentially leading to trajectories that could perform similar to those of mfDCA. We envision that, because these mfDCA-inferred parameters tend to favor the contributions of the strongest couplings and the overall trajectory explored less sequence space, the mfDCA simulations could represent an evolutionary scenario in which the selection pressure is higher, thereby restricting the evolution to variants with increased fitness. On the other hand, bmDCA simulations also resulted in many functional proteins, just some with better and some with worse antibiotic resistance than WT. Here, because it more accurately captures lower frequency couplings from the input MSA and explores further sequence divergence, the generative model obtained with bmDCA at $T = 0.75$ could represent an environment with a lower selective pressure, allowing for more freedom to explore sequence space. Moreover, the fact that mfDCA tends to explore less sequence space than bmDCA could represent the difference between a short-term or local evolutionary exploration versus a long-term evolution of distantly related sequences. As it has been recently shown, patterns of diversity observed in nature differed depending on whether the comparison set was polymorphisms within a species (short term) or fixed differences across distantly related species (long term), with there being less pairwise site diversity in the short versus long term (36). It can be expected that generative models that capture the single and pairwise marginals of the entire family would be better able to output the improbable sequence changes that would be found in more distantly related proteins. Overall, these observations highlight the fact that SEEC, as a platform to model and learn about evolution, can incorporate epistatic constraints from different sources including experimental data. In addition, the epistatic relationships need not be limited to second order: Higher-order interactions have been found through both computational (54–56) and experimental methods (57–59).

To further decipher the difference in functional activity between the two versions of the model, we looked into elements that could inform the propensity for mutation across different regions, such as catalytic regions, specific protein–protein interfaces, and how they impact which substitutions will be tolerated at specific sites. We found that the BLIP and BLIPII binding hotspots were conserved among our tested SEEC-nt variants for both mfDCA and bmDCA models. As an enzymatic inhibitor, it would make evolutionary sense that BLIPs would target a region necessary for catalytic function. We speculate that inserting multiple mutations in this region, as is the case in our bmDCA mutants, has diminished catalytic function. Simultaneous mutations at multiple sites in this region have been shown to slightly harm catalytic efficiency (60). Given that there are more of these positions mutated in the bmDCA pool, this might explain the reduced activity of these variants in comparison to their mfDCA counterparts. Further studies exploring the BLIPs' binding activity of SEEC-nt variants could clarify the picture of evolutionary and physiological constraints on the explored mutational space in our simulations.

The future of the SEEC model includes exploring the prospects of neofunctionalization; by steering evolution into specific new fitness minima of the sequence–function landscape, novel functions and contexts could be investigated. While building functional models is important work, we can also apply this knowledge of evolutionary constraints now to other subjects. There exists the potential to trace functional pathways for the evolution of viruses using the existing sequence space. With these data, we could study viruses such as HIV or SARS-CoV-2 and potentially aid the development of vaccines or therapeutics for future variants before their arrival. Awareness of this virus fitness landscape allows us to utilize proactive design against the most probable variants (61–64). In addition, one can study how the genome adapts in the presence of SEEC-modified essential genes; the tunability of the function based on local Hamiltonian fluctuations might make new compensatory pathways available that are not accessible when β-lactamase activity is completely abrogated (65). This is more consistent with the phenomenon of gradual genetic change. The ability to use sequence-based computational models of evolution such as SEEC will continue to provide us better insight into the process of neutral evolution, ancestral reconstruction of sequences, novel protein-design applications as well as critical predictive power regarding pathogen landscapes during unrelenting epidemics.

Materials and Methods

Input Sequence Datasets. *E. coli* TEM-1 is a member of the β-lactamase2 domain family (entry ID PF13354) in the Pfam database (66). To test the ability of SEEC-aa to output functional variants of TEM-1, an MSA of homologous UniProtKB database sequences was obtained from the Pfam database. Any homologues with continuous stretches of gaps totaling greater than 5% of the model length ($N = 202$) were excluded from the final alignment, which came to 15,495 UniProtKB sequences. After reweighting sequences with 80% or greater sequence identity using a pseudocount of 0.5, the effective number of sequences was $\sim$ 3, 834. For the second phase of model testing, we made several changes in order to improve coverage of the protein domain. First, we modified the domain family definition so it included the entire sequence of TEM-1 (minus the signal peptide), in case there was a problem with accumulating mutations in the domain that eventually became incompatible with the upstream and downstream portions (compare *SI Appendix*, Fig. 1*C* and Fig. 5*C*). Second, with this updated model, we obtained a revised MSA by using the TEM-1 sequence as the seed from hidden Markov model (HMM) Build to construct an HMM profile (67, 68). With this profile, we then utilized HMM search to obtain matches within the UniProt database, including entries in both Trmbl and Swiss Prot (69). After filtering to 5% continuous gaps and reweighting as before, the effective number of sequences was $\sim$ 1, 152.

Parameter Inference and Hamiltonian. The DCA method (2) was then applied to the MSAs discussed before to estimate the direct coupling between all pairwise residues as well as the residue preferences at each position. As described in ref. 2, DCA utilizes maximum entropy modeling to estimate the joint probability distribution of protein sequences ($\vec{\sigma}$):

$$P(\vec{\sigma}) = \frac{1}{Z}\exp\left(\sum_i h_i + \sum_{ij} e_{ij}\right), \quad [1]$$

where Z is the partition function, the position of residues within the aligned domain or protein sequence are denoted as i and j, and parameters e_{ij} and h_i can be numerically inferred. The e_{ij} parameters quantify the coupling strength for residues i and j for all possible amino acid occurrence pairs. The amino acid biases for independent positions are captured by the parameter h_i. The sums of the e_{ij} and h_i parameters can be characterized as an energy function, or Hamiltonian (H):

$$H(\vec{\sigma}) = -\sum_i h_i(\sigma_i) - \sum_{i<j} e_{ij}(\sigma_i, \sigma_j). \quad [2]$$

The Hamiltonian represents a sequence statistical energy and has been predictive of functional and nonfunctional effects in proteins and RNA (17, 70–72). As it so happens, the inference of the exact parameters is an intractable problem, so they are estimated, instead, using multiple approximations with a variety of complexities and accuracy. In this work, we used both the mean-field implementation (2), which places an emphasis on the identification of highly coupled sites and is minimally complex, as well as bmDCA, which unlike mfDCA produces generative protein family models but is computationally expensive (23). In our previous work introducing the SEEC algorithm, we tested the statistical properties of the evolutionary trajectories using both mfDCA and bmDCA models as input parameters and found that both captured the statistical features found across several theories of neutral evolution (1). To investigate whether both types of models were able to produce functional variants, we performed mean-field and Boltzmann machine learning DCA implementations to infer the model parameters. Original codes for e_{ij} and h_i parameter inference by DCA were written in MATLAB (The MathWorks, Natick, MA) and previously published at https://github.com/morcoslab/coevolution-compatibility (18) and https://github.com/matteofigliuzzi/bmDCA (23).

Selection Temperature. One can introduce an additional parameter T to restrict the average value of the Hamiltonian being described by Eq. **2**. Analogous with treatments in statistical mechanics, this is called a selection temperature (14, 31) and it parameterizes the DCA distribution as:

$$P(\vec{\sigma}) \propto e^{-H(\vec{\sigma})} \rightarrow P(\vec{\sigma}; T) \propto e^{-H(\vec{\sigma})/T}. \quad [3]$$

This has the effect of modifying how the overall sampling is restricted. If a change had a probability p of happening at a given evolutionary step, now this probability is scaled to $p^{1/T}$. For larger T, this distribution flattens and changes that would not normally be accepted, may now be more probable, increasing the overall Hamiltonian value for the resulting evolutionary trajectory. Conversely, a decrease in selection temperature would restrict changes to only selected mutations with more advantageous Hamiltonian scores, driving the trajectory toward more negative values.

SEEC-Amino Acid Algorithm. The e_{ij} and h_i parameters estimated by DCA are then used as input for the SEEC-aa evolutionary simulations as previously described (1). Briefly, this approach chooses a position based on a uniformly distributed random variable and then samples from a conditional probability distribution for all possible amino acids at that site, given the amino acid identities of the rest of the sequence at that step. Importantly, the model of 202 positions excludes 25 residues upstream (not including the signal peptide) and 30 residues downstream of the Pfam domain, and so these positions remain as they are in the WT sequence. For the new sequence ($\vec{\sigma}$), an energy function, or Hamiltonian (H), can be calculated from Eq. **2**. This new sequence is now the input for the next evolutionary step, and the next position is chosen as before. Finally, a Hamiltonian trajectory can be plotted, which tracks the relative fitness effects of each step in the evolutionary simulation.

SEEC-Nucleotide Algorithm. The SEEC-nt algorithm is similar to the one presented in our previous work (1) but modified to account for mutations at the nucleotide level as well as to prevent insertion and deletion mutations (37). For this, we track the nucleotide sequence which, with the standard genetic code, translates to the amino acid sequence currently being evolved. At each evolutionary step:

1. One position i of the amino acid sequence is chosen by sampling a uniform distribution over all sites that are not gaps.
2. Once chosen, we calculate the probability distribution of the amino acid identity of the site (α), conditioned to the rest of the sequence, given by:

$$P\left(\sigma_i = \alpha | \{\sigma_j\}_{1\le j\le L, j\neq i}\right) \propto \exp\left\{h_i(\alpha) + \sum_{j\neq i} e_{ij}(\alpha, \sigma_j)\right\}. \quad [4]$$

This probability distribution is then sampled. If the resulting sampling selects a gap, it is discarded and the probability distribution at the same site is sampled once again.

3. If, however, an amino acid is sampled, we check whether there exists a codon corresponding to this amino acid that is a maximum of 1 nucleotide change from the current codon identity. When this condition is met, the amino acid sequence is updated with the new residue in place, as well as the corresponding codon chosen at random from the possible neighboring codons. When the closest codon differs by 2 or more nucleotides, the site distribution is re-sampled. A site can be sampled up to 100 times before registering as a completed step and leaving both the amino acid and nucleotide sequence unaltered.

Variant Selection and Synthesis. To obtain variants, we ran the SEEC-aa and SEEC-nt simulations for several thousand steps under the bmDCA or mfDCA model parameters; we chose a sampling of variants from each trajectory for experimental testing based on the strategies described in Fig. 1. For Phase II, in addition to the changes to the algorithm represented in SEEC-nt (see *Materials and Methods* section on "SEEC-nucleotide Algorithm"), we also optimized the selection temperature. Based on the rationale given in the text, we chose $T = 1.5$ and $T = 0.75$ for the SEEC-nt simulations using mfDCA and bmDCA parameters, respectively. All variant sequences are provided in Datasets S1 and S2. Selected TEM-1 variants were cloned into a pET28a expression vector with inducible IPTG-controlled expression by the lac operon, along with an N-terminal His-tag and a kanamycin resistance gene. These expression vectors with our variant genes of interest were then each transformed into *E. coli* (BL21(DE3)) host cells. Gene synthesis of the TEM-1 sequence, mutagenesis of variants, and plasmid cloning were performed by GenScript.

Variant Strain *E. coli* Growth Assays. Individual transformed *E. coli* variant strains were grown in triplicates of cuvettes with 2 mL of culture volume to compare the rate of cell growth based on the change in optical density (OD) at 600 nm over time. To begin the assay, cuvette setup involved adding 2 mL of Luria broth media, kanamycin (final concentration of 30 μg/mL), IPTG (final concentration of 1 mM), and either MIC of ampicillin (concentration of 5 μg/mL), a standard concentration (50 μg/mL), or a challenging concentration (100 μg/mL), depending on the set of variants being tested. Positive control sets of triplicate cuvettes were grown for each strain including WT, containing all the same culture reagents except for ampicillin. The negative control for every assay was an empty vector pET strain that contained all the same plasmid elements except for an ampicillin resistance gene. This empty vector strain was grown in triplicate containing all of the same reagents including equivalent levels of ampicillin. After culture setup, individual cuvettes were inoculated with 100 uL of overnight culture of each variant strain. Immediately after inoculation, culture OD was measured with a spectrophotometer, and readings were continued approximately every hour for a period of 24 to 72 h. Final samples were collected and Sanger sequenced.

Analysis of Growth Assay Data. Culture OD at 600 nm was measured periodically for each strain over the period of 1 to 3 d to monitor cell growth and survival in the presence of ampicillin. The growth curve data points are the mean of 3 experimental replicates and error bars represent SDs. The maximum absorbance measurement of our spectrophotometer was 2. To analyze the relative fold growth during the growth assays, rationalized growth (RG) was calculated according to the equation:

$$\mathrm{RG} = \frac{\mu_{\mathrm{Exp\ OD}}}{\mu_{+\mathrm{Ctrl\ OD}}}, \qquad [5]$$

where $\mu_{\mathrm{Exp\ OD}}$ is the mean OD of the triplicate experimental cuvettes and $\mu_{+\mathrm{Ctrl\ OD}}$ is the mean OD of the triplicate positive control cuvettes for each variant. The rationalized growth values for the variants were then normalized to the rationalized growth values of the WT strain to indicate the relative fold difference of growth (or relative fold growth, $\tilde{G}$) according to the equation:

$$\tilde{G} = \frac{\mathrm{RG_{Var}}}{\mathrm{RG_{WT}}}, \qquad [6]$$

where $\mathrm{RG_{Var}}$ is the rationalized growth for each variant and RG_{WT} is the rationalized growth of the WT strain. Relative fold difference in growth can be visualized with positive peaks illustrating enhanced growth and relative fold growth less than 1 representing sub-optimal function when compared to WT.

Sanger Sequencing. After each growth assay, experimental replicates were combined, miniprepped (Qiagen), and quantitated using absorbance at 260 nm via NanoDrop spectrophotometer. Relevant samples were then sent for Sanger sequencing performed by the Genome Center at The University of Texas at Dallas (Richardson, TX).

Data, Materials, and Software Availability. Data are in Datadryad.org (https://doi.org/10.5061/dryad.n5tb2rc1c) (73). Scripts and model details are accessible at https://github.com/morcoslab/SEEC-NT (74).

ACKNOWLEDGMENTS. This research was funded by University of Texas Dallas (F.M.), the NSF Faculty Early Career Development (CAREER) Program grant number MCB-1943442 (F.M.), and the NIH R35GM133631 (for F.M., C.M.N., and S.A.).

Author affiliations: [a]Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080; [b]Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080; and [c]Center for Systems Biology, University of Texas at Dallas, Richardson, TX 75080

Author contributions: S.A., C.M.N., and F.M. designed research; S.A., C.M.N., N.M., J.A.d.l.P., and T.H. performed research; J.A.d.l.P. and F.M. contributed new reagents/analytic tools; S.A., C.M.N., N.M., T.H., and F.M. analyzed data; and S.A., C.M.N., J.A.d.l.P., and F.M. wrote the paper.

1. J. A. de la Paz, C. M. Nartey, M. Yuvaraj, F. Morcos, Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 5873–5882 (2020).
2. F. Morcos *et al*., Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
3. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
4. J. Sheng, S. Wang, Coevolutionary transitions emerging from flexible molecular recognition and eco-evolutionary feedback. *Iscience* **24**, 102861 (2021).
5. M. S. Sohail, R. H. Louie, Z. Hong, J. P. Barton, M. R. McKay, Inferring epistasis from genetic time-series data. *Mol. Biol. Evol.* **39**, msac199 (2022).
6. A. Biswas, A. Haldane, E. Arnold, R. M. Levy, Epistasis and entrenchment of drug resistance in HIV-1 subtype B. *Elife* **8**, e50524 (2019).
7. D. Ding *et al*., Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nat. Ecol. Evol.* **6**, 590–603 (2022).
8. J. L. Harman *et al*., Evolution avoids a pathological stabilizing interaction in the immune protein S100A9. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2208029119 (2022).
9. I. Choudhuri, A. Biswas, A. Haldane, R. M. Levy, Contingency and entrenchment of drug-resistance mutations in HIV viral proteins. *J. Phys. Chem. B* **126**, 10622–10636 (2022).
10. J. D. Bryngelson, P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7524–7528 (1987).
11. J. N. Onuchic, Z. Luthey-Schulten, P. G. Wolynes, Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).
12. A. Lapedes, B. Giraud, C. Jarzynski, Using sequence alignments to predict protein structure and stability with high accuracy. arXiv [Preprint] (2002). https://arxiv.org/abs/1207.2484 (Accessed 1 February 2023).
13. O. Haq, M. Andrec, A. V. Morozov, R. M. Levy, Correlated electrostatic mutations provide a reservoir of stability in HIV protease. *PLoS Comput. Biol.* **8**, e1002675 (2012).
14. F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, P. G. Wolynes, Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12408–12413 (2014).
15. R. M. Levy, A. Haldane, W. F. Flynn, Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* **43**, 55–62 (2017).
16. T. A. Hopf *et al*., Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
17. M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, M. Weigt, Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* **33**, 268–280 (2016).
18. X. L. Jiang, R. P. Dimas, C. T. Chan, F. Morcos, Coevolutionary methods enable robust design of modular repressors by reestablishing intra-protein interactions. *Nat. Commun.* **12**, 5592 (2021).
19. A. Murugan *et al*., Roadmap on biology in time varying environments. *Phys. Biol.* **18**, 041502 (2021).
20. J. Gizzio, A. Thakur, A. Haldane, R. M. Levy, Evolutionary divergence in the conformational landscapes of tyrosine vs serine/threonine kinases. *Elife* **11**, e83368 (2022).
21. M. A. Stiffler *et al*., Protein structure from experimental evolution. *Cell Syst.* **10**, 15–24 (2020).

22. M. Fantini, S. Lisi, P. De Los Rios, A. Cattaneo, A. Pastore, Protein structural information and evolutionary landscape by in vitro evolution. *Mol. Biol. Evol.* **37**, 1179–1192 (2020).
23. M. Figliuzzi, P. Barrat-Charlaix, M. Weigt, How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35**, 1018–1027 (2018).
24. F. McGee *et al.*, The generative capacity of probabilistic protein sequence models. *Nat. Commun.* **12**, 6302 (2021).
25. A. Haldane, R. M. Levy, Mi3-GPU: MCMC-based inverse ising inference on GPUs for protein covariation analysis. *Comput. Phys. Commun.* **260**, 107312 (2021).
26. J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, M. Weigt, Efficient generative modeling of protein sequences using simple autoregressive models. *Nat. Commun.* **12**, 5800 (2021).
27. A. P. Muntoni, A. Pagnani, M. Weigt, F. Zamponi, adabmDCA: Adaptive Boltzmann machine learning for biological sequences. *BMC Bioinf.* **22**, 1–19 (2021).
28. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87**, 012707 (2013).
29. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence-and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674–15679 (2013).
30. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S. I. Lee, C. J. Langmead, Learning generative models for protein fold families. *Proteins: Struct. Funct. Bioinf.* **79**, 1061–1078 (2011).
31. W. P. Russ *et al.*, An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
32. P. Tian *et al.*, Design of a protein with improved thermal stability by an evolution-based generative model. *Angew. Chem. Int. Ed.* **61**, e202202711 (2022).
33. C. Ziegler, J. Martin, C. Sinner, F. Morcos, Latent generative landscapes as maps of functional diversity in protein sequence space. *Nat. Commun.* **14**, 2222 (2023).
34. C. M. Weisman, The origins and functions of de novo genes: Against all odds? *J. Mol. Evol.* **90**, 244–257 (2022).
35. D. M. Keeling, P. Garza, C. M. Nartey, A. R. Carvunis, The meanings of 'function' in biology and the problematic case of de novo gene emergence. *Elife* **8**, e47014 (2019).
36. L. Vigué *et al.*, Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes. *Nat. Commun.* **13**, 4030 (2022).
37. M. Bisardi, J. Rodriguez-Rivas, F. Zamponi, M. Weigt, Modeling sequence-space exploration and emergence of epistatic signals in protein evolution. *Mol. Biol. Evol.* **39**, msab321 (2022).
38. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
39. N. Tokuriki, D. S. Tawfik, Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
40. S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, D. S. Tawfik, Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).
41. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
42. T. D. Schneider, R. M. Stephens, Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
43. S. Sunyaev *et al.*, Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597 (2001).
44. R. Karchin *et al.*, LS-SNP: Large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* **21**, 2814–2820 (2005).
45. Y. Bromberg, B. Rost, SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823–3835 (2007).
46. N. C. Strynadka *et al.*, Structural and kinetic characterization of a β-lactamase-inhibitor protein. *Nature* **368**, 657–660 (1994).
47. N. C. Strynadka, S. E. Jensen, P. M. Alzari, M. N. James, A potent new mode of β-lactamase inhibition revealed by the 1.7 Å X-ray crystallographic structure of the TEM-1-blip complex. *Nat. Struct. Biol.* **3**, 290–297 (1996).
48. B. G. Fryszczyn *et al.*, Role of β-lactamase residues in a common interface for binding the structurally unrelated inhibitory proteins BLIP and BLIP-II. *Prot. Sci.* **23**, 1235–1246 (2014).
49. D. H. Bryant *et al.*, Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
50. A. Hawkins-Hooker *et al.*, Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
51. J. Dauparas *et al.*, Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
52. B. Wicky *et al.*, Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
53. A. Madani *et al.*, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1–8 (2023).
54. T. D. Townsley *et al.*, A novel approach to identifying and ranking critical non-proximal interdependencies within the overall protein structure. *Bioinfor. Adv.* **2**, vbac058 (2022).
55. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein sectors: Evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
56. M. Schmidt, K. Hamacher, hoDCA: Higher order direct-coupling analysis. *BMC Bioinf.* **19**, 1–5 (2018).
57. Y. T. Tamer *et al.*, High-order epistasis in catalytic power of dihydrofolate reductase gives rise to a rugged fitness landscape in the presence of trimethoprim selection. *Mol. Biol. Evol.* **36**, 1533–1550 (2019).
58. P. E. O'maille *et al.*, Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nat. Chem. Biol.* **4**, 617–623 (2008).
59. A. Ballal *et al.*, Sparse epistatic patterns in the evolution of terpene synthases. *Mol. Biol. Evol.* **37**, 1907–1924 (2020).
60. G. W. Rudgers, T. Palzkill, Identification of residues in β-lactamase critical for binding β-lactamase inhibitory protein. *J. Biol. Chem.* **274**, 6963–6971 (1999).
61. R. H. Louie, K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty, M. R. McKay, Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E564–E573 (2018).
62. T. Zhang *et al.*, Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS Genet.* **16**, e1009009 (2020).
63. H. L. Zeng, V. Dichio, E. Rodríguez Horta, K. Thorell, E. Aurell, Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 31519–31526 (2020).
64. J. Rodriguez-Rivas, G. Croce, M. Muscat, M. Weigt, Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2113118119 (2022).
65. J. V. Rodrigues, E. I. Shakhnovich, Adaptation to mutational inactivation of an essential gene converges to an accessible suboptimal fitness peak. *Elife* **8**, e50509 (2019).
66. R. D. Finn *et al.*, Pfam: The protein families database. *Nucl. Acids Res.* **42**, D222–D230 (2014).
67. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
68. S. R. Eddy, HMMER: 3.3.2 (2020). http://hmmer.org.
69. The UniProt Consortium, Uniprot: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
70. R. R. Cheng *et al.*, Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. *Mol. Biol. Evol.* **33**, 3054–3064 (2016).
71. Q. Zhou *et al.*, Global pairwise RNA interaction landscapes reveal core features of protein recognition. *Nat. Commun.* **9**, 2511 (2018).
72. K. Ravishankar, X. Jiang, E. M. Leddin, F. Morcos, G. A. Cisneros, Computational compensatory mutation discovery approach: Predicting a PARP1 variant rescue mutation. *Biophys. J.* **121**, 3663–3673 (2022).
73. S. Alvarez *et al.*, Data from: In vivo functional phenotypes from a computational epistatic model of evolution [Dataset]. Datadryad. https://doi.org/10.5061/dryad.n5tb2rc1c. Deposited 5 January 2024.
74. S. Alvarez *et al.*, SEEC-nt. GitHub. https://github.com/morcoslab/SEEC-NT. Deposited 1 March 2023.