Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

Biophysical Journal
Article

Biophysical Society

Generative β-hairpin design using a residue-based physicochemical property landscape

Vardhan Satalkar,[1] Gemechis D. Degaga,[2] Wei Li,[1] Yui Tik Pang,[3] Andrew C. McShan,[4] James C. Gumbart,[3,4] Julie C. Mitchell,[2,*] and Matthew P. Torres[1,4,*]

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia; [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]School of Physics, Georgia Institute of Technology, Atlanta, Georgia; and [4]School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia

ABSTRACT De novo peptide design is a new frontier that has broad application potential in the biological and biomedical fields. Most existing models for de novo peptide design are largely based on sequence homology that can be restricted based on evolutionarily derived protein sequences and lack the physicochemical context essential in protein folding. Generative machine learning for de novo peptide design is a promising way to synthesize theoretical data that are based on, but unique from, the observable universe. In this study, we created and tested a custom peptide generative adversarial network intended to design peptide sequences that can fold into the β-hairpin secondary structure. This deep neural network model is designed to establish a preliminary foundation of the generative approach based on physicochemical and conformational properties of 20 canonical amino acids, for example, hydrophobicity and residue volume, using extant structure-specific sequence data from the PDB. The beta generative adversarial network model robustly distinguishes secondary structures of β hairpin from α helix and intrinsically disordered peptides with an accuracy of up to 96% and generates artificial β-hairpin peptide sequences with minimum sequence identities around 31% and 50% when compared against the current NCBI PDB and nonredundant databases, respectively. These results highlight the potential of generative models specifically anchored by physicochemical and conformational property features of amino acids to expand the sequence-to-structure landscape of proteins beyond evolutionary limits.

SIGNIFICANCE Diverse protein sequences with similar physicochemical properties can lead to identical structural folds and functions. These evolutionarily driven protein sequences, composed of combinatorial arrangements of canonical amino acid residues, have a distinct physicochemical property landscape. Machine learning models can play a significant role in understanding the complex sequence-structure relationships in the context of physicochemical property landscapes and extend their scope beyond extant sequence space. Here, we developed a generative ML model, beGAN, that encodes PDB-derived sequences by their physicochemical properties and generates nonnatural β-hairpin sequences. We demonstrate that the model accurately classifies β hairpins from helical and disordered folds and generates diverse β-hairpin sequences. These results demonstrate the utility of using physicochemical property-based architectures for protein generative models.

INTRODUCTION

De novo protein design, whereby protein structures are designed from scratch and without reference to naturally occurring sequences, is a long-standing endeavor that promises to expand the finite limits of the sequence/structure landscape. Distinct from protein re-design, wherein an extant protein sequence/structure provides a framework for the designed structure, de novo design seeks to create entirely new sequences that will adopt a fold structure of interest (1). Thus far, this challenge has been met using a variety of approaches and with increasing success over the last 20 years (2). De novo protein design is now used in an array of synthetic biology and biomedical applications for the creation of biomaterials and bioactive protein switches, in the treatment of viral infections, and in the modulation of immuno-signaling pathways (3–13).

Computational design principles underlying de novo protein design can take on many forms but generally involve

Submitted October 23, 2023, and accepted for publication January 25, 2024.

*Correspondence: mitchelljc@ornl.gov or mtorres35@gatech.edu

Editor: Alexandr Kornev.

https://doi.org/10.1016/j.bpj.2024.01.029

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

some form of backbone sampling, sequence optimization, and possibly functional site design, all of which are aimed at identifying optimal sequence/structure combinations (1,2,14). A long-standing approach is to use local structural assembly, wherein folded modules of extant protein structures are fused to achieve a final larger protein structure in which the relative position and orientation of secondary structure elements, such as side chains and backbone structure, are sampled to identify local energy minima in the sequence-structure landscape (15–17). Several research studies have also noted the importance of physicochemical properties such as hydrophobicity, residue volume, and amino acid propensities of canonical amino acids in classifying protein folds (18–20). However, to the best of our knowledge, no generative machine learning (ML) model has yet fully explored the application of physicochemical and conformational properties to design protein secondary structures.

ML models have already surged within the field of de novo molecular structure prediction and design, including high-throughput in silico molecular screening and physicochemical property prediction (18–26). Particularly, ML models based on the generative adversarial network (GAN) are emerging as a promising new mechanism for exploring the theoretical protein sequence/structure landscape (27–32). These ML approaches are not only leveraging the accuracy of predictions but also accelerating existing computational models, such as molecular dynamics (MD), that are comparatively slow and require rigorous manual monitoring (33–35). Moreover, in the case of big-data availability, researchers are shifting from the use of ML models, such as support vector machines (36–38) and random forests (39,40), for predictive and clustering studies toward more robust techniques such as attention-based deep learning (41,42) and natural language processing (43,44).

Despite extraordinary advances in protein generative artificial intelligence, inverse folding remains a challenging problem owing to the vast protein sequence space (20^n, where n is the number of amino acid residues in a sequence). Compared to the large sequence space, protein fold space is only about 10^3 (45,46). Thus, it is essential to understand the correlation between sequence space and energy-governed fold space when designing functional protein scaffolds for biological applications (47,48). An explosion of effort in this area for full-length protein design has been published in recent years (1,42,49–56). In comparison, fewer models have addressed this challenge for peptide design (27,57,58,31,32,59–61).

ML models have also begun to address the challenge of mapping the theoretical sequence-structure relationship for peptides (8–50 amino acids) (27,28,31,32,58) and small proteins (>50 amino acids) (27,58,42,62,63). For instance, Xie and coworkers formulated the HelixGAN model by combining sequence and structural features from 3 million helix fragments from the PDB, resulting in the generation of stereospecific (D/L) helix structures (31). Karimi et al. also used fold-specific sequences to train a conditional Wasserstein GAN model to obtain protein secondary structures with novel fold representations (28). Most recently, Batra and coworkers developed an autonomous search engine combining a Monte Carlo tree search and ML-based random forest model with MD simulations to discover nonintuitive self-assembling pentapeptide sequences (58). In the similar vein, Pandi and coworkers developed and explored the utility of the generative variational autoencoder model to generate ∼500,000 antimicrobial peptides (64,65).

Peptide sequences with the β-hairpin secondary fold have been shown to play a key role in several protein regulatory functions, such as protein aggregation (66–68), biomolecular recognition (69,70), and antimicrobial or anti-viral drugs (32,71). An improved understanding of the sequence-structure relationship of β-hairpin peptides could potentially lead to new insights on the protein polymers such as amyloid fibers, well known for their toxicity to eukaryotic cells and their correlation with human disease (72).

Despite their growing importance, attempts to design isolated β-hairpin peptides have been met with difficulty due to their tendency to spontaneously assemble into insoluble higher-order structures (66,73). Although manual approaches have been employed for the design of β-hairpin peptides, these efforts have yielded relatively small-scale peptidomimetic β-hairpin libraries (74–79). We note that, very recently, Dupai and coworkers reported a thorough study on all 49,000 unique β-hairpin sequences extracted from the PDB (80). Based on their comprehensive structure-based sequence data exploration, they suggested several fundamental design principles for β-hairpin scaffolds, such as the incorporation of βstrands with amphipathic faces, specificity of the structural orientation at the turn region, and selection of amino acid residues at a particular position based on their physicochemical and electronic properties (80).

Here we describe the development, performance, and validation of a ML model beta GAN (beGAN) for generating sequences that conform to a β-hairpin scaffold. The model is composed of a two-class secondary structure discriminator that classifies peptide secondary structures based on the physicochemical and structural properties of position-specific amino acid residues with a high accuracy of about 96%, coupled with a generator that designs new deep-fake sequences. We employ beGAN to design β-hairpin peptides that we subsequently validated through in silico modeling such as MD simulations, state-of-the-art protein structure prediction techniques, and experimental techniques for structure determination by circular dichroism (CD) spectroscopy and solution nuclear magnetic resonance (NMR) spectroscopy. The results provide essential feedback on using extant protein structure databases for peptide-based generative models, give new insights on the physicochemical landscape of the β-hairpin scaffold and their design

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

principles, and highlight tested avenues for the success or failure of such approaches.

MATERIALS AND METHODS

ML feature engineering

In deep-learning-based protein structure predictions, contact maps and distance matrices have been recognized as primary methods of image-like protein representations. These image-formatted contact maps are constructed based on either residue covariation in amino acid contacts within protein structures or protein sequence patterns and coevolutionary couplings, also known as multiple sequence alignment (81,82). Such representation of proteins makes it convenient to customize deep learning architectures already developed for imaging and computer vision technologies such as convolutional neural networks (NNs) and residual convolutional NNs (83,84). Inspired by deep-network models, in this work, we encoded our peptide sequences using relevant physicochemical, electronic, and structural features associated to each residue, where each residue descriptor was computed from the AAindex database using the propy package, a tool for automated descriptor generation for amino acids (85,86). Throughout this work, each residue was assigned 12 descriptors, namely hydrophobicity; α-CH chemical shift; conformational parameters for β strands, α helices, and β turns; residue volume; steric parameters; normalization frequency for β strands, α helices, and β turns; β strands indices; and α-helical indices (Figs. S1 and S4). This gave a representation of each 16-mer peptide with 192 features, which allowed our models to learn patterns in the peptides based on relevant residue level local information.

Training data exploration

The accuracy and stability of deep learning models can depend upon the amount of data on which they are trained. With the increasing number of sequences available in protein databases (87), we first parsed about 100,000 known 16-mer β-hairpin peptide sequences based on their DSSP information. Since the turn regions of these sequences were variable in the number of residues and residue positions, we filtered for hairpins where the turn region was only limited to the middle residue positions R8 and R9, as shown in Fig. S1. This reduced the number of ground-truth examples considered for generative modeling to about 16,528 sequences. Corresponding to this β-hairpin set, we also amassed about 40,000 16-mer helices and about 72,000 disordered peptides. The disordered peptide sequences were selected based on the IUPred disordered prediction of their amino acids, where all consecutive amino acids within a window of 16-mer peptide length have a disorder value of 8.5 and above (88). In both cases, a stratified train-test-split was set to 0.25 where the models were trained on 75% of the datasets and 25% of examples were used for unbiased testing. The amino acid sequence logo was created using WebLogo 3.0 webserver (89).

Model training

The unique architecture of deep convolutional discriminative NN provides a significant advantage in the protein fold-classification problem (90,91). In this work, we have developed two distinct discriminator models based on deep convolutional discriminative NN, which can be coupled with a generator NN. The two-class discriminator (β hairpin/non-β hairpin) was trained using a dataset of 33,056 16-mer peptides evenly distributed between β hairpins (16,528) and non-β hairpins (i.e., helix/intrinsically disordered peptide) (16,528). Model training convergence was achieved using a batch size of 2400, 300 epochs, and a learning rate of 0.0002. The beGAN model was trained using a learning rate of 0.0002, a batch size of 24, and 1500 epochs. To tune the hyperparameters of the beGAN model, we utilized the learning rate for the ADAM optimization algorithm due to its quick convergence (92). To avoid mode collapse in the generator NN, a dropout rate of 3% was applied at each layer for regularization, and, to improve the overall convergence of the beGAN model, a nonlinear LeakyReLU activation function was set for all three hidden layers (93). When applying the GAN model, parameter values were mapped to a sequence by applying an L1 norm to each set of 12 descriptor values to select the closest amino acid. The terms of the L1 norm were scaled by the difference in maximum and minimum parameter values for the descriptors to ensure no descriptor had undue weight in determining the closest residue in the descriptor space.

Peptide structure prediction using AlphaFold2 and ESMFold

The PDB-derived and beGAN generated peptide (GP) sequences were utilized to predict the corresponding 3D structure of β-hairpin peptides using AlphaFold2 (41) and ESMFold (94,62) algorithms, respectively. For each peptide, the structure corresponding to the highest predicted local distance difference test (pLDDT) score was selected to determine the overall percentage of β secondary structure content. The β content was defined based on the percentage of β sheet residues (E) in the 16-mer peptide sequence. The 3D peptide structure was characterized using the DSSP algorithm to obtain per-residue secondary structure annotation such as β sheet (E), random coil (C), or helix (H) (95). Based on the β content, the GPs were sorted into four quadrants of increasing β% (Q1, 0%–12.5%; Q2, 12.5%–50%; Q3, 50%–75%; and Q4, 75%–100%).

MD simulations

Each peptide was constructed in an initially extended conformation using Molefacture in VMD (96), and solvated in a $60 \times 60 \times 60$ Å^3 box of $\sim$6600 TIP3P water molecules (97). Na^+ and Cl^- ions were added to a net concentration of 150 mM ($\sim$20 of each type of ion). After a short equilibration with NAMD 2.14 (98), subsequent simulations were run using Amber16 on GPUs using the three force fields described below (99). Although equilibration was run with constant pressure and temperature, production runs used constant volume and a constant temperature of 298 K enforced with Langevin dynamics. For each sequence, between two and eight 3- to 6-μs simulations (a total of 538 μs) were conducted, and the average fraction of β-hairpin content per residue was calculated. These data were also converted into a single average β-content percentage for the entire peptide (Fig. S7.1–S7.10).

MD force fields

Three distinct biomolecular force fields were used to simulate all peptides with TIP3P water: Amber ff14SB (100), CHARMM36m (C36m) (101), and CHARMM22* (C22*) (102). Simulations using C36m or C22* had a short-range cutoff of 12 Å for Lennard-Jones interactions and a switching function starting at 10 Å; simulations using ff14SB had a cutoff of 9 Å and no switching function. All simulations used the particle-mesh Ewald method for long-range electrostatics (103). A uniform time step of 2 fs was used for the first set of simulations for each force field. The SETTLE algorithm was used to constrain water molecules, and the SHAKE algorithm constrains all other hydrogen atoms. An additional set of C22* simulations for all peptides was run using a 4-fs time step along with hydrogen mass repartitioning (HMR) (104,105). HMR increases the mass of hydrogen atoms to 3 amu, decreasing the parent atoms' masses accordingly to conserve mass. We note that others have also identified imbalances in peptide conformations using ff14SB, which have since been corrected in ff19SB (106,107).

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

Representative MD snapshots selection

For simulation trajectory snapshots in Figs. 4 and S8, three were selected from the C22* MD simulation as the most representative. To achieve this, we first divided the entire trajectory into three equal segments. Then, from each segment, we chose one frame with the maximum number of structured residues, either as α helices or β sheets, to serve as a representative.

Peptide synthesis

Synthetic peptides were synthesized with N-terminal acetylation and C-terminal amidation to greater than 95% purity by Genscript Biotech. Lyophilized peptides were reconstituted in deionized water and further diluted to desired concentrations in 10 mM phosphate buffer (pH 7).

CD spectroscopy

Far-UV (190–250 nm) CD spectra were recorded on JASCO J-815 CD spectropolarimeter with a Peltier temperature control and a quartz cuvette with 1-mm path length. All experiments were carried out under 25°C, with 1-nm bandwidth and 1-s response time, and scanned at 50 nm/min in 0.2-nm steps. For each sample, 15 scans were averaged after solvent baseline correction. The final concentration of all peptides is 142 μM in 10 mM phosphate buffer (pH 7). CD analysis (secondary structure content estimation) was obtained using the BeStSel algorithm (108).

Solution NMR structure determination of peptides

From 1 to 2 mg of peptide ($\geq$95% purity, natural isotopic abundance, chemically synthesized from GenScript) was dissolved in 300 μL of 50 mM NaCl, 20 mM sodium phosphate pH 4.86, 10% D_2O, vortexed, and transferred to a 3-mm NMR tube. The pH was lowered to 4.86 to minimize the solvent exchange of backbone amide hydrogens. Each sample was used to record a suite of experiments that enabled backbone and side-chain resonance assignments: 2D ^{1}H-^{13}C HSQC, 2D ^{1}H-^{15}N HSQC, 2D ^{1}H–^{1}H NOESY (300- and 50-ms mixing times), and 2D ^{1}H-^{1}H TOCSY spectrum (80-ms isotropic mixing time) (109). All experiments were acquired using standard parameters at a ^{1}H field of 800 MHz at 4°C with recycle delay (d1) set to 1.2 s on a Bruker AVIIIHD-800 spectrometer equipped with TCI cryoprobe. Data were processed with nmrPipe (110). Chemical shift and NOE cross-peaks assignment were performed manually in Sparky (111). Torsion angle restraints for structure calculations were performed in TALOS-N using H, N, CA, and CB chemical shift values (112). Structure calculations were performed in CS-Rosetta using TALOS-N-derived restraints together with experimentally determined intramolecular NOEs. A total of 5000 conformations were calculated and sorted by total energy. Peptide ensembles were generated from the 10 lowest energy structures and validated using MolProbity.

Basic Local Alignment Search Tool

Basic Local Alignment Search Tool (BLAST) searching of the NCBInr and PDB databases was accomplished using the web-based BLASTp search engine with the following parameters: word size = 2; expected value = 200,000; hitlist size = 10; gapcosts = 9,1; matrix = PAM30; filter string = F; genetic code = 1; window size = 40; threshold = 11; and composition-based stats = 0. BLAST searching the training dataset containing approximately 99,000 unique training set sequences was accomplished using Blast2GO (113,114) to create a BLAST searchable database, followed by BLASTp analysis using a local version of the NCBI Genome Workbench (115).

Graphics and statistical analysis

MD and BLAST data were analyzed using JMP Pro 16 (SAS) and Python software. NMR structures were visualized using UCSF ChimeraX software (116).

RESULTS

Overview and performance evaluation of a beGAN model for β-hairpin sequence design

Generative models play a significant role in statistical and ML modeling where data augmentation is needed or in cases where an imbalanced dataset is encountered. In recent years, GANs have emerged as a powerful tool for generative modeling in computer imaging technologies. A GAN is composed of two NNs, a generator and a discriminator, working in an adversarial fashion. The generator is trained to design realistic new data, in an attempt to fool the discriminator, whereas the discriminator tries to distinguish synthetic instances produced by the generator from real samples in the input dataset. As this adversarial but complementary learning process goes on, both NNs get better at their respective jobs. When convergence is achieved, the GAN can produce synthetic new data that are indistinguishable from the real samples to the discriminator. Without design constraints or conditioning and sampling algorithms, given a random sequence vector, z, the generator designs new targeted sequences through joint minmax cross-entropy-loss optimization of the function (117),

$$E_x[\log(D(x))] + E_z[\log(1 - D(G(z)))] \quad (1)$$

where $D(x)$ is the discriminator's estimate of the probability that real sequence x is real, E_x is the expected value over all real sequences, $G(z)$ is the generator's output when given random sequence vector z, and $D(G(z))$ is the discriminator's estimate of the probability that a designed sequence is real. E_z is the expected value over all random inputs to the generator.

In this work, we utilized the GAN framework to develop the ML model that can classify β hairpins from α helices and disordered peptides to generate novel peptide sequences that adopt the desired β-hairpin fold. The architecture of the ML model is summarized in Fig. 1. The polypeptide chain consisting of 16 amino acid residues was chosen based on a minimum requirement for β-hairpin structures that were a primary focus of downstream validation studies. Our ML model utilizes an input feature matrix encoded using 12 physicochemical and fold-specific conformational indices of individual amino acid residues in the 16-mer peptide sequence (Fig. S1). The discriminator NN was trained with the ground-truth feature dataset using 16,528 PDB-derived 16-mer β-hairpin sequences. On the other hand, the generator NN was provided with a random vector. After training the discriminator NN on the ground-truth feature set derived from the PDB for 2400 batches and

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

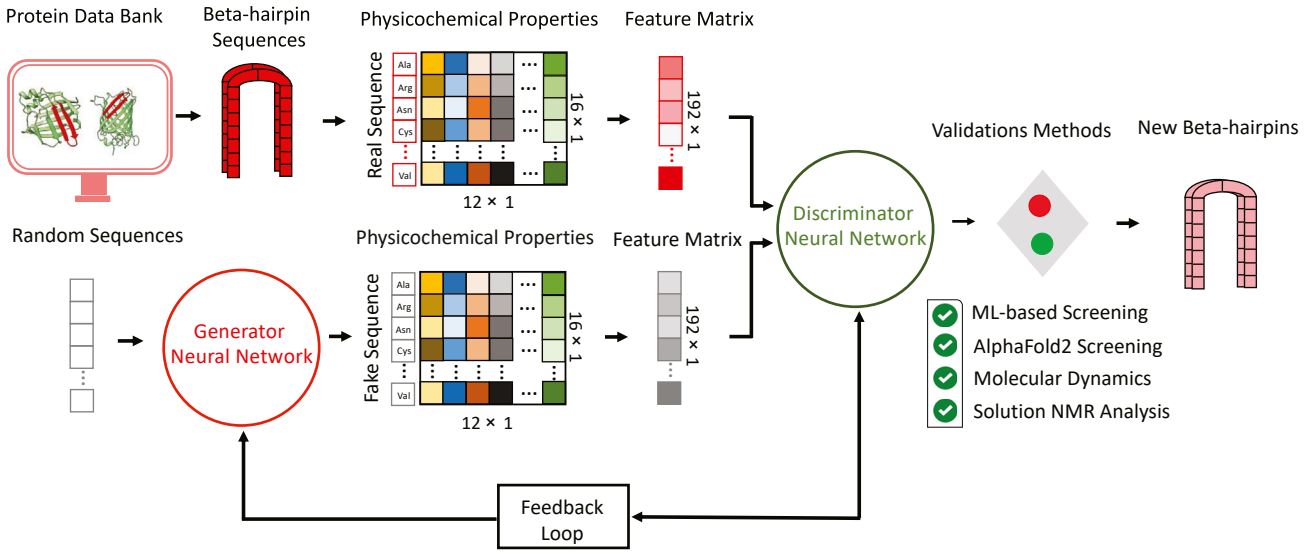


FIGURE 1 The model architecture of a GAN to design β-hairpin sequences. The input feature matrix was constructed by encoding 16-mer peptide sequences with 12 physicochemical and conformational properties of individual amino acids. The discriminator NN was trained on the ground-truth dataset of 16,528 16-mer β-hairpin sequences accrued from the PDB. The generator NN was trained to generate new target peptides from a feature matrix encoded using physicochemical and conformational properties of random 16-mer amino acid sequences. The architecture beGAN model contains an input layer with 192 features connected to three hidden layers with sizes 1024, 512, and 256 for the discriminator and these hidden layers are followed by a binary output layer with a sigmoid activation function. The generator has a similar architecture except the hidden layer is the conjugate-transpose of that of the discriminator.

300 epochs, the training loss was computed to be 0.02, and the training accuracy was 0.94 (Fig. 2 *A*).

The classification performance of the discriminator NN was determined using the area under the receiver operating characteristic curve (AUC-ROC). The value of AUC-ROC was computed to be 0.96, which suggests that the ML model was able to distinguish β hairpins from α helices and disordered peptides accurately (Fig. 2 *B*). Based on the binary classification confusion matrix, both values of recall and precision were determined to be 0.91 (Fig. 2 *C*). The beGAN model, wherein the discriminator model is coupled with the generator NN, was then trained for 2400 epochs using 24 batches. After model convergence was achieved, the generator NN was set to generate new 16-mer peptides (Fig. S2). The newly generated peptides were sorted into two distinct categories based on their classification probability, termed GP score. The first category includes the non-β-hairpin class, with a GP score ranging from 0.0 to 0.05. The second category is the β-hairpin class, which has a GP score ranging from 0.95 to 1.0.

In addition to ML validation, the amino acid conservation and relative frequencies were compared using the sequences from PDB-derived training data and newly generated binary classes (β hairpin and non-β hairpin) across 16-mer peptides using sequence logos (Fig. 2 *D*) and residue frequency distribution plots (Fig. S3). Naturally occurring β-hairpin sequences from the PDB are enriched with glycine (G), aspartic acid (D), and asparagine (N) at the turn region (residue positions 8 and 9) (Figs. 2 *D* and S3 *A*). This observation is consistent with a previously reported systematic data exploration of β-hairpin datasets, wherein G, D, and N are found to have higher propensities than other amino acid residues at the turn regions of the β-hairpin scaffold (80). We found that β-hairpin sequences generated by the beGAN model closely resemble the natural order of amino acid occurrences at both turn and β strand regions with a few exceptions (Figs. 2 *E* and S3 *B*). For example, the generated 16-mer β-hairpin sequences include G residue at the turn region (R9) with about 25.9% higher residue frequencies than PDB-derived 16-mer sequences (Figs. 2 *E* and S3 *C*).

As expected, the sequence conservation and amino acid frequencies from the non-β-hairpin dataset are significantly different compared to the β-hairpin training and generated datasets (Fig. 2 *F*). For example, the sequence conservation and frequencies suggest the enrichment of histidine (H), methionine (M), and phenylalanine (F) residues at R8 and R9 positions.

Most interestingly, the beGAN model was able to learn and generate the various physicochemical and conformational trends of the amino acid residues across the peptide sequence. For instance, the hydrophobicity feature distributions of sequences from the β-hairpin training and generated β-hairpin datasets showed a significant variation in median hydrophobicity values from 0.0 up to 1.4 among turn and middle β strand regions (Fig. 2 *H* and *I*). Moreover, at turn position (R9), the upper quartile value of hydrophobicity was as low as 0.6 and 0.0 for the β-hairpin training set and generated β-hairpin class, respectively. On the other hand, for the generated non-β-hairpin class, the hydrophobicity values did not show such significant variation among turn and strand

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

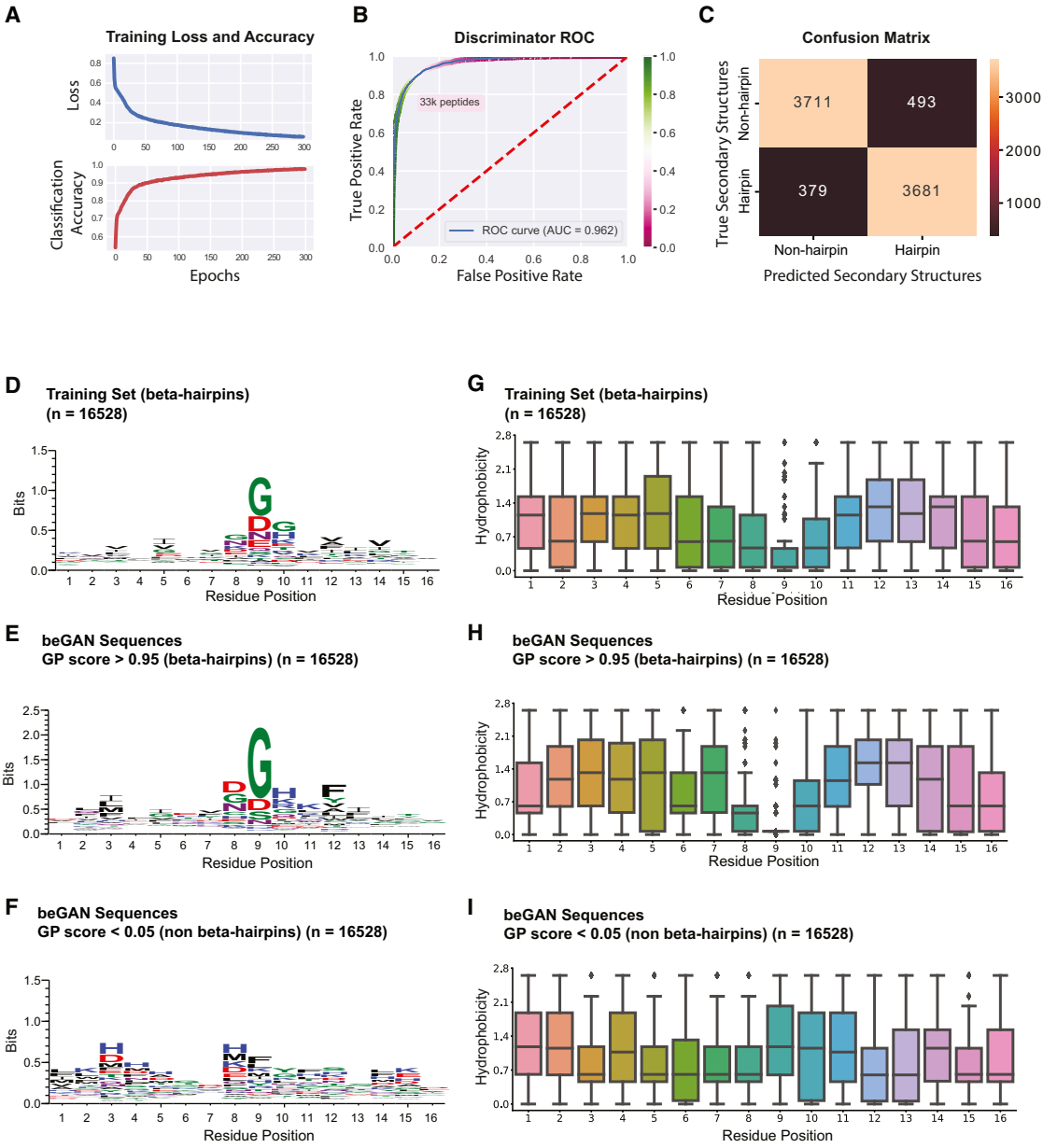


FIGURE 2 Model evaluation. (*A*) Computed training loss and classification accuracy trends are shown for the two-class discriminator NN (β-hairpin/non-β-hairpin class) during model training using 300 epochs. (*B*) The classification performance of the two-class discriminator NN (β-hairpin/nonβ-hairpin class) was measured with the AUC-ROC of the discriminator NN using 25% of the overall dataset ($n = 33{,}054$). (*C*) Confusion matrix for binary classification between hairpin and nonhairpin classes. (*D–F*) Sequence logos of 16-mer peptides are shown to indicate amino acid conservation and their relative frequencies. The data for the three logos consist of 16-mer amino acid sequences from PDB-derived β-hairpin training set, generated β-hairpin class, and generated non-β-hairpin class, respectively. (*H–J*) Hydrophobicity feature distribution across 16 residue positions shown for sequences from PDB-derived β-hairpin training dataset, generated β-hairpin class, and generated non-β-hairpin class, respectively.

regions, wherein the median of hydrophobicity values only changed from 0.6 to 1.2 across all residue positions (Fig. 2 *J*). Further analysis of the other 11 model features with respect to amino acid position in the β hairpin showed that other physicochemical properties captured from natural sequences were recapitulated in generated sequences (Fig. S4.1 and S4.2). Based on these trends, we conclude that the beGAN model learned the complex feature correlations between the selected feature set and was able to generate a physicochemical and conformational property landscape that was effectively decoded into novel amino acid sequences most suitable for a β-hairpin scaffold.

In addition to 16-mers, the beGAN model architecture was also extended to generate alternative peptide lengths including 14-mer, 18-mer, and 20-mer models. Similar to the 16-mer model, the extended models were trained on

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

the available β-hairpin sequences extracted from PDB and their performance was evaluated (Fig. S4). All new models were also highly accurate (above 95%) and generated unique 14-mer, 18-mer, and 20-mer peptide sequences that are based on the position-specific residue distribution of the corresponding PDB-derived peptide sequence dataset (Fig. S5). However, training data available from the PDB for β hairpins become a limiting factor and can affect the generative performance of the model. For example, with an extensive training dataset (n = 15,402), the 14-mer model was able to recapitulate the training sequence distribution and better mimic hydrophobicity trends compared to the 14-mer model trained on the relatively smaller dataset (n = 4631) (Fig. S6).

beGAN performance evaluation by AlphaFold2 and ESMFold protein structure prediction tools

Recent protein structure prediction methods have been demonstrated to predict short β-hairpin structures with high accuracy (118). In this work, we utilized AlphaFold2 (41) and ESMFold (94,119) models to predict the structures of 500 beGAN-GP sequences and used this to designate a success rate of our model. For each peptide sequence, the peptide structure corresponding to the highest pLDDT score (41) was selected to determine the overall percentage of β secondary structure content (95,120). Based on this, the GPs were sorted into four quadrants of increasing β content percentage (β%) (Q1, 0%–12.5%; Q2, 12.5%–50%; Q3, 50%–75%; and Q4, 75%–100%), wherein the range from Q2 through Q4 represent hairpins of increasing structural quality (Fig. 3 *A*). Of the 500 16-mer-peptides screened using AlphaFold2, 345 (69%) were classified as β hairpin (i.e., sorting into Q2–Q4), whereas 155 (31%) were classified as non-β-hairpin structures (i.e., sorting into Q1). Of the 345 β-hairpin structures, 244 were characterized as well-structured β hairpins involving 12 or more amino acid residues (Q4) in the β sheet structural fold. In comparison, 87 peptide sequences adopted moderately well-structured β-hairpin structures with 8–11 amino acid residues in β-hairpin fold (Q3), whereas only 14 were determined to be lower-quality β-hairpin structures with as few as two to eight amino acid residues involved in a β sheet structure (Q2). Compared to the predictive performance of AlphaFold2, ESMFold yielded 198 β-hairpin structures (Q2–Q4) and 301 non-β-hairpin structures (Q1) (Fig. 3 *B*).

To verify the ability of AlphaFold2 to predict the structures of experimentally resolved 16-mer β hairpins, we also used the algorithm to predict structures of PDB-derived β-hairpin sequences. Out of 500 PDB β-hairpin structures, AlphaFold2 predicted 445 (89%) β hairpins (Q2–Q4) and 55 (11%) non-β hairpins (Q1) (Fig. 3 *D*). Moreover, the structural comparison between 100 PDB β-hairpin structures and corresponding AlphaFold2 structures yielded Cα-root-mean-square deviation (RMSD) values with a median of 1.13 Å and standard deviation of 0.3 Å (Fig. 3 *E*). The lowest Cα-RMSD value of 0.283 Å represented excellent matching between AlphaFold2 and the extracted PDB β-hairpin structure. In contrast, the highest variation in Cα-RMSD was found to be 19.15 Å, where AlphaFold2 does not correctly predict the β-hairpin structure.

Based on the performance of the AlphaFold2 model to predict structures of the generated sequences, the overall success rate of the beGAN model was evaluated across GP scores ranging from 0.95 to 1.0 (Fig. 3 *C*). The success rate was determined as the ratio of the total number of positive β hairpins from Q2 to Q4 to the total number of peptides in all quadrants in the GP score range. The success rate of the beGAN model ranged from 88.9% above the highest threshold of 0.995 down to 77% at a score threshold of 0.95. In contrast to AlphaFold2, evaluating the success rate based on ESMFold resulted in a lower range of success rates from 59.0% to 44.4%, corresponding to the sequences having a GP score range from 0.95 to 1.0 (Fig. 3 *C*).

The performance of AlphaFold2 was further validated on 500 β-hairpin sequences generated by an alternative model, ProteinMPNN (Fig. 3 *F*). We found that, out of 500 ProteinMPNN-generated and AlphaFold2-predicted peptides, 397 peptides (79.4%) were predicted to be β hairpins and 103 peptides (20.6%) were predicted as non-β hairpins.

Evaluation of beGAN-generated sequences by MD simulation

In addition to using AlphaFold2 and ESMFold models to validate beGAN-generated sequences, we also opted to evaluate our GAN model performance using MD simulations. We selected nine sequences for evaluation: four beGAN-generated sequences from the nonhairpin class, exhibiting GP scores less than 0.09; four from the β-hairpin class with GP scores greater than 0.9; and one control sequence obtained from the well-folded β scaffold from PDB: 1A70 (Table S1) (121). Test sequences were determined to be nonredundant and significantly distinct from one another using CD-HIT (see section "materials and methods"). Since MD force fields are calibrated based on protein structure data, we refrained from making a priori assumptions about their effectiveness. Instead, we tested three distinct force fields: CHARMM22* (C22*), CHARMM36m (C36m), Amber ff14SB (ff14SB); and a variant of the first, C22* with hydrogen mass repartitioning (C22*hmr) (see section "materials and methods") (101,102,104,105). The β-hairpin and non-β hairpin classes were statistically distinguishable ($p < 0.05$) regardless of the force field used. Particularly, the MD data obtained from C22*, C36m, and C22*hmr force-field simulations showed better class separation than data from ff14SB simulations (Fig. 3 *G–J*; Table S1). Secondary structure analysis of C22* MD data was performed using the STRIDE algorithm (122), which

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

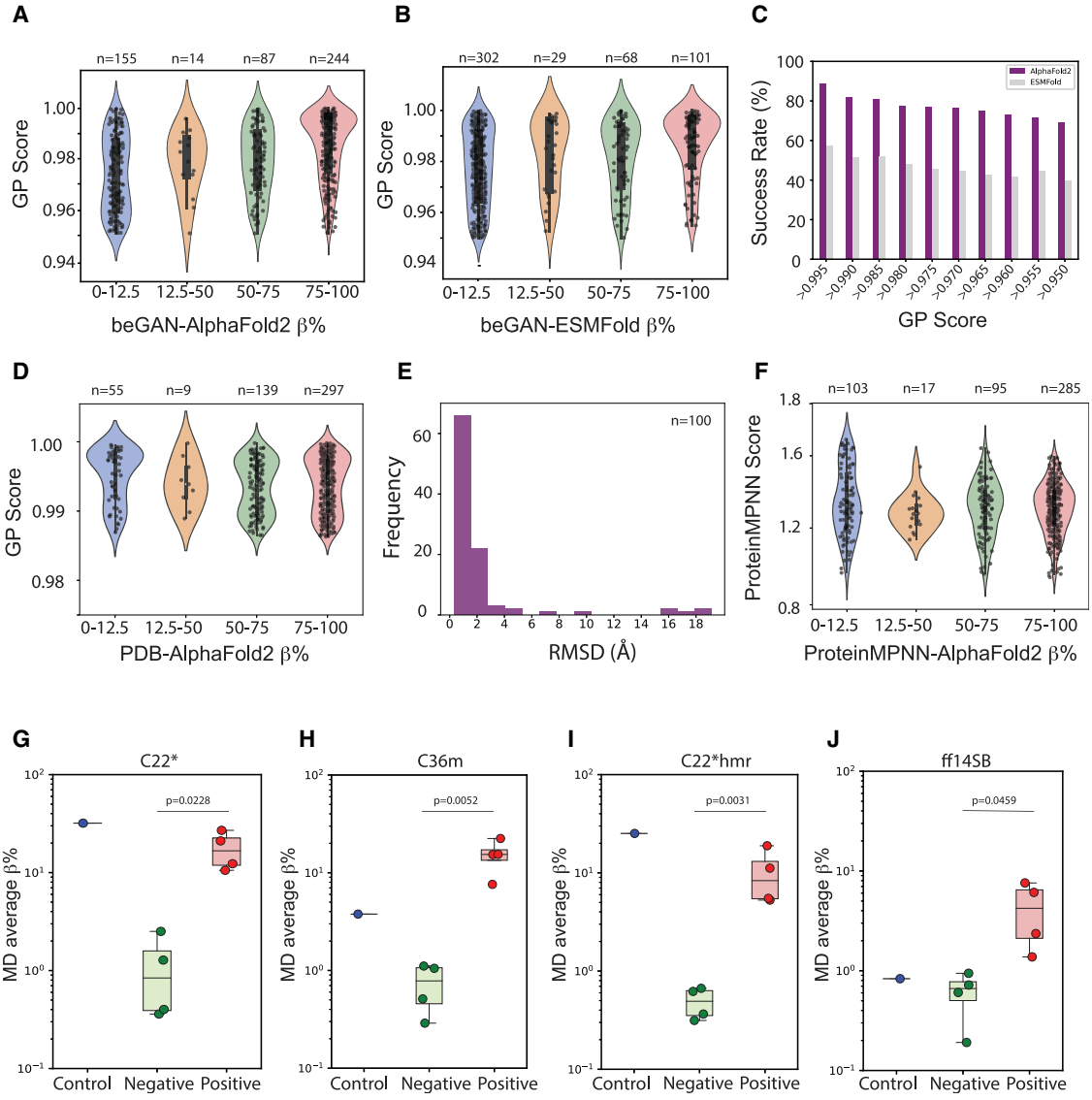
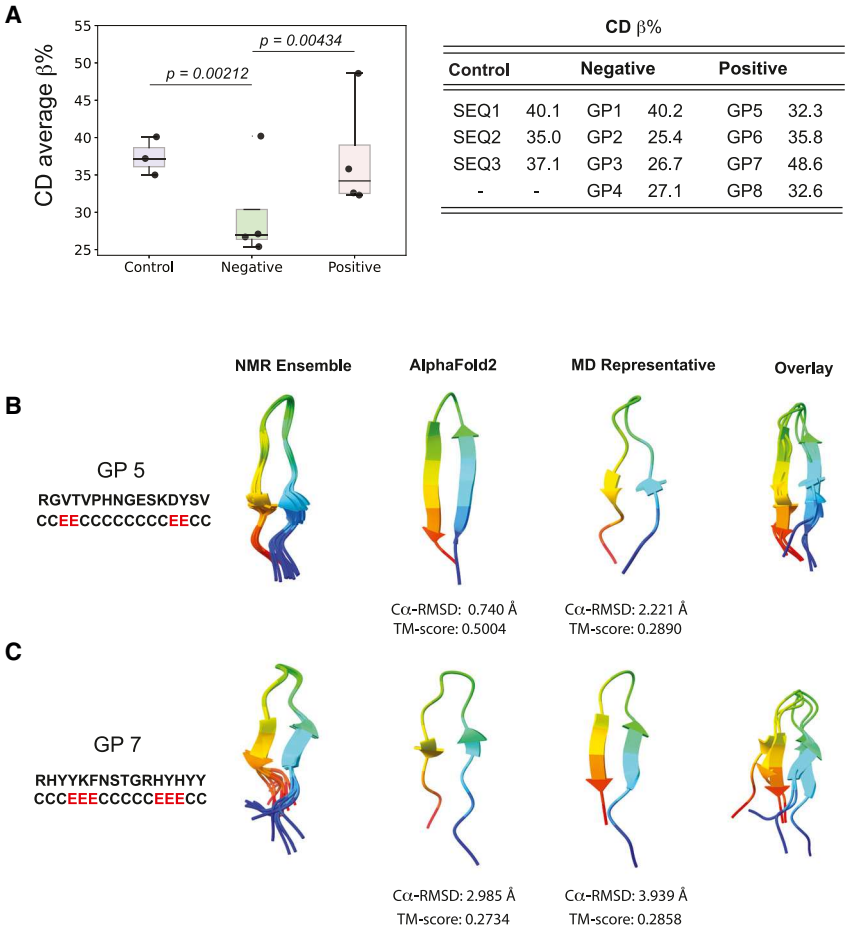


FIGURE 3 Screening and validation of beGAN-generated sequences using in silico and in vitro methods. (*A and B*) AlphaFold2 and ESMFold models were utilized to predict the secondary structures of the 500 beGAN-generated sequences with GP scores higher than 0.95. The 500 GP peptides were sorted into four quadrants (Q1, Q2, Q3, and Q4) based on the total β% content calculated by DSSP algorithm. (*C*)The success rate of the beGAN model was computed across 10 windows of GP score ranging from 0.95 to 1.0 using AlphaFold2 (*magenta*) and ESMFold (*gray*) models. (D) AlphaFold2 model was utilized to predict the secondary structures of the 500 PDB-derived β-hairpin sequences with computed GP scores. The 500 PDB peptides were sorted into four quadrants (Q1, Q2, Q3, and Q4) based on the total β% content calculated by DSSP algorithm. (*E*) Structural comparison between 100 PDB-extracted β-hairpin structures and their AlphaFold2 structures were performed and the corresponding Cα-RMSD values were plotted in the histogram. (*F*) AlphaFold2 was utilized to predict the secondary structures of the 500 ProteinMPNN-generated β-hairpin sequences, which are sorted into four quadrants (Q1, Q2, Q3, and Q4) based on the total β% content calculated by the DSSP algorithm. (*G–I*) β-hairpin content established by the average maximum per-residue β content quantified from replicate MD simulations using four force fields revealing distinct classification of positive β (*red circles*) from the negative class (non-β) GP peptides (*green circles*). Control peptide corresponds to a sequence retrieved from the PDB that adopts an individual hairpin that is not part of a larger β sheet secondary structure (PDB: 1A70). Data plotted on a $\log_{10}$ scale for improved visualization. Two-tailed tests (p) using unequal sample sizes were performed for each of the four forcefields using positive (*red circles*) and negative (*green circles*) clusters.

revealed a median difference of 15.8% in the β-hairpin content between the positive and negative class ($p = 0.0228$), wherein the positive class ranged from 10% to 27%, and the negative class ranged from 0% to 3%. The control β scaffold exhibited β-hairpin structure for 32% of the C22* trajectories. In the case of MD simulations using C36m and ff14SB force fields, the control peptide yielded lower β-content values than the corresponding generated positive class peptides, whereas the C22*hmr simulation yielded similar results to C22*.

Simulated peptide structures sampled from the MD simulations provided further insight into the dynamics of beGAN peptides in each class (see section "materials and methods") (Figs. S7.1–S7.10 and S8). As expected,

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029



CD β%

Control		Negative		Positive	
SEQ1	40.1	GP1	40.2	GP5	32.3
SEQ2	35.0	GP2	25.4	GP6	35.8
SEQ3	37.1	GP3	26.7	GP7	48.6
-	-	GP4	27.1	GP8	32.6

FIGURE 4 Structural comparison of select beGAN-GPs using experimental and computational methods. (*A*) β-hairpin content derived from CD spectral fitting revealing classification between four positive-class (β) peptides (*red box*) from the four negative-class (non-β) peptides (*green box*) (p (two-tailed) = 0.00434 with *Z* score $\geq$1.5). The CD spectral fitting of three control peptides (*purple box*) corresponding to sequences retrieved from the PDB can also be distinguished from the four negative (non-β) peptides (p (two-tailed) = 0.00212 with *Z* score $\geq$1.5). (*B and C*) An experimentally determined solution NMR structure ensemble of two representative peptides (*left*) is shown with their sequences and secondary structure assignments per residue using highest-ranked NMR structure. Secondary structure assignments were determined using the DSSP algorithm. The highest-ranked NMR structure of peptide was used to compare against the corresponding AlphaFold2-predicted structure with the highest pLDDT score (*middle*) and a representative snapshot sampled from MD simulations (C22*, *right*) using both Cα-RMSD values and TM score. All three structures are shown overlayed on the far right.

MD-generated configurations of the beGAN β-hairpin sequences showed a trend toward the canonical β-hairpin secondary structure. The secondary structures of non-β-class sequences were classified into a mixed class that is either α helix or intrinsically disordered peptide. Overall, β-hairpin sequences tended to adopt their target structure at greater frequency than did non-β-hairpin sequences throughout the 4-μs simulations, often remaining stable for 1–2 μs. Additionally, the middle residue (R9) of the β-hairpin peptides consistently participated in the turn among all positive test cases.

beGAN-generated sequencesadopt a β-hairpin structure in vitro

To experimentally evaluate model performance, we selected eight potentially soluble, generated sequences: four from the nonhairpin class, exhibiting GP scores less than 0.09, and four from the β-hairpin class with GP scores greater than 0.9 (Table S1). Peptide solubility was estimated using a hydrophilic residue count, which we validated with an ML-based solubility model based on a recurrent NN (123) (Fig. S10). Each peptide was synthesized, purified commercially, and then analyzed independently by CD spectroscopy. Spectral fitting of CD data for each synthetic peptide revealed only a 7.2% difference between the β-hairpin content of the positive and negative class (p = 0.004), wherein the positive class ranged from 32% to 48% and the negative class from 25% to 40% (Figs. 4 *A* and S9). In comparison, control peptides selected directly from the PDB that appeared to form β hairpins in isolation from larger β sheet structures showed a narrower range in β content spanning from 35% to 40% that was statistically different (p = 0.002) than all negative class GP peptides.

We successfully resolved NMR structures of two β-hairpin peptide sequences, namely GP 5 and GP 7, generated by beGAN (Fig. 4 *B* and *C*). These specific peptide sequences were selected to survey high GP-scored peptides (above 0.90) that exhibit high water solubility and high β% content based on MD simulation results that have also passed through AlphaFold2 screening. Both peptides adopted the intended β-hairpin structure comprising strongly bound anti-parallel βstrands and a two-residue hairpin loop (Fig. 4 *B* and *C*). Top-ranked peptide structures obtained from AlphaFold2 and MD simulations were compared with the corresponding NMR structure using

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

RMSD in Cα atoms (Cα-RMSD) and TM-score metrics to quantitively determine a topological similarity between the two structures (124,125). The top-ranked NMR structure of GP 5 closely resembled its corresponding AlphaFold2 predicted structures with a Cα-RMSD value of 0.740 Å and TM score of 0.5004, indicating excellent agreement. In the case of GP 7, the NMR structure was found to be similar to the AlphaFold2 predicted structure with a lower Cα-RMSD value of 2.985 Å and a TM-score of 0.2734. Moreover, the solved NMR structures of GP 5 and GP 7 showed similar β propensities to the corresponding representative structures from the MD simulations with Cα-RMSD values of 2.221 and 3.939 Å and TM scores of 0.2890 and 0.2858, respectively. The beGAN peptides with higher MD β% (Fig. 3 *G–J*) and CD β% (Fig. 4 *A*) also adopted the β-hairpin NMR ensemble in the aqueous phase (Fig. 4 *B* and *C*). Overall, we found that there was a positive trend in the β% content observed by CD and corroborated with MD simulation and NMR spectroscopy, further supporting the generative performance of the beGAN model.

beGAN-generated sequences are unique and diverse from naturally evolved sequences

beGAN-generated sequences are created from random seeds using PDB-derived evolutionary training data. Therefore, we sought to establish their uniqueness among sequences observed in nature. We used the NCBI BLAST algorithm to quantify sequence homology between the AlphaFold2-screened 500 GP peptide sequences (see Fig. 3 *A*) and the comprehensive PDB and nonredundant (NR) databases. Results from this search demonstrated that beGAN sequences matched both databases to varying degrees of query coverage and sequence identity (Fig. 5). When the complete range of query coverage is selected (100% query coverage; full 16-mer), the homology search of GP peptides against the NCBI PDB database resulted in a minimum sequence identity of 31.25% for Q1 (non-β class) and 37.50%, 31.25%, and 31.25% for Q2–Q4 (β-class), respectively. In some cases, beGAN was found to generate sequences that already existed in the NCBI PDB database, but these were relatively rare. For example, five out of 244 sequences in Q4 (2%) were found to be a perfect match with extant sequences in the NCBI PDB database (100% query coverage and 100% sequence identity). Moreover, homology searching of the same peptides against the NCBInr database resulted in the lowest minimum sequence identity of 50.00%, 56.25%, 56.25%, and 50.00% ranging from Q1 to Q4 quadrants, respectively (Fig. S11). From the last Q4 quadrant, seven out of 244 sequences in Q4 (2.8%) were found to be a perfect match with extant sequences in the NCBI NR database. In contrast, there were no perfect matches found in any of Q1–Q3 ($n = 256$) in both NCBI PDB and NR databases (Figs. 5 *E–G* and S11 *E–G*).

To estimate the ability of beGAN-GPs with minimum sequence identity to adopt a β-hairpin structure, we evaluated beGAN with sequence identities ranging from 50% to 56% calculated using the NCBInr dataset (Fig. S12). The AlphaFold2-predicted β-hairpin structures of the 10 GP peptides (GP 9–18) with the lowest sequence identity are shown (Fig. S12 *A*), along with their percentage β content and beGAN-generated GP score (Fig. S12 *B*). Each of the 10 peptides exhibits a high β content of 75% with 12 out of 16 residues involved in the β-hairpin fold. Taken together, these results demonstrate that sequences produced by beGAN, which relies on physicochemical properties rather than sequence or structural homology, are unique from natural sequences and properly fold into their intended fold structures.

Benchmarking against a current state-of-the-art generative model

The current ProteinMPNN model meets several challenges in protein design (42). However, unlike our beGAN model that relies predominantly on physicochemical features of amino acids, ProteinMPNN is designed to take existing static protein or peptide structures as its input. Here, we compared the performance of the beGAN model with ProteinMPNN to generate novel peptide sequences. Using 30 16-mer β-hairpin structures extracted from the PDB (Fig. 6 *A*), we generated 300 sequences using the ProteinMPNN algorithm (Fig. 6 *B*). These 300 sequences were compared against 300 beGAN sequences in terms of positional residue frequency and peptide sequence diversity.

We compared the amino acid propensity at each position in the generated sequences (Fig. 6 *B* and *C*). Results from both ProteinMPNN and beGAN models reveal a wide array of possible amino acid combinations, suggesting that not one minimal sequence motif defines the physicochemical landscape of well-structured β hairpins. Beyond this, both models favored amino acids with lower residue volumes in the turn regions such as G, D, and N. Both models also favored hydrophobic amino acids in the β strand region, although beGAN peptide sequences were notably more diverse. For example, ProteinMPNN favored valine (V) (13%–42%), isoleucine (I) (0.6%–9.3%), and tyrosine (Y) (0.3%–9.3%) in the βstrands of many peptides, whereas beGAN generated a more diverse distribution of amino acids such as I (0%–24.6%), Y (2.6%–12%), F (2.3%–29.6%), and V (0.6%–21.7%). Interestingly, for both models, we also found a smaller proportion of peptides with βstrands containing complementary charged amino acid pairs (Fig. 6 *D* and *E*). ProteinMPNN generated sequences that were enriched with lysine (K) at positions 1 (31.0%) and 10 (35.0%) and glutamic acid (E) at positions 7 (9.6%) and 16 (30.3%). Meanwhile, beGAN generated sequences enriched with K at position 11 (26.6%) and E at position 6 (14.0%) or H at position 10 (32.0%) and D at position 8 (25.6%). Upon further investigation of beGAN

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

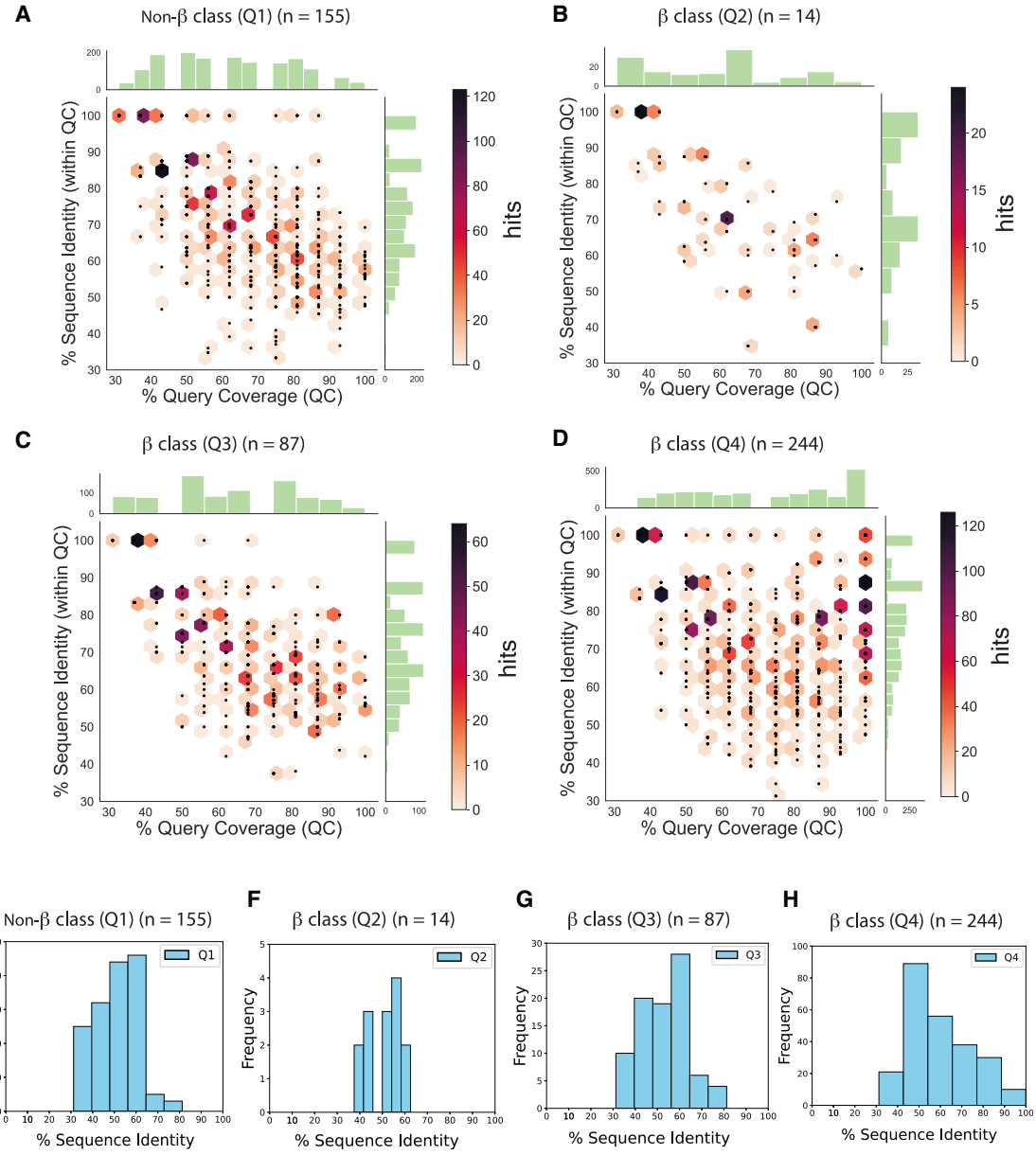


FIGURE 5 Homology search of 500 beGAN-generated sequences in the NCBI PDB database. (*A–D*) NCBI PDB BLAST results for 500 beGAN-generated 16-mer peptides (GP) sequences across four quadrants (Q1–Q4) were determined using calculated β% content of AlphaFold2-predicted structures with the highest pLDDT score. Percentage identity (within query coverage (QC)) corresponds to the percentage of the query sequence that shares an identical residue(s) with the target sequence per target length. QC corresponds to the percentage of residues in the 16-mer that are involved in the match. Hits (*black circles*) correspond to individual matches between GP peptides and the PDB sequence dataset across query coverage range from 0% to 100% from the top 10 matches obtained from the protein BLAST search. Density of the hits is also shown for each quadrant in colored hexagonal bins. (*E–H*) Percentage sequence identity values of the top match from the BLASTP search (*blue*) and frequencies of the individual beGAN sequences from four (Q1–Q4) quadrants are shown. Percentage sequence identity was calculated as the percentage of the query sequence that shares an identical residue(s) with the target sequence per length of the GP.

peptides, we found the residue-residue contact distance between these charge pairs was within the cutoff distance (4–6 Å) for a potential electrostatic interaction (Fig. S13) (126). It is also noteworthy to mention that Batra and coworkers reported similar trends of complementary charged residues in their β-hairpin self-assembly peptide design model (58).

Additional differences in the amino acid frequency datasets generated by the two ML models can be summarized based on the amino acid frequency distributions. The beGAN model was found to generate 28% more G, 32% more H, and 30% more F at positions 9, 10, and 12, respectively, compared to the ProteinMPNN dataset. ProteinMPNN generated 30% more V and 26% more of

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

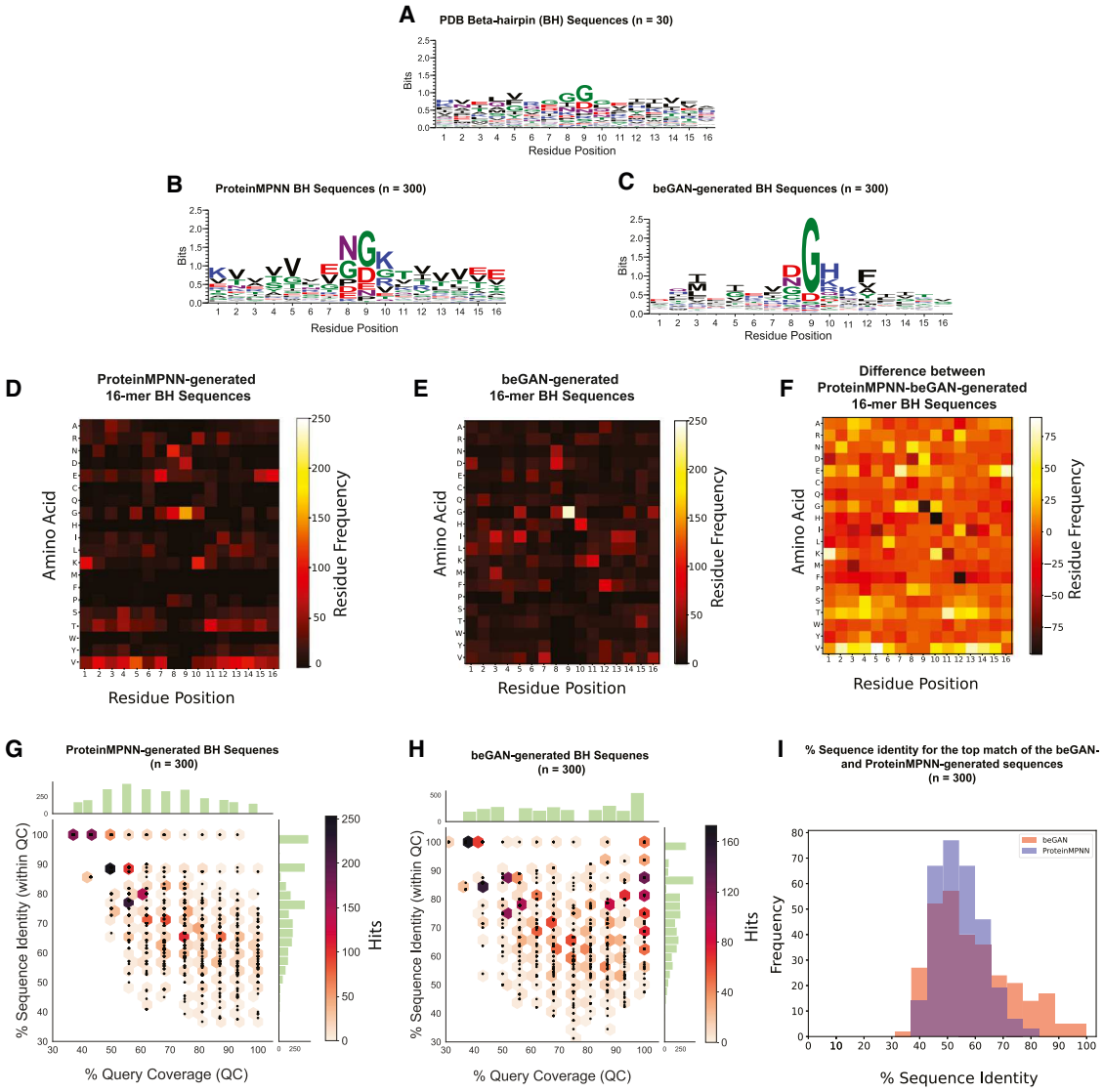


FIGURE 6 Benchmarking beGAN model against ProteinMPNN. (*A*) *Sequence logos of 30 PDB-derived 16-mer peptides which were used as the indiviual input sequences for the ProteinMPNN model.* (*B*) Sequence logos of 16-mer peptides are shown to indicate amino acid conservation and their relative frequencies using 300 ProteinMPNN-generated β-hairpin sequences (*left*). (*C*) Sequence logos of 16-mer peptides are shown to indicate amino acid conservation and their relative frequencies using 300 beGAN-generated β-hairpin sequences. (*D*) Amino acid residue frequency distribution plots for 300 sequences from ProteinMPNN-generated sequence dataset, (*E*) beGAN-generated dataset, and (*F*) the difference in amino acid residue frequencies from ProteinMPNN and beGAN sequence datasets. (*G*)) NCBI PDB BLAST results for 300 ProteinMPNN-generated, and (*H*) beGAN-generated sequences are shown using a full query coverage range. (*I*) Comparison between percentage sequence identity values of the top match from the BLAST search of the individual beGAN-generated sequences and ProteinMPNN-generated sequences is shown for 300 generated sequences. Percentage sequence identity was calculated using the percentage of the query sequence that shares identical residue(s) with the target sequence per length of the generated peptide.

both K and E at residue positions of 5, 1, and 16, respectively.

To compare diversity for the generated sequences, we performed a BLAST homology search against the NCBI PDB database for 300 beGAN- and 300 ProteinMPNN-generated sequences. When the complete range of query coverage is compared (100% query coverage; full 16-mer), beGAN performed slightly better in generating sequences with lower percentage identities (compare 31.2% for beGAN to 37.5% for ProteinMPNN) (Fig. 6 *G–I*). However, ProteinMPNN exhibited a lower overall median percentage identity (compare 56.2% for beGAN to 50% for ProteinMPNN). Moreover, although beGAN produced a small number of identically matched sequences (five out of 300), ProteinMPNN did not produce any identical sequences compared to the NCBI PDB dataset. In addition to the BLAST homology search, we also performed a clustering analysis to quantitatively evaluate the sequence diversity generated by beGAN and identify sequence redundancy. To do this, we used CD-HIT to cluster 300 beGAN

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

β-hairpin sequences from quadrants Q2 to Q4 that were validated by AlphaFold and had classification scores of 0.95 or higher. Depending on the word size and identity threshold used, a range between 185 and 295 unique sequence clusters ws observed, indicating a high level of sequence diversity in beGAN sequences. In comparison, identical analysis of 300 ProteinMPNN sequences generated using 50 reference β-hairpin scaffolds yielded a range between 34 and 130 unique clusters, indicating lower sequence diversity (Fig. S14). These results indicate that the generator and discriminator performance of the beGAN model produce diverse sequences.

DISCUSSION

Here, we have demonstrated the potential of a generative ML model that relies on the physicochemical and conformational properties of amino acids to create novel peptide sequences designed to adopt a β-hairpin secondary structure. We show that the model is able to uncover key trends and fingerprints of 12 physicochemical and conformational properties essential for β-hairpin folding and generates peptide sequences that adopt their target structure in silico and in vitro. The beGAN model is able to learn amino acid-encoded physicochemical feature space and classify with high accuracy (96%), precision (91%), and recall (91%) β-hairpin folds from other fold structures such as helix/random coil features. Moreover, the GAN model architecture is extendable to generate variable peptide lengths ranging from 14-mer, 18-mer, and 20-mer amino acid sequences primarily dependent upon existing input data.

Our model also provides insights into the design principles of the β-hairpin peptide scaffold based on the physicochemical feature analysis of the generated sequence datasets. For example, the median hydrophobicity of β-turn residues (R8/R9) was found to be significantly lower compared to the median hydrophobicity of the residues at β strand positions (R10–R14) and (R3–R7), indicating incorporation of hydrophilic residues such as G and N at the turn regions (Fig. 2). A similar observation was found for residue volume, where the median of the residue volume at the turn region (R9) was substantially lower by at least 30 $Å^2$ or more than the median residue volume at the β strand region (Fig. S4.2). The electrostatic pairing observed between oppositely charged amino acid pairs on opposite ends of the β strands, such as in positions 8–10 and 6–11, may also stabilize the β-hairpin structure in some cases (Figs. S3 and S13).

The beGAN model has been validated through both computational and experimental evidence. Computational evidence was provided largely through AlphaFold2 predictions, which show that beGAN achieves up to 88.9% success rate in designing sequences that adopt a β-hairpin fold. An important extension of this work is in establishing the utility of AlphaFold2 in predicting secondary structures of short peptides at reasonable accuracy and lower overall computational and experimental cost. Given that the training set for AlphaFold2 is composed of naturally occurring sequences, we hypothesize that its ability to make accurate predictions will be reduced for peptides with lower sequence identity. In support of this hypothesis, we found that AlphaFold2 successfully predicted β-hairpin structures for 89% of a PDB-extracted β-hairpin test set in which sequence identities are 100% (Fig. 3). In contrast, for beGAN-generated sequences, which are significantly diverse from naturally occurring ones, we found that peptides with a high degree of sequence identity to naturally occurring sequences (between 80% and 100% identical) were only found in the Q4 quadrant as classified as well structured by β hairpins AlphaFold2 (Fig. 5 *E–H*). Thus, in general, given the current training dataset of AlphaFold2, it may perform better on naturally occurring rather than designed peptide sequences.

We also validated beGAN peptides with in vitro experiments and MD simulations. Indeed, a test set of AlphaFold2-predicted peptides were also found to adopt the expected secondary structure during microsecond-scale MD simulations, and these results were further supported by both CD and NMR experimental evidence (Figs. 3, 4, S7 and S8). We further highlight that the beGAN model designed fold-competent sequences that were diverse from naturally occurring ones (Fig. 5). These promising results show the potential of generative ML models for designing novel targeted peptides based on the physicochemical and conformational properties of amino acid residues. Moreover, relying on the physicochemical properties of amino acids instead of the sequence allows potential flexibility in extending the model to noncanonical amino acid residues.

Some assumptions were necessary when initiating the development of the beGAN model, one being that peptides isolated from whole protein structures in the PDB would preserve their secondary structure when isolated in silico or in vitro. However, this is likely not often true as, in many cases, β hairpins require stabilizing forces provided by surrounding residues in the whole folded protein (as in a β sheet structure) or by nonprotein structures, such as membranes, that are lost in our experiments (127). Beyond this, there is an apparent discrepancy between what conforms to an isolated hairpin as measured by crystallography and whether it would also form a well-structured hairpin as a peptide in vitro. Thus, by virtue of the model's reliance on crystal structure data for training, data for the classifier are not perfectly labeled for the intended outcome of a soluble hairpin. Indeed, isolated β-hairpin peptides have a documented tendency to exhibit low solubility due to aggregation (66,128). Experimental validation of β-hairpin peptides requires that they are soluble, which can be predicted by hydrophilic residue count (Fig. S10). We recommend this type of filter to identify soluble β hairpins from beGAN or other similar models.

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

The beGAN ML or similar peptide generative models may be useful for applications that can benefit from an expansion of the sequence-structure landscape of peptide libraries. Such applications may include antimicrobial peptide generation, materials science, or in the study of proteinopathies such as Alzheimer's disease. Indeed, many antimicrobial peptides adopt β-hairpin-like structures designed and screened for cell-killing properties (32,60,71,129). In materials science, aggregation in the form of amyloids is a desirable property for engineered biomaterials that rely on β-hairpin secondary structures (58,66,73), and the GAN-based ML model could provide particular utility in this area by expanding the sequence/structure possibilities for such materials. In the study of proteinopathies that rely on β-hairpin peptides, the model may provide similar advantages (66,73). Utilizing generative models to establish detoxifying hairpin structures that prevent fibril formation, for example, could be enhanced by generative ML models that allow the generation of peptide structures without being sequence constrained. Alternatively, the beGAN model may be useful in exploring the physicochemical nature of β-hairpin aggregation and amyloid formation that can expand beyond the landscape of naturally occurring sequences. Although beGAN is not designed for these applications in particular, it may be tailored to achieve such objectives if given functional features that inform the model on desired functional outcomes.

DATA AND CODE AVAILABILITY

The codes used to run the generative model are available at https://github.com/juliecmitchell/beGAN. Training data and other codes are available by request to mitchelljc@ornl.gov.

The assigned chemical shifts of GP 5 (RGVTVP HNGESKDYSV), and GP 7 (RHYYKFNSTGRHYHYY) peptides have been deposited to the Biological Magnetic Resonance Data Bank under accession codes 31101 and 31094, respectively. The atomic coordinates of GP 5 and GP 7 peptides have been deposited to the RCSB PDB under accession codes 8TXS and 8T61, respectively.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2024.01.029.

AUTHOR CONTRIBUTIONS

M.T., J.M., and J.G. designed experiments. V.S., M.T., G.D., J.G., and J.M. wrote the manuscript. G.D. and J.M. developed the initial ML model. V.S. tested and extended ML models and generated ML figures. W.L. and M.T. conducted bioinformatics and MD comparisons, including figure generation. Y.P. carried out MD simulations and generated figures. A.C.M. designed and performed NMR experiments as well as NMR structure calculations.

ACKNOWLEDGMENTS

We would like to acknowledge the Georgia Tech (GT) Southeast Center for Math and Biology (SCMB) for ongoing scientific feedback on this project. Thanks to Gary Newman, Bettina Bommarius, Nicholas Hud, and the GT Institute for Bioengineering and Bioscience (IBB) core facilities for support. This work was funded by NSF-Simons grant 1764406 (to J.M. and M.T.) and by NIH R01-GM148586 (to J.C.G.). G.D.,V.S., and J.M. acknowledge the CADES and Summit computational resources provided through the Oak Ridge Leadership Computing Facility. A.C.M. acknowledges start-up funds from the Georgia Institute of Technology. MD simulations were run using resources provided through the Extreme Science and Engineering Discovery Environment (XSEDE, TG-MCB130173), which is supported by NSF grant ACI-1548562, as well as the Hive cluster, which is supported by NSF grant 1828187 and is managed by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Pan, X., and T. Kortemme. 2021. Recent advances in de novo protein design: Principles, methods, and applications. *J. Biol. Chem.* 296, 100558.
2. Korendovych, I. V., and W. F. DeGrado. 2020. *De novo* protein design, a retrospective. *Q. Rev. Biophys.* 53:e3.
3. Shen, H., J. A. Fallas, …, D. Baker. 2018. De novo design of self-assembling helical protein filaments. *Science.* 362:705–709.
4. Gonen, S., F. DiMaio, …, D. Baker. 2015. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science.* 348:1365–1368.
5. Chen, Z., M. C. Johnson, …, F. Dimaio. 2019. Self-assembling 2D arrays with de novo protein building blocks. *J. Am. Chem. Soc.* 141:8891–8895.
6. Feng, J., B. W. Jester, …, D. Baker. 2015. A general strategy to construct small molecule biosensors in eukaryotes. *Elife.* 4, 10606.
7. Bick, M. J., P. J. Greisen, …, D. Baker. 2017. Computational design of environmental sensors for the potent opioid fentanyl. *Elife.* 6, 28909.
8. Glasgow, A. A., Y.-M. Huang, …, T. Kortemme. 2019. Computational design of a modular protein sense-response system. *Science.* 366:1024–1028.
9. Chen, Z., R. D. Kibler, …, D. Baker. 2020. De novo design of protein logic gates. *Science.* 368:78–84.
10. Cao, L., I. Goreshnik, …, D. Baker. 2020. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science.* 370:426–431.
11. Silva, D. A., S. Yu, …, D. Baker. 2019. De novo design of potent and selective mimics of IL-2 and IL-15. *Nature.* 565:186–191.
12. Mohan, K., G. Ueda, …, K. C. Garcia. 2019. Topological control of cytokine receptor signaling induces differential effects in hematopoiesis. *Science.* 364, eaav7532.
13. Chevalier, A., D. A. Silva, …, D. Baker. 2017. Massively parallel de novo protein design for targeted therapeutics. *Nature.* 550:74–79.
14. Huang, P.-S., S. E. Boyken, and D. Baker. 2016. The coming of age of de novo protein design. *Nature.* 537:320–327.
15. Anfinsen, C. B. 1973. Principles that Govern the Folding of Protein Chains. *Science.* 181:223–230.
16. Saven, J. G. 2002. Combinatorial protein design. *Curr. Opin. Struct. Biol.* 12:453–458.

Please cite this article in press as: Satalkar et al., Generative β-hairpin design using a residue-based physicochemical property landscape, Biophysical Journal (2024), https://doi.org/10.1016/j.bpj.2024.01.029

17. Norn, C., B. I. M. Wicky, …, V. Christian. 2021. Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. USA.* 118, e2017228118.
18. Zakharov, A. V., M. L. Peach, …, M. C. Nicklaus. 2014. QSAR modeling of imbalanced high-throughput screening data in PubChem. *J. Chem. Inf. Model.* 54:705–712.
19. Baskin, I. I., D. Winkler, and I. V. Tetko. 2016. A renaissance of neural networks in drug discovery. *Expet Opin. Drug Discov.* 11:785–795.
20. Segler, M. H. S., T. Kogej, …, M. P. Waller. 2018. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4:120–131.
21. Wu, Z., B. Ramsundar, …, V. Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9:513–530.
22. Ain, Q. U., A. Aleksandrova, …, P. J. Ballester. 2015. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 5:405–424.
23. Lima, A. N., E. A. Philot, …, K. M. Honorio. 2016. Use of machine learning approaches for novel drug discovery. *Expet Opin. Drug Discov.* 11:225–239.
24. Varnek, A., and I. Baskin. 2012. Machine learning methods for property prediction in chemoinformatics: quo vadis? *J. Chem. Inf. Model.* 52:1413–1437.
25. Ramakrishnan, R., P. O. Dral, …, O. A. von Lilienfeld. 2015. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *J. Chem. Theor. Comput.* 11:2087–2096.
26. Mitchell, J. B. O. 2014. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 4:468–481.
27. Anand, N., and P. Huang. 2018. Generative modeling for protein structures. *Adv. Neural Inf. Process. Syst.* 31.
28. Karimi, M., S. Zhu, …, Y. Shen. 2020. De Novo Protein Design for Novel Folds Using Guided Conditional Wasserstein Generative Adversarial Networks. *J. Chem. Inf. Model.* 60:5667–5681.
29. Janson, G., G. Valdes-Garcia, …, M. Feig. 2023. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* 14:774.
30. Repecka, D., V. Jauniskis, …, A. Zelezniak. 2021. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* 3:324–333.
31. Xie, X., P. A. Valiente, and P. M. Kim. 2023. HelixGAN a deep-learning methodology for conditional *de novo* design of α-helix structures. *Bioinformatics.* 39:btad036.
32. Randall, J. R., C. D. DuPai, …, B. W. Davies. 2023. Designing and identifying β-hairpin peptide macrocycles with antibiotic potential. *Sci. Adv.* 9, eade0008.
33. Wang, Y., J. M. L. Ribeiro, and P. Tiwary. 2019. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat. Commun.* 10:3573.
34. Gao, X., F. Ramezanghorbani, …, A. E. Roitberg. 2020. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* 60:3408–3415.
35. Zhou, G., B. Nebgen, …, S. Tretiak. 2020. Graphics Processing Unit-Accelerated Semiempirical Born Oppenheimer Molecular Dynamics Using PyTorch. *J. Chem. Theor. Comput.* 16:4951–4962.
36. Cai, Y.-D., X.-J. Liu, …, G.-P. Zhou. 2001. Support vector machines for predicting protein structural class. *BMC Bioinf.* 2:3.
37. Bhasin, M., and G. P. S. Raghava. 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* 32:W414–W419.
38. Cai, C. Z., L. Y. Han, …, Y. Z. Chen. 2003. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31:3692–3697.
39. Busch, J. R., P. A. Ferrari, …, F. Leonardi. 2009. Testing statistical hypothesis on random trees and applications to the protein classification problem. *Ann. Appl. Stat.* 3:542–563.
40. Chen, X.-W., and M. Liu. 2005. Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics.* 21:4394–4400.
41. Jumper, J., R. Evans, …, D. Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature.* 596:583–589.
42. Dauparas, J., I. Anishchenko, …, D. Baker. 2022. Robust deep learning–based protein sequence design using ProteinMPNN. *Science.* 378:49–56.
43. Madani, A., R. Socher, …, N. Naik. 2023. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol.* 41:1099–1106.
44. Nijkamp, E., J. Ruffolo, …, A. Madani. 2022. ProGen2: Exploring the Boundaries of Protein Language Models.
45. Andreeva, A., D. Howorth, …, A. G. Murzin. 2014. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* 42:D310–D314.
46. Andreeva, A., E. Kulesha, …, A. G. Murzin. 2020. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* 48:D376–D382.
47. Park, H., P. Bradley, …, F. DiMaio. 2016. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theor. Comput.* 12:6201–6212.
48. Xiong, P., M. Wang, …, H. Liu. 2014. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nat. Commun.* 5:5330.
49. Bradley, P., K. M. S. Misura, and D. Baker. 2005. Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science.* 309:1868–1871.
50. Boyken, S. E., Z. Chen, …, D. Baker. 2016. De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity. *Science.* 352:680–687.
51. Dou, J., A. A. Vorobieva, …, D. Baker. 2018. De novo design of a fluorescence-activating β-barrel. *Nature.* 561:485–491.
52. Anishchenko, I., T. M. Chidyausiku, …, D. Baker. 2020. De Novo Protein Design by Deep Network Hallucination. Bioengineering.
53. Vorobieva, A. A., P. White, …, D. Baker. 2021. De novo design of transmembrane β barrels. *Science.* 371, eabc8182.
54. Anand, N., R. Eguchi, …, P.-S. Huang. 2022. Protein sequence design with a learned potential. *Nat. Commun.* 13:746.
55. Ni, B., D. L. Kaplan, and M. J. Buehler. 2023. Generative design of de novo proteins based on secondary-structure constraints using an attention-based diffusion model. *Chem.* 9:1828–1849, S2451929423001390.
56. Wang, J., H. Cao, …, Y. Qi. 2018. Computational Protein Design with Deep Learning Neural Networks. *Sci. Rep.* 8:6349.
57. Miao, J., M. L. Descoteaux, and Y.-S. Lin. 2021. Structure prediction of cyclic peptides by molecular dynamics + machine learning. *Chem. Sci.* 12:14927–14936.
58. Batra, R., T. D. Loeffler, …, S. K. R. S. Sankaranarayanan. 2022. Machine learning overcomes human bias in the discovery of self-assembling peptides. *Nat. Chem.* 14:1427–1435.
59. Bhardwaj, G., V. K. Mulligan, …, D. Baker. 2016. Accurate de novo design of hyperstable constrained peptides. *Nature.* 538:329–335.
60. Tucs, A., D. P. Tran, …, K. Tsuda. 2020. Generating Ampicillin-Level Antimicrobial Peptides with Activity-Aware Generative Adversarial Networks. *ACS Omega.* 5:22847–22851.
61. Zhang, H., K. M. Saravanan, …, J. Z. H. Zhang. 2023. Deep Learning-Based Bioactive Therapeutic Peptide Generation and Screening. *J. Chem. Inf. Model.* 63:835–845.
62. Hsu, C., R. Verkuil, …, A. Rives. 2022. Learning inverse folding from millions of predicted structures. *Proceedings of the 39th International Conference on Machine Learning.* 162:8946–8970.

63. Strokach, A., D. Becerra, …, P. M. Kim. 2020. Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Syst.* 11:402–411.e4.

64. Pandi, A., D. Adam, …, T. J. Erb. 2023. Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. *Nat. Commun.* 14:7197.

65. Das, P., T. Sercu, …, A. Mojsilovic. 2021. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* 5:613–623.

66. Larini, L., and J.-E. Shea. 2012. Role of β-Hairpin Formation in Aggregation: The Self-Assembly of the Amyloid-β(25–35) Peptide. *Biophys. J.* 103:576–586.

67. Naldi, M., J. Fiori, …, V. Andrisano. 2012. Amyloid β-Peptide 25–35 Self-Assembly and Its Inhibition: A Model Undecapeptide System to Gain Atomistic and Secondary Structure Details of the Alzheimer's Disease Process and Treatment. *ACS Chem. Neurosci.* 3:952–962.

68. Maity, S., M. Hashemi, and Y. L. Lyubchenko. 2017. Nano-assembly of amyloid β peptide: role of the hairpin fold. *Sci. Rep.* 7:2344.

69. Athanassiou, Z., R. L. A. Dias, …, J. A. Robinson. 2004. Structural Mimicry of Retroviral Tat Proteins by Constrained β-Hairpin Peptidomimetics: Ligands with High Affinity and Selectivity for Viral TAR RNA Regulatory Elements. *J. Am. Chem. Soc.* 126:6906–6913.

70. Butterfield, S. M., and M. L. Waters. 2003. A Designed β-Hairpin Peptide for Molecular Recognition of ATP in Water. *J. Am. Chem. Soc.* 125:9580–9581.

71. Huan, Y., Q. Kong, …, H. Yi. 2020. Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Front. Microbiol.* 11, 582779.

72. Hoyer, W., C. Grönwall, …, T. Härd. 2008. Stabilization of a β-hairpin in monomeric Alzheimer's amyloid-β peptide inhibits amyloid formation. *Proc. Natl. Acad. Sci. USA.* 105:5099–5104.

73. Di Natale, C., S. La Manna, …, D. Marasco. 2020. Engineered β-hairpin scaffolds from human prion protein regions: Structural and functional investigations of aggregates. *Bioorg. Chem.* 96, 103594.

74. Chen, P.-Y., B. G. Gopalacushina, …, P. A. Evans. 2001. The role of a β-bulge in the folding of the β-hairpin structure in ubiquitin. *Protein Sci.* 10:2063–2074.

75. Cochran, A. G., R. T. Tong, …, N. J. Skelton. 2001. A Minimal Peptide Scaffold for β-Turn Display: Optimizing a Strand Position in Disulfide-Cyclized β-Hairpins. *J. Am. Chem. Soc.* 123:625–632.

76. Robinson, J. A. 2008. β-Hairpin Peptidomimetics: Design, Structures and Biological Activities. *Acc. Chem. Res.* 41:1278–1288.

77. Mahalakshmi, R. 2019. Aromatic interactions in β-hairpin scaffold stability: A historical perspective. *Arch. Biochem. Biophys.* 661:39–49.

78. Batalha, I. L., I. Lychko, …, A. C. A. Roque. 2019. β-Hairpins as peptidomimetics of human phosphoprotein-binding domains. *Org. Biomol. Chem.* 17:3996–4004.

79. Pace, J. R., B. J. Lampkin, …, J. A. Kritzer. 2021. Stapled β-Hairpins Featuring 4-Mercaptoproline. *J. Am. Chem. Soc.* 143:15039–15044.

80. DuPai, C. D., B. W. Davies, and C. O. Wilke. 2021. A systematic analysis of the beta hairpin motif in the Protein Data Bank. *Protein Sci.* 30:613–623.

81. Adhikari, B., J. Hou, and J. Cheng. 2018. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics.* 34:1466–1472.

82. Wang, S., S. Sun, …, J. Xu. 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* 13, e1005324.

83. LeCun, Y., B. Boser, …, L. D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1:541–551.

84. He, K., X. Zhang, …, J. Sun. 2016. Proceedings of the IEEE conference on computer vision and pattern recognition. *Convolutional Pose Mach* 4724–4732.

85. Kawashima, S., and M. Kanehisa. 2000. AAindex: amino acid index database. *Nucleic Acids Res.* 28:374.

86. Cao, D.-S., Q.-S. Xu, and Y.-Z. Liang. 2013. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics.* 29:960–962.

87. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

88. Dosztányi, Z. 2018. Prediction of protein disorder based on IUPred. *Protein Sci.* 27:331–340.

89. Crooks, G. E., G. Hon, …, S. E. Brenner. 2004. WebLogo: A Sequence Logo Generator: Figure 1. *Genome Res.* 14:1188–1190.

90. Hou, J., B. Adhikari, and J. Cheng. 2018. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics.* 34:1295–1303.

91. Liu, Y., Y.-H. Zhu, …, D.-J. Yu. 2021. Why can deep convolutional neural networks improve protein fold recognition? A visual explanation by interpretation. *Briefings Bioinf.* 22, bbab001.

92. Kingma, D. P., and J. Ba. 2014. Adam: A method for stochastic optimization. Preprint at ArXiv. https://doi.org/10.48550/ArXiv14126980.

93. Srivastava, N., G. Hinton, …, R. Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15:1929–1958.

94. Verkuil, R., O. Kabeli, …, A. Rives. 2022. Language models generalize beyond natural proteins. Preprint at bioRxiv.

95. Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22:2577–2637.

96. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: Visual molecular dynamics. *J. Mol. Graph.* 14:33.

97. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

98. Phillips, J. C., R. Braun, …, K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26:1781–1802.

99. Case, D. A., T. E. Cheatham, …, R. J. Woods. 2005. The Amber biomolecular simulation programs. *J. Comput. Chem.* 26:1668–1688.

100. Maier, J. A., C. Martinez, …, C. Simmerling. 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theor. Comput.* 11:3696–3713.

101. Huang, J., S. Rauscher, …, A. D. MacKerell. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods.* 14:71–73.

102. Piana, S., K. Lindorff-Larsen, and D. E. Shaw. 2011. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* 100:L47–L49.

103. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

104. Hopkins, C. W., S. Le Grand, …, A. E. Roitberg. 2015. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theor. Comput.* 11:1864–1874.

105. Balusek, C., H. Hwang, …, J. C. Gumbart. 2019. Accelerating Membrane Simulations with Hydrogen Mass Repartitioning. *J. Chem. Theor. Comput.* 15:4673–4686.

106. Kamenik, A. S., P. H. Handle, …, K. R. Liedl. 2020. Polarizable and non-polarizable force fields: Protein folding, unfolding, and misfolding. *J. Chem. Phys.* 153, 185102.

107. Tian, C., K. Kasavajhala, …, C. Simmerling. 2020. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theor. Comput.* 16:528–552.

108. Micsonai, A., F. Wien, …, J. Kardos. 2018. BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res.* 46:W315–W322.

109. Bax, A. 1989. Two-dimensional NMR and protein structure. *Annu. Rev. Biochem.* 58:223–256.

110. Delaglio, F., S. Grzesiek, …, A. Bax. 1995. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR.* 6:277–293.

111. Lee, W., M. Tonelli, and J. L. Markley. 2015. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinforma. Oxf. Engl.* 31:1325–1327.

112. Shen, Y., and A. Bax. 2015. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol. Biol.* 1260:17–32.

113. Götz, S., J. M. García-Gómez, …, A. Conesa. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36:3420–3435.

114. Shen, Y., O. Lange, …, A. Bax. 2008. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. USA.* 105:4685–4690.

115. Kuznetsov, A., and C. J. Bollin. 2021. NCBI Genome Workbench: Desktop Software for Comparative Genomics, Visualization, and GenBank Data Submission. *In* Multiple Sequence Alignment: Methods and Protocols. K. Katoh, ed Springer US, New York, NY, pp. 261–295.

116. Pettersen, E. F., T. D. Goddard, …, T. E. Ferrin. 2021. UCSF ChimeraX : Structure visualization for researchers, educators, and developers. *Protein Sci.* 30:70–82.

117. Radford, A., L. Metz, and S. Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. Preprint at ArXiv. https://doi.org/10.48550/ArXiv151106434.

118. McDonald, E. F., T. Jones, …, A. Gulsevin. 2023. Benchmarking AlphaFold2 on peptide structure prediction. *Structure.* 31:111–119.e2.

119. Hie, B., S. Candido, …, A. Rives. 2022. A high-level programming language for generative protein design. *Pripritn at bioRxiv.*

120. Joosten, R. P., T. A. H. Te Beek, …, G. Vriend. 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Res.* 39:D411–D419.

121. Binda, C., A. Coda, …, A. Mattevi. 1998. Structure of the Mutant E92K of [2Fe–2S] Ferredoxin I from *Spinacia oleracea* at 1.7 Å Resolution. *Acta Crystallogr. D Biol. Crystallogr.* 54:1353–1358.

122. Heinig, M., and D. Frishman. 2004. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* 32:W500–W502.

123. Ansari, M., and A. D. White. 2023. Serverless Prediction of Peptide Properties with Recurrent Neural Networks. *J. Chem. Inf. Model.* 63:2546–2553.

124. Zhang, Y., and J. Skolnick. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins.* 57:702–710.

125. Xu, J., and Y. Zhang. 2010. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics.* 26:889–895.

126. Ciani, B., M. Jourdan, and M. S. Searle. 2003. Stabilization of β-Hairpin Peptides by Salt Bridges: Role of Preorganization in the Energetic Contribution of Weak Interactions. *J. Am. Chem. Soc.* 125:9038–9047.

127. Reid, K. A., C. M. Davis, …, J. T. Kindt. 2018. Binding, folding and insertion of a β-hairpin peptide at a lipid bilayer surface: Influence of electrostatics and lipid tail packing. *Biochim. Biophys. Acta Biomembr.* 1860:792–800.

128. D'Ursi, A. M., M. R. Armenante, …, D. Picone. 2004. Solution Structure of Amyloid β-Peptide (25−35) in Different Media. *J. Med. Chem.* 47:4231–4238.

129. Li, B., X. Ouyang, …, J. Ni. 2022. Novel β-Hairpin Antimicrobial Peptides Containing the β-Turn Sequence of -RRRF- Having High Cell Selectivity and Low Incidence of Drug Resistance. *J. Med. Chem.* 65:5625–5641.