ACS Synthetic Biology

pubs.acs.org/synthbio Letter

Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning

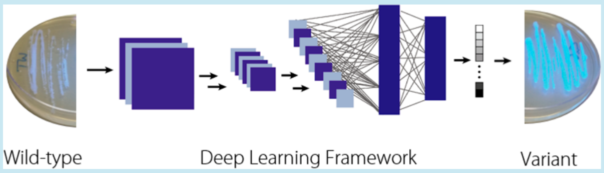
Raghav Shroff,*,[‖] Austin W. Cole,[‖] Daniel J. Diaz, Barrett R. Morrow, Isaac Donnell, Ankur Annapareddy, Jimmy Gollihar, Andrew D. Ellington, and Ross Thyer*

Cite This: https://dx.doi.org/10.1021/acssynbio.0c00345 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

Downloaded via UNIV OF NEW ENGLAND on October 16, 2020 at 22:07:48 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ABSTRACT: Despite the promise of deep learning accelerated protein engineering, examples of such improved proteins are scarce. Here we report that a 3D convolutional neural network trained to associate amino acids with neighboring chemical microenvironments can guide identification of novel gain-of-function mutations that are not predicted by energetics-based approaches. Amalgamation of these mutations improved protein function *in vivo* across three diverse proteins by at least 5-fold. Furthermore, this model provides a means to interrogate the chemical space within protein microenvironments and identify specific chemical interactions that contribute to the gain-of-function phenotypes resulting from individual mutations.



KEYWORDS: *computational protein design, neural networks, machine learning, protein engineering*

Protein engineering is a transformative approach in biotechnology and biomedicine commonly used to alter natural proteins to tolerate non-native environments,[1] modify substrate specificity,[2] and improve catalytic activity.[3] Underpinning these properties is a protein's ability to fold and adopt a stable active configuration. This property is currently engineered either from sequence[4] or energetic simulations.[5] Machine learning approaches have been reported to rationalize mutation effects; however, validations of these algorithms are presently restricted to benchmark data sets containing only thousands of annotated observations.[6−11] Efforts to functionally improve a target protein are restricted to *post hoc* analysis of existing deep mutational studies on that target protein.[8] Furthermore, these data sets are predominantly comprised of deleterious mutations that bias models toward negative predictions, rather than improvement of protein function. These limitations have thwarted the use of deep learning in *de novo* protein design and engineering.

Recently, a 3D CNN was trained to associate the spatial orientation of carbon, oxygen, nitrogen, and sulfur atoms present within a protein microenvironment with a central amino acid.[12] Given structural data and resulting local chemistry, this model was able to predict wild-type amino acids at residues where destabilizing mutations had been experimentally confirmed. We hypothesized that the converse might also be true: gain-of-function mutations could be introduced at positions where the wild-type residue is incongruent with the local spatial orientation of functional groups leading to disfavored intermolecular interactions. In this report, we detail how a deep learning algorithm trained on amino acid-structure relationships across the entirety of the observed proteome can direct protein mutagenesis and ultimately improve protein function *in vivo* by between 6- and 30-fold in three disparate proteins.

RESULTS AND DISCUSSION

Before investigating residues where expectations of the neural network and nature differed, we sought to strengthen the deep learning framework's association between amino acids and their local chemistry. First, we rebuilt the neural network architecture published by Torng and Altman with minor modifications (Figure 1a, see Supporting Methods for details), and replicated the reported classification accuracy of 41.2% (Figure 1b) in original training and testing sets (32 760 and 1601 structures, respectively).[12] Then we made several discrete changes to enrich our computational representation of protein chemistry. First, we incorporated hydrogen atoms to explicitly recapitulate local hydrogen bonding networks, which increased accuracy to 43.4%. Next, we added biophysical annotations such as partial charge and solvent accessibility to each atom, improving the accuracy of predicting wild-type to 52.4%.

The selection methodology for both protein structures and amino acid residues introduced several biases to the training data. The data set contained multiple structures of closely related proteins which biased training toward overrepresented

Received: June 29, 2020

ACS Publications © XXXX American Chemical Society

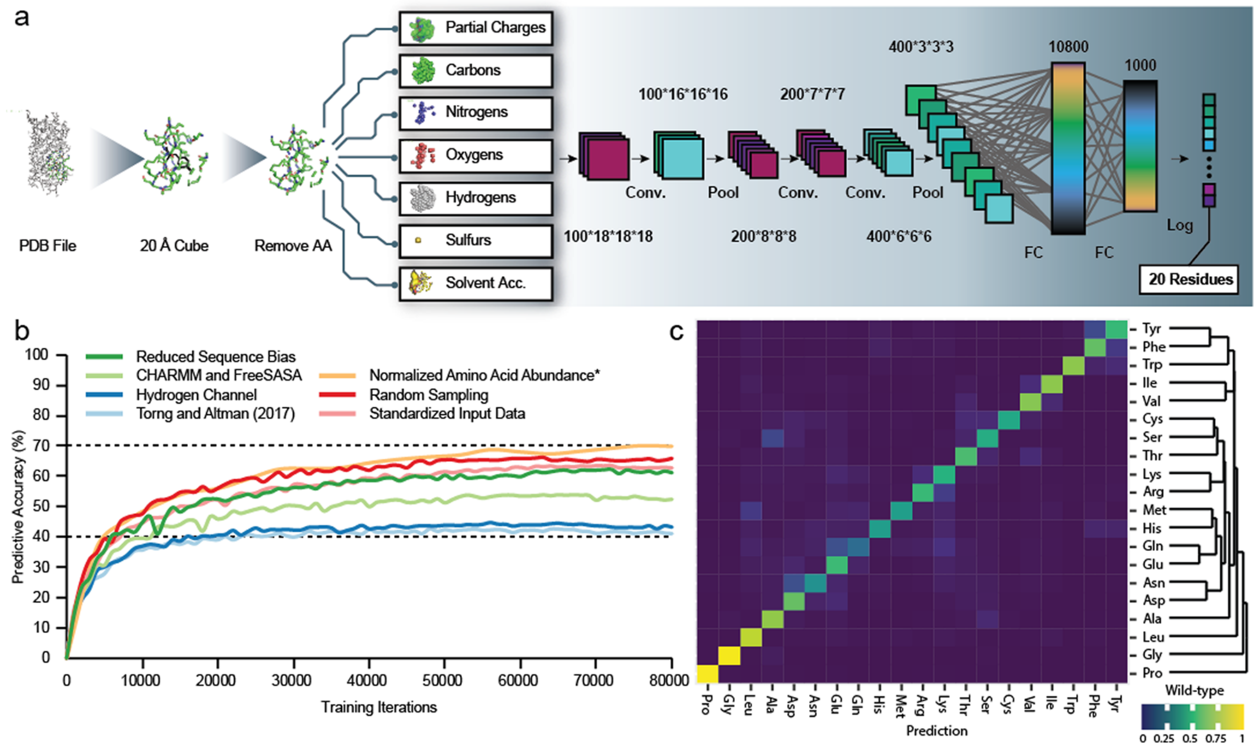


Figure 1. Design and performance of a deep learning program capable of classifying wild-type amino acids with improved accuracy. (a) Schematic of the model depicting the data pipeline and neural network architecture. An amino acid local environment is first extracted from a protein structure. After isolating the neighboring atoms and biophysical characteristics into individual channels, the data is processed through a series of convolution and pooling steps to find relevant structural patterns. Integration of these feature extraction steps occurs through fully connected layers (FC) into a final softmax layer to calculate amino acid probabilities. (b) Cumulative changes made to the neural net framework described by Torng and Altman[10] and their effect on classification accuracy. *Normalizing the amino acid abundance of the training data increased the size of the data set by roughly 4-fold. While the number of epochs decreased, the number of training iterations needed for convergence remained similar to the other versions. (c) Confusion matrix showing bias of wild-type amino acid classification. Structurally unique amino acids Gly and Pro are assigned as wild-type with very high probability.

protein structures, where the 32 760 PDB IDs map to only 11 418 UniProtKB IDs. Additionally, deposited crystallographic structures are refined by algorithms of their time which are not necessarily the current state of the art. To improve data set composition and uniformity, we gathered all PDB structures with less than 2.5 Å resolution and at most 50% sequence similarity and drew from structures in the PDB-REDO database, where existing protein structures are refined in a uniform manner.[13] These two changes in data consistency resulted in 19 436 structures for training with 300 of these structures held for out-of-sample testing and increased wild-type prediction accuracy to 63%. While the training and test structures were controlled for sequence similarity, we find some test proteins that have structural similarity to proteins in the training set; however, the high similarity results in little to no effect on predicting wild type accuracy (SI Figure S1). In the original data set, an unintended consequence of the sampling methodology was that the population of local chemistries surrounding amino acids were heavily biased toward surface residues. We removed this bias by sampling residues randomly throughout the protein sequence, and thereby increased wild-type classification accuracy to 66%.

Despite dramatic improvements in classification accuracy, we observed that cysteine and methionine were predicted by the model at a higher frequency than expected given their abundance in the proteome (SI Figure S2). The original data set included equal frequencies of each amino acid, potentially biasing the expectation of the neural network toward rare amino acids. Sampling was altered so that amino acid frequencies mirrored their natural abundance in the PDB database and normalized relative to the least abundant amino acid (Cys). This increased the size of the training set by approximately 4-fold, to 1.6 million amino acid environments, and further improved classification accuracy to nearly 70%. Following this change, classification accuracy more closely resembled expected amino acid abundance. Altogether, the series of steps generated an amino acid-structure model with unprecedented accuracy when wild-type is most congruent with the local protein chemistry.

In assessing a confusion matrix (Figure 1c), amino acids with similar properties were commonly misclassified, most evident by the disparity in accuracy between the acid and amide amino acids. While aspartic acid and glutamic acid are both classified with high accuracy, asparagine and glutamine are among the least correctly called residues at 48% and 33%, respectively. Furthermore, the misclassification of asparagine for aspartic acid and the glutamine for glutamic acid are both 23% while the converse misclassifications are both 5%. This trend in misclassification is explainable by acid−base chemistry, while simultaneously illustrating a chemical phenomenon the model has learned. Aspartic acid and glutamic acid partition their activity between the acid and their conjugate base. The conjugate bases have a delocalized negative charge and are hydrogen bond acceptors while the acids, like the amides, are polar neutral and are both hydrogen bond donors and acceptors. Thus, the net can classify acidic amino acids where the local chemistry demands a negative charge but struggles to differentiate between the acid and the amide when the local chemistry is in need of a polar-neutral, pi-system functional group that is both a hydrogen bond donor and acceptor. Additionally, proline, being the only amino acid with a secondary amine, and glycine, being the only achiral amino acid, can result in atypical protein microenvironments and our net is capable of associating these two amino acids with their unique local protein chemistry with over 96% accuracy. In comparing the performance of our algorithm spatially within a protein, we find that surface residues are misclassified (predicting a residue other than wild-type) much more frequently than core residues (SI Figure S3), although residues with extreme surface exposure represent a very small fraction of the overall data set (SI Figure S3). This observation raises an interesting point of contrast between our deep learning model and traditional energetics-based tools which tend to perform better at identifying mismatched core residues. Altogether, the series of modifications to the model linked amino acids to their congruent local chemistry with unprecedented accuracy.

We next investigated the performance of the model using empirical data from deep mutational scanning (DMS) experiments for the proteins TEM-1 β-lactamase, immunoglobulin binding domain of protein G (gb1), Aminoglycoside-3′-Phosphotransferase-IIa, ubiquitin, and Hsp90.[14] In this aggregate data set, the effects of all possible single substitutions were quantified with a ceiling for activity set at wild-type function; *i.e.,* no beneficial mutations were observable. We identified 292 positions where any substitution incurred a measurable fitness cost and benchmarked classification accuracy on this subset, the presumption being that the model's classification accuracy on amino acids that are demonstrably best suited for a defined environment should exceed its overall classification accuracy. Consistent with this, the final version of the model achieves a recall of 87.0%, which is 25.4% higher than the starting model (SI Figure S4), and 17% higher than its baseline out-of-sample accuracy. Similarly, precision recall curves for this task also confirm improvement over the initial 4-channel model (SI Figure S5). Taken together, these data confirm our model classifies wild-type amino acids with unprecedented accuracy compared to previously reported structure-based deep learning approaches.[12,15−17]

Having shown that probability profiles capture wild-type residues that are optimal fits with their environment, we attempted to identify gain-of-function mutations at wild-type residues that were discordant with their environment's probability profile. These mismatched wild-type residues might be substituted to improve fit within the panorama of natural protein structures and similarly improve folding and function of a protein. To investigate this hypothesis we selected three model proteins amenable to quantitative high-throughput screening: BFP (PDB: 3M24), phosphomannose isomerase (PDB: 1PMI), and TEM-1 β-lactamase (PDB: 1BTL). Neither BFP or PMI were present in the training data set (a circularly permuted TEM-1 β-lactamase was present) and homologues with similar sequence were rare for all three (SI Figure S6).

We initially tested this hypothesis empirically in an engineered blue fluorescent protein secBFP2.1[18] (SI Table S1) by building saturating libraries at residues assigned either the lowest (disfavored) or highest (favored) wild-type probabilities by our model. We also mutagenized ten residues selected at random to serve as a control. Six of nine disfavored residues, one of ten random residues, and zero of ten favored residues could be substituted to improve fluorescence of secBFP2.1 (p = 0.01 by a Fisher's exact test for disfavored *versus* random subsets; Figure 2a and SI Figures S7−S9). We amalgamated the beneficial substitutions into a single variant, designated BFP-Bluebonnet (BB), which improved florescence in *E. coli* by more than 6-fold (Figure 2b,c). Interestingly, two of the mutations (T127P and N173T) either reverted or changed mutations made in mTagBFP and TagRFP, respectively. Furthermore, purified BFP-Bluebonnet remained monomeric (N173T changes one of the mutations made at the dimer interface in TagRFP) (SI Figure S10) and exhibited minor improvements to thermal tolerance and slower denaturation in guanidinium relative to secBFP2.1 and mTagBFP2 (SI Figure S11). Blue fluorescent proteins are used less frequently than their counterparts across the visual spectrum for localization studies, *e.g.,* GFP and RFP, in large part due to their maturation kinetics and solubility *in vivo*. Bluebonnet addresses these limitations and offers improved *in vivo* performance while retaining the advantageous properties of its ancestor.

To verify that our model was generalizable to catalytically active proteins, we built site-saturation libraries at the ten most disfavored residues (excluding active site residues) in each of two structurally and functionally unrelated enzymes, TEM-1 β-lactamase and *Candida albicans* phosphomannose isomerase (*Ca*PMI). TEM-1 β-lactamase is a model protein for deep mutational scanning and protein evolution, thereby providing a rich benchmark of cross-validated mutational annotations for the 3D CNN's predictions. *Ca*PMI is poorly soluble in *E. coli* and lacks an easily screened readout for directed evolution, and thereby serves as an exemplar of how the model can be applied in conjunction with a generalizable high-throughput protein engineering workflow (see Methods). Seven of the ten residues in TEM-1 β-lactamase could be substituted to rescue antibiotic resistance and six of the ten residues in *Ca*PMI yielded mutants with improved fluorescence, respectively, phenotypes in these assays which are associated with folding and stability (Figure 2d,e). While beneficial mutations have previously been identified at most of the residues identified by the model in TEM-1 β-lactamase (N52, F60, Q88, Q99, T114, and M182),[19] to the best of our knowledge E197P is first reported here. In *Ca*PMI, aggregating the individual mutations improved folding in *E. coli* by 5-fold without abolishing catalytic activity (SI Figure S12). Across all three test proteins, only four residues identified by our final model as candidates for mutagenesis were also identified by the 4-Channel model, and the assigned probabilities for the wild-type amino acid differed significantly (SI Figure S19 and SI Table S2).

In many situations, protein properties are not amenable to measurement in large scale screens. We explored whether the 3D CNN probability profile could guide specific mutagenesis by substituting wild-type residues poorly matched with their local chemistry with the residue suggested to best fit with local chemistry. Using this approach, we recovered several mutations which improved the functional readout for each protein (3 of 17 for secBFP2.1, 5 of 17 for TEM-1 and 9 of 22 for *Ca*PMI). As before, mutational effects were independent and in combination improved each of the three distinct

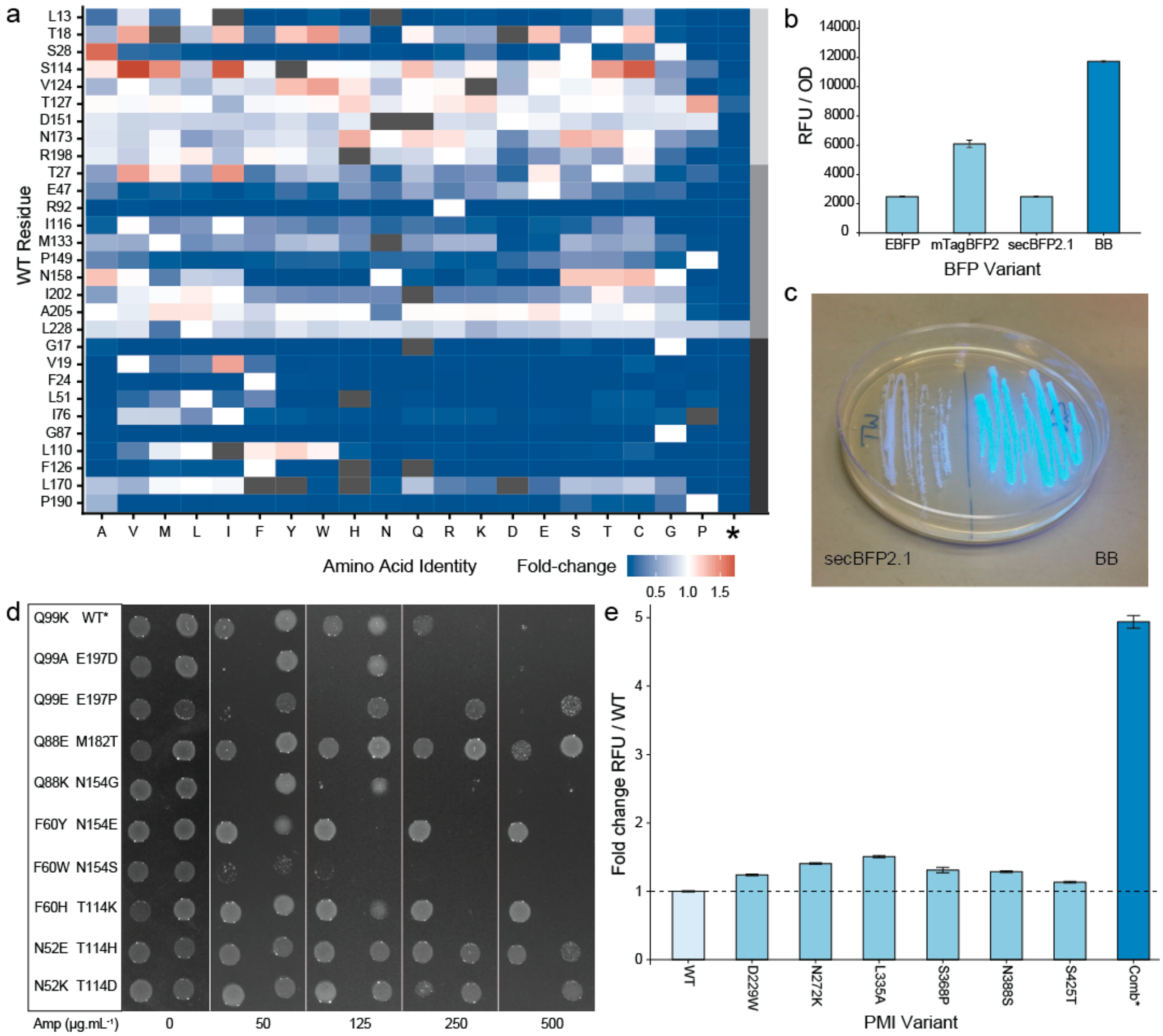


Figure 2. Empirical validation of the model as a tool for protein engineering using three model proteins. (a) Heatmap showing fold-change over wild-type for site-saturation mutants of secBFP2.1. The light gray, dark gray, and black bars on the right indicate the series of disfavored, random, and favored residues, respectively. Note, L228 is only five residues away from the C-terminus. Substitutions at this position, including stop codons, have minimal impact on fluorescence. (b) An improved variant of secBFP2.1 containing mutations T18W, S28A, S114 V, V124T, T127P, D151G, N173T, and R198L was ~6-fold more fluorescent *in vivo* than the parental protein. This variant was named BFP-Bluebonnet (BB). (c) Plate assay showing increased *in vivo* fluorescence of BFP-Bluebonnet compared to secBFP2.1. (d) Gain-of-function mutations were identified in TEM-1 β-lactamase at N52, F60, Q88, Q99, T114, M182, and E197. WT* contains the destabilizing mutation L250Q. Residue Q88 was ranked as the 11th least favorable in TEM-1 β-lactamase and was included in place of D214 which lies in the active site. (e) Beneficial mutations were identified in *Ca*PMI at residues D229, N272, L335, S368, N388, and S425. A combined mutant containing D229W, N272 K, L335A, N388S, and S425T was 5-fold more fluorescent than wild-type using the split-GFP assay. While S368P was beneficial by itself, it was deleterious in combination.

phenotypes by at least 5-fold (see Supporting Results and SI Figures S13−S15). To examine any potential influence of homologous proteins in the training data set we aligned each of the three model proteins to the five closest sequence homologues (secBFP2.1 and TEM-1) or in the case of *Ca*PMI which had no close homologues by sequence, to the two closest structural homologues (an archaeal and a bacterial phosphomannose isomerase) (SI Figure S16−18). While some gain-of-function mutations could be found in close homologues, the presence (or absence) of closely related proteins in the training set did not obviously influence the residues identified by the model.

Although we focused on the protein engineering applications of our model, it also has considerable potential as a tool to parse chemical biology. We performed input masking on TEM-1 β-lactamase to gain a deeper understanding of the contributing factors influencing mutational prediction. We systematically deleted each atom in a given microenvironment and cataloged that atom's effect on the probability profile. To test whether our model is learning known chemical phenomena, we probed the microenvironment surrounding the conserved disulfide bond formed between Cys 77 and Cys 123. For both residues (which the model classified as wild-type), the atom which most decreased the wild-type probability mass was the sulfur atom on the neighboring cysteine (Figure 3, left), which conforms with the expected biochemistry.

https://dx.doi.org/10.1021/acssynbio.0c00345

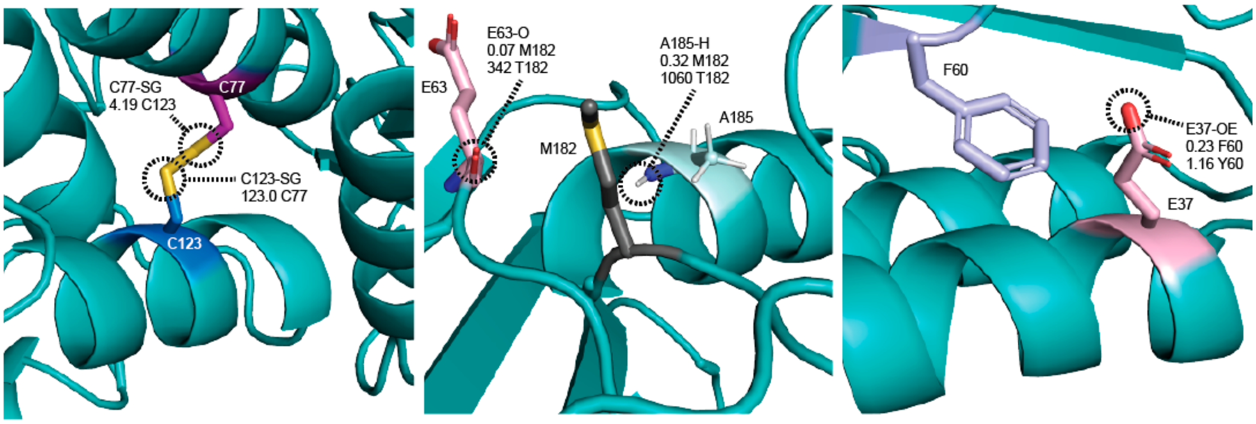


Figure 3. Input masking through systematic deletion of atoms within a microenvironment can be used to identify interacting atoms and propose biochemical mechanisms for stabilizing mutations. TEM-1 β-lactamase was used as a model protein to investigate how proximal atoms influence the prediction of wild-type residues. Each atom is scored by the fold change in the prediction probability of the wild-type amino acid and the amino acid with the highest assigned probability. Scores are calculated by dividing the probabilities assigned in the native microenvironment by those from microenvironments in which a single surrounding atom is removed. Atoms that are more influential decrease the prediction probability when deleted, which is reflected as an increase in the fold change. Values <1 indicate the atom decreases the probability of the associated amino acid. The neighboring atoms with the largest fold changes are annotated for three different microenvironments: (left) the conserved disulfide bond between C77−C123 and two stabilizing mutations, (middle) M182T, and (right) F60Y.

We next sought to explain the model's assignment of a very low wild-type probability to M182 (0.009) and a high probability (0.534) of threonine at this location. The M182T mutation in TEM-1 β-lactamase is a global suppressor mutation that has been identified in many clinical isolates. Despite its identification decades ago, a mechanistic explanation for stabilization remains under debate. One model proposes that the threonine hydroxyl forms an N-cap H bond with the backbone amide nitrogen of Ala 185 as determined through crystallographic analysis,[20] while a competing explanation determined through molecular modeling suggests a stabilizing hydrogen bond with Glu 63 and/or Glu 64.[21−23] Our method identified two atoms with disproportionately large effects on the probabilities for methionine and threonine, the backbone oxygen of Glu 63 and the amide hydrogen on Ala 185. Removal of either atom decreased the probability of observing a threonine by over 300 fold while simultaneously increasing the chance of observing a methionine (Figure 3, middle). We also used input masking to interrogate another mutation flagged by the model, F60Y. This has been identified as a stabilizing mutation, but a biochemical mechanism has yet to be proposed. A similar analysis of the amino acid microenvironment revealed that the neighboring ε-oxygen atom on Glu 37 decreases the wild-type probability by more than 4-fold while slightly increasing the probability of tyrosine, potentially through the formation of a hydrogen bond with the hydroxyl group (Figure 3, right). These input masking experiments suggest that our deep learning model can help to elucidate and quantify competing chemical phenomena responsible for residue stabilization in addition to identifying candidate residues for mutagenesis.

Two well-documented computational approaches to guide protein stabilization are Rosetta pmut_scan and FoldX PositionScan, both of which rely on energetics simulations to identify new energy minima. If our model learned inferences accessible by energetics calculations in either of these programs, we would expect significant overlap between the residues it identified and those predicted by these programs to occupy local energy maxima. Only three of 30 positions identified by the model were also identified by either Rosetta or FoldX, which also largely identified separate residues, and only four were shared between our final model and the original 4-channel model (SI Figure S19). A complete comparison of the probabilities assigned by our model and the starting point at the ten least favored wild-type residues in the three model proteins is shown in SI Table S2. Furthermore, in TEM-1 β-lactamase, each of these three methods uniquely identified beneficial mutations reported elsewhere in the literature[19,24,25] (SI Figure S19). Therefore, our model can identify novel residues for mutagenesis not captured by other commonly used programs.

Here we report a modified 3D CNN architecture that associates local chemistries with the wild-type amino acids at state-of-the-art accuracy. Where native residues deviate from their structural and chemical expectation or are otherwise discordant with their local environment, we demonstrate that these positions are excellent targets for site-saturation mutagenesis and yield gain-of-function mutants at frequencies that exceed random selection. Combining the individual mutations identified in each of three model proteins improved variant phenotypes several fold relative to their ancestor. It is important to note that given the model's learned ruleset is unknown, the screening assays used to empirically validate the model were intentionally chosen to be both agnostic to different types of beneficial mutations, and more importantly, capture as wide a selection of beneficial mutations as possible. As a result, the phenotypes we chose to measure serve as broad proxies for gain-of-function rather than specific benchmarks of any one biophysical property. Ongoing research aims to unravel the biological basis for the model's predictions. This work is the first demonstration of using deep learning to empirically improve protein function and opens new avenues for protein engineering. This model complements existing protein design tools by identifying sets of mutations that do not overlap with those derived from energetics simulations. We anticipate that our algorithm will find broad utility in the protein engineering community and have made it available *via* a web-tool which requires no computational experience.

■ METHODS

Computational Methods. *Data Set.* To reduce any bias resulting from the differential abundance of protein families in the PDB, we sought to build a data set of protein structures with balanced phylogeny. To achieve this, we took all structures in the PDB database and clustered to 50% similarity to avoid oversampling toward certain protein classes. We further reduced the variability in the data set by cross-referencing the structures to the PDB-redo database,[13] which uses a consistent algorithm to refine, rebuild, and validate structures from raw crystallographic data. Within each clustered set of sequences, we identified the structure with the lowest resolution. If no structure existed below a resolution of 2.5 Å the entire cluster was discarded. Chains within the structure files were selected if the sequence similarity was less than 90% identical with any other chain within the same file to eliminate oligomerization artifacts arising from crystallization. This process yielded 19 436 structures, of which 300 were set aside for out of sample testing and the remainder used to generate the training set.

Box Extraction. In addition to atomic annotations, our model adds additional channels for the partial charges and solvent accessibility associated with each atom. While all structure files label oxygen, carbon, nitrogen, and sulfur, hydrogens may be missing depending on the resolution of the structure. Using the program PDB: 2PQR (v2.2.1),[26] hydrogens were placed into the structure and optimized while partial charges were assigned with the CHARMM force field. Solvent accessibility was calculated with the program FreeSASA (v2.0.2).[27] To avoid oversampling residues from larger proteins, we limited the number of sampled environments from an individual protein to either half of the length of the protein or 100 amino acids, whichever number was less. Atomic environments consisting of a 20 Å cube centered around a single residue were generated as described in Torng and Altman.[12]

Neural Network Training. The convolutional neural network was built using theano (v1.0.3) and consists of six layers, all with ReLu (rectified linear unit) activations. The first two convolutions were performed with a filter size of 3 × 3 × 3 with no padding and increased the depth to 200 channels. We then performed a max pooling step, followed by two additional convolutions with a filter size of 2 × 2 × 2 and increasing the depth to 400. Max pooling was used again before flattening and feeding into two successive fully connected layers with dropout rates of 0.5 and 0.2, respectively. Softmax activation was applied to the logits to obtain probability scores for each of the 20 amino acids.

Neural network training was performed on TACC's Maverick cluster with a NVIDIA Tesla K40 GPU. 1.6 million amino acid environments were generated with the abundance of individual amino acids mirroring the natural frequency observed in the PDB. As the data set was too large to load entirely into memory, we split the data into 20 000 samples and randomly shuffled the order after loading. Batch sizes of 20 samples were used and the loss was calculated through RMSprop. Training was performed with an adaptive learning rate and lowered by 10% if validation accuracy did not decrease within 8000 training iterations. Four epochs were run, at which point overfitting was observed. Test and validation accuracy were measured in 6000 amino acid environments with equal representation of each residue.

Confusion Matrix and Regression Bias. To calculate the frequency at which wild-type residues were correctly predicted, 20 000 amino acid environments were generated from out of sample PDBs (*i.e.*, structures not seen during training) with an amino acid distribution mirroring natural frequencies. Regressions highlighting amino acid bias were created by plotting the sum of the predicted probability values against the frequency in the test set. The confusion matrix was generated by plotting the single amino acid assigned the highest probability at each microenvironment sampled compared to the wild-type amino acid.

Rosetta/FoldX Calculations. The pmut_scan program within the Rosetta software suite (v3.9)[28] was used to calculate the computational effect of mutations with a large ΔΔG cutoff value to output both stabilizing and destabilizing mutations, using the following command:

```
pmut_scan_parallel.cxx11mpi.linuxiccrelease -database <database> -s <pdb> -ex1 -ex2 -extrachi_cutoff 1 -use_input_sc -ignore_unrecognized_res -no_his_his_pairE -multi_cool_annealer 10 -DDG_cutoff 999
```

To perform the analogous operation in FoldX (ver. 4), the PositionScan module was used, with the following command:

```
foldx --command=PositionScan --pdb=Optimized_1btl.pdb --positions=<all residues>
```

In either program, the least favorable sites were found by summing values less than zero (the sign change of a stabilizing mutation) and identifying the ten sites with the most negative value.

Deep Mutational Scanning (DMS) Analysis. Normalized fitness values were derived from Gray *et al.* (2018)[14] with a threshold of 1.02 to determine if a variant greater than wild-type exists. Each computational method was assessed using deep mutational scanning data sets paired with the corresponding structures: TEM-1 β-lactamase, PDB: 1BTL; protein G, PDB: 2QMT; aminoglycoside-3′-phosphotransferase-IIa, PDB: 1ND4; ubiquitin, PDB: 4XOF; and Hsp90, PDB: 2BRC. While 12 DMS experiments are amalgamated in Gray *et al.* (2018), each data set had to meet four criteria to be included in our benchmarking data set; (1) data sets need to cover all 20 amino acids, (2) have a high-resolution crystal structure available for the target protein, (3) measure the protein's native function and have clear thresholds for variants with different activity, (4) provide mutagenesis and fitness data in a format which allows automated extraction. Only DMS experiments for the five proteins above contained data which met these criteria. For the aph(3′)-IIa DMS experiment, we only used the data set corresponding to activity against the natural substrate. Within this subset, a positive result was defined if no other variant empirically exhibited better fitness than wild-type and, for our model, the wild-type amino acid was assigned the largest probability, or, for Rosetta and FoldX calculations, the minimum ΔΔG value (*i.e.*, the most stabilizing value) was greater than zero.

Molecular Methods. *Molecular Biology.* Experiments described in this manuscript were performed using standard molecular biology techniques. Unless otherwise indicated, all plasmids, single point mutations in reporter genes and site-saturation mutagenesis libraries were constructed using Gibson assembly. For site-saturation mutagenesis libraries, 2 μL of the reaction mixture was transformed into 50 μL of chemically competent *E. coli* cells. Transformations were required to exceed 10-fold library coverage (>320 single colonies).

BFP Fluorescence Assay. SecBFP2.1 was cloned into a kanamycin resistant derivative of plasmid pQE flanked by a T7 promoter and terminator. Site-saturation libraries were transformed into *E. coli* strain BL21 DE3 and a series of 10-fold dilutions (spanning 2 orders of magnitude) were plated on solid media to ensure sufficient discrete single colonies. 96-well deep-well plates were inoculated with 92 individual library transformants and four wild-type controls. Two plates were assayed for each library. Cells were cultured ON at 37 °C in plate shakers at 850 rpm. Twenty μL of the ON cultures were diluted into 880 μL LB and incubated for 2 h. Cells were induced by the addition of 100 μL media containing 0.5 mM IPTG, resulting in a final concentration of 50 μM. After a 4 h induction, cells were harvested by centrifugation and resuspended in 1 mL PBS. Fluorescence was measured on a Tecan M200 Pro using 400 nm for excitation and 460 nm for emission. A maximum of 12 individuals at each library site exhibiting fluorescence/OD600 values greater than wild-type were sequenced. Candidate mutations were recloned into the pQE plasmid and rephenotyped. Rephenotyping was performed in biological and technical triplicate.

Protein Purification. To purify secBFP2.1 mutants, a 6xHis tag was appended to the C-terminus *via* a Gly-Ser-Gly linker. BL21 DE3 cells were cultured in Superior Broth to mid log phase (~OD600 0.6) and induced with 1 mM IPTG for 16 h at 18 °C. Following induction, cells were harvested by centrifugation and lysed by sonication in 50 mM sodium phosphate, 300 mM NaCl, 20 mM Imidazole pH 7.4 buffer containing protease inhibitor (Pierce Protease Inhibitor) and Benzonase Nuclease (EMD Millipore). Cell lysate was clarified by centrifugation (40 000*g*) and BFP variants purified using HisPur Ni-NTA Resin. Purified protein was dialyzed into 50 mM sodium phosphate pH 7.4 buffer and analyzed by SDS PAGE to assess purity.

Thermal Melt Assay. Purified blue fluorescent proteins were diluted to 0.01 mg·mL^{-1} in PBS pH 7.4 and 100 μL aliquots were heat treated for 10 min in PCR strips on a thermal gradient using a thermal cycler. Fluorescence of thermally challenged variants and controls incubated at room temperature was assayed using excitation and emission wavelengths of 402 and 457 nm, respectively. Fluorescence readings were normalized to the mean of solutions incubated at room temperature, *e.g.*, a measurement of 0.8 indicates that a heat treated protein retained 80% of its untreated fluorescence.

Guanidinium Denaturation Assay. Purified blue fluorescent proteins were diluted to 0.01 mg·mL^{-1} in 6 M guanidinium hydrochloride. 100 μL aliquots in technical triplicate were added to wells of a 96-well clear-bottom black-walled plate and incubated at 25 °C for 23 h. These purified fluorescent proteins were assayed at 30 min intervals using excitation and emission wavelengths of 402 and 457 nm, respectively. Plates were agitated preceding each measurement. Fluorescence values measured at time zero were used to normalize fluorescence through the remainder of the assay; *e.g.*, a measurement of 0.8 indicates that the protein retained 80% of its initial fluorescence.

TEM-1 Assay. The blaTEM-1 gene encoding TEM-1 β-lactamase, including the native promoter, was amplified from pETDuet-1 and cloned into pCDFDuet-1 immediately upstream of the second T7 terminator, replacing both T7 promoters and both polylinkers. The L250Q mutation was introduced into TEM-1 to destabilize the protein and enable easy identification of compensatory stabilizing mutations.[29] Site-saturation libraries were transformed into *E. coli* strain DH10B and recovered ON in liquid medium supplemented with spectinomycin. ON cultures were diluted and plated on a range of different carbenicillin concentrations (0, 50, 125, 250, and 500 μg·mL^{-1}). For each library, 12 single colonies from the plate containing the highest concentration of carbenicillin were isolated and the blaTEM-1 gene sequenced. Beta-lactamase variants identified by library screening were recloned into pCDFDuet-1 and rephenotyped. Rephenotyping was performed by diluting overnight cultures, in biological triplicate, 100-fold and spotting 5 μL onto solid media containing a gradient of carbenicillin concentrations.

PMI Stability Assay. Improved variants of CaPMI were identified using the split GFP reporter system described by Cabantous *et al.* (2008) with minor modifications.[30] Briefly, a fusion protein consisting of residues 173–238 of folding reporter GFP, a (GGGS)2 linker, residues 2–440 of CaPMI, a (GGGS)2 linker and residues 2–172 of superfolder GFP was assembled in a derivative of pACYCDuet-1 (SI Table S1). Site-saturation libraries were transformed into *E. coli* strain BL21 DE3 and a series of dilutions plated on solid media supplemented with 0.25 mM IPTG. Following ON incubation at 37 °C, plates were further incubated at 4 °C for 8 h at which point highly fluorescent colonies were manually selected. PMI variants were subcloned and the fusion protein ORF fully sequenced prior to rephenotyping to ensure that increased fluorescence was not the result of mutations in the GFP fragments, linker regions, or plasmid backbone. Transformants were screened as described for BFP in 96-well deep-well plates in biological and technical triplicate. Fluorescence was measured on a Tecan M200 Pro using 475 nm for excitation and 535 nm for emission.

PMI Functional Assay. The manA gene encoding phosphomannose isomerase was disrupted in *E. coli* strain BL21 DE3 using lambda red recombineering to introduce a kanamycin resistance marker. Successful deletions were confirmed by colony PCR of KanR colonies using primers which flanked the manA locus. The wild-type *C. albicans* manA gene or variants containing combinations of stabilizing point mutations were cloned into a derivative of pACYCDuet-1 using Gibson assembly. BL21 DE3 ΔmanA::kan cells were transformed with PMI expression plasmids and plated on LB agar with appropriate antibiotics. Single transformants, in biological triplicate, were transferred to liquid M9 minimal medium with 0.4% glucose and cultured ON. Cells were washed in a 1:1 volume of M9 medium without any carbon source and 2 μL streaked on M9 minimal medium plates supplemented with 0.4% mannose and 0.25 mM IPTG. Wild-type BL21 DE3 cells and BL21 DE3 ΔmanA::kan cells containing an empty expression plasmid were used as positive and negative controls, respectively. Plates were incubated at 37 °C for 24 h.

NGS. Purified plasmids encoding BFP variants were quantified using the QuantIT dsDNA Assay Kit (Thermo Scientific) and 50 ng of each plasmid was pooled by well position, resulting in 192 samples each with a different variant for each tested position. Samples were prepped for sequencing by amplifying two discrete ~350 bp regions of the BFP gene with primers containing Illumina adapters and dual indexes. Sequencing was performed using Illumina MiSeq with paired end 2 × 300 bp reads.

Variant Calling. Following sequencing, paired end reads were joined together using fastq-join (v1.3.1). Raw sequences were aligned to the original gene sequence with bwa (v0.7.17). Read counts across a single position were normalized to the observed fraction of each codon. Variants that contained at least 100 read counts and exceeding 1% of wild-type counts were labeled as such, while variants that failed were left unlabeled. Outliers were identified through the OPTICS algorithm in the scikit-learn package (v0.21.3). Each sample was normalized by dividing the fluorescence by the OD600 and as well as to the average wild-type value per plate. Significance was determined using a two sided Fisher's exact test where success of a position was defined as the mean fluorescence of the most fluorescent variant being at least three standard deviations higher than the wild-type mean.

Statistical Methods and Data Presentation. All data in the manuscript are displayed as mean ± s.e.m. unless specifically indicated. Bar graphs, regressions, confusion matrix, NGS variant graphs were plotted in R 3.4.1 using the package ggplot2 (v2.2.1).

This tool is freely available for academic use at www.mutcompute.com.

The data sets and plasmids generated in this study are available upon request from the corresponding author.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.0c00345.

Supporting Results, Figures S1−S21, Table S1; Sequence and structure similarity in data set, correcting amino acid abundance in training set, performance at surface and core residues, classification accuracy in published benchmark data sets, comparison of sequence and structure of experimental proteins with training data set, fluorescence data at site saturated positions, gel confirmation of monomeric BFP, thermal and guanidinium challenge of improved BFP variant, retention of catalytic activity in improved *Ca*PMI mutants, experimental assays of neural net guided mutations, multiple sequence alignment of experimental proteins to closet homologue in training data set, comparison to other computational protein design tools (PDF)

Table S2 (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Raghav Shroff − *Center for Systems and Synthetic Biology, The Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, United States;* orcid.org/0000-0002-7331-2799; Email: rshroff@utexas.edu

Ross Thyer − *Center for Systems and Synthetic Biology, The Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, United States;* orcid.org/0000-0002-0356-5790; Email: ross.thyer@utexas.edu

Authors

Austin W. Cole − *Center for Systems and Synthetic Biology, The Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, United States;* orcid.org/0000-0002-2828-811X

Daniel J. Diaz − *The Department of Chemistry, The University of Texas at Austin, Austin, Texas 78712, United States*

Barrett R. Morrow − *Center for Systems and Synthetic Biology, The Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, United States*

Isaac Donnell − *Center for Systems and Synthetic Biology, The Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, United States*

Ankur Annapareddy − *US Army Research Laboratories − South, Austin, Texas 78712, United States*

Jimmy Gollihar − *US Army Research Laboratories − South, Austin, Texas 78712, United States*

Andrew D. Ellington − *Center for Systems and Synthetic Biology, The Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, United States;* orcid.org/0000-0001-6246-5338

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.0c00345

Author Contributions

[‖]RS and AC contributed equally.

Author Contributions

RS, AC, JG, AE, and RT designed the experiments. RS and DD wrote the code for the deep learning algorithm. RS, AC, BM, ID, AA, and RT performed the experimental work. RS, AC, AE, and RT wrote the manuscript.

Notes

The authors declare the following competing financial interest(s): RS, AC, AE and RT are named inventors on an IP filing relating to methods and compositions described in this manuscript. RS, AC, and DD hold an equity stake in Aperiam Bio, a company which licenses methods described in this manuscript from the University of Texas at Austin.

ACKNOWLEDGMENTS

Funding from the Welch Foundation (F-1654), the ARO (SP0036191-PROJ0009952 *via* MURI), and the NSF (1541244) is acknowledged.

REFERENCES

(1) Jensen, P. F., Kadziola, A., Comamala, G., Segura, D. R., Anderson, L., Poulsen, J. N., Rasmussen, K. K., Agarwal, S., Sainathan, R. K., Monrad, R. N., Svendsen, A., Nielsen, J. E., Lo Leggio, L., and Rand, K. D. (2019) Structure and Dynamics of a Promiscuous Xanthan Lyase from Paenibacillus nanensis and the Design of Variants with Increased Stability and Activity. *Cell Chem. Bio.l 26*, 191−202.

(2) Windle, C. L., Simmons, K. J., Ault, J. R., Trinh, C. H., Nelson, A., Pearson, A. R., and Berry, A. (2017) Extending enzyme molecular recognition with an expanded amino acid alphabet. *Proc. Natl. Acad. Sci. U. S. A. 114*, 2610−2615.

(3) Studer, S., Hansen, D. A., Pianowski, Z. L., Mittl, P. R. E., Debon, A., Guffy, S. L., Der, B. S., Kuhlman, B., and Hilvert, D. (2018) Evolution of a highly active and enantiospecific metalloenzyme from short peptides. *Science 362*, 1285−1288.

(4) Trudeau, D. L., Kaltenbach, M., and Tawfik, D. S. (2016) On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol. Biol. Evol. 33*, 2633−2641.

(5) Potapov, V., Cohen, M., and Schreiber, G. (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng., Des. Sel. 22*, 553−560.

(6) Jia, L., Yarlagadda, R., and Reed, C. C. (2015) Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. *PLoS One 10*, No. e0138022.

(7) Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J., and Arnold, F. H. (2019) Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A. 116*, 8852−8858.

(8) Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods 16*, 1315.

(9) Riesselman, A. J., Ingraham, J. B., and Marks, D. S. (2018) Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods 15*, 816−822.

(10) Saito, Y., Oikawa, M., Nakazawa, H., Niide, T., Kameda, T., Tsuda, K., and Umetsu, M. (2018) Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. *ACS Synth. Biol. 7*, 2014−2022.

(11) Yoshida, M., Hinkley, T., Tsuda, S., Abul-Haija, Y. M., McBurney, R. T., Kulikov, V., Mathieson, J. S., Galiñanes Reyes, S., Castro, M. D., and Cronin, L. (2018) Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides. *Chem. 4*, 533−543.

(12) Torng, W., and Altman, R. B. (2017) 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinf. 18*, 302.

(13) Joosten, R. P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A. C., Blanchet, C., Bongcam-Rudloff, E., Combet, C., Da Costa, A. L., Deleage, G., Diarena, M., Fabbretti, R., Fettahi, G., Flegel, V., Gisel, A., Kasam, V., Kervinen, T., Korpelainen, E., Mattila, K., Pagni, M., Reichstadt, M., Breton, V., Tickle, I. J., and Vriend, G. (2009) PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J. Appl. Crystallogr. 42*, 376−384.

(14) Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., and Fowler, D. M. (2018) Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst. 6*, 116−124.

(15) Wang, J., Cao, H., Zhang, J. Z. H., and Qi, Y. (2018) Computational Protein Design with Deep Learning Neural Networks. *Sci. Rep. 8*, 6349.

(16) Li, Z., Yang, Y., Faraggi, E., Zhan, J., and Zhou, Y. (2014) Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins: Struct., Funct., Genet. 82*, 2565−2573.

(17) O'Connell, J., Li, Z., Hanson, J., Heffernan, R., Lyons, J., Paliwal, K., Dehzangi, A., Yang, Y., and Zhou, Y. (2018) SPIN2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Struct., Funct., Genet. 86*, 629−633.

(18) Costantini, L. M., Subach, O. M., Jaureguiberry-Bravo, M., Verkhusha, V. V., and Snapp, E. L. (2013) Cysteineless non-glycosylated monomeric blue fluorescent protein, secBFP2, for studies in the eukaryotic secretory pathway. *Biochem. Biophys. Res. Commun. 430*, 1114−1119.

(19) Salverda, M. L. M., De Visser, J. A. G. M., and Barlow, M. (2010) Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiol. Rev. 34*, 1015−1036.

(20) Kather, I., Jakob, R. P., Dobbek, H., and Schmid, F. X. (2008) Increased folding stability of TEM-1 beta-lactamase by in vitro selection. *J. Mol. Biol. 383*, 238−251.

(21) Zimmerman, M. I., Hart, K. M., Sibbald, C. A., Frederick, T. E., Jimah, J. R., Knoverek, C. R., Tolia, N. H., and Bowman, G. R. (2017) Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. *ACS Cent. Sci. 3*, 1311−1321.

(22) Farzaneh, S., Chaibi, E. B., Peduzzi, J., Barthelemy, M., Labia, R., Blazquez, J., and Baquero, F. (1996) Implication of Ile-69 and Thr-182 residues in kinetic characteristics of IRT-3 (TEM-32) beta-lactamase. *Antimicrob. Agents Chemother. 40*, 2434−2436.

(23) Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P., and Stevens, R. C. (2001) Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nat. Struct. Biol. 8*, 238−242.

(24) Bratulic, S., Gerber, F., and Wagner, A. (2015) Mistranslation drives the evolution of robustness in TEM-1 beta-lactamase. *Proc. Natl. Acad. Sci. U. S. A. 112*, 12758−12763.

(25) Guthrie, V. B., Allen, J., Camps, M., and Karchin, R. (2011) Network models of TEM beta-lactamase mutations coevolving under antibiotic selection show modular structure and anticipate evolutionary trajectories. *PLoS Comput. Biol. 7*, No. e1002184.

(26) Dolinsky, T. J., et al. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res. 35*, W522−525.

(27) Mitternacht, S. (2016) FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research 5*, 189.

(28) Leaver-Fay, A., et al. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol. 487*, 545−574.

(29) Jacquier, H., et al. (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl. Acad. Sci. U. S. A. 110*, 13067−13072.

(30) Cabantous, S., Rogers, Y., Terwilliger, T. C., and Waldo, G. S. (2008) New molecular reporters for rapid protein folding assays. *PLoS One 3*, No. e2387.