JOINT GENERATION OF PROTEIN SEQUENCE AND STRUCTURE WITH ROSETTAFOLD SEQUENCE SPACE DIFFUSION

Sidney Lyayuga Lisanza[‡1,2,3], Jake Merle Gershon[‡2,4], Sam Tipps[‡1,2], Lucas Arnoldt[‡1,2,5], Samuel Hendel[‡1,2], Jeremiah Nelson Sims[6], Xinting Li[1,2], David Baker[*1,2,7]

‡Equal contribution
*To whom correspondence should be addressed

1. Department of Biochemistry, University of Washington, Seattle, WA 98105, USA
2. Institute for Protein Design, University of Washington, Seattle, WA 98105, USA
3. Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA
4. Department of Molecular Engineering, University of Washington, Seattle, WA 98105, USA
5. University of Heidelberg, Heidelberg, Germany
6. Molecular & Cellular Biology, Medical Scientist Training Program, University of Washington, Seattle, WA 98105, USA
7. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA

May 8, 2023

ABSTRACT

Protein denoising diffusion probabilistic models (DDPMs) show great promise in the *de novo* generation of protein backbones but are limited in their inability to guide generation of proteins with sequence specific attributes and functional properties. To overcome this limitation, we develop ProteinGenerator, a sequence space diffusion model based on RoseTTAfold that simultaneously generates protein sequences and structures. Beginning from random amino acid sequences, our model generates sequence and structure pairs by iterative denoising, guided by any desired sequence and structural protein attributes. To explore the versatility of this approach, we designed proteins enriched for specific amino acids, with internal sequence repeats, with masked bioactive peptides, with state dependent structures, and with key sequence features of specific protein families. ProteinGenerator readily generates sequence-structure pairs satisfying the input conditioning (sequence and/or structural) criteria, and experimental validation showed that the designs were monomeric by size exclusion chromatography (SEC), had the desired secondary structure content by circular dichroism (CD), and were thermostable up to 95°C. By enabling the simultaneous optimization of both sequence and structure, ProteinGenerator allows for the design of functional proteins with specific sequence and structural attributes, and paves the way for protein function optimization by active learning on sequence-activity datasets.

1 Main

Protein function arises from a complex interplay of sequence and structural features, hence designing new protein functions requires reasoning over both sequence and structure space. Many protein design methods sample structures and sequences in separate steps, typically by generating protein backbones first and using inverse folding methods to generate sequences. Traditional methods like Rosetta flexible backbone protein design[1] alternate between structure and sequence design, while recent deep learning based approaches typically generate backbones first and then use sequence design methods such as ProteinMPNN to identify sequences that fold into a given backbone[2–5]. Among the latter class of approaches, denoising diffusion probabilistic models[6] (DDPMs), which have shown considerable promise in continuous data domains allow for the generation of protein backbones subject to a wide range of structural constraints[7–9]. DDPMs approximate the probability density function over a data distribution by learning to denoise samples corrupted with Gaussian noise, enabling the generation of high-quality samples from a Gaussian prior; they have been explored less in categorical domains such as text and protein sequences[2,3,5,10]. Simultaneous generation of sequence and structure could have advantages over methods that alternate between optimization in the two domains

independently by enabling coordinated guidance with both sequence and structural features. Hallucination approaches that apply activation-maximization to structure prediction networks[11–13] can generate sequence-structure pairs without additional training, but these solutions can be adversarial, require a large number of steps to converge, and robust experimental success requires subsequent sequence design on the hallucinated backbones[3].

We reasoned that diffusion approaches could be powerful for simultaneous generation of sequence and structure while avoiding the adversarial solutions of activation maximization, and set out to develop a diffusion model which jointly generates sequence-structure pairs and can be guided by constraints in both domains. We hypothesized that RoseTTAfold's ability to simultaneously generate protein sequences and structures, as illustrated by RoseTTAFold Joint Inpainting[11], could be adapted for diffusive generation of coherent sequence-structure pairs by finetuning to recover noised native protein sequences while imposing a loss on structure prediction accuracy, and that such a DDPM could be readily guided by constraints in both domains.

1.1 DDPM Implementation

We chose to implement diffusion in sequence space by representing amino acid sequences as scaled one hot tensors where true values are set to 1 and all other values set to -1, allowing progressive corruption with Gaussian noise $N(\mu = 0, \sigma = 1)$ [14,15]. This approach is advantageous over other categorical diffusion methods, where diffusion occurs within a learned embedding space of text[16,17], because it simplifies the use of raw sequence based classifiers for guidance. To finetune RoseTTAFold we input the protein sequences progressively noised according to a square root schedule[16], the corresponding time step, and optional structural information. We task the model to generate ground truth sequence-structure pairs by applying a categorical cross entropy loss to the predicted sequence (relative to the ground truth sequence) and FAPE structure loss on the predicted structure. Self-conditioning[14], which allows the model to condition on its previous prediction, was employed to improve training and inference performance. Protein generation begins with an Lx20 dimensional sequence of Gaussian noise, and at each timestep ($\mathbf{x_t}$) the model predicts $\mathbf{x_0}$ from $\mathbf{x_t}$, after which $\mathbf{x_0}$ is noised to $\mathbf{x_{t-1}}$ (Figure 1A, top panel). Conditioning information (guidance) can be combined with $\mathbf{x_0}$ to guide the model towards a constrained sequence space using activity data, sequence specific potentials, secondary structure features, and more (Figure 1A, bottom)[18].

1.2 Unconditional generation

Starting with a sequence of Gaussian noise, the model generates sequence-structure pairs with amino acid compositions similar to those of native proteins (Figure 1B, left). The generated sequences and structures are internally consistent: AlphaFold2 and ESMFold predictions of the structures adopted by the generated sequences are very close to the generated structures (Figure 1C, S1B) and confident (Figure S1A). Sampling from different noise distributions resulted in different amino acid frequencies and secondary structure compositions in the generated outputs[19] (Figure S1A, S2, S3, S4). Samples of unconditionally generated designs with 100aa, 200aa or 300aa length can be found in Figures S5, S6 and S7. For the longer lengths the success rate of ProteinGenerator in generating sequences that fold to the designed structures is lower than that of the RoseTTAfold based structure diffusion method RFdiffusion[8] followed by ProteinMPNN[3]; this may reflect intrinsic differences between diffusion in sequence and structure space, or arise from differences in model training.

For experimental characterization, we unconditionally generated 70-80 residue proteins, filtered for high AF2 confidence (pLDDT > 90) and AF2 RMSD to design < 2Å (Table S4). A second subset with high ProteinGenerator confidence (model pLDDT > 90) and AF2 RMSD to design < 2Å, but low AF2 confidence (AF2 pLDDT < 80) was tested as well (Table S4). Synthetic genes encoding the designs were transformed into E. coli, and the proteins were expressed and purified using nickel-NTA chromatography. Of the 42 proteins tested, 34 were soluble and monomeric by size exclusion chromatography (SEC) and circular dichroism (CD) experiments showed they had the anticipated secondary structure and were stable up to 95°C (Figure 1D).

1.3 Conditioning on Single or Multi-state Structural Information

ProteinGenerator can be conditioned on either explicit 3D coordinates or on secondary structure as described by per-residue DSSP features. *In silico* tests show that when conditioned on 3D structural motif information the model generates proteins accurately recapitulating these motifs (Figure S8). During training coordinates for structural motifs are provided 40% of the time either as continuous spans of 4-9 residues or as 5-10 sparse residues, distant in sequence space. For lower resolution secondary structure guidance, DSSP[20] features were specified on a per-residue level 25% of the time and masked between 0% and 90% (randomly sampled). This allows guidance of a part or all of a structure towards a specific secondary structure type or fold, which the ProteinGenerator does quite well (Figure 1F, 3A).

Designing an amino acid sequence that can adopt distinct structural conformations upon an external trigger is a challenging task, as the energy landscape must contain two discrete minima with free energy differences small enough for a trigger to induce state switching[21]. We reasoned that ProteinGenerator was well equipped for this task because of its understanding of sequence-structure relationships and its ability to apply constraints in both domains.

We experimented with going beyond single-state structural specification by seeking to condition on distinct structural features of two different states. We applied multistate conditioning to design fold switching proteins by inputting the same sequence with two (or more) input sets of structural constraints at each step and averaging the output logits (Figure 1E). This allows the model to search sequence space for high-confidence solutions that satisfy all constraints. We used this approach to generate designs consisting of two fragments separated by a protease cleavage site which adopt different secondary structures following: in the intact parent state beta strand conditioning is used in the region flanking the cleavage site, while alpha helical conditioning is used for the two resulting subsequences. Logits from the parent and children sequences are averaged together at each step to arrive at a single sequence. As designed, AF2 structure predictions for the intact parent sequence have beta sheets in this region, whereas the two fragments (predicted independently) are entirely helical; the 3D structures of both the intact parent and the two children are very close to the design models (Figure 1F). A similar multistate approach can be applied to design monomers that adopt multiple oligomeric states and other conformationally switching systems.

1.4 Sequence guidance with amino-acid based potentials

An advantage of diffusion in sequence space is that sequence-based guiding functions can be readily implemented and applied. As a first test of this, we sought to design proteins with high frequencies of specific amino acids conferring structural or functional properties (cysteines can form disulfide bonds to make stable proteins, tryptophans possess spectroscopic properties, and histidines can confer pH sensitivity). Given a specification of the desired fraction of a given amino acid, at each denoising step positions are ranked based on the extent to which the output logits favor the amino acid, and the desired fraction are biased further in this direction (Figure 2A). We found this allowed more fine grained control in generating sequences than imposing a global bias towards a particular amino acid. We used this procedure to generate proteins with high frequencies (20%) of tryptophan, cysteine, valine, histidine, and methionine one at a time (Figure 2B, S9). We obtain compositionally biased protein sequences composed of nearly 20% of the desired amino acids that are strongly predicted to adopt the corresponding structures.

To evaluate the compositionally biased designs experimentally, we generated 70 to 80 residue proteins with different amino acids upweighted, filtered on AF2 pLDDT $>$ 90 and AF2 RMSD to design $<$ 2Å, and experimentally characterized the top 96 designs (Table S4). Of the characterized designs, SEC traces indicated the proteins were monomeric for 4/5 upweighted cysteine proteins, 8/19 upweighted tryptophan proteins, 19/22 upweighted valine proteins, 10/12 upweighted histidine proteins, and 10/10 upweighted methionine proteins. CD spectra were obtained for a subset of the monomeric designs, and in all cases indicated secondary structure was consistent with the designed structure (Figure 2D,E). Guiding for high cysteine content at the sequence level resulted in the formation of 3 to 5 disulfide bonds per protein without any structural conditioning as indicated by mass spectrometry in the presence and absence of the reducing agent TCEP at 50mM (Figure 2D, Table S2). Proteins designed with upweighted tryptophans exhibited high absorbance at 280 nm, and proteins with upweighted valine exhibited higher beta sheet content by CD (Figure 2D middle, right). These results indicate the model understands general sequence to structure relationships beyond the typical sequence space of native proteins (Figure 2C).

We next explored the generation of proteins with prespecified charge composition, isoelectric points and hydrophobicity which can influence solubility, activity, subcellular location[22], pharmacokinetic clearance, and retention[23]. Biasing away from hydrophobic amino acids can lead to better expression and solubility[24], and designing towards hydrophobic interfaces is advantageous for protein-protein interactions[25]. We implemented sequence based potentials to guide[18] the diffusive process towards these characteristics to enable fine-tuned control over physical properties of the output sequence. This approach enabled the design of proteins with a range of user-defined hydrophobicities (Figure 2F) and isoelectric points (Figure 2G).

1.5 Scaffolding bioactive sequences

The design of proteins with activities conditional on an outside input is of considerable general interest, and could enable generation of therapeutics with spatial and temporal control[26]. As a first exploration of the use of ProteinGenerator for such proteins, we sought to scaffold bioactive peptide sequences within an inert protein cage. Unlike our previous LOCKR[27,28] sensor system, in which the bioactive sequence must be in a helical conformation and make specific interactions with the caging scaffold, the generality of ProteinGenerator requires only that the sequence of the bioactive peptide be specified–neither the structure this adopts nor the structure of the overall cage need be decided on in advance.

ProteinGenerator is able to generate structures containing peptide sequences corresponding to known lytic peptides and the designed sequences are confidently predicted to adopt the designed structures (Figure 3B). We used this approach to scaffold a bioactive peptide in the terminus of a protein that can be conditionally released upon proteolytic cleavage of a terminal loop (Figure 3C). We specified the sequence, not the structure, of the bioactive segment, and used DSSP conditioning to force the cleavage motif to be in a loop. We chose to scaffold the pore forming peptide melittin[29] currently being explored as a cancer therapy[30]. Starting with the melittin sequence and a flanking cleavage site, we generated an additional 125 residues to scaffold the peptide into a globular protein. Melittin-scaffolded proteins generated by the model were in agreement with AlphaFold2 models (AF2 pLDDT > 85, AF2 RMSD < 2Å) (Figure 3C). We obtained synthetic genes encoding 12 proteins scaffolding melittin and found that 9/12 were monodisperse by SEC and had the correct secondary structure by CD (Figure 3C).

1.6 Generation of sequence repeat proteins

Repeat proteins containing tandem copies of a sequence-structure unit are ubiquitous in nature and play central roles in molecular recognition and signaling[31]. Previous work in designing repeat proteins has required extensive pre-specification of structural features[32]. We reasoned ProteinGenerator could be used to readily generate repeat proteins given only the sequence length of the repeat unit and number of repeats desired. At each timestep we symmetrize the noised sequence distribution accordingly (Figure 3D). Unconditional generation with this approach yielded largely beta solenoid structures which AF2 corroborated. We added helical caps to a subset of designs to promote stability and reduce aggregation[33,34]. To encourage further exploration of the repeat protein universe we specified the secondary structure for a small percent (2%-10%) of residues, which yielded a wide range of all alpha, all beta, and mixed alpha-beta designs (Figure 3E). We generated 165-185 residue repeat proteins, filtered them (AF2 pLDDT>85 and RMSD to design < 2), and experimentally characterized 74 repeat proteins with helical caps and 86 repeat proteins without helical caps. Of these, 27 repeats with caps and 10 repeats without helical caps were soluble and monomeric by SEC, and 7/8 proteins evaluated using circular dichroism had the expected secondary structure (Figure 3E, Table S3[35]).

1.7 Guidance with sequence only classifiers

Designing proteins with a desired biological activity is a long standing goal of *de novo* protein design. An advantage of our approach is that diffusion can be directly guided by function classifiers that operate in sequence space. We first sought to guide the network with the DeepGOPlus Gene Ontology (GO) classifier[36] to generate proteins with specific characteristics and functions. Although GO classification scores increased with guidance for nitrogen compound metabolic process (GO:0006807) and membrane (GO:0016020), we found the classifier had a high false positive rate often assigning high scores to native sequences outside the GO domain (Figure S10). In a separate approach, we trained a simple transformer encoder and single linear layer to discriminate unconditionally generated sequences from nanobody sequences and immunoglobulin (IG) folds aggregated from Integrated Nanobody Database for Immunoinformatics[37] and Structural Classification of Proteins database[38,39]. We generated 125 residue proteins, roughly the length of a nanobody, and found when classifier guidance or strand bias (1%) was used alone the classifier scores increased; when used in combination classifier scores increased (Figure 4A) along with the fraction of beta-strand containing proteins (Figure S11). 14% of designs made with the classifier alone were found to be beta sandwiches, which increased to 45% when applying a strand bias to 1% of the residues. Of the designs made with the combination of strand bias and classifier guidance 68.7% matched with tm-align > 0.5 to IG folds. AlphaFold models of sequences with high classifier scores matched the design models well (Figure 4B).

1.8 Guidance using protein family sequence information

Protein families often have specific residues important for function that are conserved throughout the family. Position-specific scoring matrices (PSSMs) capture this information and have been previously used to generate active enzymes with corresponding sequence composition using consensus sequence design[40], but these approaches do not consider sequence-structure coherence, while Rosetta PSSM guided flexible backbone structure based sequence design[41] can require expensive MCMC calculations. We sought to use ProteinGenerator to design new members of protein sequence families, conditioning on both family multiple sequence alignments and key structural features associated with function. We generated a PSSM for the GFP fluorescent protein family[42] (all sequences in uniprot having greater than 30% sequence identity to GFP)[43]. At each step in denoising, we used the PSSM to bias the sequence distribution towards that of the family, and the calculations were conditioned on the coordinates of the residues contacting the chromophore (this cannot be done using purely sequence based methods) (Figure 4C). To tune guidance the PSSM was scaled by a factor of 0.25, 0.5, 0.75, and 1.0, and we found that sequences clustered closer together became more similar to GFP family members as scaling increased (Figure 4D, E, S12). AlphaFold2 and ESM-Fold are unable to predict native GFP from a single sequence (Figure S13) but predict sequences generated by the model with high accuracy to the design

(Figure 4F). Active site coordinates provided as conditioning are in close agreement between the AF2 models and the native model, and demonstrate the model's ability to condition on both sequence and structural features (Figure 4F).

2 Discussion

By taking advantage of the ability of RoseTTAFold to jointly model protein sequences and structures, ProteinGenerator is able to directly sample in sequence space while ensuring that the 3D structure is coherent and satisfies any desired constraints. While RFdiffusion[8], Chroma[9], and other protein backbone diffusion models have demonstrated success in generating complex backbones with precise control over structural features, ProteinGenerator designs not only protein backbones but also sequences, enabling the design of proteins with any combination of desired sequence and structural attributes. We anticipate that our sequence space diffusion approach could also be employed with large language models that also generate structures, such as ESMfold[44]. A particularly attractive application is active learning-based protein design/engineering optimization: given a sequence-activity predictor, iterating between ProteinGenerator design of structurally coherent proteins predicted to have high activity, experimental testing, and updating the activity predictor could be a powerful path to achieving high activity.

3 Methods

3.1 Sequence representation

To apply the diffusion framework in sequence space, a continuous representation of the categorical sequence data is needed. To implement this we represented the sequence, $\mathbf{x_0}$, with dimensions Lx20 where L corresponds to the protein length with 20 possibilities for each amino acid type. This takes the form of a one-hot encoded vector that is centered at zero by multiplying the Lx20 tensor by 2 and subtracting 1. Each logit within the tensor is a real number, with higher values corresponding to a higher probability for that specific amino acid at that position. With this representation we noise $\mathbf{x_0}$ to obtain $\mathbf{x_t}$ with the below equation following Ho et al. formulation for a standard forward process sampling from Gaussian noise with mean at 0 and standard deviation of 1.

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{a}_t}x_0, (1 - \bar{a}_t)\mathrm{I}) \tag{1}$$

A critical part of the forward diffusion process is selecting the noising schedule. Determining the correct bin of a categorical distribution is trivial at low time steps by argmaxing the input sequence. Therefore more noise should be present at low timesteps to increase the difficulty of the task during training. The square root noise schedule[16] satisfies this requirement and was employed in this study.

3.2 Training

To train the model we began by sampling t uniformly from [0,T], where t=0 is an un-noised sequence and t=T is pure Gaussian noise. We then noise $\mathbf{x_0}$ to $\mathbf{x_t}$ with equation (1) and tasked the model to predict the un-noised sequence $\mathbf{x_0}$ and its corresponding structure $\mathbf{y}$. The timestep feature was added to the sequence template passed to the model. We applied a categorical cross entropy loss to $\mathbf{x_0}$ and structure losses to $\mathbf{y}$ (FAPE, bond angle, bond length, distogram, lddt). An additional KL loss[16] was applied to the calculated $\mathbf{x_{t\text{-}1}}$. Self conditioning[14] was implemented to allow the model to condition on the previous $\mathbf{x_0}$ prediction and the back calculated $\mathbf{x_{t\text{-}1}}$ during both training and inference. To self condition in practice the model was used with gradients turned off to first predict $\mathbf{x_0}$ from $\mathbf{x_{t+1}}$, which was then passed in as a sequence template to the model. During training, RoseTTAFold was allowed 1 to 3 uniformly sampled "recycle" steps to refine structure predictions via multiple passes through the model[45]. Pseudo training and inference code is available in the supplementary information (Pseudocode S1, S2). In later training iterations secondary structure conditioning was provided to the model by concatenating a tensor representing DSSP features onto the sequence template. These features were provided 25% of the time and masked uniformly between 0% and 90% when provided.

Along with the standard diffusion task (40% of the time), the model was also challenged with structure prediction (seq2str) and fixed backbone sequence design (30% of the time each). Incorporating these additional tasks during training helped maintain the agreement of sequence-structure pairs diffused by the model. Training examples were conditioned on sequence or structure by either unmasking 1 to 4 spans of residues, each 4 to 8 amino acids in length to simulate motif scaffolding, or unmasking randomly selected residues for the model to scaffold as an active site scaffolding problem. Unmasked structure conditioning information was supplied to the input for RoseTTAFold as templates in the 1D sequence track as well as the 2D and 3D structural information tracks.

3.3 Inference

During inference starting from $\mathbf{x_t}$ the model predicts $\mathbf{x_0}$ and simultaneously decodes it to $\mathbf{y}$. $\mathbf{x_0}$ is then back calculated to $\mathbf{x_{t-1}}$ with equation (1) and passed through the network with the previously predicted $\mathbf{x_0}$ to apply self conditioning. Benchmarking against conditioning on $\mathbf{x_t}$ as done in Ho et al[6] with equation (2), shows this approach performs better (SF 1C), as seen in other categorical diffusion methods[15,16].

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t \mathrm{I}) \tag{2}$$

$$\text{where } \tilde{\mu}_t(x_t|x_0) := \frac{\sqrt{\bar{a}_{t-1}}\beta_t}{1-\bar{a}_t} x_0 + \frac{\sqrt{a_t}(1-\bar{a}_{t-1})}{1-\bar{a}_t} x_t \text{ and } \tilde{\beta}_t := \frac{1-\bar{a}_{t-1}}{1-\bar{a}_t}\beta_t$$

This is done for T steps, but T can be varied, and does not have to be what was used during training. The model finds solutions to some problems in as little as 10 steps (Figure 1C). Furthermore, clamping the model's output logits from -3,3 gives better agreement with AF2 predictions (Fig S1B). $\mathbf{x_{t-1}}$ is sampled from either a zero-mean Normal distribution or a Non-Bayesian Gaussian Mixture distribution with equal mixing probabilities. For the Non-Bayesian Gaussian Mixture models we defined a mixture with two Normals centered at [-1, 1] (GMM2) and a mixture with three Normals centered at [-1, 0, 1] (GMM3).

3.4 DSSP guidance

For constructing the DSSP features we calculated each training example's DSSP based on the structure with helix, strand, loop, and masked labels[20]. During training, the calculated per-residue secondary structure features were appended to RoseTTAFold's t1d features, and were one-hot encoded for 25% or 50% of the time and masked for 30% or 80% of the time. During inference, DSSP features are appended to the t1D features as necessary and masked when not.

3.5 Classifier guidance

For classifier guidance we utilized the DeepGOPlus model[36] and trained a vanilla transformer model (2 Multihead Attention heads with each 2 layers, Embedding dimension: 64, Hidden Layer dimension: 64) on the INDI database for nanobodies[37]. Classifier guidance was implemented as described by Dhariwal and Nichol, 2021[18].

3.6 Multistate guidance

Parent and child protein pairings were generated in 25 steps using DSSP features "XH-HHHHHHHHHHHHLLLHHHHHHHHHHHHHHLLLEEEEEELLLEEEEEXXXXXXXEEEEEL-LLEEEEEELLLHHHHHHHHHHHHHHLLLHHHHHHHHHHHHHX" for parent, "XHHHHHHHHHHHHH-LLLHHHHHHHHHHHHHHLLLHHHHHHHHHHHHHHHHHHX" for child A, and "XHHHHHHHHHHHHH-LLLHHHHHHHHHHHHHHLLLHHHHHHHHHHHHHX" for child B. A mixing coefficient of 0.25 was used to combine parent and child sequences together at each step. Multistate design pseudocode implementation is available in the supplements (Pseudocode S3).

3.7 Single Sequence Prediction with AlphaFold2

All designs used in ESMFold benchmarks were modeled by using the curl command to predict single sequence structures. All designs used in AlphaFold2 (AF2) benchmarks and ordered for experimental characterization were predicted in single-sequence structure prediction mode with model 4. Pairwise backbone RMSDs between the design model and AF2 model were calculated for each design.

3.8 Sequence Identity calculations

Blast alignment was used to examine sequence alignment and similarity using query coverage >90% and target coverage >50%. Alignment to natives was done against Uniref-90.

3.9 Unconditional Protein Generation

Unconditionally generated proteins were assessed against a set of thousand native proteins with a length deviating up to five residues randomly sampled from the RCSB database. For experimental verification proteins ranging from 70-80 amino acids in length with no conditioning information were generated in 25 steps. Designs were filtered by AF2

pLDDT > 90 and AF2 RMSD to design < 2 Å for ordering final constructs. Additionally, proteins with high model confidence but moderate AF2 confidence were ordered by filtering on design pLDDT > 90, AF2 pLDDT < 80, and AF2 RMSD to design < 5 Å.

3.10 Compositionally Biased Protein Generation

Proteins ranging from 70-80 amino acids in length with an amino acid compositional potential were generated in 25 steps. Designs were filtered by AF2 pLDDT > 90, AF2 RMSD to design < 2 Å, and SAP score[46] < 30. The top 10 to 22 designs were ordered for each upweighted amino acid type (tryptophan, cysteine, valine, histidine, and methionine). Pseudocode for the implementation of the amino acid compositional potential is provided in the supplements (Pseudocode S4).

3.11 Charge Biased Protein Generation

Proteins of 50 amino acids in length with charge potentials applied were generated in 25 steps with charge conditioning information. The ground truth charge for each protein was calculated at pH 7.4 by using the Henderson-Hasselbach equation. Pseudocode for the implementation of the charge potential is provided in the supplements (Pseudocode S5A, S5B).

3.12 Hydrophobic Biased Protein Generation

Proteins of 50 amino acids in length with hydrophobic potentials applied were generated in 25 steps with hydrophobicity conditioning information. The ground truth hydropathy index for each design was calculated by summing the hydropathy index for each residue and dividing by the sequence length[47]. Pseudocode for the implementation of the hydrophobic potential is provided in the supplements (Pseudocode S6A, S6B).

3.13 PSSM guidance

We formulate guidance with a PSSM by simply adding a precalculated PSSM to the output $\mathbf{x_0}$ prediction. This can be further scaled with a scaling factor, λ, to either promote stronger agreement with the PSSM, or lower, to promote more diversity. PSSMs were calculated from MSAs generated by mmseqs2[48] with a 30-90% sequence identity cutoff to the query GFP sequence. Designs were filtered on AF2 pLDDT > 80 and AF2 RMSD to design < 2 Å.

3.14 Repeat protein generation

Repeat proteins ranging from 125-150 amino acids in length were generated in 50 steps with and without DSSP conditioning information. Designed proteins contained 5 repeat units using one of the one of the following DSSP strings, where X represents mask, E represents strand, and H represents helix:
"XXXXEEEEEXXXXXXXXXXXXXXXHHHHHXXXX",
"XXXXEEEEEXXXXXHHHHHXXXXXEEEEEXXXX",
"XXXXHHHHHXXXXXEEEEEXXXXXHHHHHXXXX",
"XXXXHHHHHXXLXXHHHHHXXLXXHHHHHXXXX",
"XXXXEEEEEXXHXXEEEEEXXHXXEEEEEXXXX",
"XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX",
"XXXXEXXXXXXXXXEXXXXXXXXXEXXXXXXXX".
Designs were filtered on AF2 pLDDT > 80 and AF2 RMSD to design < 2 Å.

3.15 Scaffolding Bioactive peptides

Proteins of 155 amino acids in length were generated in 25 steps with spans of helical DSSP conditioning to encourage the model to generate helical bundles and cleavage loops. Additional DSSP features were provided to scaffold the furin protease cleavage site "GRRKR". The sequence for melittin was provided without DSSP conditioning as N-terminal 26 amino acids "GIGAVLKVLTTGLPALISWIKRKRQQ". Designs were filtered by AF2 pLDDT > 85, AF2 RMSD to design < 2 Å, and SAP scores < 40.

3.16 Plasmid construction

Protein designs were cloned into plasmids as in Watson et al[8]. Briefly, designs were ordered as synthetic genes (eBlocks, Integrated DNA Technologies) with BsaI overhangs compatible with a ccdB-encoding expression vector

LM0627[13]. Genes cloned into LM0627 result in the following sequence: MSG-design-GSGSHHWGST**HHHHHH**, (SNAC cleavage tag and **6XHis affinity tag** are indicated). We used the NEBridge® Golden Gate Assembly Kit (New England Biolabs) with a total reaction volume of 5 µL and a ratio of 1:2 by mass of LM0627 plasmid DNA to design. We then incubated the reaction mixture at 37 C for 30 minutes, halted the reaction by incubating the reaction mixture at 60 C for 5 minutes, and transformed 1 µL of the reaction mixture into 6 µL of BL21 competent cells (New England Biolabs). After heat shock and recovery in SOC media, transformed BL21 cells were grown overnight in 1.0 mL of LB from which glycerol stock were created and small-scale expression cultures were inoculated.

3.17 1 mL-scale protein purification

Initially, proteins were expressed with small-scale expression screens as previously reported[13] with small adaptations. Briefly, designs were inoculated with 100 uL of overnight growths and 900 uL of auto-induction media (sterile-filtered TBII media supplemented with 50 µg/mL kanamycin, 2 mM MgSO4, 1X 5052) in deep-well 96-well plates. 16 hours post-inoculation, cells were harvested and lysed in lysis buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM imidazole supplemented with 1X BugBuster, 1 mM PMSF, 0.1 mg/mL lysozyme, 0.1 mg/mL DNase). Clarified lysates were added to a 50 µL bed of Ni-NTA agarose resin in a 96-well fritted plate equilibrated with wash buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM Imidazole). After sample application and flow through, the resin was washed three times with wash buffer, and samples were eluted in 200 µL of elution buffer (50 mM Tris-HCl (pH 8), 0.3 M NaCl, 0.5 M imidazole, 5 mM EDTA (pH 8)). All eluates were sterile filtered with a 96-well 0.22µm filter plate (Agilent 203940-100) prior to size exclusion chromatography (SEC). Protein designs were then screened via SEC using an AKTA FPLC outfitted with an autosampler capable of running samples from a 96-well source plate. Samples were run on a SuperdexS75 Increase 5/150 GL column (Cytiva 29148722; 3,000 to 70,000 Da separation range) in a running buffer (20 mM Tris pH 8, 150 mM NaCl). To improve peak resolution, the SEC column was connected directly in line from the autosampler to the UV detector. 0.25 mL fractions were collected from each run. Absorption spectra were collected by the AKTA U9-M at 230 nm and 280 nm.

3.18 50 mL-scale protein purification

Proteins selected for further downstream characterization were expressed in 50 mL of auto-induction media. 16 hours post-inoculation, cells were harvested and lysed in lysis buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM imidazole, 1 mM PMSF, 0.1 mg/mL lysozyme, 0.1 mg/mL DNase) through sonication. Clarified lysates were added to a 2 mL bed of Ni-NTA agarose resin in a 20 mL column (Bio-Rad 7321010) equilibrated with wash buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM Imidazole). After sample application and flow through, the resin was washed 3 times with 10 mL wash buffer, and samples were eluted in 2 mL elution buffer (50mM Tris-HCl (pH 8), 0.5M NaCl, 200mM Imidazole). All eluates were sterile filtered with a 3 mL 0.22uM filter plate prior to SEC. Protein designs were then screened via SEC using an AKTA FPLC outfitted with an autosampler capable of running samples from a 96-well source plate. Samples were run on a SuperdexS75 Increase 10/300 GL column (Cytiva 29148721; 3,000 to 70,000 Da separation range) in a running buffer (20 mM Tris pH 8, 150 mM NaCl). 1 mL fractions were collected from each run. Absorption spectra were collected by the AKTA U9-M at 230 nm and 280 nm.

3.19 Cysteine bias protein expression

Proteins guided towards high cysteine content were transformed into and expressed in Rosetta-gami B(DE3) Competent Cells (Novagen 71137). The 1 mL and 50 mL scale protein purification protocols were otherwise followed.

3.20 Circular Dichroism

Circular dichroism (CD) spectra were collected on a Jasco J-1500 CD Spectrometer with 1 nm bandwidth, 50 nm permanent scan rate, and data integration time of 4 seconds per read. Sample cuvettes stored in 2% Hellmanex (Hellma 9-307-011-4-507) were washed with deionized water, 2% Hellmanex, deionized water, then 20% ethanol, after which 300 uL SEC-purified protein was added for CD spectra measurements. Thermal melts were performed at 25°C and 95°C.

3.21 Mass Spectrometry

To identify the molecular mass of each protein, intact mass spectra were obtained via reverse-phase LC/MS on an Agilent G6230B TOF on an AdvanceBio RP-Desalting column, and subsequently deconvoluted by way of Bioconfirm using a total entropy algorithm. Disulfide formation was determined by injecting protein at 1.5 mg/mL in the presence and absence of 50 mM TCEP-HCl (Millipore Sigma 646547-10X1ML) and detecting the mass shift.

Acknowledgements

We would like to acknowledge Lisa Li, Saana Mansoor, Dimitry Zorine, Ian Humphreys, Harley Pyles, Brian Trippe, DéJenaé Ray, Abbas Idris, Xiaochuang Han, Meerit Said, Florence Dou, Linna Ann, Kejia Wu, Derrick Hicks, Hao Nguyen, Elias Kinfu, Adam Chazin-Gray, Quoc Tran, Marlo Zorman, Namrata Anand, and Naveen Jasti for helpful discussions and support. Chris Norn for PSSM scripts. Sergey Ovchinnikov for DSSP scripts. David Chmielewski for help with experimental procedures. Nate Ennist for help with CD. Doug Tischer for developing "contig" class for processing user inputs when running inference. Ivan Anishchenko for scripts to run TM - align, sequence similarity, and multidimensional scaling plots. Jue Wang and Justas Dauparas for benchmarking scripts. Minkyung Baek and Frank DiMaio for training scripts and RoseTTAFold code base. Joe Watson, David Juergens, and Nate Bennett for helpful scripts and conversations. Ian Haydon, Lance Stewart, Luki Goldschimdt, Adam Sadowski, Kandise Van Wormer, and Lauren Carter for general operations.

This work was supported by the Defense Threat Reduction Agency Grant HDTRA1-19-1-0003 (X.L.), by funding from the DARPA program Harnessing Enzymatic Activity for Lifesaving Remedies (HEALR) under award HR0011-21-2-0012 (X.L.), the Juvenile Diabetes Research Foundation International (JDRF) grant # 2-SRA-2018-605-Q-R (X.L.), AMGEN (S.L.), the Helmsley Charitable Trust Type 1 Diabetes (T1D) Program Grant # 2019PG-T1D026 (X.L.), the Bill and Melinda Gates Foundation Grant #OPP1156262 (X.L.), the Audacious Project at the Institute for Protein Design (J.S.), the Howard Hughes Medical Institute (J.S.).

Code Availability

The code for this projects is available here (with the exception of training scripts): `https://github.com/RosettaCommons/protein_generator`. For greater accessibility, thank you to Simon Dürr and HuggingFace who supplied a GPU grant to run the model interactively in your browser: `https://huggingface.co/spaces/merle/PROTEIN_GENERATOR`.

Bibliography

1. Huang, P.-S. et al. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. PLOS ONE 6, e24109 (2011). 2. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. (2022). 3. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. Science 378, 49–56 (2022). 4. Hsu, C. et al. Learning inverse folding from millions of predicted structures. 2022.04.10.487779 Preprint at https://doi.org/10.1101/2022.04.10.487779 (2022). 5. Anand, N. et al. Protein sequence design with a learned potential. Nat. Commun. 13, 746 (2022). 6. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. Preprint at http://arxiv.org/abs/2006.11239 (2020).
7. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. arXiv.org https://arxiv.org/abs/2205.15019v1 (2022).
8. Watson, J. L. et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. 2022.12.09.519842 Preprint at https://doi.org/10.1101/2022.12.09.519842 (2022).
9. Ingraham, J. et al. Illuminating protein space with a programmable generative model. 2022.12.01.518682 Preprint at https://doi.org/10.1101/2022.12.01.518682 (2022).
10. Brown, T. B. et al. Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).
11. Wang, J. et al. Scaffolding protein functional sites using deep learning. Science 377, 387–394 (2022).
12. Frank, C. et al. Efficient and scalable de novo protein design using a relaxed sequence space. 2023.02.24.529906 Preprint at https://doi.org/10.1101/2023.02.24.529906 (2023).
13. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. Science 378, 56–61 (2022).
14. Chen, T., Zhang, R. & Hinton, G. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. Preprint at http://arxiv.org/abs/2208.04202 (2022).
15. Han, X., Kumar, S. & Tsvetkov, Y. SSD-LM: Semi-autoregressive Simplex-based Diffusion Language Model for Text Generation and Modular Control. Preprint at http://arxiv.org/abs/2210.17432 (2022).
16. Li, X. L., Thickstun, J., Gulrajani, I., Liang, P. & Hashimoto, T. B. Diffusion-LM Improves Controllable Text Generation. Preprint at https://doi.org/10.48550/arXiv.2205.14217 (2022).
17. Dieleman, S. et al. Continuous diffusion for categorical data. Preprint at http://arxiv.org/abs/2211.15089 (2022).
18. Dhariwal, P. & Nichol, A. Diffusion Models Beat GANs on Image Synthesis. arXiv.org https://arxiv.org/abs/2105.05233v4 (2021).
19. Nachmani, E., Roman, R. S. & Wolf, L. Non Gaussian Denoising Diffusion Models. Preprint at http://arxiv.org/abs/2106.07582 (2021).
20. Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J. P. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput. Appl. Biosci. CABIOS 13, 291–295 (1997).
21. Wei, K. Y. et al. Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. Proc. Natl. Acad. Sci. 117, 7208–7215 (2020).
22. Tokmakov, A. A., Kurotani, A. & Sato, K.-I. Protein pI and Intracellular Localization. Front. Mol. Biosci. 8, 775736 (2021).
23. Boswell, C. A. et al. Effects of Charge on Antibody Tissue Distribution and Pharmacokinetics. Bioconjug. Chem. 21, 2153–2163 (2010).
24. March, D., Bianco, V. & Franzese, G. Protein Unfolding and Aggregation near a Hydrophobic Interface. Polymers 13, 156 (2021).
25. Rego, N. B., Xi, E. & Patel, A. J. Identifying hydrophobic protein patches to inform protein interaction interfaces. Proc. Natl. Acad. Sci. 118, e2018234118 (2021).
26. Zeng, Z. et al. Customized Reversible Stapling for Selective Delivery of Bioactive Peptides. J. Am. Chem. Soc. 144, 23614–23621 (2022).
27. Lajoie, M. J. et al. Designed protein logic to target cells with precise combinations of surface antigens. Science 369, 1637–1643 (2020).
28. Quijano-Rubio, A. et al. De novo design of modular and tunable protein biosensors. Nature 591, 482–487 (2021).
29. Lee, M.-T., Sun, T.-L., Hung, W.-C. & Huang, H. W. Process of inducing pores in membranes by melittin. Proc. Natl. Acad. Sci. 110, 14243–14248 (2013).
30. Duffy, C. et al. Honeybee venom and melittin suppress growth factor receptor activation in HER2-enriched and triple-negative breast cancer. Npj Precis. Oncol. 4, 1–16 (2020).
31. Parmeggiani, F. & Huang, P.-S. Designing repeat proteins: a modular approach to protein design. Curr. Opin. Struct. Biol. 45, 116–123 (2017).
32. Brunette, T. J. et al. Exploring the repeat protein universe through computational protein design. Nature 528, 580–584 (2015).
33. Peralta, M. D. R. et al. Engineering Amyloid Fibrils from β-Solenoid Proteins for Biomaterials Applications. ACS Nano 9, 449–463 (2015).

34. MacDonald, J. T. et al. Synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension. Proc. Natl. Acad. Sci. 113, 10346–10351 (2016).
35. Micsonai, A. et al. BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. Nucleic Acids Res. 50, W90–W98 (2022).
36. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36, 422–429 (2020).
37. Deszyński, P. et al. INDI—integrated nanobody database for immunoinformatics. Nucleic Acids Res. 50, D1273–D1281 (2022).
38. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A. G. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. 42, D310–D314 (2014).
39. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 48, D376–D382 (2020).
40. Sternke, M., Tripp, K. W. & Barrick, D. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. Proc. Natl. Acad. Sci. 116, 11275–11284 (2019).
41. Khersonsky, O. et al. Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. Mol. Cell 72, 178-186.e5 (2018).
42. Matz, M. V. et al. Fluorescent proteins from nonbioluminescent Anthozoa species. Nat. Biotechnol. 17, 969–973 (1999).
43. Ormö, M. et al. Crystal Structure of the Aequorea victoria Green Fluorescent Protein. Science 273, 1392–1395 (1996).
44. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).
45. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876 (2021).
46. Voynov, V., Chennamsetty, N., Kayser, V., Helk, B. & Trout, B. L. Predictive tools for stabilization of therapeutic proteins. mAbs 1, 580–582 (2009).
47. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132 (1982).
48. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028 (2017).

Figures

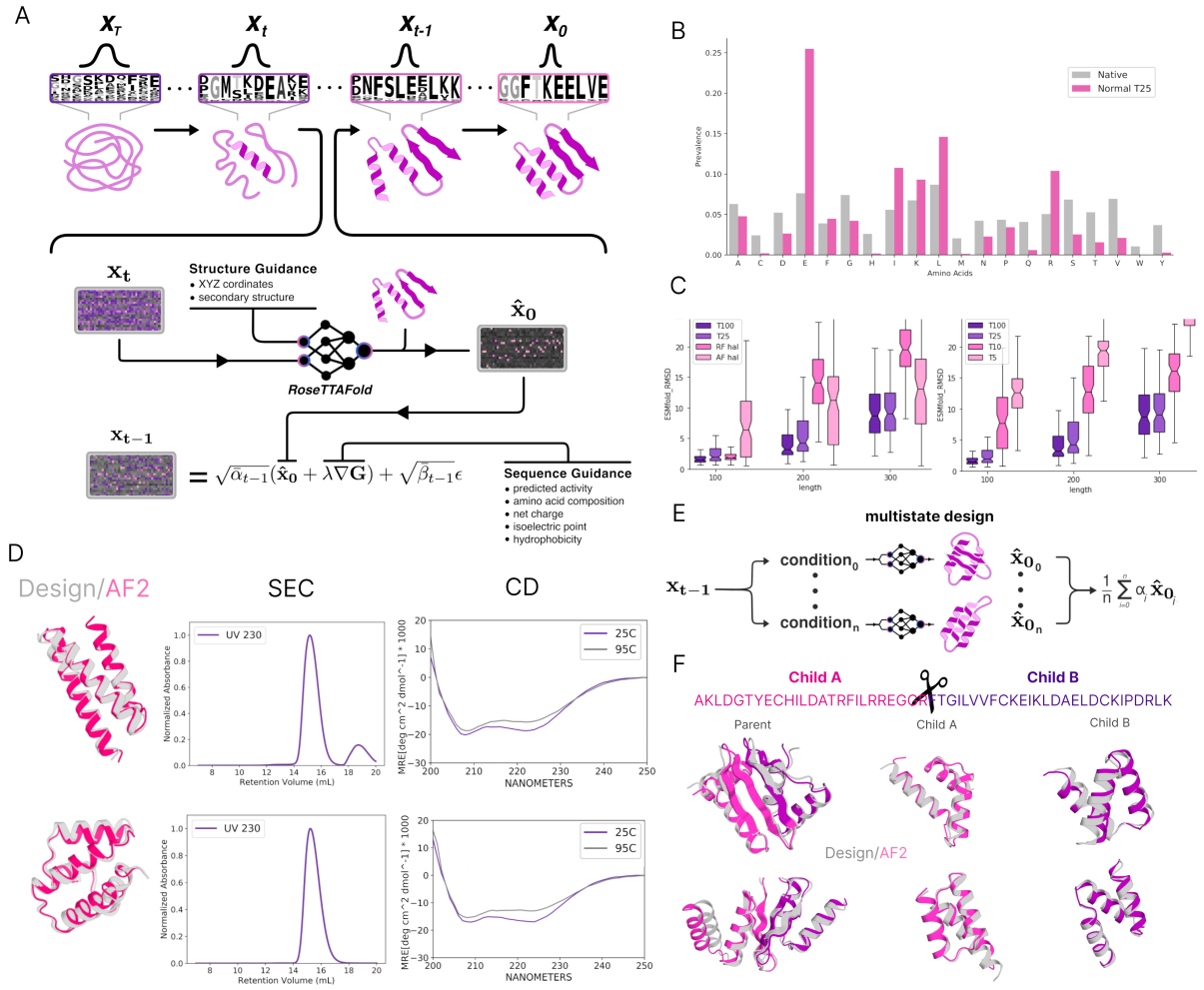

Figure 1: Overview of ProteinGenerator. (A) Inference schematic indicating how a noised sequence $\mathbf{x_t}$ is passed through the model with structural conditioning and updated to $\mathbf{x_{t-1}}$ for the next pass. At each step in the diffusion process $\mathbf{x_0}$ is predicted from $\mathbf{x_t}$ and guidance can be added to the predicted $\mathbf{x_0}$ prior to scaling with noise. This process is repeated for T steps as the sequence-structure pair converges on a high confidence solution. (B) Sampling from a Gaussian distribution yields sequences that approach that of native sequences. (C) Unconditional designs have higher ESMfold pLDDT and lower ESMfold RMSD to design compared other joint models: RosettaFold (RF) and Alphafold2 (AF) hallucination. (D) Experimental validation of unconditional designs: Design (grey) and AlphaFold2 (pink) models of unconditionally generated proteins. Size exclusion chromatography and circular dichroism experiments show these designs are soluble, monodispersed, and thermostable to 95°C. (E) Multistate guidance allows for the design of a single sequence with a variety of structural conditioning states to converge on a single sequence predicted to adopt multiple states. (F) Use of multistate guidance to generate sequences which upon fragmentation switch from alpha/beta to all alpha secondary structure. Parent design (left) switches secondary structure when split into two child proteins (right). Designed structures of parent and children are shown in grey and AlphaFold2 predictions are overlaid in pink/purple.

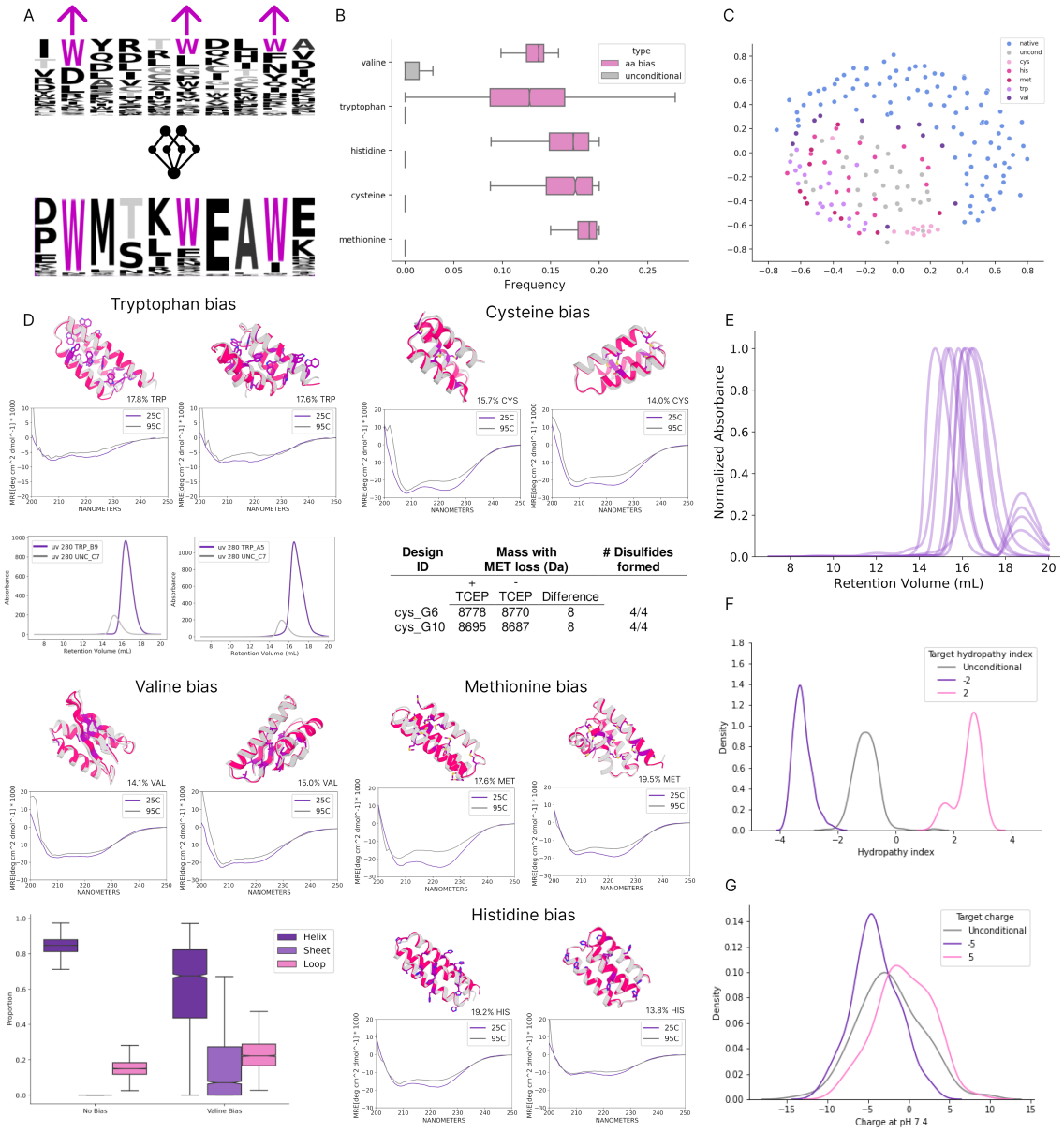

Figure 2: Generation of folded proteins with target sequence compositions. (A) Amino acid compositional bias: Applying an amino acid compositional bias generates proteins with desired sequence composition (tryptophan bias shown as an example). (B) Generation of compositionally biased proteins using ProteinGenerator. Bar plot displaying the average amino acid composition of designed proteins when guiding for 20% amino acid of interest compared to unconditional generation. (C) Sequence diversity representation: Multidimensional scaling of compositionally biased protein sequences occupy distinct spaces. (D) Experimentally validated amino acid compositionally biased proteins: Proteins generated with upweighted tryptophan, cysteine, valine, methionine, and histidine are predicted with high confidence by AlphaFold2 (pink), match the design model (grey), and are thermostable up to 95°C by circular dichroism. Proteins designed with high tryptophan bias show five-fold higher absorbance at 280nm compared to unconditionally generated proteins. Experimental validation of cysteine-rich proteins under reducing and non-reducing conditions confirms the presence of designed disulfide bonds by mass spectrometry. Proteins designed with high valine bias show increased beta strand propensity compared to unconditionally generated designs. (E) Size exclusion chromatography overlay: Proteins designed with sequence potentials are soluble and monodisperse by size exclusion chromatography. (F) Hydrophobic sequence potential: Biasing the sequence away or toward hydrophobic amino acids results in a shift in the distribution of hydrophobicity scores for the output sequences. (G) Net charge sequence potential: Resulting distribution of net charges when guidance is used to bias sequences toward +/- 5 net charge.

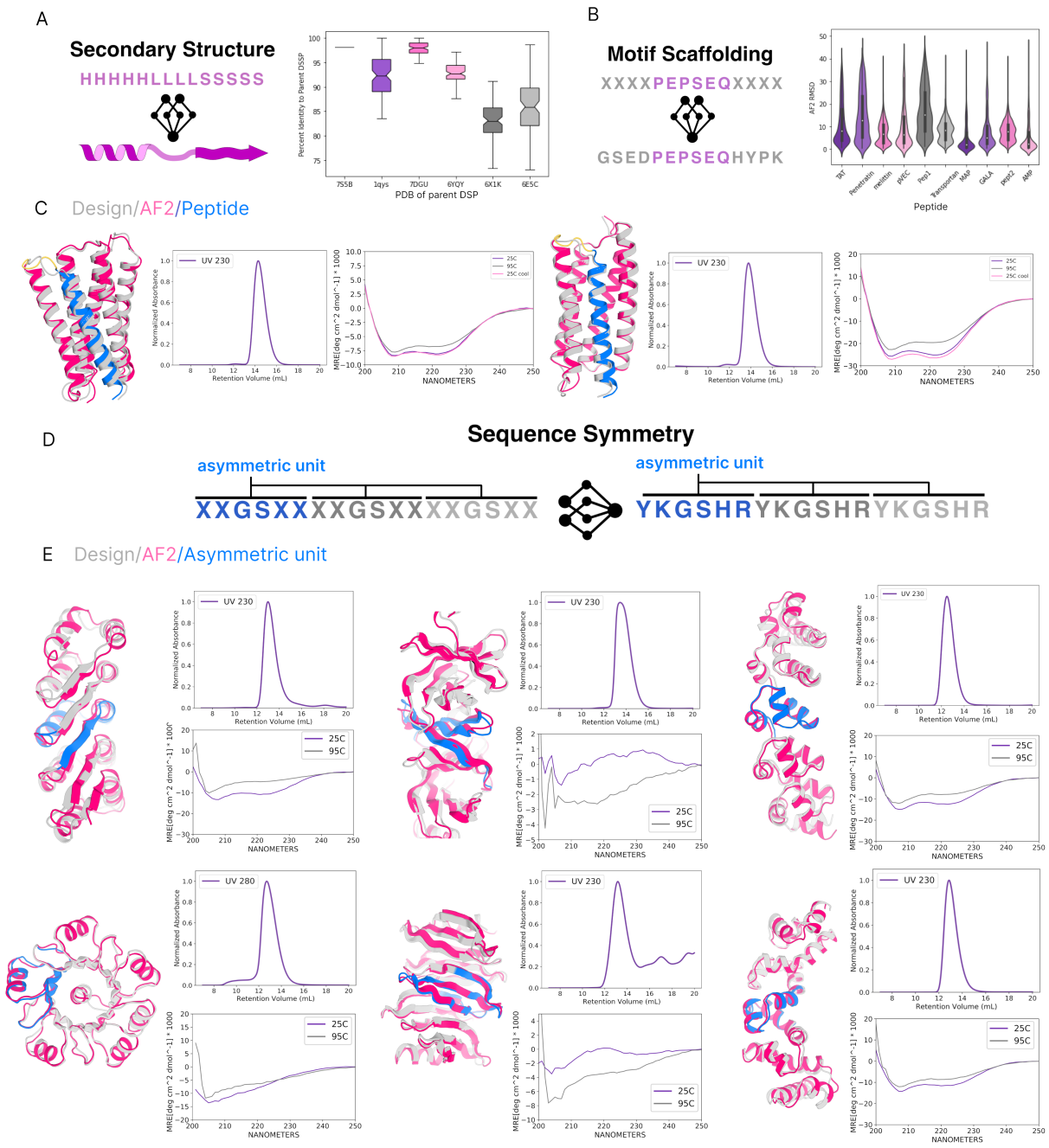

Figure 3: Scaffolding bioactive peptides and designing repeat proteins with ProteinGenerator. (A) Secondary structure conditioning: Protein generation with secondary structure conditioning recapitulates the DSSP of the target protein. (B) Unstructured (sequence only) motif scaffolding: Sequence motif scaffolding of unstructured bioactive peptides yields designs with low AlphaFold2 RMSD to design. (C) Melittin scaffold designs: Scaffolding melittin yields designs (grey) that are corroborated by AlphaFold2 (pink), are soluble and monodispersed by size exclusion chromatography, and thermostable to 95°C by circular dichroism. (D) Sequence symmetry: Sequence repeat symmetry is applied at inference time by symmetrizing update (xt-1) sequences to generate tandem repeat proteins. (E) Experimentally characterized symmetric repeat proteins: Designed repeat proteins (grey) with secondary structure conditioning are corroborated by AlphaFold2 (pink), are soluble and monodispersed by size exclusion chromatography, and are thermostable up to 95°C by circular dichroism.

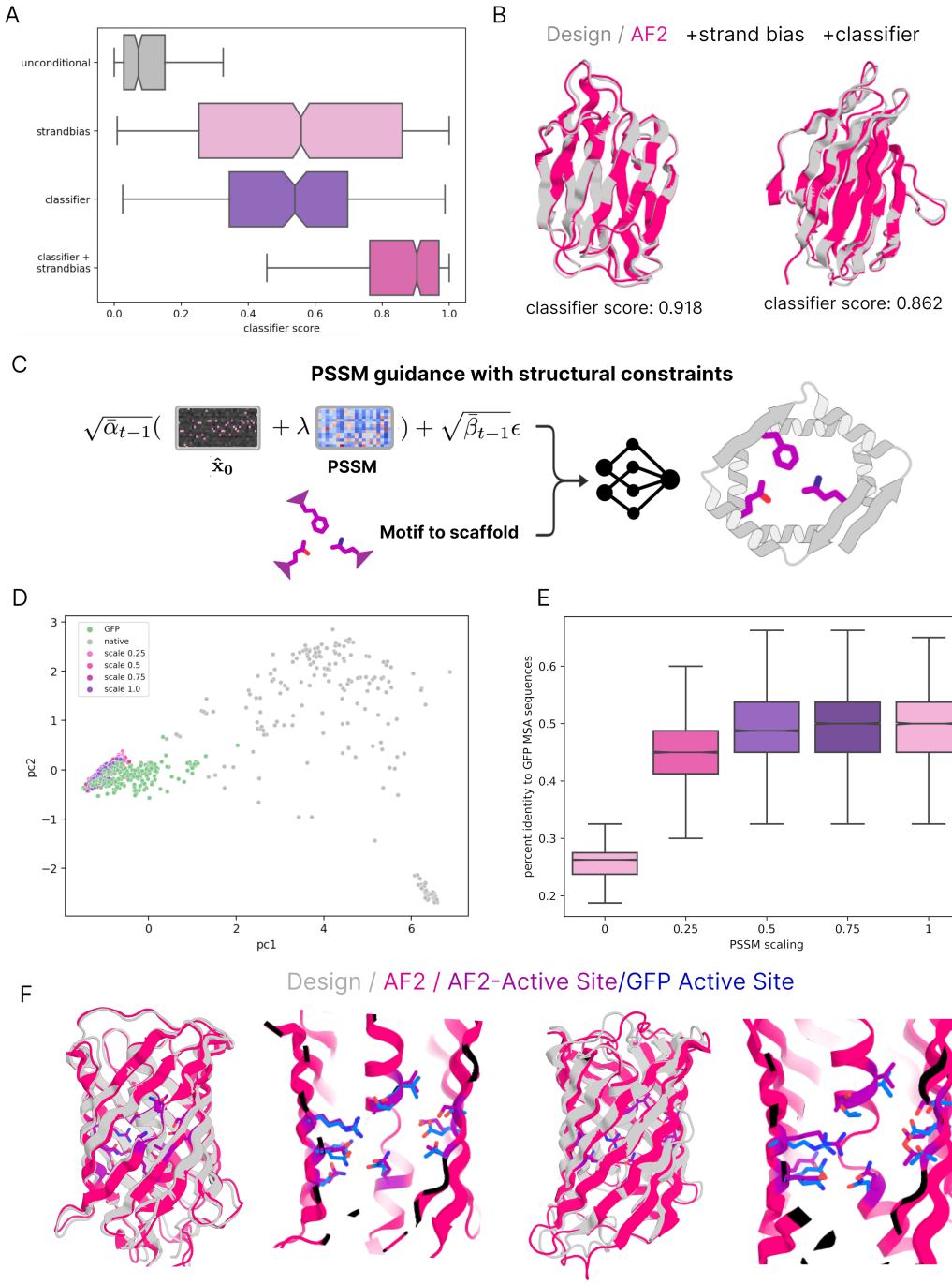

Figure 4: Fold and function guided protein generation. (A) immunoglobulin (IG) sequence based classifier score distributions from outputs generated unconditionally, with strand DSSP conditioning, with gradients from IG classifier, and with combination of IG classifier gradients and secondary structure conditioning. (B) Design model (grey) and AF2 model (pink) of proteins generated with classifier guidance and strand DSSP conditioning. (C) Schematic of PSSM guidance with scaling and motif scaffolding. (D) Guidance by position specific score matrix (PSSM) generates sequence-structure pairs sampled from a similar embedding space as native GFP designs. Designs and natives were embedded with ESM and first and second principle components derived from embeddings are plotted. (E) Sequence identity to native GFPs increase with increasing guidance scaling: Sequence similarity of PSSM designs at varying guide scales to native GFPs. (F) PSSM guided design examples: Design models (grey) and AF2 models (pink) of proteins generated with a PSSM guidance scale of 0 (left) or 8 (right) and 30% DSSP masking. Active site residues on AF2 predicted structures (purple) over-layed with wildtype GFP side chains in blue.

Supplements

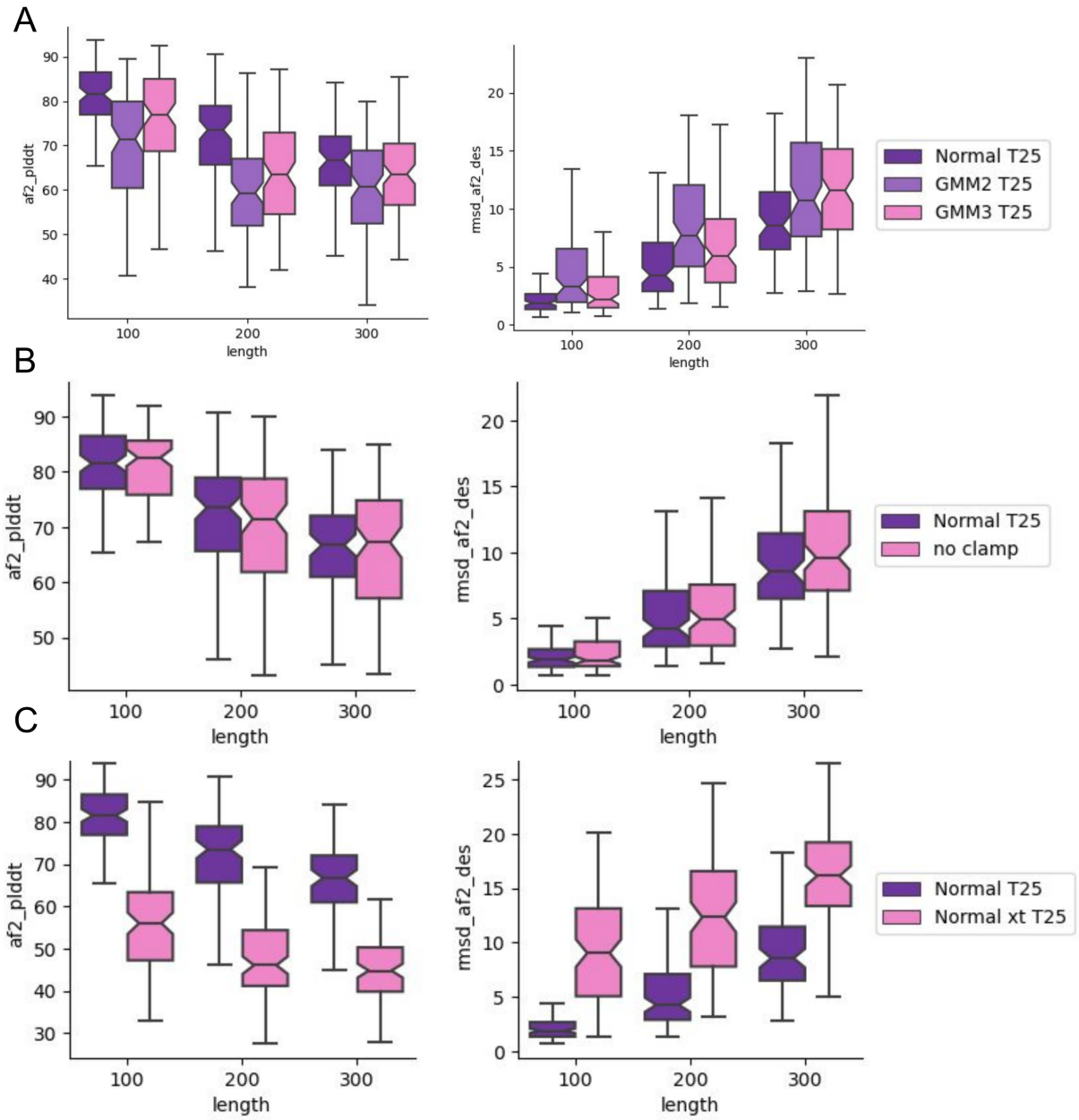

Figure S1: Inference benchmarks. (A) Boxplot of AF2 pLDDT of sequences from model clustered by length on left, right RMSD of AF2 model to design. (B) Boxplot AF2 pLDDT with clamp (-3,3) applied post sampling $\mathbf{x_{t-1}}$ and no clamp on left, AF2 RMSD to design. (C) Comparison of with and without conditioning on $\mathbf{x_t}$ when sampling $\mathbf{x_{t-1}}$, AF2 pLDDT right, AF2 RMSD to design left.

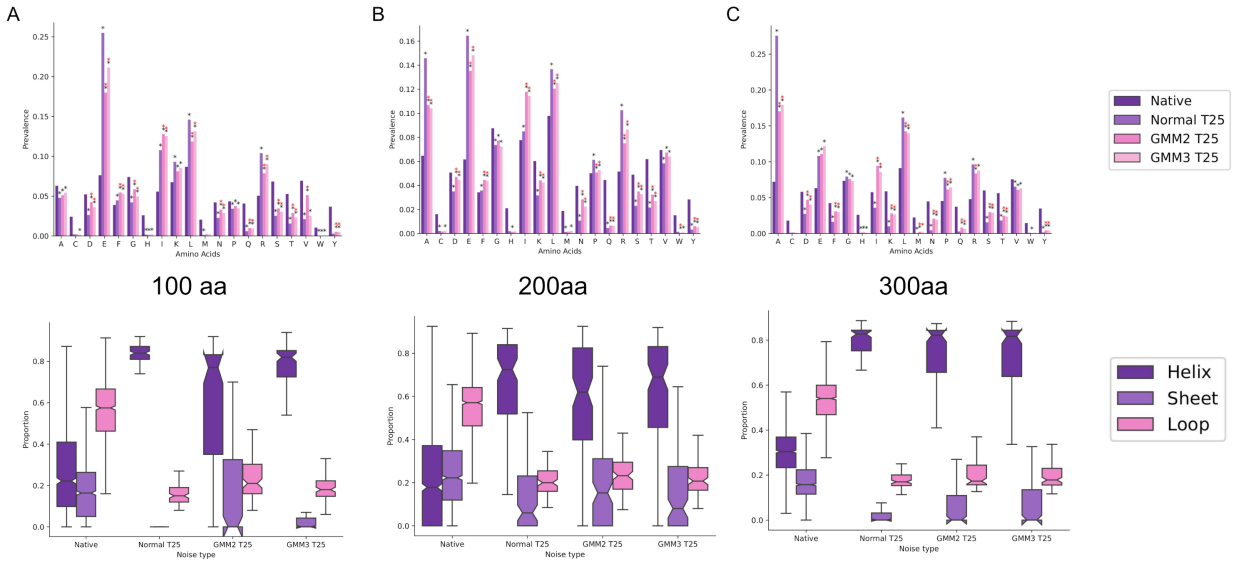

Figure S2: Amino acid distributions and secondary structure propensities for (A) 100AA, (B) 200AA, and (C) 300AA length proteins when sampling from normally distributed noise, GMM2, or GMM3. Significant amino prevalence changes between native and unconditional designs as well as between unconditional designs sampled from normal noise and sampled from GMM2 or GMM3 noise are displayed via black asterisk and a red asterisk.

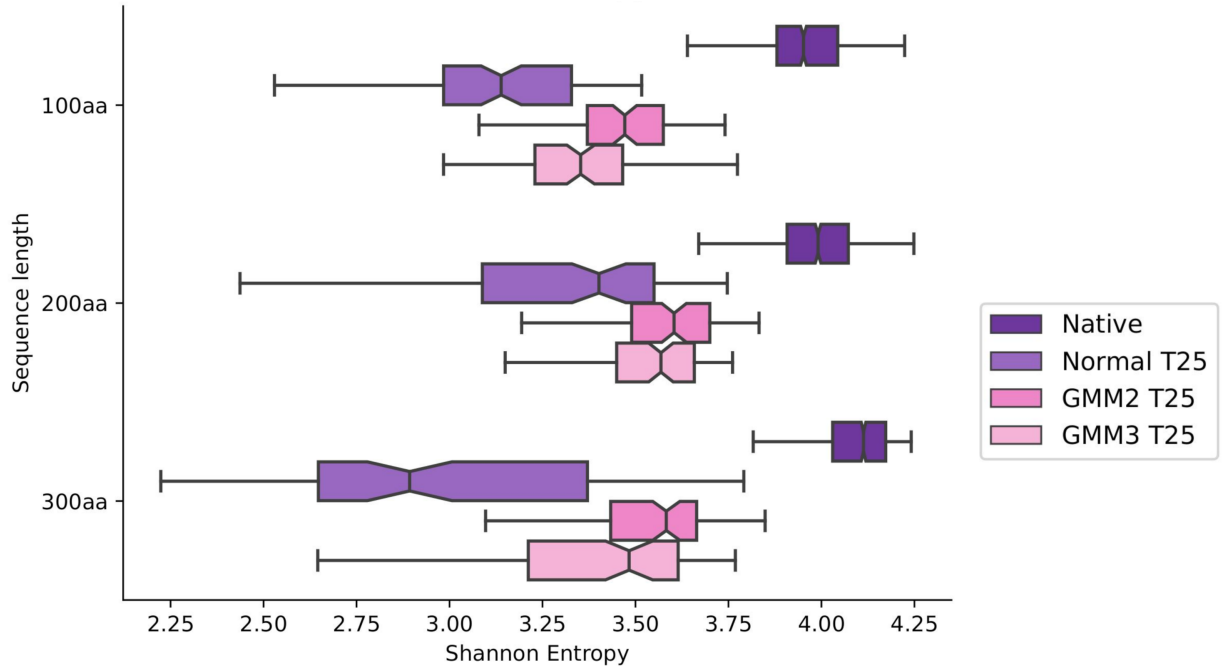

Figure S3: Sequence entropy of native proteins, normally sampled sequences, GMM2, and GMM3 for 100AA, 200AA, and 300AA proteins.

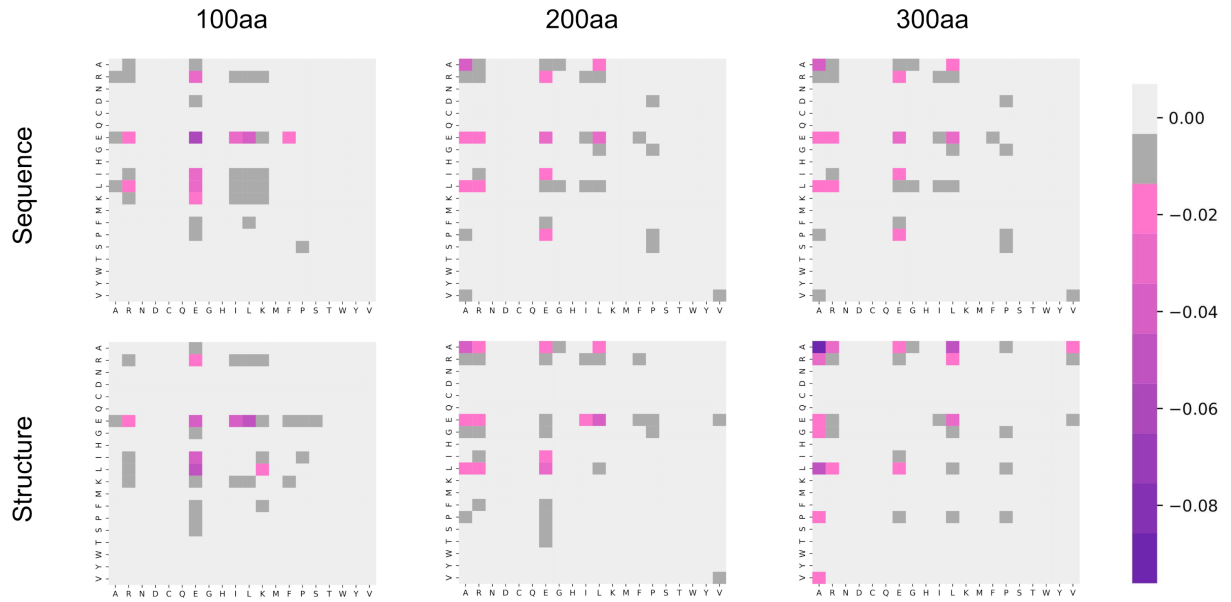

Figure S4: Sampling from different noise distributions generates proteins with different sequence and structural neighbors. Frequencies of amino acid neighbors in generated sequences (top row) and nearest structure neighbors (bottom row). Values are calculated as the difference between native sequences and unconditionally generated sequences for 100AA (left), 200AA (middle) and 300AA (right). Unconditional designs of 200AA and 300AA are characterized by more frequent alanine-leucine, alanine-glutamic acid, and alanine-alanine sequence and structure contacts. In structure space unconditional proteins exhibit lower frequencies of glutamic acid-glutamic acid and glutamic acid-proline neighbors.

Figure S5: Sampling from different noise distributions generates proteins with more diverse secondary structure. Representative 100AA unfiltered and unconditionally generated proteins from normal distribution, GMM2, and GMM3. Colored by model pLDDT (red → high confidence).

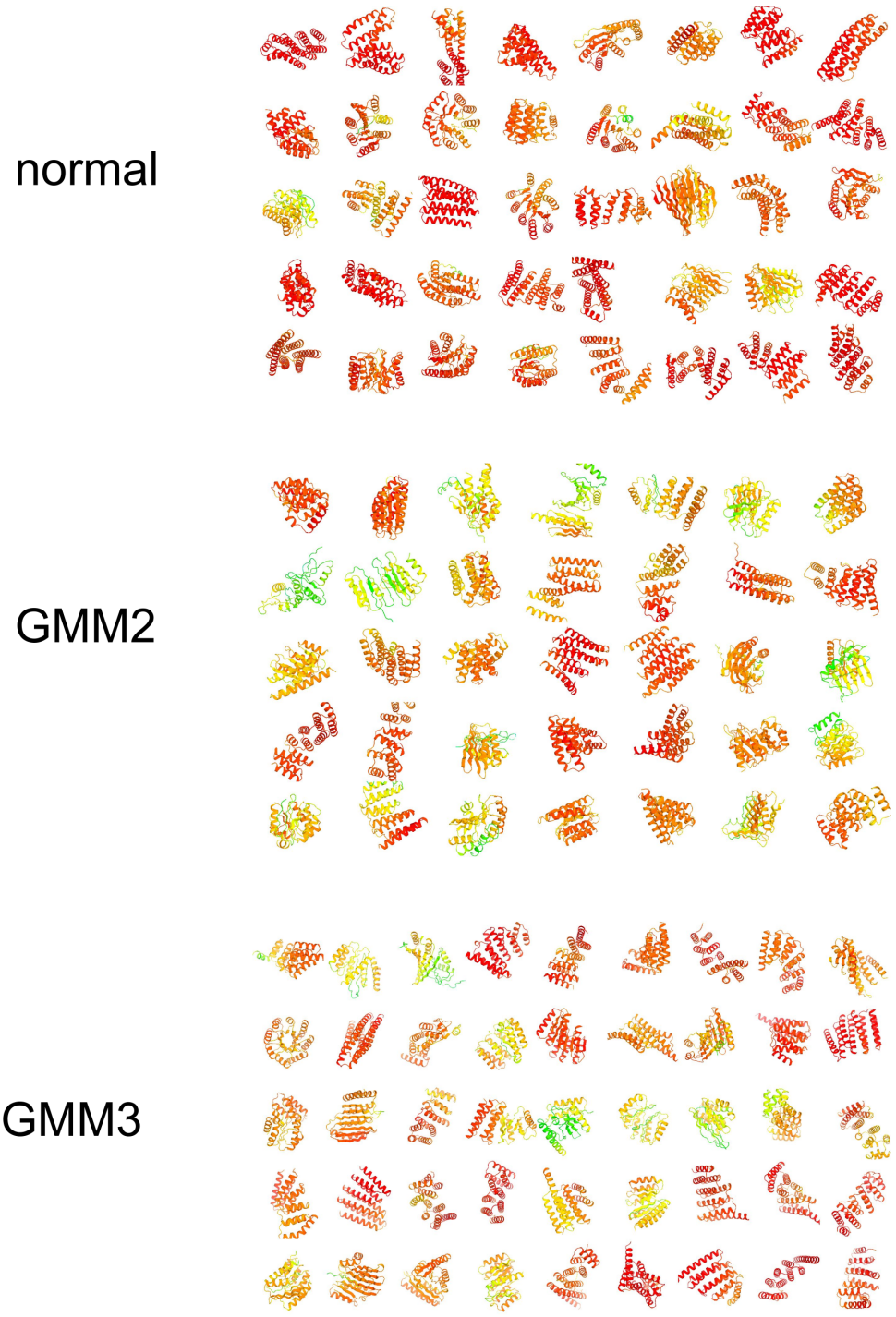

Figure S6: Sampling from different noise distributions generates proteins with more diverse secondary structure. Representative 200AA unfiltered and unconditionally generated proteins from normal distribution, GMM2, and GMM3. Colored by model pLDDT (red → high confidence).

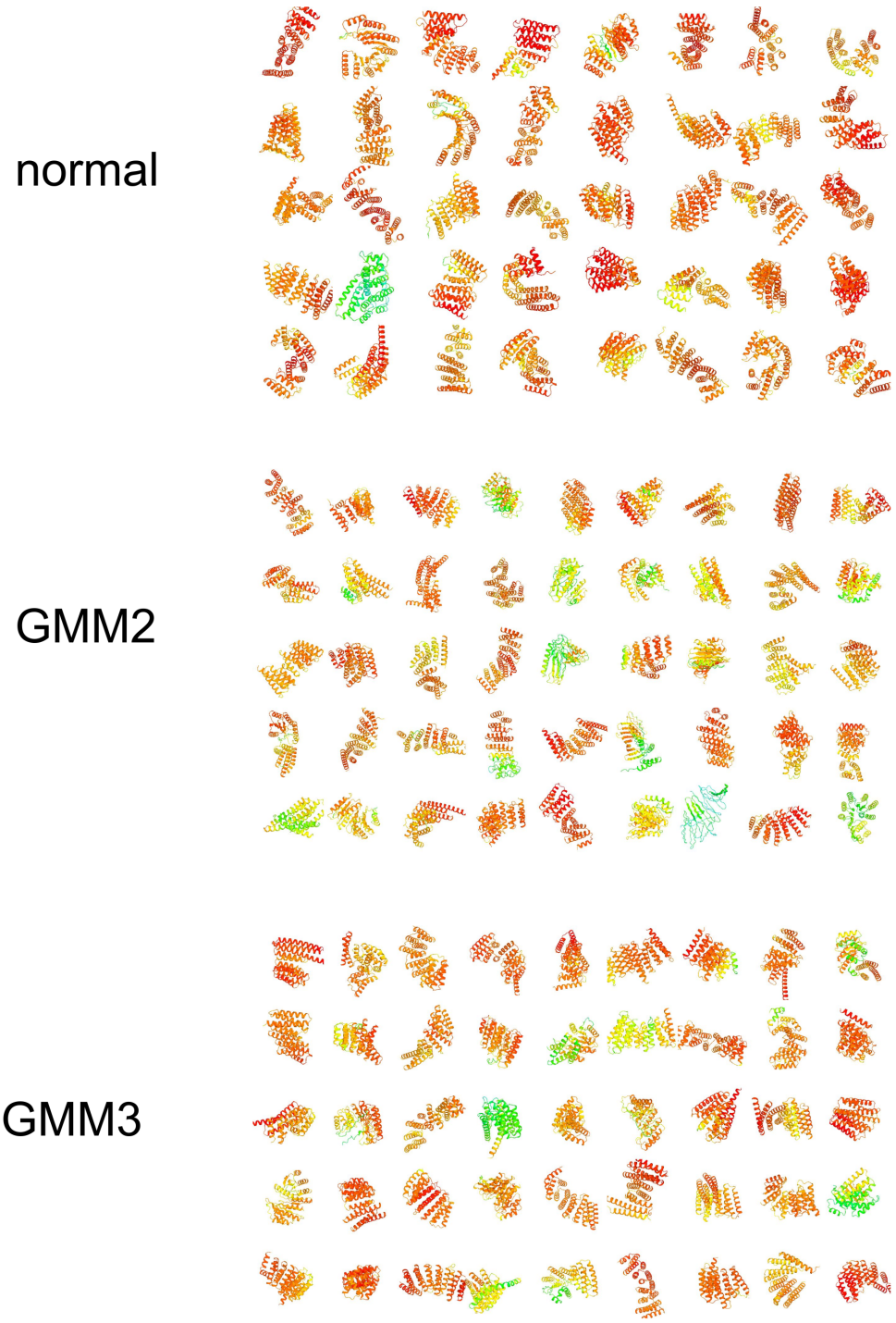

Figure S7: Sampling from different noise distributions generates proteins with more diverse secondary structure. Representative 300AA unfiltered and unconditionally generated proteins from normal distribution, GMM2, and GMM3. Colored by model pLDDT (red → high confidence).

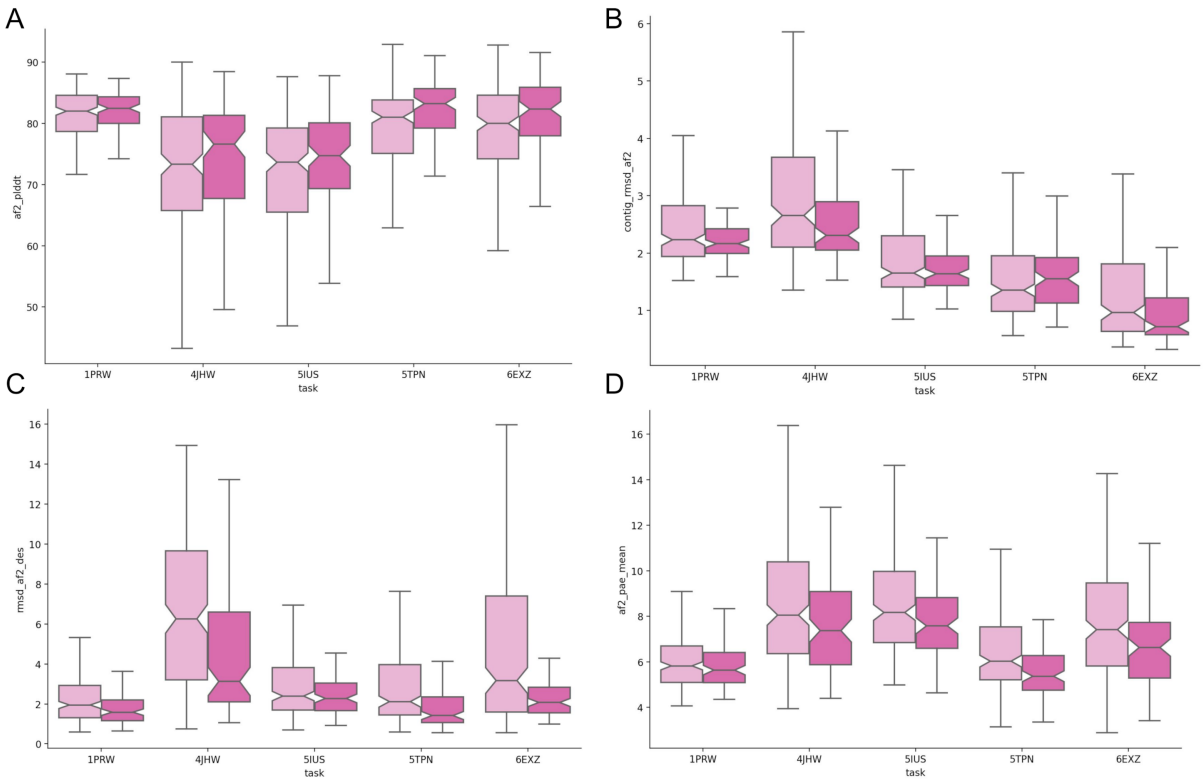

Figure S8: AF2 metrics for scaffolding of structure-sequence motifs in the PDB IDs listed in 25 (light pink) and 100 (dark pink) time steps. (A) AF2 pLDDT for designs, (B) RMSD of motif predicted by AF2 to design, (C) RMSD of AF2 to design for whole structure, (D) predicted alignmed error (pAE) of designs from AF2. The following contig arguments were used to run motif scaffolding benchmark: 1PRW - contigs 8-20,A21-31,16-25,A56-67,8-20, 6EXZ - contigs 0-95,A28-42,0-95, 5TPN - contigs 10-40,A163-181,10-40, 5IUS - contigs 0-30,A119-140,15-40,A63-82,0-30, 4JHW - contigs 0-25,F196-212,15-30,F63-69,10-25

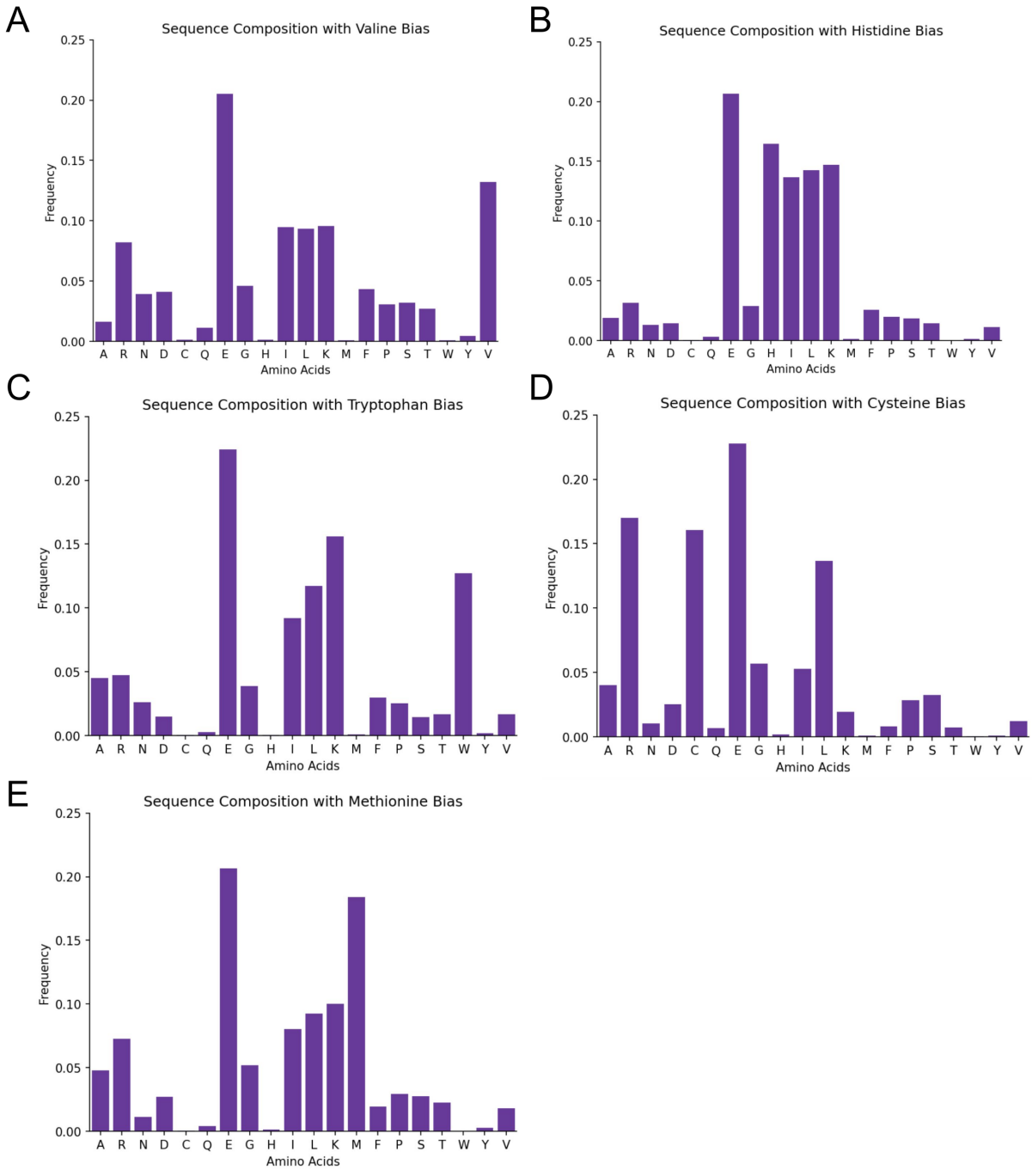

Figure S9: Biasing for specific amino acids results in increased frequency of the specified residue in generated proteins. Amino acid distributions of proteins generated with amino acid compositional bias for (A) valine, (B) histidine, (C) tryptophan, (D) cysteine, and (E) methionine.

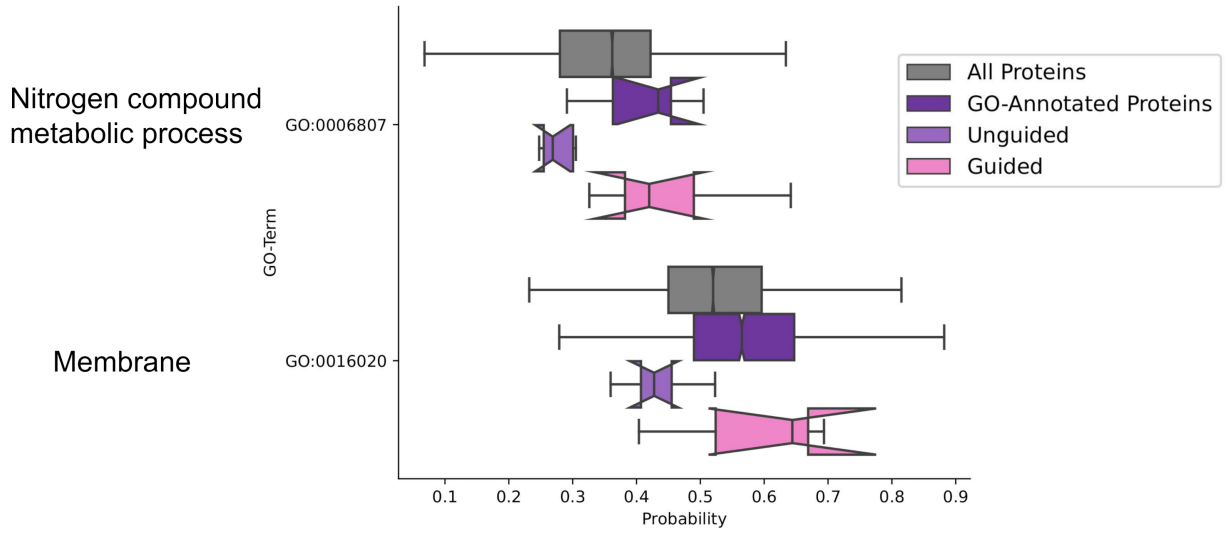

Figure S10: GO-guidance. The network has been guided with the DeepGOPlus Gene Ontology (GO) classifier to generate proteins with specific characteristics and functions. Exemplary, the classifier GO probability scores for all UniProt proteins, all proteins annotated with the chosen GO term, unconditionally unguided proteins generated with our model and guided proteins generated with our models for the GO terms nitrogen compound metabolic process (GO:0006807) and membrane (GO:0016020) are shown. The classifier has a high false positive rate due to a high mean probability as well as for all UniProt proteins including proteins not annotated with this specific GO term. For both GO terms a shift in the probabilities can be shown for guided proteins in comparison to unguided proteins.

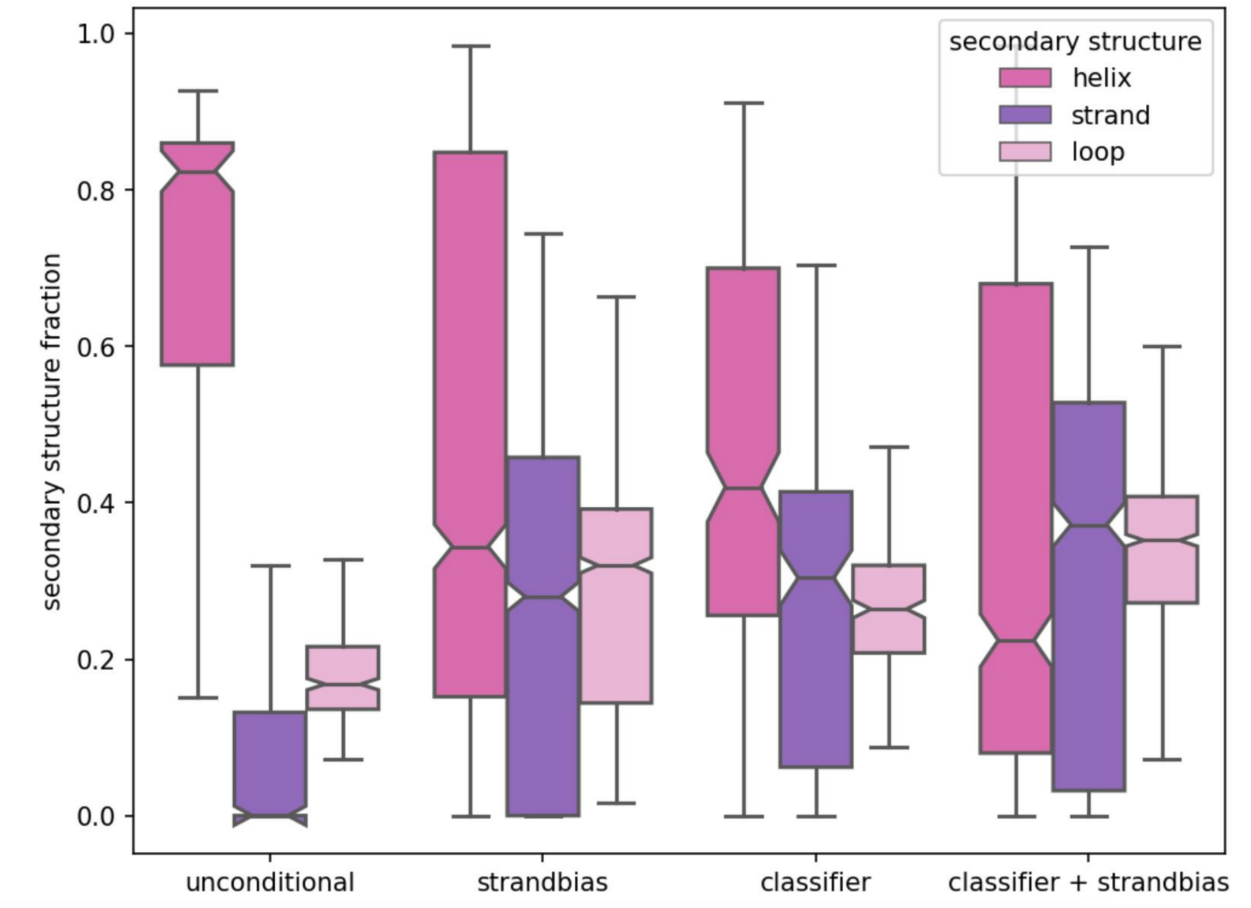

Figure S11: Secondary structure composition comparison when generation unconditional designs, designs with strand bias, designs with classifier guidance, and combination of classifier guidance and strand bias.

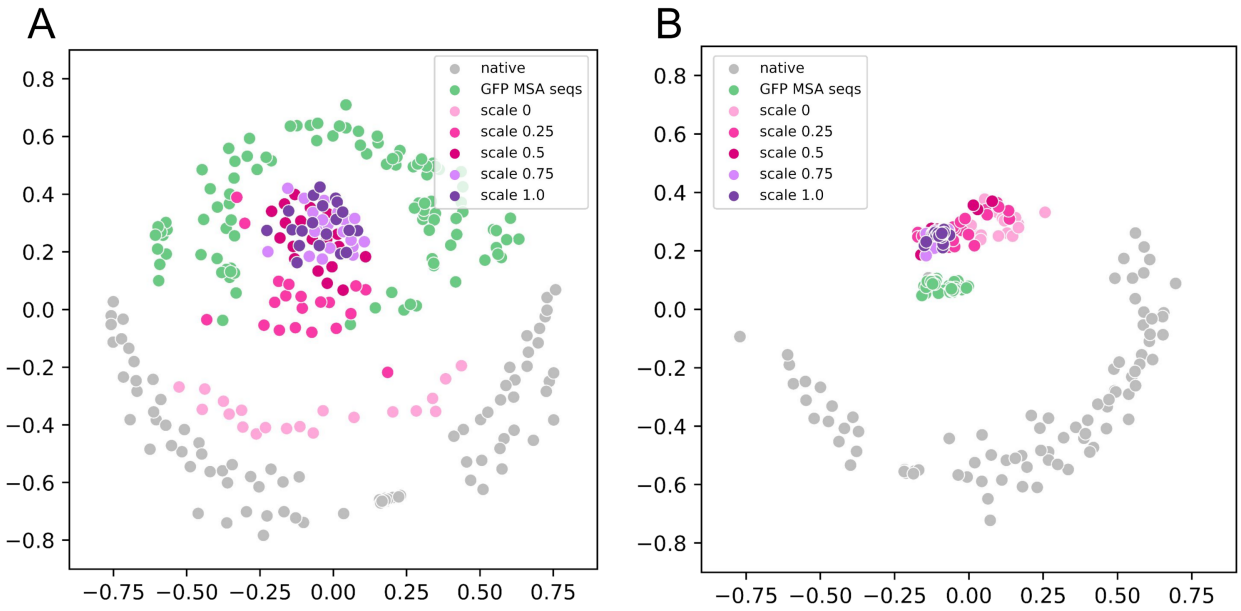

Figure S12: Multidimensional scaling plots of proteins generated with increasing GFP PSSM guidance scales. (A) Higher PSSM scaling increases sequence clustering to native GFPs. Distance metric is percent sequence identity. Green dots are native GFP sequences derived from a GFP MSA with sequence identity cutoffs 30-90% to the query sequence. Grey are randomly sampled native sequences from Uniprot90 (B) Low PSSM scaling results in increased structural diversity and samples of more diverse beta barrels. Higher PSSM scaling reduces structural diversity and clusters closer to native GFPs. Distance metric is TM score. Green dots are structures derived from the same MSA as (A) and grey dots are structures derived from the same set as (A).

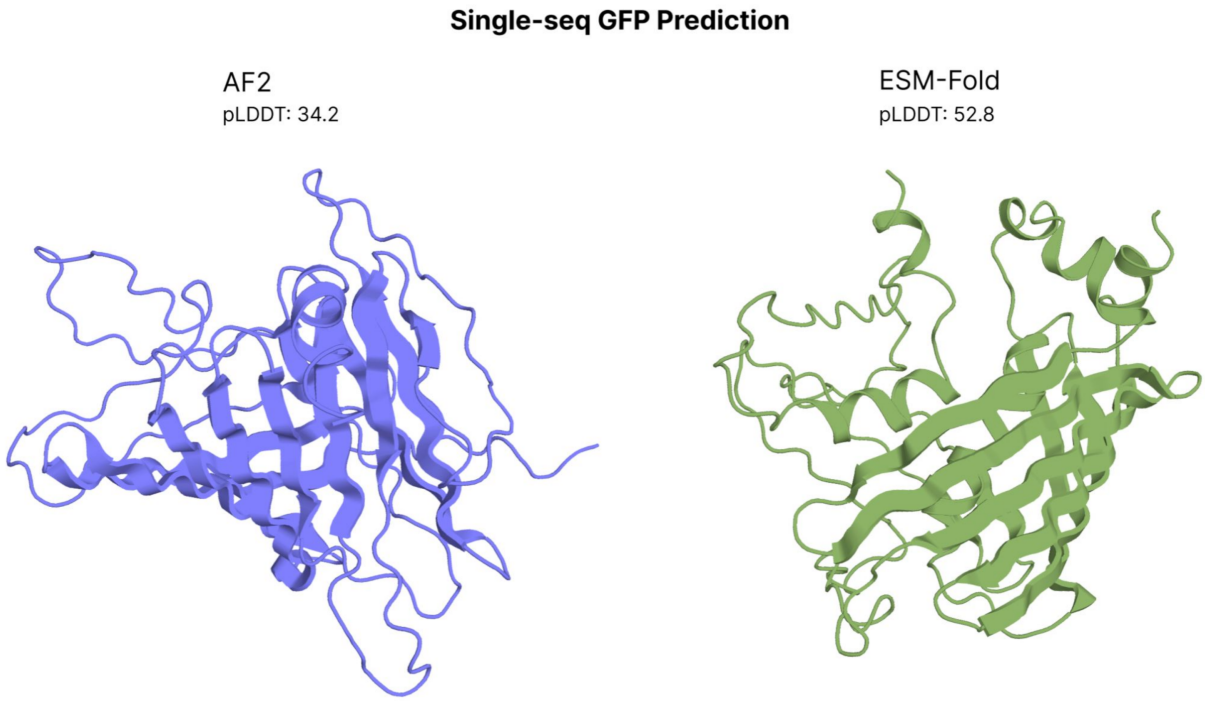

Figure S13: Single-sequence structure predictions of Green Fluorescent Protein (GFP), (PDB 1EMA). AlphaFold2 (left) and ESM-Fold (right) predictions fail to recover the tertiary structure of GFP when run in single-sequence mode. Both models return structures with low confidence (pLDDT). Models were run with 6 recycles.

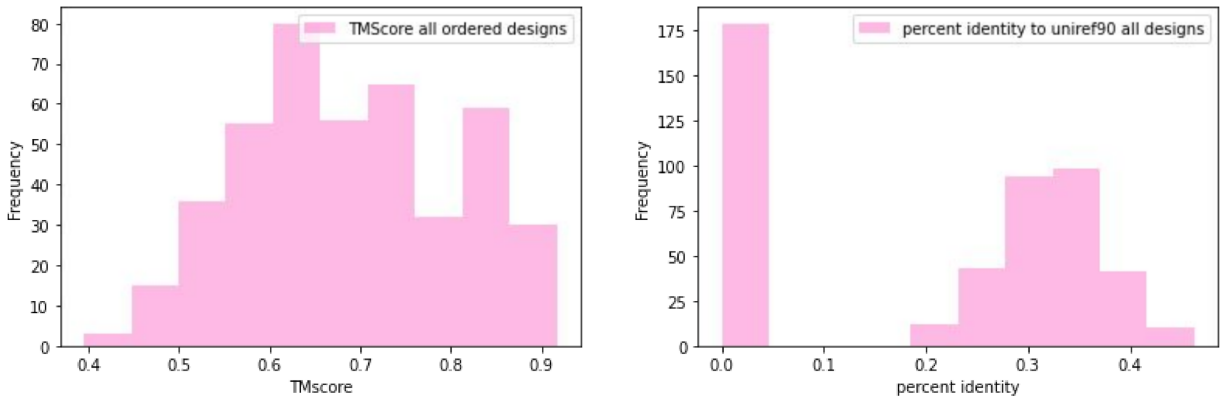

Figure S14: TMScore (left) and sequence identity (right) distributions of all ordered designs against PDB.

```
def train(seq,xyz,steps,cond,mask_struc,mask_seq):

  #one hot encode
  x=one_hot(seq)*2-1
  t=uniform(0,steps)/steps

  #noise
  eps = normal(mean=0, std=1)
  x_t = sqrt(gamma(t)) * x + sqrt(1 - gamma(t)) * eps

  #masking
  xyz_t=full_like(xyz,Nan)
  xyz_t[~mask_struc]=xyz[~mask_struc]
  x_t[~mask_seq] = x[~mask_seq]

  #self condition
  with stop_gradient:
    x_prev = None
    x_t_1 = sqrt(gamma(t+1)) * x + sqrt(1 - gamma(t+1)) * eps
    x_prev = model(x_t_1,x_prev,t,cond)

  #predict and calc seq and struc loss
  x_pred,struc_pred = model(x_t,x_prev,t,cond,xyz_t)
  loss  = CCE(x_pred,one_hot(seq))
  loss += Structure_Losses(struc_pred,xyz)

  #KL loss
  x_pred_t1 = sqrt(gamma(t-1)) * x_pred + sqrt(1 - gamma(t-1)) * eps
  seq_t1 = sqrt(gamma(t-1)) * one_hot(seq) + sqrt(1 - gamma(t-1)) * eps
  loss += KL(x_pred_t1,seq_t1)

  return loss.mean()
```

Pseudocode S1: Training.

```
def generate(L,steps,cond,seq,xyz, mask_struc,mask_seq, classifier):

  #setup
  x_prev=None
  seq_start=full(1,L,20)
  x_pred=one_hot(seq_start)*2-1
  seq=one_hot(seq)

  for step in range(steps):
    #noise
    t=(steps-step)/steps
    eps = normal(mean=0, std=1)
    x_t = sqrt(gamma(t)) * x_pred + sqrt(1 - gamma(t)) * eps

    #masking
    xyz_t=full_like(xyz,Nan)
    xyz_t[~mask_struc]=xyz[~mask_struc]
    x_t[~mask_seq] = seq[~mask_seq]

    #predict x_0
    x_pred=model(x_t,x_prev,t,cond,xyz_t)

    #classfifier guidance
    if classifier is not None:
      x_pred += grad(classifier(x_pred))

  return argmax(x_pred)
```

Pseudocode S2: Inference.

```
def generate(L, steps, conditionings, seq, xyz, mask_struc, mask_seq, classifier):

  #setup
  x_prev=None
  seq_start=full(1,L,20)
  x_pred=one_hot(seq_start)*2-1
  seq=one_hot(seq)

  for step in range(steps):
    #list of output seqs
    pred_Xo = list()
    for cond in conditionings:
      #noise
      t=(steps-step)/steps
      eps = normal(mean=0, std=1)
      x_t = sqrt(gamma(t)) * x_pred + sqrt(1 - gamma(t)) * eps

      #masking
      xyz_t=full_like(xyz,Nan)
      xyz_t[~mask_struc]=xyz[~mask_struc]
      x_t[~mask_seq] = seq[~mask_seq]

      #predict x_0
      x_pred=model(x_t,x_prev,t,cond,xyz_t)
      pred_Xo.append(x_pred)

    #updated x_pred with average output seqs
    x_pred = mean(pred_Xo)

  return argmax(x_pred)
```

Pseudocode S3: Multistate design.

```
def aa_bias(seq):

  #softmaxed probability distribution for seq [L,21]
  soft_seq = softmax(seq)

  #stack gradients in list for each AA
  grad_stack = []

  #iterate through each aa and fraction to bias sequence toward:
  for aa, fraction_to_bias in aa_bias_list:

  #set up aa bias by initializing seq of zeros
  potential = zeros_like(seq)

  #set residue type of interest to 1
    potential[:,aa] = 1

    #get mean squared error between soft_seq and potential
    dist = MSE(potential - soft_seq)

    #get gradients of soft_seq w.r.t. dist
    gradients = get_grads(soft_seq, dist)

  #set update gradients
  update_grads = zeros_like(seq)

  #find top-k residues closest to desired aa
  top-k_resi_list = get_topk(soft_seq[:,aa], (L * frac_to_bias))

  #iterate over each residue in the sequence
  for resi in num_residues:

      #neg gradient to bias toward aa of interest
      if resi in top-k_resi_list:
        update_grads[resi,:] = -gradients[resi,:]

      #pos gradient to bias away aa_of_interest
   else:
        update_grads[resi,:] = gradients[resi,:]

  #pos gradient to bias away
  grad_stack.append(update_grads)


  #average over multiple gradients when biasing for more than one aa
  update_grads = mean(grad_stack)

  return update_grads
```

Pseudocode S4: Amino acid composition potential.

```
def charge_at_pH(seq, pH, target_charge):

  #softmax
  soft_seq = softmax(seq)

  #get table of AA partial charges at pH
  #based on Henderson Hasselbach equation
  pos_charges = (1.0 / (10 ** (pH - pos_pKs_matrix)) + 1.0)
  neg_charges = (1.0 / (10 ** (neg_pKs_matrix - pH)) + 1.0)

  #make table based on sequence length of all combinations of
  #positive, negative, and neutral charges that sum to target_charge
  table = make_table(seq.shape[0])

  #classify each position of soft_seq as
  #positive, negative, or neutral
  charge_classification = classify_resis(soft_seq)

  #find closest table entry that sums to target_charge
  #based on the currently classified soft_seq
  target_charge_ratios = get_target_charge_ratios(table, charge_classification)

  #determine gradients at each position based on target_charge_ratio
  #and the current charge_classification
  #return +1 for positions that should have positive gradients, -1 for positions
  #that should have negative gradients, else 0
  Guided_charge_classification = draft_resis(target_charge_ratios,
    charge_classification)

  #sum of soft charges at each position
  soft_charge = sum(soft_seq * (pos_charges - neg_charges), dim = -1)

  #calculate MSE loss with respect to soft_charge
  loss = mean(((guided_charge_classification - soft_charge)**2)**0.5)
  loss.backward()

  #get gradients
  gradients = soft_seq.grad

  return gradients
```

Pseudocode S5A: Net charge potential.

```
# pKa lists to account for every residue.
pos_pKs = [[0.0, 12.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 5.98, 0.0, 0.0, 10.0, 0.0,
    0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]]
neg_pKs = [[0.0, 0.0, 0.0, 4.05, 9.0, 0.0, 4.45, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
    0.0, 0.0, 0.0, 0.0, 0.0, 10.0, 0.0, 0.0]]
cterm_pKs = [[0.0, 0.0, 0.0, 4.55, 0.0, 0.0, 4.75, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
    0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]]
nterm_pKs = [[7.59, 0.0, 0.0, 0.0, 0.0, 0.0, 7.7, 0.0, 0.0, 0.0, 0.0, 0.0, 7.0,
    0.0, 8.36, 6.93, 6.82, 0.0, 0.0, 7.44, 0.0]]

# Repeat charged pKs L - 2 times to populate in all non-terminal residue indices
pos_pKs_repeat = self.pos_pKs.repeat(seq.shape[0] - 2, 1)
neg_pKs_repeat = self.neg_pKs.repeat(seq.shape[0] - 2, 1)

# Concatenate all pKs tensors with N-term and C-term pKas to get full L X 21
    charge matrix
self.pos_pKs_matrix = cat((zeros_like(nterm_pKs), pos_pKs_repeat, self.nterm_pKs))
self.neg_pKs_matrix = cat((cterm_pKs, neg_pKs_repeat, zeros_like(self.cterm_pKs)))
```

Pseudocode S5B: Net charge potential data structures and tables.

```
def hydropathy_index(seq, target_score):

  #get table of hydropathy values
  hydropathy_matrix = hydropathy_list.repeat(seq.shape[0])

  #softmax
  soft_seq = torch.softmax(seq)

  #sum of (softmax * hydropathy values)
  hydropathy_score = sum(soft_seq * hydropathy_matrix, dim = -1)

  #calculate MSE loss with respect to soft_seq
  loss = ((hydropathy_score - target_score)**2)**0.5
  loss.backward()

  #get gradients
  gradients = soft_seq.grad

  return gradients
```

Pseudocode S6A: Hydrophobicity potential.

```
# AA conversion
conversion_list = list("ARNDCQEGHILKMFPSTWYVX")

# Dictionary to convert amino acids to their hyropathy index
gravy_dict = {'C': 2.5, 'D': -3.5, 'S': -0.8, 'Q': -3.5, 'K': -3.9,
        'I': 4.5, 'P': -1.6, 'T': -0.7, 'F': 2.8, 'N': -3.5,
        'G': -0.4, 'H': -3.2, 'L': 3.8, 'R': -4.5, 'W': -0.9,
        'A': 1.8, 'V':4.2, 'E': -3.5, 'Y': -1.3, 'M': 1.9, 'X': 0, '-': 0}

gravy_list = [self.gravy_dict[a] for a in self.conversion_list]
```

Pseudocode S6B: Hydrophobicity potential data structures and tables.

Table S1: Observed and predicted mass for designs used in mass spectrometry experiments.

Design	Observed Mass (Da)	Predicted Mass (Da)	Difference
TrpA5	11955	12086.18	-131.18
TrpA9	11592	11722.61	-130.61
TrpB9	11312	11765.34	-453.34
ValF3	10470	10600.58	-130.58
ValF12	11203	11333.68	-130.68
ValF11	10850	10981.23	-131.23
UncC7	11690	11820.99	-130.99
UncC4	10851	10981.47	-130.47
UncD5	11191	11321.68	-130.67
RCC12	23072	23202.83	-130.83
RCE4	22880	23010.88	-130.88
RCE8	22575	22705.43	-130.43
RCF12	22997	22996.3	0.7
NCG1	21162	21292.13	-130.13
NCA1	20779	20910.78	-131.78
CysG4	8427	8561.43	-134.43
CysH6	8675	8813.75	-138.75
CysH11	8633	8642.57	-9.57
TEV A1	14905	15036	-131
TEV A2	14713	14844	-131
TEV A3	14005	14136	-131
TEV A4	14044	14175	-131
TEV A5	14265	14396	-131
TEV A6	14178	14309	-131
TEV A7	14714	14845	-131
FUR A9	22894	23025	-131
FUR A10	23556	23687	-131
TEV B1	14864	14995	-131
TEV B2	14312	14443	-131
TEV B3	14561	14692	-131
TEV B6	14547	14661	-114
FUR B10	19733	19916	-183
FUR C2	18934	19065	-131
FUR D2	18853	18984	-131
FUR D3	18869	19000	-131

Table S2: Mass spec data of experimentally validated cysteine-rich designs. The mass of each design is reported in the presence and absence of the reducing agent TCEP. The mass difference between reduced and non-reduced designs is used to calculate the number of disulfides formed and compared to the number of designed disulfides.

Design ID	**Mass with MET loss (Da)**			**# Disulfides formed**
	+ TCEP	- TCEP	Difference	
cys_G4	8431	8425	6	3/3
cys_G6	8778	8770	8	4/4
cys_G10	8695	8687	8	4/4
cys_G11	8368	8363	5	3/3
cys_H6	8683	8675	8	4/5

Table S3: Secondary structure prediction of CD data (200-250 nm) of designs RC_E8 Fig 3E middle top, and RC_F11 Fig 3E middle bottom with BeStSel server indicating high percentage of beta content.

	CD RC_E8	**CD RC_F11**
Helix	1.0	2.7
regular	0.2	1.2
distorted	0.8	1.5
Antiparallel	40.7	26.8
left-twisted	0.0	0.0
relaxed	24.3	12.2
right-twisted	16.4	14.6
Parallel	0.0	0.0
Turn	14.0	17.9
Others	44.3	52.7

Table S4: Ordered protein sequences, design confidence, and AF2 confidence metrics.

	design_name	design_aa_seq	design_plddt	af2_plddt
0	230119 his scale2.0 batch7 185	SHFHEIHKEIEELKKEFKKLLHEGPSLE HLHHLIEHIEKLIEKIKHLGLHHLHKEI EELLHEIHKLIEELEKSG	0.959	91.464
2	230119 his scale2.0 batch7 197	LAHEVEHIRNKIHELEKEAEHLGDEEIH HLIAEAHHLIHEAFELFEAGDHSKAHEL HKLAKHLIEHAEERIHKIREGG	0.979	92.502
4	230119 his scale1.5 batch8 41	NEKHAKILELLHKLREHLKKQGHKEIVE KLDHILEKIRHGASKEEIIAEIKEIHDH MKKHGLISPEIRHHIREIIELLKN	0.968	91.583
6	230119 his scale1.25 batch9 181	GLRERHEEIEHLLHELNRIYKEIKKAKE KGDEDKVHEALKHAQEIHERIEHLGNKE AVEHAKEHLKEIEALH	0.973	91.815
7	230119 his scale1.5 batch4 58	GEHEIHELIHRLFEEILAILHHVKHLIH HGGDEEEIHHLIHEIKEKLHHLRRLGAD PHKIKHIEKKLEEIEKEIKER	0.98	92.67
8	230119 his scale1.25 batch0 446	NEHFKKLLEKIEKIREMIHEGFSEEEIH EEIEHIKKELEELHESGHFDPEDREIIF EKLKELHRELEKHKG	0.969	91.309
10	230119 his scale1.75 batch8 319	GHIEKLHHHFEELKHLFHKGHISHKEII KKLNELIKEIEHLIKKHKHNKELVEELE KLIKHIEEFIEELHKES	0.974	91.814
11	230119 his scale1.25 batch5 160	GEDERLLREILEELDELIHHLERHGHHE LIHEIEKVIELIHAGDHEKALEELHHLI DKLRHAGIDPKDIEKVVEHLRKL	0.972	92.247
12	230119 his scale1.5 batch1 86	NEHKLIHELLNEIEQEHEDIKKLHKEHG SEELHELLHKIHEVRHELKSGNLEEALK KFEKIHKHYKEIKHKLG	0.968	91.191
14	230119 his scale1.5 batch8 108	NEQAEAHAKEIHKELEKLKHLIEEGNLE EALKIIKHLEEIIAEARHHLDPAELHRH LHELHEFHKEIEAK	0.977	92.224
16	230119 his scale1.25 batch9 71	GKEHRFETIIHKINEIEKHHLELKEKIE RLEKHKGKLNKEEHEDKIHEVETEIKET HENIKELEHELKEIE	0.968	91.461
18	230119 his scale1.25 batch5 342	EHLKERIDHIRHEFEEIKELHKKGKISR EEVLKKLERIIEETDELVKHLKENGFDP ELIKHTEHIIEEIKEHIREIKEH	0.966	91.159
20	230119 his scale1.25 batch1 219	SPEKIEELKEEIKKHLEKLHEIIKRIPP DHPHELKHFLEEIDRIKEHFEKAHSVEE IKNIIHHTEELIHRIEKHIEEN	0.976	91.956
27	230119 trp scale1.5 batch7 369	SWRDKRIAFWKEEIKKFEEKWEKLKKEG TKEEIEKWLEWIEKWLKEVDELWKETGN EEIREIWLKLKNIRKELKEWKEER	0.963	90.862
28	230119 trp scale1.75 batch1 430	GRKEWEERIWEWLKRIRELLEKGDWEEI EELIKELWEWFKGWPEEIKKFIEELRKC KDPEKLRELLERLEEWLEREW	0.952	90.26
29	230119 trp scale1.5 batch7 28	NLRELWEEIRECIEKGDEEKLRELFDRI KEWAWKNPEEAIEWLKELWEWIEERGKK PWAEEIEEWWKRIKG	0.959	90.771
31	230119 trp scale1.75 batch0 462	GRERKWWEWRLRKIREEIKRWLENPKNI SWEEIEKLKEWLREIWENISPEEWEKWR KELKELKEEIEELKEEWWRRREG	0.941	90.499
33	230119 trp scale1.5 batch1 113	NPEERRRELAEWILEKWEAGWSWEELLR EALEWLERWGITLEEFWEIWEEWCRARG LLEWWKRFKKEFLAEK	0.971	91.921

35	230119 trp scale1.75 batch3 444	GEREWWIKKLEEKIEKLIAGKWTKEELR EWLKWWLERYPDWRKEWIRTIERYWEKV KDEKERKWLKEIWEFLKEWG	0.978	90.595
36	230119 trp scale2.0 batch1 470	EKREKWLKEIREELSELWELWKEGKITR EELENFIERLEELGIKITWEEFWEWLKE AKERGWTPEELWKKIKEWWKN	0.971	90.569
38	230119 trp scale1.5 batch9 29	EWREKIKEWEKEWKELWEKFSEWKKNLT PEEREEWEKLRSEWKELMKLFENISEEE KKEILKKIEEILKELKELWKKWKS	0.972	90.725
39	230119 trp scale1.25 batch4 328	PTREEWLKWIERWIEEIREWLEKLLEWG KEWRKKAKELEEIIEKLEEWIKECKKRG WTPEEIREEFKEWRKRIEEILK	0.966	90.671
41	230119 trp scale1.25 batch2 394	NPEERFEEILEWLSKPIDEQEWKELIQE IEKWLEENGWEDWLEELKKWIEEWLSNP DITREEFKEKLKEWIREFLKKI	0.975	91.377
42	230119 trp scale1.75 batch6 253	NKEEFLKELEKAIEAIEKGDWEKAFKWL KKLIELLKEAGLKEEIEEIEKWWELWKN GKFSRDEWLEWLKKWREEWKARG	0.986	90.786
43	230119 trp scale1.5 batch4 224	NWEKLWEELKEWLEKNGIEDPDEWINLI KEWIEEKLRKGWITKEEALKLLEEIIKE WEERGNEEEIERLKEIFAELEKW	0.976	90.647
45	230119 trp scale1.75 batch3 65	ISWEDLIREMRELFGWTEWEIEEWRKEW EEAKARGDWWELLEKLWEWAKENGSKEE WRILLERWRELREKLG	0.957	90.573
46	230119 trp scale1.25 batch1 306	GPTREEFIKKLIDLLWKGASWEKIRKLI LEWAKRWGWTPEEIERILRIIDEWPGMS PEEIRRRLEEWWEEWERAKG	0.968	91.183
48	230119 met scale1.25 batch7 16	NPEDLLRKLEEMREEFRRMVEAGDKEGI EELEMQFEELMEEMEELMAAGKITPEMR EEMEMAREEMREMLAEARRMH	0.949	90.913
50	230119 met scale1.5 batch2 331	MPEKMREMLEELERRIEEAKGMTPEELR ELFERMESLRMMMEEMRRAGKISEEELR EMLERIERMLEELKRLMM	0.973	90.923
51	230119 met scale1.25 batch8 0	MMTREEFERLLERMRARGMEEAAEMLEM AMEMMEAGRPPEVRAAMREVERALEAAG APPELRAEMERMREMLEREMRG	0.977	90.589
52	230119 met scale2.0 batch6 254	SKIEESMKEMLEMLDGSPEDLRKMRRRM EEMLEMMKKSGMSMETIAMIEKILMMLK EGEKESMKESIEEMLEQLG	0.962	90.708
53	230119 met scale1.5 batch2 274	MKDRLENIRKKLEMIMKEMENMGLNDPE IKEMINEIMEEMKKIRKGNMTEEEREEM IMKTEKKMMEIKEMIEEMKKG	0.97	91.022
57	230119 met scale1.25 batch5 24	NGSIKKIMEEVEMIIEKIEEFEKKMGGG EDLMKEIEEIKEMILEDPENITEEELKE IQERLRKAMEKIEEMMENG	0.933	91.033
59	230119 met scale1.25 batch1 417	MDMETKRKFLEEMEDMMRKIEELMESGR LTKEEMDEIMAKMEEMMELIEKGEKEEI ERLLKKIQAEIKKMMG	0.984	91.235
64	230119 met scale1.25 batch0 407	SMKEAMRREREELLKDMERMRKELMEMR KAGMTPEEIDEMLEEMERVLKMIEAGMS PEELREEIERIRERIEEMQKRL	0.95	92.533
67	230119 met scale1.25 batch2 85	GSRDEMMAQMRERAMEMLEEARRRGDQA EIMRIMKEFREEMEEVGISREEAIEMLM EMMRRSGMSEAEAKEMIDRMG	0.984	90.451
69	230119 met scale2.0 batch6 42	MLMEKMEELIMEIEGLSPEEIVRRMREM IEENREEMSDEMIEEMREMLMDIEKLMG KKTPEEIREFMREKIEEMLKRMRE	0.985	90.349

75	230126 uncond batch2 269	NADEALREIERISEKLREIEELIRAGSP EELREAIETLEKIEEELKRLMKESEGPL KELKKLLEEAQRLDRRIKEIRG	0.942	92.36
76	230126 uncond batch9 56	SNEEIEKQLEEIKKEIEKIKKLFESGFN REEIINELEKIKEKIEELAKKYGPSDEI RKILKEIEELLKELQ	0.967	92.275
80	230126 uncond batch5 166	SPKEEERLARIREAEKELEEVMRLLKSG NREEAIELLKKIREILEELKGLSPEEKE KINKILEEIEKAEKELES	0.966	92.806
81	230126 uncond batch2 262	SPAARRARLRAELREFAERAERLAEEFR RAGEEDLAREAEALAREIRRLAELGPSE EEIEEIRERIEQLREEAEKLLG	0.984	94.848
86	230126 uncond batch5 155	EEYEQLLAELEELIAELEELLALLGGDP EERRQIEEILAEIRELIKAFEAGKITPE EVRERLEELSREIRELRERLG	0.938	92.523
87	230126 uncond batch0 309	NLSEELQELKKKYREDFEEALELAEENG NKAKLEELEELDKEFNELLKNGNISEIE ETLKELESLKKELKES	0.974	94.455
88	230126 uncond batch8 451	ELEELLKKLEEIKKEIEELREKGINITE EIIKQISEIEKEIKELKAAGKPDKEEIE RIENQIEEIREEIEKLRS	0.971	93.078
89	230126 uncond batch9 421	GKQKQIRELIEEIRELLKRIQELLKSGK PEEAEEELEKLEERIEEIRKLCKEENIP LPEELEEIEEEREELRELLKN	0.977	93.558
90	230126 uncond batch0 316	GSPEQLIARVEREIAELERLIAAGGATR EEIQAAMAETEKLIEELRALGDADEAER LAKRIEELARECEARLKG	0.953	92.376
96	repeat protein dssp0 D0.2 ps1 000004 A	DKPESVDIELQPGSTISEDDARKLADAL RDANIDKPESVDIELQPGSTISEDDARK LADALRDANIDKPESVDIELQPGSTISE DDARKLADALRDANIDKPESVDIELQPG STISEDDARKLADALRDANIDKPESVDI ELQPGSTISEDDARKLADALRDANI	0.892	nan
97	repeat protein dssp0 D0.2 ps3 c1 6 000009 A	GKAKSISLHLDPGTLDDLKKRDISLREF FDHLAGKAKSISLHLDPGTLDDLKKRDI SLREFFDHLAGKAKSISLHLDPGTLDDL KKRDISLREFFDHLAGKAKSISLHLDPG TLDDLKKRDISLREFFDHLAGKAKSISL HLDPGTLDDLKKRDISLREFFDHLA	0.76	nan
98	repeat protein dssp0 D0.2 ps3 c1 8 000000 A	YKPEVKLVADASSITEEQRDDIIKFLRE ARDDGYKPEVKLVADASSITEEQRDDII KFLREARDDGYKPEVKLVADASSITEEQ RDDIIKFLREARDDGYKPEVKLVADASS ITEEQRDDIIKFLREARDDGYKPEVKLV ADASSITEEQRDDIIKFLREARDDG	0.712	nan
99	repeat protein dssp0 H0.2 ps2 c1 3 000002 A	KKHARAIIHIHHHKGRTREEIVEEARHI LEELGKKHARAIIHIHHHKGRTREEIVE EARHILEELGKKHARAIIHIHHHKGRTR EEIVEEARHILEELGKKHARAIIHIHHH KGRTREEIVEEARHILEELGKKHARAII HIHHHKGRTREEIVEEARHILEELG	0.766	nan
100	repeat protein dssp0 H0.3D0.3E0.3 ps3 c1 9 000005 A	AHIPHVEIHIDADGLTEEEIDEKIEKAI EEAKEAHIPHVEIHIDADGLTEEEIDEK IEKAIEEAKEAHIPHVEIHIDADGLTEE EIDEKIEKAIEEAKEAHIPHVEIHIDAD GLTEEEIDEKIEKAIEEAKEAHIPHVEI HIDADGLTEEEIDEKIEKAIEEAKE	0.8	nan

101	repeat protein dssp0 H0.3D0.3 ps1 c1 7 000002 A	APVDRIVIDMPDPADIKADDLQRARRQA REAGLAPVDRIVIDMPDPADIKADDLQR ARRQAREAGLAPVDRIVIDMPDPADIKA DDLQRARRQAREAGLAPVDRIVIDMPDP ADIKADDLQRARRQAREAGLAPVDRIVI DMPDPADIKADDLQRARRQAREAGL	0.723	nan
102	repeat protein dssp0 H0.3D0.3 ps1 c1 9 000000 A	IDVDKVTLTKNIGSGDKNIDELIADIKK LKAAGIDVDKVTLTKNIGSGDKNIDELI ADIKKLKAAGIDVDKVTLTKNIGSGDKN IDELIADIKKLKAAGIDVDKVTLTKNIG SGDKNIDELIADIKKLKAAGIDVDKVTL TKNIGSGDKNIDELIADIKKLKAAG	0.769	nan
103	repeat protein dssp0 H0.3D0.3 ps2 c1 5 000001 A	DKDGATLEFEVQPDDDPEDVAEKIQDIL DKNHLDKDGATLEFEVQPDDDPEDVAEK IQDILDKNHLDKDGATLEFEVQPDDDPE DVAEKIQDILDKNHLDKDGATLEFEVQP DDDPEDVAEKIQDILDKNHLDKDGATLE FEVQPDDDPEDVAEKIQDILDKNHL	0.867	nan
104	repeat protein dssp0 H0.3 ps3 c1 1 000006 A	HHLHHLVIRFGAGHTPEHLAHAFNRIEH MIAAGHHLHHLVIRFGAGHTPEHLAHAF NRIEHMIAAGHHLHHLVIRFGAGHTPEH LAHAFNRIEHMIAAGHHLHHLVIRFGAG HTPEHLAHAFNRIEHMIAAGHHLHHLVI RFGAGHTPEHLAHAFNRIEHMIAAG	0.788	nan
105	repeat protein dssp1 H0.2 ps2 c1 4 000005 A	VEPGAHAILPPGHTAAHARAHGFRKVYV HHPDKVEPGAHAILPPGHTAAHARAHGF RKVYVHHPDKVEPGAHAILPPGHTAAHA RAHGFRKVYVHHPDKVEPGAHAILPPGH TAAHARAHGFRKVYVHHPDKVEPGAHAI LPPGHTAAHARAHGFRKVYVHHPDK	0.847	nan
106	repeat protein dssp2 D0.2 ps1 c1 1 000000 A	GEIDALDLLEKHPNAKLIIRAGDTPQDV RDRAGGEIDALDLLEKHPNAKLIIRAGD TPQDVRDRAGGEIDALDLLEKHPNAKLI IRAGDTPQDVRDRAGGEIDALDLLEKHP NAKLIIRAGDTPQDVRDRAGGEIDALDL LEKHPNAKLIIRAGDTPQDVRDRAG	0.848	nan
107	repeat protein dssp2 D0.2 ps1 c1 6 000009 A	GPLSLEDLKDAGIKSLRFDGDRYSVDDI RRLFGGPLSLEDLKDAGIKSLRFDGDRY SVDDIRRLFGGPLSLEDLKDAGIKSLRF DGDRYSVDDIRRLFGGPLSLEDLKDAGI KSLRFDGDRYSVDDIRRLFGGPLSLEDL KDAGIKSLRFDGDRYSVDDIRRLFG	0.853	nan
108	repeat protein dssp2 D0.2 ps2 c1 9 000003 A	EVPTIQDLIDKGAKTLEFDLSGMDKDDI DRFLEEVPTIQDLIDKGAKTLEFDLSGM DKDDIDRFLEEVPTIQDLIDKGAKTLEF DLSGMDKDDIDRFLEEVPTIQDLIDKGA KTLEFDLSGMDKDDIDRFLEEVPTIQDL IDKGAKTLEFDLSGMDKDDIDRFLE	0.82	nan
109	repeat protein dssp2 D0.3 ps1 000006 A	KDVTARDAIAAGAKSISFTGDDLTADDI KDLLGKDVTARDAIAAGAKSISFTGDDL TADDIKDLLGKDVTARDAIAAGAKSISF TGDDLTADDIKDLLGKDVTARDAIAAGA KSISFTGDDLTADDIKDLLGKDVTARDA IAAGAKSISFTGDDLTADDIKDLLG	0.871	nan

110	repeat protein dssp2 H0.2D0.2 ps1 c1 5 000003 A	AITREEAKALIKSGKLHIHIPAGVSLDE VAKRYAITREEAKALIKSGKLHIHIPAG VSLDEVAKRYAITREEAKALIKSGKLHI HIPAGVSLDEVAKRYAITREEAKALIKS GKLHIHIPAGVSLDEVAKRYAITREEAK ALIKSGKLHIHIPAGVSLDEVAKRY	0.895	nan
111	repeat protein dssp2 H0.2D0.2 ps1 c1 8 000000 A	LHATLEDLIDRGVDLTIVIKPDMTEAEI KDRYELHATLEDLIDRGVDLTIVIKPDM TEAEIKDRYELHATLEDLIDRGVDLTIV IKPDMTEAEIKDRYELHATLEDLIDRGV DLTIVIKPDMTEAEIKDRYELHATLEDL IDRGVDLTIVIKPDMTEAEIKDRYE	0.82	nan
112	repeat protein dssp2 H0.2D0.2 ps2 c1 0 000008 A	ALPTAEEIRAAGIKELTITISNATDEEI QAALDALPTAEEIRAAGIKELTITISNA TDEEIQAALDALPTAEEIRAAGIKELTI TISNATDEEIQAALDALPTAEEIRAAGI KELTITISNATDEEIQAALDALPTAEEI RAAGIKELTITISNATDEEIQAALD	0.841	nan
113	repeat protein dssp2 H0.2D0.2 ps2 c1 5 000006 A	HMTLAELAKEHGIKSLHIDADGYSIDEI RALIGHMTLAELAKEHGIKSLHIDADGY SIDEIRALIGHMTLAELAKEHGIKSLHI DADGYSIDEIRALIGHMTLAELAKEHGI KSLHIDADGYSIDEIRALIGHMTLAELA KEHGIKSLHIDADGYSIDEIRALIG	0.87	nan
114	repeat protein dssp2 H0.2D0.2 ps2 c1 7 000001 A	HHMTIDELLERGVEIRIILDGDDDFEEF QRRTGHHMTIDELLERGVEIRIILDGDD DFEEFQRRTGHHMTIDELLERGVEIRII LDGDDDFEEFQRRTGHHMTIDELLERGV EIRIILDGDDDFEEFQRRTGHHMTIDEL LERGVEIRIILDGDDDFEEFQRRTG	0.832	nan
115	repeat protein dssp2 H0.2D0.2 ps3 c1 9 000002 A	GKTIAEVLEEENIKDFDLVVNNDSEEKV DDILGGKTIAEVLEEENIKDFDLVVNND SEEKVDDILGGKTIAEVLEEENIKDFDL VVNNDSEEKVDDILGGKTIAEVLEEENI KDFDLVVNNDSEEKVDDILGGKTIAEVL EEENIKDFDLVVNNDSEEKVDDILG	0.801	nan
116	repeat protein dssp2 H0.2 ps1 c1 1 000006 A	KPLTLEELKAAGIKTLCLEGESITPEEA EHLFGKPLTLEELKAAGIKTLCLEGESI TPEEAEHLFGKPLTLEELKAAGIKTLCL EGESITPEEAEHLFGKPLTLEELKAAGI KTLCLEGESITPEEAEHLFGKPLTLEEL KAAGIKTLCLEGESITPEEAEHLFG	0.849	nan
117	repeat protein dssp2 H0.2 ps1 c1 5 000002 A	GKTLRELIHEHKPKEFEISFHGQTPEEI RRALGGKTLRELIHEHKPKEFEISFHGQ TPEEIRRALGGKTLRELIHEHKPKEFEI SFHGQTPEEIRRALGGKTLRELIHEHKP KEFEISFHGQTPEEIRRALGGKTLRELI HEHKPKEFEISFHGQTPEEIRRALG	0.833	nan
118	repeat protein dssp2 H0.2 ps2 c1 0 000009 A	HIPTVAEIKAAGLTHLSLHLENGSEEEI DEFAKHIPTVAEIKAAGLTHLSLHLENG SEEEIDEFAKHIPTVAEIKAAGLTHLSL HLENGSEEEIDEFAKHIPTVAEIKAAGL THLSLHLENGSEEEIDEFAKHIPTVAEI KAAGLTHLSLHLENGSEEEIDEFAK	0.811	nan

119	repeat protein dssp2 H0.2 ps2 c1 9 000007 A	HKLTLEEALKKGLEISIHIHPGETFKEV LHRHNHKLTLEEALKKGLEISIHIHPGE TFKEVLHRHNHKLTLEEALKKGLEISIH IHPGETFKEVLHRHNHKLTLEEALKKGL EISIHIHPGETFKEVLHRHNHKLTLEEA LKKGLEISIHIHPGETFKEVLHRHN	0.838	nan
120	repeat protein dssp2 H0.2 ps3 c1 0 000006 A	HAPTIEHLAHKGVKHVTINFHNATKEEI EHFFKHAPTIEHLAHKGVKHVTINFHNA TKEEIEHFFKHAPTIEHLAHKGVKHVTI NFHNATKEEIEHFFKHAPTIEHLAHKGV KHVTINFHNATKEEIEHFFKHAPTIEHL AHKGVKHVTINFHNATKEEIEHFFK	0.819	nan
121	repeat protein dssp2 H0.3D0.3E0.3 ps2 c1 1 000005 A	RLPTIEELKEAGIKELDIEIENPTAEEL KELFDRLPTIEELKEAGIKELDIEIENP TAEELKELFDRLPTIEELKEAGIKELDI EIENPTAEELKELFDRLPTIEELKEAGI KELDIEIENPTAEELKELFDRLPTIEEL KEAGIKELDIEIENPTAEELKELFD	0.839	nan
122	repeat protein dssp2 H0.3D0.3E0.3 ps2 c1 8 000003 A	ELTLEEIREMVEKGVRLELTITGDEFRE LIERGELTLEEIREMVEKGVRLELTITG DEFRELIERGELTLEEIREMVEKGVRLE LTITGDEFRELIERGELTLEEIREMVEK GVRLELTITGDEFRELIERGELTLEEIR EMVEKGVRLELTITGDEFRELIERG	0.905	nan
123	repeat protein dssp2 H0.3D0.3E0.3 ps3 c1 6 000008 A	LDVADVRALIEKGRIYVTVDGDDSADEA AERFGLDVADVRALIEKGRIYVTVDGDD SADEAAERFGLDVADVRALIEKGRIYVT VDGDDSADEAAERFGLDVADVRALIEKG RIYVTVDGDDSADEAAERFGLDVADVRA LIEKGRIYVTVDGDDSADEAAERFG	0.862	nan
124	repeat protein dssp2 H0.3D0.3 ps1 c1 0 000003 A	PSKADLFALLKAGKVIIHLQPEDTRDEI IKRYGPSKADLFALLKAGKVIIHLQPED TRDEIIKRYGPSKADLFALLKAGKVIIH LQPEDTRDEIIKRYGPSKADLFALLKAG KVIIHLQPEDTRDEIIKRYGPSKADLFA LLKAGKVIIHLQPEDTRDEIIKRYG	0.897	nan
125	repeat protein dssp2 H0.3D0.3 ps2 c1 5 000009 A	GKTVGELAEENGIKDLDLHFGGMSIEEI HRLLGGKTVGELAEENGIKDLDLHFGGM SIEEIHRLLGGKTVGELAEENGIKDLDL HFGGMSIEEIHRLLGGKTVGELAEENGI KDLDLHFGGMSIEEIHRLLGGKTVGELA EENGIKDLDLHFGGMSIEEIHRLLG	0.825	nan
126	repeat protein dssp2 H0.3D0.3 ps2 c1 7 000000 A	LDAEEVKRLIEDGKLHIHIDPNETIDDL CDRYDLDAEEVKRLIEDGKLHIHIDPNE TIDDLCDRYDLDAEEVKRLIEDGKLHIH IDPNETIDDLCDRYDLDAEEVKRLIEDG KLHIHIDPNETIDDLCDRYDLDAEEVKR LIEDGKLHIHIDPNETIDDLCDRYD	0.86	nan
127	repeat protein dssp2 H0.3D0.3 ps2 c1 9 000006 A	MTVDEVKKAIEAGKLKIHFHGSDTAEEI ADRFGMTVDEVKKAIEAGKLKIHFHGSD TAEEIADRFGMTVDEVKKAIEAGKLKIH FHGSDTAEEIADRFGMTVDEVKKAIEAG KLKIHFHGSDTAEEIADRFGMTVDEVKK AIEAGKLKIHFHGSDTAEEIADRFG	0.873	nan

128	repeat protein dssp2 H0.3D0.3 ps3 c1 7 000001 A	KHLTLKELVARGVKTITIHGDGITADDI RDLLDKHLTLKELVARGVKTITIHGDGI TADDIRDLLDKHLTLKELVARGVKTITI HGDGITADDIRDLLDKHLTLKELVARGV KTITIHGDGITADDIRDLLDKHLTLKEL VARGVKTITIHGDGITADDIRDLLD	0.833	nan
129	repeat protein dssp2 H0.3D0.3 ps3 c1 8 000006 A	EKITLPDLVKAGKHIHIDIKADDSKDEI DDKIDEKITLPDLVKAGKHIHIDIKADD SKDEIDDKIDEKITLPDLVKAGKHIHID IKADDSKDEIDDKIDEKITLPDLVKAGK HIHIDIKADDSKDEIDDKIDEKITLPDL VKAGKHIHIDIKADDSKDEIDDKID	0.807	nan
130	repeat protein dssp2 H0.3 ps1 c1 0 000006 A	DEISIKEAIEQGVKTIHFPGHMTAEEIE ALLKKDEISIKEAIEQGVKTIHFPGHMT AEEIEALLKKDEISIKEAIEQGVKTIHF PGHMTAEEIEALLKKDEISIKEAIEQGV KTIHFPGHMTAEEIEALLKKDEISIKEA IEQGVKTIHFPGHMTAEEIEALLKK	0.878	nan
131	repeat protein dssp2 H0.3 ps1 c1 0 000007 A	HRLSLEEAVAAGIRVTLVLHPGESLEEV LARHGHRLSLEEAVAAGIRVTLVLHPGE SLEEVLARHGHRLSLEEAVAAGIRVTLV LHPGESLEEVLARHGHRLSLEEAVAAGI RVTLVLHPGESLEEVLARHGHRLSLEEA VAAGIRVTLVLHPGESLEEVLARHG	0.869	nan
132	repeat protein dssp2 H0.3 ps3 c1 7 000004 A	HHLTLEEAVARGIDVHITIRPHHTFKAV FEAHGHHLTLEEAVARGIDVHITIRPHH TFKAVFEAHGHHLTLEEAVARGIDVHIT IRPHHTFKAVFEAHGHHLTLEEAVARGI DVHITIRPHHTFKAVFEAHGHHLTLEEA VARGIDVHITIRPHHTFKAVFEAHG	0.854	nan
133	repeat protein dssp3 D0.2 ps1 c1 4 000008 A	MTLKDFLKKEDLSPLDLAKKTGKTLDEI LDKLNMTLKDFLKKEDLSPLDLAKKTGK TLDEILDKLNMTLKDFLKKEDLSPLDLA KKTGKTLDEILDKLNMTLKDFLKKEDLS PLDLAKKTGKTLDEILDKLNMTLKDFLK KEDLSPLDLAKKTGKTLDEILDKLN	0.932	nan
134	repeat protein dssp3 D0.2 ps2 c1 1 000004 A	SADDIRDALQAGISIEDLIRAGVDEDEI ADTLGSADDIRDALQAGISIEDLIRAGV DEDEIADTLGSADDIRDALQAGISIEDL IRAGVDEDEIADTLGSADDIRDALQAGI SIEDLIRAGVDEDEIADTLGSADDIRDA LQAGISIEDLIRAGVDEDEIADTLG	0.918	nan
135	repeat protein dssp3 D0.2 ps2 c1 5 000003 A	MDILDFGKKEGKTVDDIIDKFDISAKYI AKDTGMDILDFGKKEGKTVDDIIDKFDI SAKYIAKDTGMDILDFGKKEGKTVDDII DKFDISAKYIAKDTGMDILDFGKKEGKT VDDIIDKFDISAKYIAKDTGMDILDFGK KEGKTVDDIIDKFDISAKYIAKDTG	0.907	nan
136	repeat protein dssp3 D0.2 ps2 c1 8 000003 A	MTLREFLKDEDLSLDDFIKREGKDIDDV IDKYNMTLREFLKDEDLSLDDFIKREGK DIDDVIDKYNMTLREFLKDEDLSLDDFI KREGKDIDDVIDKYNMTLREFLKDEDLS LDDFIKREGKDIDDVIDKYNMTLREFLK DEDLSLDDFIKREGKDIDDVIDKYN	0.916	nan

137	repeat protein dssp3 D0.2 ps3 c1 4 000007 A	KSALDIIDEEKMSFKKFLDDNKLSLDDF IDLTGKSALDIIDEEKMSFKKFLDDNKL SLDDFIDLTGKSALDIIDEEKMSFKKFL DDNKLSLDDFIDLTGKSALDIIDEEKMS FKKFLDDNKLSLDDFIDLTGKSALDIID EEKMSFKKFLDDNKLSLDDFIDLTG	0.936	nan
138	repeat protein dssp3 D0.2 ps3 c1 5 000009 A	AEEIVELIREGKDADDIAKILDIDKDEV KARITAEEIVELIREGKDADDIAKILDI DKDEVKARITAEEIVELIREGKDADDIA KILDIDKDEVKARITAEEIVELIREGKD ADDIAKILDIDKDEVKARITAEEIVELI REGKDADDIAKILDIDKDEVKARIT	0.914	nan
139	repeat protein dssp3 D0.3 ps1 c1 1 000007 A	IPDIVDLIRDGKTIDEIADELGKSRDEI VDDIDIPDIVDLIRDGKTIDEIADELGK SRDEIVDDIDIPDIVDLIRDGKTIDEIA DELGKSRDEIVDDIDIPDIVDLIRDGKT IDEIADELGKSRDEIVDDIDIPDIVDLI RDGKTIDEIADELGKSRDEIVDDID	0.923	nan
140	repeat protein dssp3 D0.3 ps1 c1 2 000009 A	GNLTLDDLKDAGISLKELFDKGDNDIIR ELIDAGNLTLDDLKDAGISLKELFDKGD NDIIRELIDAGNLTLDDLKDAGISLKEL FDKGDNDIIRELIDAGNLTLDDLKDAGI SLKELFDKGDNDIIRELIDAGNLTLDDL KDAGISLKELFDKGDNDIIRELIDA	0.91	nan
141	repeat protein dssp3 D0.3 ps1 c1 4 000009 A	ITIKELIDYNGLTFAELAEKNGVSLLDL AERDGITIKELIDYNGLTFAELAEKNGV SLLDLAERDGITIKELIDYNGLTFAELA EKNGVSLLDLAERDGITIKELIDYNGLT FAELAEKNGVSLLDLAERDGITIKELID YNGLTFAELAEKNGVSLLDLAERDG	0.911	nan
142	repeat protein dssp3 H0.2D0.2 ps1 c1 2 000005 A	AGDIRRLILAGITVEELQKRYDLSKEDI FHKITAGDIRRLILAGITVEELQKRYDL SKEDIFHKITAGDIRRLILAGITVEELQ KRYDLSKEDIFHKITAGDIRRLILAGIT VEELQKRYDLSKEDIFHKITAGDIRRLI LAGITVEELQKRYDLSKEDIFHKIT	0.911	nan
143	repeat protein dssp3 H0.2D0.2 ps1 c1 4 000006 A	MTLRDYLASEKLTLDELIKKNGLTIDDI LSKFNMTLRDYLASEKLTLDELIKKNGL TIDDILSKFNMTLRDYLASEKLTLDELI KKNGLTIDDILSKFNMTLRDYLASEKLT LDELIKKNGLTIDDILSKFNMTLRDYLA SEKLTLDELIKKNGLTIDDILSKFN	0.919	nan
144	repeat protein dssp3 H0.2D0.2 ps1 c1 7 000003 A	VTIEELAKELGLSKEELARRLKPEDIIR YLDRGVTIEELAKELGLSKEELARRLKP EDIIRYLDRGVTIEELAKELGLSKEELA RRLKPEDIIRYLDRGVTIEELAKELGLS KEELARRLKPEDIIRYLDRGVTIEELAK ELGLSKEELARRLKPEDIIRYLDRG	0.904	nan
145	repeat protein dssp3 H0.2D0.2 ps1 c1 9 000005 A	GLTIKDLREEGISLEEVLDTLTGDDIIK FLDLHGLTIKDLREEGISLEEVLDTLTG DDIIKFLDLHGLTIKDLREEGISLEEVL DTLTGDDIIKFLDLHGLTIKDLREEGIS LEEVLDTLTGDDIIKFLDLHGLTIKDLR EEGISLEEVLDTLTGDDIIKFLDLH	0.92	nan

146	repeat protein dssp3 H0.2D0.2 ps3 c1 7 000001 A	DKDHIISLYHEGVSIDELIALGFSKADI RAAFKDKDHIISLYHEGVSIDELIALGF SKADIRAAFKDKDHIISLYHEGVSIDEL IALGFSKADIRAAFKDKDHIISLYHEGV SIDELIALGFSKADIRAAFKDKDHIISL YHEGVSIDELIALGFSKADIRAAFK	0.918	nan
147	repeat protein dssp3 H0.2 ps1 000008 A	MSIEEFAERNGLTLVEIADKFNLTLEEL LKQSGMSIEEFAERNGLTLVEIADKFNL TLEELLKQSGMSIEEFAERNGLTLVEIA DKFNLTLEELLKQSGMSIEEFAERNGLT LVEIADKFNLTLEELLKQSGMSIEEFAE RNGLTLVEIADKFNLTLEELLKQSG	0.929	nan
148	repeat protein dssp3 H0.2 ps2 c1 1 000002 A	ITFAELLKHENLSLAEFLKHHNLSIEEI HRHHGITFAELLKHENLSLAEFLKHHNL SIEEIHRHHGITFAELLKHENLSLAEFL KHHNLSIEEIHRHHGITFAELLKHENLS LAEFLKHHNLSIEEIHRHHGITFAELLK HENLSLAEFLKHHNLSIEEIHRHHG	0.91	nan
149	repeat protein dssp3 H0.2 ps3 c1 1 000008 A	HLHQILHEIGAGATEEELLKRGFSPHHI HAAHGHLHQILHEIGAGATEEELLKRGF SPHHIHAAHGHLHQILHEIGAGATEEEL LKRGFSPHHIHAAHGHLHQILHEIGAGA TEEELLKRGFSPHHIHAAHGHLHQILHE IGAGATEEELLKRGFSPHHIHAAHG	0.914	nan
150	repeat protein dssp3 H0.3D0.3E0.3 ps2 c1 1 000002 A	AITLKDLKEAGLTIEDVLEEHGLTIEEL QELGIAITLKDLKEAGLTIEDVLEEHGL TIEELQELGIAITLKDLKEAGLTIEDVL EEHGLTIEELQELGIAITLKDLKEAGLT IEDVLEEHGLTIEELQELGIAITLKDLK EAGLTIEDVLEEHGLTIEELQELGI	0.904	nan
151	repeat protein dssp3 H0.3D0.3E0.3 ps2 c1 1 000008 A	PRELERLLREGITAEELAEKLEISKEEV KDKITPRELERLLREGITAEELAEKLEI SKEEVKDKITPRELERLLREGITAEELA EKLEISKEEVKDKITPRELERLLREGIT AEELAEKLEISKEEVKDKITPRELERLL REGITAEELAEKLEISKEEVKDKIT	0.909	nan
152	repeat protein dssp3 H0.3D0.3E0.3 ps3 c1 0 000003 A	KVEEAIRRLDEGKSHEELLKEGFTDEEI EEARKKVEEAIRRLDEGKSHEELLKEGF TDEEIEEARKKVEEAIRRLDEGKSHEEL LKEGFTDEEIEEARKKVEEAIRRLDEGK SHEELLKEGFTDEEIEEARKKVEEAIRR LDEGKSHEELLKEGFTDEEIEEARK	0.907	nan
153	repeat protein dssp3 H0.3D0.3E0.3 ps3 c1 0 000005 A	MTPEDLAKKLGKSIEELIDEGEISAEEI LKEEEMTPEDLAKKLGKSIEELIDEGEI SAEEILKEEEMTPEDLAKKLGKSIEELI DEGEISAEEILKEEEMTPEDLAKKLGKS IEELIDEGEISAEEILKEEEMTPEDLAK KLGKSIEELIDEGEISAEEILKEEE	0.931	nan
154	repeat protein dssp3 H0.3D0.3E0.3 ps3 c1 1 000000 A	MTLRELLEAGELSAEELIERHDLTIEEL IEHTGMTLRELLEAGELSAEELIERHDL TIEELIEHTGMTLRELLEAGELSAEELI ERHDLTIEELIEHTGMTLRELLEAGELS AEELIERHDLTIEELIEHTGMTLRELLE AGELSAEELIERHDLTIEELIEHTG	0.906	nan

155	repeat protein dssp3 H0.3D0.3 ps1 c1 1 000008 A	SLSILDALKAGLTLEDIAALYNKTVDEV RSELKSLSILDALKAGLTLEDIAALYNK TVDEVRSELKSLSILDALKAGLTLEDIA ALYNKTVDEVRSELKSLSILDALKAGLT LEDIAALYNKTVDEVRSELKSLSILDAL KAGLTLEDIAALYNKTVDEVRSELK	0.935	nan
156	repeat protein dssp3 H0.3D0.3 ps2 c1 7 000007 A	VKLFDVLKANNLTLDELAQLLGKSIDEI LKDHNVKLFDVLKANNLTLDELAQLLGK SIDEILKDHNVKLFDVLKANNLTLDELA QLLGKSIDEILKDHNVKLFDVLKANNLT LDELAQLLGKSIDEILKDHNVKLFDVLK ANNLTLDELAQLLGKSIDEILKDHN	0.903	nan
157	repeat protein dssp3 H0.3D0.3 ps3 c1 5 000008 A	MTLDELIDKNNITIDEFLKKNNISHLDL IKDYNMTLDELIDKNNITIDEFLKKNNI SHLDLIKDYNMTLDELIDKNNITIDEFL KKNNISHLDLIKDYNMTLDELIDKNNIT IDEFLKKNNISHLDLIKDYNMTLDELID KNNITIDEFLKKNNISHLDLIKDYN	0.915	nan
158	repeat protein dssp4 D0.2 ps1 c1 5 000002 A	AEDVDVTIERDDDGATIASAVVDGKRYS FTFPDAEDVDVTIERDDDGATIASAVVD GKRYSFTFPDAEDVDVTIERDDDGATIA SAVVDGKRYSFTFPDAEDVDVTIERDDD GATIASAVVDGKRYSFTFPDAEDVDVTI ERDDDGATIASAVVDGKRYSFTFPD	0.78	nan
159	repeat protein dssp4 D0.3 ps1 c1 3 000009 A	NVDNVTITKDDDSGKVKVTADFDGKDFT ATFPDNVDNVTITKDDDSGKVKVTADFD GKDFTATFPDNVDNVTITKDDDSGKVKV TADFDGKDFTATFPDNVDNVTITKDDDS GKVKVTADFDGKDFTATFPDNVDNVTIT KDDDSGKVKVTADFDGKDFTATFPD	0.832	nan
160	repeat protein dssp4 H0.2D0.2 ps2 c1 2 000005 A	DDDVSVLVHRTEDGHDDVSLHIHGKTYR VHVNPDDDVSVLVHRTEDGHDDVSLHIH GKTYRVHVNPDDDVSVLVHRTEDGHDDV SLHIHGKTYRVHVNPDDDVSVLVHRTED GHDDVSLHIHGKTYRVHVNPDDDVSVLV HRTEDGHDDVSLHIHGKTYRVHVNP	0.746	nan
161	repeat protein dssp4 H0.3D0.3E0.3 ps3 c1 2 000000 A	EDREYRITLHPEFPEAEIELDEDGRLEI TVRDEEDREYRITLHPEFPEAEIELDED GRLEITVRDEEDREYRITLHPEFPEAEI ELDEDGRLEITVRDEEDREYRITLHPEF PEAEIELDEDGRLEITVRDEEDREYRIT LHPEFPEAEIELDEDGRLEITVRDE	0.898	nan
162	repeat protein dssp0 H0.2D0.2 ps1 c2 9 000004 A	GADTVTLHFESDDGLTEEEIQRLRELIA EAIKQGADTVTLHFESDDGLTEEEIQRL RELIAEAIKQGADTVTLHFESDDGLTEE EIQRLRELIAEAIKQGADTVTLHFESDD GLTEEEIQRLRELIAEAIKQGADTVTLH FESDDGLTEEEIQRLRELIAEAIKQ	0.784	nan
163	repeat protein dssp0 H0.2D0.2 ps2 c2 0 000000 A	MKHRVILRFHSPKRHGLTDDDIIKFAED LRNAGMKHRVILRFHSPKRHGLTDDDII KFAEDLRNAGMKHRVILRFHSPKRHGLT DDDIIKFAEDLRNAGMKHRVILRFHSPK RHGLTDDDIIKFAEDLRNAGMKHRVILR FHSPKRHGLTDDDIIKFAEDLRNAG	0.792	nan

164	repeat protein dssp0 H0.2D0.2 ps2 c2 7 000005 A	HATIDLHIDISRILHEGHSHDEIIDLIK RLKDDHATIDLHIDISRILHEGHSHDEI IDLIKRLKDDHATIDLHIDISRILHEGH SHDEIIDLIKRLKDDHATIDLHIDISRI LHEGHSHDEIIDLIKRLKDDHATIDLHI DISRILHEGHSHDEIIDLIKRLKDD	0.927	nan
165	repeat protein dssp0 H0.2D0.2 ps3 c2 5 000000 A	SHGLTVVLTIHLRDDDLDEDEFSHALDK AKHLKSHGLTVVLTIHLRDDDLDEDEFS HALDKAKHLKSHGLTVVLTIHLRDDDLD EDEFSHALDKAKHLKSHGLTVVLTIHLR DDDLDEDEFSHALDKAKHLKSHGLTVVL TIHLRDDDLDEDEFSHALDKAKHLK	0.799	nan
166	repeat protein dssp0 H0.2 ps1 c2 2 000009 A	GIKAVAHIHLGEAVKAGISPEEAIKLAK ELEKEGIKAVAHIHLGEAVKAGISPEEA IKLAKELEKEGIKAVAHIHLGEAVKAGI SPEEAIKLAKELEKEGIKAVAHIHLGEA VKAGISPEEAIKLAKELEKEGIKAVAHI HLGEAVKAGISPEEAIKLAKELEKE	0.896	nan
167	repeat protein dssp0 H0.2 ps1 c2 4 000002 A	AGVKTLHLHFHSPEHVKAFGLEEFEKLI KEAAAAGVKTLHLHFHSPEHVKAFGLEE FEKLIKEAAAAGVKTLHLHFHSPEHVKA FGLEEFEKLIKEAAAAGVKTLHLHFHSP EHVKAFGLEEFEKLIKEAAAAGVKTLHL HFHSPEHVKAFGLEEFEKLIKEAAA	0.838	nan
168	repeat protein dssp0 H0.2 ps1 c2 4 000003 A	HGVEVKIIVEVKPGEDLSGQIGHIKDLI EHLKKHGVEVKIIVEVKPGEDLSGQIGH IKDLIEHLKKHGVEVKIIVEVKPGEDLS GQIGHIKDLIEHLKKHGVEVKIIVEVKP GEDLSGQIGHIKDLIEHLKKHGVEVKII VEVKPGEDLSGQIGHIKDLIEHLKK	0.767	nan
169	repeat protein dssp0 H0.2 ps2 c2 8 000005 A	MGIDCVNISAHHPHMTAEEAEQQLLEFI EKAAEMGIDCVNISAHHPHMTAEEAEQQ LLEFIEKAAEMGIDCVNISAHHPHMTAE EAEQQLLEFIEKAAEMGIDCVNISAHHP HMTAEEAEQQLLEFIEKAAEMGIDCVNI SAHHPHMTAEEAEQQLLEFIEKAAE	0.798	nan
170	repeat protein dssp0 H0.2 ps2 c2 9 000005 A	HHKDEVHIHLSHPHLSIEEVREIIKEGI KEARKHHKDEVHIHLSHPHLSIEEVREI IKEGIKEARKHHKDEVHIHLSHPHLSIE EVREIIKEGIKEARKHHKDEVHIHLSHP HLSIEEVREIIKEGIKEARKHHKDEVHI HLSHPHLSIEEVREIIKEGIKEARK	0.855	nan
171	repeat protein dssp0 H0.2 ps3 c2 0 000001 A	QGIKSVHIRLTPPDLTAEEVKDIHLELI AKAAKQGIKSVHIRLTPPDLTAEEVKDI HLELIAKAAKQGIKSVHIRLTPPDLTAE EVKDIHLELIAKAAKQGIKSVHIRLTPP DLTAEEVKDIHLELIAKAAKQGIKSVHI RLTPPDLTAEEVKDIHLELIAKAAK	0.773	nan
211	repeat protein hb0.1sb0.4lb0.2 sym5 H0.3D0.3E0.3 ps1 c1 run3 99 000004 model 4 ptm seed 0 unrelaxed A	IEVKEVEGTPLYELEIDGKLYVFDPKTG EFFEAIEVKEVEGTPLYELEIDGKLYVF DPKTGEFFEAIEVKEVEGTPLYELEIDG KLYVFDPKTGEFFEAIEVKEVEGTPLYE LEIDGKLYVFDPKTGEFFEAIEVKEVEG TPLYELEIDGKLYVFDPKTGEFFEA	0.89	nan

212	repeat protein hb0.1sb0lb0.4 sym5 H0.2 ps1 c1 run3 41 000004 model 4 ptm seed 0 unrelaxed A	PTPDAIIAAIKAALQALPPGVKKVVIHL TGDDVPTPDAIIAAIKAALQALPPGVKK VVIHLTGDDVPTPDAIIAAIKAALQALP PGVKKVVIHLTGDDVPTPDAIIAAIKAA LQALPPGVKKVVIHLTGDDVPTPDAIIA AIKAALQALPPGVKKVVIHLTGDDV	0.796	nan
213	repeat protein hb0.1sb0lb0 sym5 H0.3D0.3 ps2 c1 run3 62 000000 model 4 ptm seed 0 unrelaxed A	DKIEVTIHGGDDDDLLDRLGALIKAGVF KLEDIDKIEVTIHGGDDDDLLDRLGALI KAGVFKLEDIDKIEVTIHGGDDDDLLDR LGALIKAGVFKLEDIDKIEVTIHGGDDD DLLDRLGALIKAGVFKLEDIDKIEVTIH GGDDDDLLDRLGALIKAGVFKLEDI	0.842	nan
214	repeat protein hb0.2sb0.2lb0.2 sym5 H0.2 ps3 c1 run3 50 000004 model 4 ptm seed 0 unrelaxed A	HHPYSLSELARRHGLSVEEIRKHIEAGH LIIHAHHPYSLSELARRHGLSVEEIRKH IEAGHLIIHAHHPYSLSELARRHGLSVE EIRKHIEAGHLIIHAHHPYSLSELARRH GLSVEEIRKHIEAGHLIIHAHHPYSLSE LARRHGLSVEEIRKHIEAGHLIIHA	0.835	nan
215	repeat protein hb0.2sb0.4lb0 sym5 H0.3D0.3E0.3 ps1 c1 run3 58 000004 model 4 ptm seed 0 unrelaxed A	GTLTLTITDRSPDGRLYEGEATVSVPPS ALEPGGTLTLTITDRSPDGRLYEGEATV SVPPSALEPGGTLTLTITDRSPDGRLYE GEATVSVPPSALEPGGTLTLTITDRSPD GRLYEGEATVSVPPSALEPGGTLTLTIT DRSPDGRLYEGEATVSVPPSALEPG	0.725	nan
216	repeat protein hb0.2sb0lb0.2 sym5 H0.3D0.3 ps3 c1 run3 66 000003 model 4 ptm seed 0 unrelaxed A	HKLVITGDDIIDLLRDGKSLDEIKDFLD RHGDDHKLVITGDDIIDLLRDGKSLDEI KDFLDRHGDDHKLVITGDDIIDLLRDGK SLDEIKDFLDRHGDDHKLVITGDDIIDL LRDGKSLDEIKDFLDRHGDDHKLVITGD DIIDLLRDGKSLDEIKDFLDRHGDD	0.78	nan
217	repeat protein hb0.2sb0lb0.2 sym5 H0.3 ps3 c1 run3 52 000002 model 4 ptm seed 0 unrelaxed A	HHHHLLKQHPDLTFEELQHFLEEHAEQG HIVHIHHHHLLKQHPDLTFEELQHFLEE HAEQGHIVHIHHHHLLKQHPDLTFEELQ HFLEEHAEQGHIVHIHHHHLLKQHPDLT FEELQHFLEEHAEQGHIVHIHHHHLLKQ HPDLTFEELQHFLEEHAEQGHIVHI	0.799	nan
218	repeat protein hb0.2sb0lb0 sym5 H0.2 ps3 c1 run3 31 000004 model 4 ptm seed 0 unrelaxed A	MTFAEIERLLHAGHKLSAAELHSLLVHL HEDGHMTFAEIERLLHAGHKLSAAELHS LLVHLHEDGHMTFAEIERLLHAGHKLSA AELHSLLVHLHEDGHMTFAEIERLLHAG HKLSAAELHSLLVHLHEDGHMTFAEIER LLHAGHKLSAAELHSLLVHLHEDGH	0.91	nan
219	repeat protein hb0.4sb0.2lb0 sym5 H0.2 ps1 c1 run3 14 000001 model 4 ptm seed 0 unrelaxed A	VLTKSELHKLAHEHGLTPEELIRLLVKL KKAGHVLTKSELHKLAHEHGLTPEELIR LLVKLKKAGHVLTKSELHKLAHEHGLTP EELIRLLVKLKKAGHVLTKSELHKLAHE HGLTPEELIRLLVKLKKAGHVLTKSELH KLAHEHGLTPEELIRLLVKLKKAGH	0.869	nan
220	repeat protein hb0.4sb0.2lb0 sym5 H0.2 ps2 c1 run3 21 000009 model 4 ptm seed 0 unrelaxed A	QELHRLLAHLIHAGKLTVAELHHFLEHH PDLTPQELHRLLAHLIHAGKLTVAELHH FLEHHPDLTPQELHRLLAHLIHAGKLTV AELHHFLEHHPDLTPQELHRLLAHLIHA GKLTVAELHHFLEHHPDLTPQELHRLLA HLIHAGKLTVAELHHFLEHHPDLTP	0.862	nan

221	repeat protein hb0.4sb0lb0.2 sym5 H0.3D0.3 ps1 c1 run3 88 000003 model 4 ptm seed 0 unrelaxed A	KIKKDMTADELIAFAEGQNLSANQIVEL LEALPKIKKDMTADELIAFAEGQNLSAN QIVELLEALPKIKKDMTADELIAFAEGQ NLSANQIVELLEALPKIKKDMTADELIA FAEGQNLSANQIVELLEALPKIKKDMTA DELIAFAEGQNLSANQIVELLEALP	0.852	nan
222	repeat protein hb0.4sb0lb0.4 sym5 H0.3D0.3 ps2 c1 run3 73 000004 model 4 ptm seed 0 unrelaxed A	DHKDVLDKLIEKLVAGKISLEEIEAFLK KHPDIDHKDVLDKLIEKLVAGKISLEEI EAFLKKHPDIDHKDVLDKLIEKLVAGKI SLEEIEAFLKKHPDIDHKDVLDKLIEKL VAGKISLEEIEAFLKKHPDIDHKDVLDK LIEKLVAGKISLEEIEAFLKKHPDI	0.892	nan
223	repeat protein hb0.4sb0lb0.4 sym5 H0.3 ps1 c1 run3 0 000005 model 4 ptm seed 0 unrelaxed A	MTLEELVELARAGIRLTIHLGHLPPHVH ELARRMTLEELVELARAGIRLTIHLGHL PPHVHELARRMTLEELVELARAGIRLTI HLGHLPPHVHELARRMTLEELVELARAG IRLTIHLGHLPPHVHELARRMTLEELVE LARAGIRLTIHLGHLPPHVHELARR	0.781	nan
224	repeat protein hb0.4sb0lb0 sym5 H0.3 ps1 c1 run3 71 000003 model 4 ptm seed 0 unrelaxed A	RAEEVRHFIEKALELLLAGKLTVEELHK LLSHLRAEEVRHFIEKALELLLAGKLTV EELHKLLSHLRAEEVRHFIEKALELLLA GKLTVEELHKLLSHLRAEEVRHFIEKAL ELLLAGKLTVEELHKLLSHLRAEEVRHF IEKALELLLAGKLTVEELHKLLSHL	0.88	nan
225	repeat protein hb0.6sb0.2lb0.2 sym5 H0.2D0.2 ps1 c1 run3 19 000000 model 4 ptm seed 0 unrelaxed A	SLLERIESGKLTFEELTPRDIRNLVKSG YLTYKSLLERIESGKLTFEELTPRDIRN LVKSGYLTYKSLLERIESGKLTFEELTP RDIRNLVKSGYLTYKSLLERIESGKLTF EELTPRDIRNLVKSGYLTYKSLLERIES GKLTFEELTPRDIRNLVKSGYLTYK	0.886	nan
226	repeat protein hb0.6sb0.2lb0.2 sym5 H0.2D0.2 ps2 c1 run3 7 000005 model 4 ptm seed 0 unrelaxed A	AEKKDIIKDLIEDLTPEEKIRLIEELLH HDHIDAEKKDIIKDLIEDLTPEEKIRLI EELLHHDHIDAEKKDIIKDLIEDLTPEE KIRLIEELLHHDHIDAEKKDIIKDLIED LTPEEKIRLIEELLHHDHIDAEKKDIIK DLIEDLTPEEKIRLIEELLHHDHID	0.858	nan
227	repeat protein hb0.6sb0.4lb0.2 sym5 H0.2 ps1 c1 run3 21 000004 model 4 ptm seed 0 unrelaxed A	KLSAEELVELLEKSHLVHHLTFEEILKL LKAGVKLSAEELVELLEKSHLVHHLTFE EILKLLKAGVKLSAEELVELLEKSHLVH HLTFEEILKLLKAGVKLSAEELVELLEK SHLVHHLTFEEILKLLKAGVKLSAEELV ELLEKSHLVHHLTFEEILKLLKAGV	0.882	nan
228	repeat protein hb0.6sb0.4lb0 sym5 H0.3 ps2 c1 run3 9 000006 model 4 ptm seed 0 unrelaxed A	HVHLHGSHIVRMLELGIDLPELIRHLRE RGIRIHVHLHGSHIVRMLELGIDLPELI RHLRERGIRIHVHLHGSHIVRMLELGID LPELIRHLRERGIRIHVHLHGSHIVRML ELGIDLPELIRHLRERGIRIHVHLHGSH IVRMLELGIDLPELIRHLRERGIRI	0.864	nan
229	repeat protein hb0.6sb0lb0 sym5 H0.2 ps1 c1 run3 58 000002 model 4 ptm seed 0 unrelaxed A	VDADLVVAFFHHIPPPSIEELITLLEKL VELGFVDADLVVAFFHHIPPPSIEELIT LLEKLVELGFVDADLVVAFFHHIPPPSI EELITLLEKLVELGFVDADLVVAFFHHI PPPSIEELITLLEKLVELGFVDADLVVA FFHHIPPPSIEELITLLEKLVELGF	0.757	nan

230	repeat protein hb0.6sb0lb0 sym5 H0.3 ps2 c1 run3 48 000001 model 4 ptm seed 0 unrelaxed A	HLSHRQIRRLIHLVHKGKISGEELLEHL HEHKVHLSHRQIRRLIHLVHKGKISGEE LLEHLHEHKVHLSHRQIRRLIHLVHKGK ISGEELLEHLHEHKVHLSHRQIRRLIHL VHKGKISGEELLEHLHEHKVHLSHRQIR RLIHLVHKGKISGEELLEHLHEHKV	0.891	nan
231	repeat protein dssp0 H0.2D0.2 ps2 c1 4 sym5 cap10 000003 model 4 ptm seed 0 unrelaxed A	SKEELKERIEAGHRVLSFHYDDGEDPLE RHGLTLRDAIEEAGRAGHRVLSFHYDDG EDPLERHGLTLRDAIEEAGRAGHRVLSF HYDDGEDPLERHGLTLRDAIEEAGRAGH RVLSFHYDDGEDPLERHGLTLRDAIEEA GRAGHRVLSFHYDDGEDPLERHGLTLRD AIEEAGRIDHVAVAFSD	0.857	nan
232	repeat protein dssp0 H0.2D0.2 ps2 c1 6 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	KLKELLDYIIDHPGVRFILDAHELVKDL NLTFDEIKQFIRSLSDHPGVRFILDAHE LVKDLNLTFDEIKQFIRSLSDHPGVRFI LDAHELVKDLNLTFDEIKQFIRSLSDHP GVRFILDAHELVKDLNLTFDEIKQFIRS LSDHPGVRFILDAHELVKDLNLTFDEIK QFIRSLSDDERALLEES	0.882	nan
233	repeat protein dssp0 H0.2 ps1 c1 6 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	MRLREAIEAEIASPIAYIHIDSPADLEA AGITGEEIRAFIERAIASPIAYIHIDSP ADLEAAGITGEEIRAFIERAIASPIAYI HIDSPADLEAAGITGEEIRAFIERAIAS PIAYIHIDSPADLEAAGITGEEIRAFIE RAIASPIAYIHIDSPADLEAAGITGEEI RAFIERAVTLLEAILAR	0.832	nan
234	repeat protein dssp0 H0.2 ps1 c1 9 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	HHIEEIKRAVAAGAKVIVHLEGAHVLRL LERGVDIVELVRELVAAGAKVIVHLEGA HVLRLLERGVDIVELVRELVAAGAKVIV HLEGAHVLRLLERGVDIVELVRELVAAG AKVIVHLEGAHVLRLLERGVDIVELVRE LVAAGAKVIVHLEGAHVLRLLERGVDIV ELVRELVAIHEEEEKKA	0.963	nan
235	repeat protein dssp0 H0.3D0.3E0.3 ps3 c1 9 sym5 cap10 000008 model 4 ptm seed 0 unrelaxed A	HLEELEAELEAGEKEITLHFDFGKDGHS IEELEESLELLDEHNAGEKEITLHFDFG KDGHSIEELEESLELLDEHNAGEKEITL HFDFGKDGHSIEELEESLELLDEHNAGE KEITLHFDFGKDGHSIEELEESLELLDE HNAGEKEITLHFDFGKDGHSIEELEESL ELLDEHNIEAEVEIDEG	0.902	nan
236	repeat protein dssp1 H0.2D0.2 ps1 c1 1 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	GEPLREALRAARPGDFLIVKGGLTADEA RELAPGAVLIAHAKEARPGDFLIVKGGL TADEARELAPGAVLIAHAKEARPGDFLI VKGGLTADEARELAPGAVLIAHAKEARP GDFLIVKGGLTADEARELAPGAVLIAHA KEARPGDFLIVKGGLTADEARELAPGAV LIAHAKEPEAARAAAAA	0.863	nan
237	repeat protein dssp1 H0.2D0.2 ps2 c2 0 sym5 cap10 000005 model 4 ptm seed 0 unrelaxed A	SKLDQLVAFCKKHGARIILESGIDAEEF RAAGVDIFLHHHADAKKHGARIILESGI DAEEFRAAGVDIFLHHHADAKKHGARII LESGIDAEEFRAAGVDIFLHHHADAKKH GARIILESGIDAEEFRAAGVDIFLHHHA DAKKHGARIILESGIDAEEFRAAGVDIF LHHHADADVEEVLARLE	0.823	nan

238	repeat protein dssp1 H0.2D0.2 ps3 c1 9 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	SEDDLKDFVKKHPDIILHVRKGDDIEEY RRAGCKHILAHSDHPKHPDIILHVRKGD DIEEYRRAGCKHILAHSDHPKHPDIILH VRKGDDIEEYRRAGCKHILAHSDHPKHP DIILHVRKGDDIEEYRRAGCKHILAHSD HPKHPDIILHVRKGDDIEEYRRAGCKHI LAHSDHPPGDILRLLDR	0.86	nan
239	repeat protein dssp1 H0.2D0.2 ps3 c2 3 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	DLDDIRDAARHGVDIVVFDAPTDEDVAW ARRHGLKIIADHDLHHGVDIVVFDAPTD EDVAWARRHGLKIIADHDLHHGVDIVVF DAPTDEDVAWARRHGLKIIADHDLHHGV DIVVFDAPTDEDVAWARRHGLKIIADHD LHHGVDIVVFDAPTDEDVAWARRHGLKI IADHDLHPDHIAAIAKH	0.857	nan
240	repeat protein dssp1 H0.3D0.3E0.3 ps1 c1 0 sym5 cap10 000000 model 4 ptm seed 0 unrelaxed A	SAEEAIRALLAEPELIVALGEGADVERF RAAGFRVIVHGEDPPAEPELIVALGEGA DVERFRAAGFRVIVHGEDPPAEPELIVA LGEGADVERFRAAGFRVIVHGEDPPAEP ELIVALGEGADVERFRAAGFRVIVHGED PPAEPELIVALGEGADVERFRAAGFRVI VHGEDPPLEELLAARRR	0.833	nan
241	repeat protein dssp1 H0.3D0.3E0.3 ps2 c1 5 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	AEREIEAAKAAGADALIVEEGGITVEEA RAAGIDFVIVEEDAPAGADALIVEEGGI TVEEARAAGIDFVIVEEDAPAGADALIV EEGGITVEEARAAGIDFVIVEEDAPAGA DALIVEEGGITVEEARAAGIDFVIVEED APAGADALIVEEGGITVEEARAAGIDFV IVEEDAPEVEAAIERAE	0.857	nan
242	repeat protein dssp1 H0.3D0.3 ps1 c2 0 sym5 cap10 000004 model 4 ptm seed 0 unrelaxed A	SLDELRAAARRGVDFLVIPADYASDELI RRLEGYRLILVGGPVRGVDFLVIPADYA SDELIRRLEGYRLILVGGPVRGVDFLVI PADYASDELIRRLEGYRLILVGGPVRGV DFLVIPADYASDELIRRLEGYRLILVGG PVRGVDFLVIPADYASDELIRRLEGYRL ILVGGPVTPEELRRFLA	0.864	nan
243	repeat protein dssp2 H0.2D0.2 ps1 c2 1 sym5 cap10 000006 model 4 ptm seed 0 unrelaxed A	VLSFEELRRRGFTKDEIIALARAGVRIE FGPGVTAAELRAFFEGFTKDEIIALARA GVRIEFGPGVTAAELRAFFEGFTKDEII ALARAGVRIEFGPGVTAAELRAFFEGFT KDEIIALARAGVRIEFGPGVTAAELRAF FEGFTKDEIIALARAGVRIEFGPGVTAA ELRAFFEDEVIAALEAA	0.882	nan
244	repeat protein dssp2 H0.2D0.2 ps2 c1 4 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	DIKKAILSDEEDLIKDAAEKGIKTVDID IHDPKLTAEDMKHIAEDLIKDAAEKGIK TVDIDIHDPKLTAEDMKHIAEDLIKDAA EKGIKTVDIDIHDPKLTAEDMKHIAEDL IKDAAEKGIKTVDIDIHDPKLTAEDMKH IAEDLIKDAAEKGIKTVDIDIHDPKLTA EDMKHIAHDLDKIKEKL	0.873	nan
245	repeat protein dssp2 H0.2D0.2 ps3 c1 5 sym5 cap10 000000 model 4 ptm seed 0 unrelaxed A	HHDHLDELKAKDITIDELINKGAKIEIH IHGDNVDHIRKFLDDKDITIDELINKGA KIEIHIHGDNVDHIRKFLDDKDITIDEL INKGAKIEIHIHGDNVDHIRKFLDDKDI TIDELINKGAKIEIHIHGDNVDHIRKFL DDKDITIDELINKGAKIEIHIHGDNVDH IRKFLDDHRDELRDLHA	0.852	nan

246	repeat protein dssp2 H0.2D0.2 ps3 c1 7 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	HDDLIEYLEKHNLSLRDLIAKGVDIHIH VDGHDLERLRDFLKDHNLSLRDLIAKGV DIHIHVDGHDLERLRDFLKDHNLSLRDL IAKGVDIHIHVDGHDLERLRDFLKDHNL SLRDLIAKGVDIHIHVDGHDLERLRDFL KDHNLSLRDLIAKGVDIHIHVDGHDLER LRDFLKDVIEELERHKK	0.882	nan
247	repeat protein dssp2 H0.2 ps1 c1 0 sym5 cap10 000004 model 4 ptm seed 0 unrelaxed A	SAPERLLKRHGLTVKEAIEAGFPHLHVT IKPNDTAEEIRARLGGLTVKEAIEAGFP HLHVTIKPNDTAEEIRARLGGLTVKEAI EAGFPHLHVTIKPNDTAEEIRARLGGLT VKEAIEAGFPHLHVTIKPNDTAEEIRAR LGGLTVKEAIEAGFPHLHVTIKPNDTAE EIRARLGPEDFAFLKAI	0.878	nan
248	repeat protein dssp2 H0.2 ps1 c1 5 sym5 cap10 000008 model 4 ptm seed 0 unrelaxed A	GLELERRIEAGEISIPELIEAGVEIILT LHPSDTRAHIRHRLGGEISIPELIEAGV EIILTLHPSDTRAHIRHRLGGEISIPEL IEAGVEIILTLHPSDTRAHIRHRLGGEI SIPELIEAGVEIILTLHPSDTRAHIRHR LGGEISIPELIEAGVEIILTLHPSDTRA HIRHRLGPAEIAALEAE	0.862	nan
249	repeat protein dssp2 H0.2 ps2 c1 5 sym5 cap10 000005 model 4 ptm seed 0 unrelaxed A	GRHLAEFLQKHKITIEELIHLGVHIHIH FKAGDTLEELHERLGHKITIEELIHLGV HIHIHFKAGDTLEELHERLGHKITIEEL IHLGVHIHIHFKAGDTLEELHERLGHKI TIEELIHLGVHIHIHFKAGDTLEELHER LGHKITIEELIHLGVHIHIHFKAGDTLE ELHERLGLSEEEIRRIL	0.891	nan
250	repeat protein dssp2 H0.2 ps2 c2 3 sym5 cap10 000003 model 4 ptm seed 0 unrelaxed A	SLAHLRHLKRHGYTIEELKAAGIHIHLH IHPGDSLAKITKALGHGYTIEELKAAGI HIHLHIHPGDSLAKITKALGHGYTIEEL KAAGIHIHLHIHPGDSLAKITKALGHGY TIEELKAAGIHIHLHIHPGDSLAKITKA LGHGYTIEELKAAGIHIHLHIHPGDSLA KITKALGSPEEVKRLIH	0.871	nan
251	repeat protein dssp2 H0.2 ps3 c1 2 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	GERAKKLAEELNLSIEELIEKGAHVVII LHPHDTEEEIKAHYGLNLSIEELIEKGA HVVIILHPHDTEEEIKAHYGLNLSIEEL IEKGAHVVIILHPHDTEEEIKAHYGLNL SIEELIEKGAHVVIILHPHDTEEEIKAH YGLNLSIEELIEKGAHVVIILHPHDTEE EIKAHYGEEHLAALHQA	0.918	nan
252	repeat protein dssp2 H0.3D0.3E0.3 ps2 c1 8 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	EEEIVKRIEEGDLTLEEIKEAGIHELHF AGEELTAEEVRAIFGGDLTLEEIKEAGI HELHFAGEELTAEEVRAIFGGDLTLEEI KEAGIHELHFAGEELTAEEVRAIFGGDL TLEEIKEAGIHELHFAGEELTAEEVRAI FGGDLTLEEIKEAGIHELHFAGEELTAE EVRAIFGESDEEIEEAE	0.872	nan
253	repeat protein dssp2 H0.3D0.3E0.3 ps3 c1 3 sym5 cap10 000002 model 4 ptm seed 0 unrelaxed A	IHDLKEYLEKNNISLREFLDENPDHKFE IHPGDLEELKELLEENNISLREFLDENP DHKFEIHPGDLEELKELLEENNISLREF LDENPDHKFEIHPGDLEELKELLEENNI SLREFLDENPDHKFEIHPGDLEELKELL EENNISLREFLDENPDHKFEIHPGDLEE LKELLEEEKEKEEAEKE	0.918	nan

254	repeat protein dssp2 H0.3D0.3 ps1 c1 8 sym5 cap10 000004 model 4 ptm seed 0 unrelaxed A	SLDELERAFAEGKTLEELKELGIPITID VSGDEIDRLRDFLEREGKTLEELKELGI PITIDVSGDEIDRLRDFLEREGKTLEEL KELGIPITIDVSGDEIDRLRDFLEREGK TLEELKELGIPITIDVSGDEIDRLRDFL EREGKTLEELKELGIPITIDVSGDEIDR LRDFLERDEELAAAIEI	0.932	nan
255	repeat protein dssp2 H0.3D0.3 ps2 c1 2 sym5 cap10 000002 model 4 ptm seed 0 unrelaxed A	HLTDDDLRAAGLTLKEAIRLGIEGITLT VRPDDSADDFRDRFGGLTLKEAIRLGIE GITLTVRPDDSADDFRDRFGGLTLKEAI RLGIEGITLTVRPDDSADDFRDRFGGLT LKEAIRLGIEGITLTVRPDDSADDFRDR FGGLTLKEAIRLGIEGITLTVRPDDSAD DFRDRFGDEDIEAMQAA	0.898	nan
256	repeat protein dssp2 H0.3D0.3 ps2 c2 0 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	GEIIDEAKKENLSLPDLLKKNNIKTIHF DGDDLDKLLEFLKENNLSLPDLLKKNNI KTIHFDGDDLDKLLEFLKENNLSLPDLL KKNNIKTIHFDGDDLDKLLEFLKENNLS LPDLLKKNNIKTIHFDGDDLDKLLEFLK ENNLSLPDLLKKNNIKTIHFDGDDLDKL LEFLKENNEEDLLKLIE	0.918	nan
257	repeat protein dssp2 H0.3 ps1 c1 3 sym5 cap10 000006 model 4 ptm seed 0 unrelaxed A	IEQVEEHLKKHGKSLEELLALGVKLEIT VHGHELQKLEDFLERHGKSLEELLALGV KLEITVHGHELQKLEDFLERHGKSLEEL LALGVKLEITVHGHELQKLEDFLERHGK SLEELLALGVKLEITVHGHELQKLEDFL ERHGKSLEELLALGVKLEITVHGHELQK LEDFLERPRVREAIKAY	0.944	nan
258	repeat protein dssp2 H0.3 ps1 c1 9 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	GLEELKKLLEQNLTLDELIERGIRIELH FKPGDSLEEFTEVLNQNLTLDELIERGI RIELHFKPGDSLEEFTEVLNQNLTLDEL IERGIRIELHFKPGDSLEEFTEVLNQNL TLDELIERGIRIELHFKPGDSLEEFTEV LNQNLTLDELIERGIRIELHFKPGDSLE EFTEVLNEEELEALIKA	0.888	nan
259	repeat protein dssp3 H0.2D0.2 ps1 c1 2 sym5 cap10 000004 model 4 ptm seed 0 unrelaxed A	GKRELIASIRSLTKEEIKALGISIDELA KKHGFSKDELIELLKSLTKEEIKALGIS IDELAKKHGFSKDELIELLKSLTKEEIK ALGISIDELAKKHGFSKDELIELLKSLT KEEIKALGISIDELAKKHGFSKDELIEL LKSLTKEEIKALGISIDELAKKHGFSKD ELIELLKDEEKRAIEEA	0.872	nan
260	repeat protein dssp3 H0.2D0.2 ps1 c1 3 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	SKKKLERKLIARRIDELILRGKSAEEIA HILNLSIEEIKRYISARRIDELILRGKS AEEIAHILNLSIEEIKRYISARRIDELI LRGKSAEEIAHILNLSIEEIKRYISARR IDELILRGKSAEEIAHILNLSIEEIKRY ISARRIDELILRGKSAEEIAHILNLSIE EIKRYISDEEILKLKAI	0.956	nan
261	repeat protein dssp3 H0.2D0.2 ps1 c1 7 sym5 cap10 000005 model 4 ptm seed 0 unrelaxed A	AREAKEKRRRAHKLIAELGSGKLTVEEL RAMNIDGRELIKEGGAHKLIAELGSGKL TVEELRAMNIDGRELIKEGGAHKLIAEL GSGKLTVEELRAMNIDGRELIKEGGAHK LIAELGSGKLTVEELRAMNIDGRELIKE GGAHKLIAELGSGKLTVEELRAMNIDGR ELIKEGGEDAKAKLRLI	0.925	nan

262	repeat protein dssp3 H0.2D0.2 ps1 c2 3 sym5 cap10 000003 model 4 ptm seed 0 unrelaxed A	AAKREKRKELREDIDRLIQAGYTAEEIA KKLGLTVEEIKHFISREDIDRLIQAGYT AEEIAKKLGLTVEEIKHFISREDIDRLI QAGYTAEEIAKKLGLTVEEIKHFISRED IDRLIQAGYTAEEIAKKLGLTVEEIKHF ISREDIDRLIQAGYTAEEIAKKLGLTVE EIKHFISDDKLAKYERQ	0.951	nan
263	repeat protein dssp3 H0.2D0.2 ps2 c1 1 sym5 cap10 000006 model 4 ptm seed 0 unrelaxed A	HKEERELKERKDEIYDLIKQGKDADELA DLLKLSVEEIIKLISKDEIYDLIKQGKD ADELADLLKLSVEEIIKLISKDEIYDLI KQGKDADELADLLKLSVEEIIKLISKDE IYDLIKQGKDADELADLLKLSVEEIIKL ISKDEIYDLIKQGKDADELADLLKLSVE EIIKLISSDEKDDLERD	0.95	nan
264	repeat protein dssp3 H0.2D0.2 ps2 c1 3 sym5 cap10 000005 model 4 ptm seed 0 unrelaxed A	SPEDVKRLIKAGASIRDLVDAGITKEDI EAAGVHIADILKHDPAGASIRDLVDAGI TKEDIEAAGVHIADILKHDPAGASIRDL VDAGITKEDIEAAGVHIADILKHDPAGA SIRDLVDAGITKEDIEAAGVHIADILKH DPAGASIRDLVDAGITKEDIEAAGVHIA DILKHDPPSDEDLKALD	0.931	nan
265	repeat protein dssp3 H0.2D0.2 ps2 c1 5 sym5 cap10 000005 model 4 ptm seed 0 unrelaxed A	ADKKKKEDEYKELHDEFIAGKLTIEDLA KKLDKTKDEIIDHFRKELHDEFIAGKLT IEDLAKKLDKTKDEIIDHFRKELHDEFI AGKLTIEDLAKKLDKTKDEIIDHFRKEL HDEFIAGKLTIEDLAKKLDKTKDEIIDH FRKELHDEFIAGKLTIEDLAKKLDKTKD EIIDHFRDEDRDKLKKL	0.946	nan
266	repeat protein dssp3 H0.2D0.2 ps2 c2 0 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	MADAKKKFEDGELSPREIIDEHSADDLK KLGITADDIHHFIEKGELSPREIIDEHS ADDLKKLGITADDIHHFIEKGELSPREI IDEHSADDLKKLGITADDIHHFIEKGEL SPREIIDEHSADDLKKLGITADDIHHFI EKGELSPREIIDEHSADDLKKLGITADD IHHFIEKREEDREAAEK	0.951	nan
267	repeat protein dssp3 H0.2D0.2 ps3 c1 2 sym5 cap10 000002 model 4 ptm seed 0 unrelaxed A	KKDAKIKEHIAAIDHLIDKGATIDEIIR FYKHLDEDVKEKAADAAIDHLIDKGATI DEIIRFYKHLDEDVKEKAADAAIDHLID KGATIDEIIRFYKHLDEDVKEKAADAAI DHLIDKGATIDEIIRFYKHLDEDVKEKA ADAAIDHLIDKGATIDEIIRFYKHLDED VKEKAADDVYKHIRKIK	0.917	nan
268	repeat protein dssp3 H0.2D0.2 ps3 c1 2 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	HDEKKRIEDLKDRIDDLIDDGLSAEEIA AHFGLSVEDIKQFISKDRIDDLIDDGLS AEEIAAHFGLSVEDIKQFISKDRIDDLI DDGLSAEEIAAHFGLSVEDIKQFISKDR IDDLIDDGLSAEEIAAHFGLSVEDIKQF ISKDRIDDLIDDGLSAEEIAAHFGLSVE DIKQFISEDKIERLKKH	0.951	nan
269	repeat protein dssp3 H0.2D0.2 ps3 c1 3 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	DDDLRKREELKDKIHELIDAGKTAHDIA RKLDLSVDEIADLIDKDKIHELIDAGKT AHDIARKLDLSVDEIADLIDKDKIHELI DAGKTAHDIARKLDLSVDEIADLIDKDK IHELIDAGKTAHDIARKLDLSVDEIADL IDKDKIHELIDAGKTAHDIARKLDLSVD EIADLIDDEAHKRLHAL	0.918	nan

270	repeat protein dssp3 H0.2D0.2 ps3 c1 9 sym5 cap10 000003 model 4 ptm seed 0 unrelaxed A	DERAKLDEAQDLFKRAHEEGARDIEDVA HITGMSHDDVEHHLHDLFKRAHEEGARD IEDVAHITGMSHDDVEHHLHDLFKRAHE EGARDIEDVAHITGMSHDDVEHHLHDLF KRAHEEGARDIEDVAHITGMSHDDVEHH LHDLFKRAHEEGARDIEDVAHITGMSHD DVEHHLHDEAHREHLHA	0.932	nan
271	repeat protein dssp3 H0.2D0.2 ps3 c2 0 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	DEHKAKDHELHDKIHDLIDAGKDADEIA KILGLTKDDIKHHISHDKIHDLIDAGKD ADEIAKILGLTKDDIKHHISHDKIHDLI DAGKDADEIAKILGLTKDDIKHHISHDK IHDLIDAGKDADEIAKILGLTKDDIKHH ISHDKIHDLIDAGKDADEIAKILGLTKD DIKHHISKHHRAKIEKK	0.941	nan
272	repeat protein dssp3 H0.2D0.2 ps3 c2 2 sym5 cap10 000006 model 4 ptm seed 0 unrelaxed A	VDDHERRHHLHKQIDDLIKSGYTADEIA DKLHLSVDEIKHLISHKQIDDLIKSGYT ADEIADKLHLSVDEIKHLISHKQIDDLI KSGYTADEIADKLHLSVDEIKHLISHKQ IDDLIKSGYTADEIADKLHLSVDEIKHL ISHKQIDDLIKSGYTADEIADKLHLSVD EIKHLISDEDIRDIEAD	0.948	nan
273	repeat protein dssp3 H0.2 ps1 c1 6 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	AKKALRQRGRLRAEALIKSNKPITKDEL EALGFSEEEIKHFGRLRAEALIKSNKPI TKDELEALGFSEEEIKHFGRLRAEALIK SNKPITKDELEALGFSEEEIKHFGRLRA EALIKSNKPITKDELEALGFSEEEIKHF GRLRAEALIKSNKPITKDELEALGFSEE EIKHFGRLREAVKRRRQ	0.954	nan
274	repeat protein dssp3 H0.2 ps1 c2 1 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	HLHIKRKLALHREIEHLIEAGKTGAEIA KELNLSLHEIKALITHREIEHLIEAGKT GAEIAKELNLSLHEIKALITHREIEHLI EAGKTGAEIAKELNLSLHEIKALITHRE IEHLIEAGKTGAEIAKELNLSLHEIKAL ITHREIEHLIEAGKTGAEIAKELNLSLH EIKALITPEAVKRLEEL	0.945	nan
275	repeat protein dssp3 H0.2 ps3 c2 3 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	VHLIRSLHKLGITVEHLKHHGISLRELM ERHKLTISELHHHLGGITVEHLKHHGIS LRELMERHKLTISELHHHLGGITVEHLK HHGISLRELMERHKLTISELHHHLGGIT VEHLKHHGISLRELMERHKLTISELHHH LGGITVEHLKHHGISLRELMERHKLTIS ELHHHLGPEEIKRLEAI	0.904	nan
276	repeat protein dssp3 H0.3D0.3E0.3 ps1 c1 4 sym5 cap10 000003 model 4 ptm seed 0 unrelaxed A	SEAREQAERERIERVIEKLAEGITVEEL KEEGFTVEELEAAQKRIERVIEKLAEGI TVEELKEEGFTVEELEAAQKRIERVIEK LAEGITVEELKEEGFTVEELEAAQKRIE RVIEKLAEGITVEELKEEGFTVEELEAA QKRIERVIEKLAEGITVEELKEEGFTVE ELEAAQKEKLQKRVLAK	0.958	nan
277	repeat protein dssp3 H0.3D0.3E0.3 ps2 c1 1 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	LHEAKEREELRREVRELIEEGKTAEEIA EILGLSVDEIKELIGRREVRELIEEGKT AEEIAEILGLSVDEIKELIGRREVRELI EEGKTAEEIAEILGLSVDEIKELIGRRE VRELIEEGKTAEEIAEILGLSVDEIKEL IGRREVRELIEEGKTAEEIAEILGLSVD EIKELIGEEEIKKIEEK	0.95	nan

278	repeat protein dssp3 H0.3D0.3E0.3 ps3 c1 6 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	HEEEERRRRREKLEERIEAGDHSIEELA HELGLSVEEVRDLLHEKLEERIEAGDHS IEELAHELGLSVEEVRDLLHEKLEERIE AGDHSIEELAHELGLSVEEVRDLLHEKL EERIEAGDHSIEELAHELGLSVEEVRDL LHEKLEERIEAGDHSIEELAHELGLSVE EVRDLLHPEDIAELERR	0.901	nan
279	repeat protein dssp3 H0.3D0.3E0.3 ps3 c1 8 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	KEHEKAERAERHEEIEEAIAHGVSKEEL LQAGFPEDLIDEAHERHEEIEEAIAHGV SKEELLQAGFPEDLIDEAHERHEEIEEA IAHGVSKEELLQAGFPEDLIDEAHERHE EIEEAIAHGVSKEELLQAGFPEDLIDEA HERHEEIEEAIAHGVSKEELLQAGFPED LIDEAHEEILEIKEAEA	0.943	nan
280	repeat protein dssp3 H0.3D0.3E0.3 ps3 c1 9 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	EEEERRKLEKEQQIHDALLEGRTAEEIA EELGLTVEEIKDEISEQQIHDALLEGRT AEEIAEELGLTVEEIKDEISEQQIHDAL LEGRTAEEIAEELGLTVEEIKDEISEQQ IHDALLEGRTAEEIAEELGLTVEEIKDE ISEQQIHDALLEGRTAEEIAEELGLTVE EIKDEISDEEFEAIEER	0.939	nan
281	repeat protein dssp3 H0.3D0.3E0.3 ps3 c2 1 sym5 cap10 000005 model 4 ptm seed 0 unrelaxed A	SDDEKDKNELADLIEDLIENGKTIEEIA EELNLSVEEIKHLISADLIEDLIENGKT IEEIAEELNLSVEEIKHLISADLIEDLI ENGKTIEEIAEELNLSVEEIKHLISADL IEDLIENGKTIEEIAEELNLSVEEIKHL ISADLIEDLIENGKTIEEIAEELNLSVE EIKHLISEKELEEFEKL	0.942	nan
282	repeat protein dssp3 H0.3D0.3 ps1 c1 7 sym5 cap10 000003 model 4 ptm seed 0 unrelaxed A	MQEAIAKFIAGELTLRDALALGVSLPEL IRAGIGPEDIRDHITGELTLRDALALGV SLPELIRAGIGPEDIRDHITGELTLRDA LALGVSLPELIRAGIGPEDIRDHITGEL TLRDALALGVSLPELIRAGIGPEDIRDH ITGELTLRDALALGVSLPELIRAGIGPE DIRDHITAVELEEEARL	0.938	nan
283	repeat protein dssp3 H0.3D0.3 ps1 c1 8 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	AEKAARDKLADKAIDEIIAGKISIDEAA AITGLSKSEIKHRAADKAIDEIIAGKIS IDEAAAITGLSKSEIKHRAADKAIDEII AGKISIDEAAAITGLSKSEIKHRAADKA IDEIIAGKISIDEAAAITGLSKSEIKHR AADKAIDEIIAGKISIDEAAAITGLSKS EIKHRAAERAAKKAAKA	0.973	nan
284	repeat protein dssp3 H0.3D0.3 ps2 c1 3 sym5 cap10 000006 model 4 ptm seed 0 unrelaxed A	SDEEDKRDRLASLVHDAIEAGKTAEEIA DDFGLTVDEIKELIPASLVHDAIEAGKT AEEIADDFGLTVDEIKELIPASLVHDAI EAGKTAEEIADDFGLTVDEIKELIPASL VHDAIEAGKTAEEIADDFGLTVDEIKEL IPASLVHDAIEAGKTAEEIADDFGLTVD EIKELIPDEDFDDFLKR	0.95	nan
285	repeat protein dssp3 H0.3D0.3 ps2 c1 5 sym5 cap10 000004 model 4 ptm seed 0 unrelaxed A	DKDDEDDKLDKEIVEEIKAGKASVEDLA KKYGLTKDDILHHLKKEIVEEIKAGKAS VEDLAKKYGLTKDDILHHLKKEIVEEIK AGKASVEDLAKKYGLTKDDILHHLKKEI VEEIKAGKASVEDLAKKYGLTKDDILHH LKKEIVEEIKAGKASVEDLAKKYGLTKD DILHHLKDDDKDELKKI	0.916	nan

286	repeat protein dssp3 H0.3D0.3 ps2 c1 9 sym5 cap10 000000 model 4 ptm seed 0 unrelaxed A	GDDEAERDDLHEDIAIALHNGLTFEDIA RELGLSVKELADHISHEDIAIALHNGLT FEDIARELGLSVKELADHISHEDIAIAL HNGLTFEDIARELGLSVKELADHISHED IAIALHNGLTFEDIARELGLSVKELADH ISHEDIAIALHNGLTFEDIARELGLSVK ELADHISPEQFEAFEDA	0.94	nan
287	repeat protein dssp3 H0.3D0.3 ps2 c1 9 sym5 cap10 000003 model 4 ptm seed 0 unrelaxed A	DEEDEKKDAIKDKIHDLIDKGHTAEEVA DILNLSVDDIKAFISKDKIHDLIDKGHT AEEVADILNLSVDDIKAFISKDKIHDLI DKGHTAEEVADILNLSVDDIKAFISKDK IHDLIDKGHTAEEVADILNLSVDDIKAF ISKDKIHDLIDKGHTAEEVADILNLSVD DIKAFISDEDIKKYEED	0.949	nan
288	repeat protein dssp3 H0.3D0.3 ps2 c2 1 sym5 cap10 000000 model 4 ptm seed 0 unrelaxed A	DDKKKHDDKIDEKIDKLIDEGKTAEEIA KILGLSVDEVKDHISDEKIDKLIDEGKT AEEIAKILGLSVDEVKDHISDEKIDKLI DEGKTAEEIAKILGLSVDEVKDHISDEK IDKLIDEGKTAEEIAKILGLSVDEVKDH ISDEKIDKLIDEGKTAEEIAKILGLSVD EVKDHISDKDKDRLDEI	0.946	nan
289	repeat protein dssp3 H0.3D0.3 ps3 c2 1 sym5 cap10 000008 model 4 ptm seed 0 unrelaxed A	HDLHRRKHDLKDEIDKLIDAGMSADEIA DILGLTVDEIKDHIDKDEIDKLIDAGMS ADEIADILGLTVDEIKDHIDKDEIDKLI DAGMSADEIADILGLTVDEIKDHIDKDE IDKLIDAGMSADEIADILGLTVDEIKDH IDKDEIDKLIDAGMSADEIADILGLTVD EIKDHIDDEERDRIDKI	0.948	nan
290	repeat protein dssp3 H0.3 ps1 c1 8 sym5 cap10 000004 model 4 ptm seed 0 unrelaxed A	PLIEAIKKQGLSIPEFLEHNNLSIEELL ELTGKSLVEILKEHNLSIPEFLEHNNLS IEELLELTGKSLVEILKEHNLSIPEFLE HNNLSIEELLELTGKSLVEILKEHNLSI PEFLEHNNLSIEELLELTGKSLVEILKE HNLSIPEFLEHNNLSIEELLELTGKSLV EILKEHNPEEILEARKL	0.949	nan
291	repeat protein dssp3 H0.3 ps1 c2 0 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	AKHARYLREARHRFERLIEAGATLREIV EALGPGKLHKIIVEDRHRFERLIEAGAT LREIVEALGPGKLHKIIVEDRHRFERLI EAGATLREIVEALGPGKLHKIIVEDRHR FERLIEAGATLREIVEALGPGKLHKIIV EDRHRFERLIEAGATLREIVEALGPGKL HKIIVEDEEARRHIERH	0.918	nan
292	repeat protein dssp3 H0.3 ps2 c1 1 sym5 cap10 000002 model 4 ptm seed 0 unrelaxed A	HVHHLRRHHLHHAIGHLIHEGATGAELA ARFHLTASEIHHLIPHHAIGHLIHEGAT GAELAARFHLTASEIHHLIPHHAIGHLI HEGATGAELAARFHLTASEIHHLIPHHA IGHLIHEGATGAELAARFHLTASEIHHL IPHHAIGHLIHEGATGAELAARFHLTAS EIHHLIPEVERRAWHKA	0.921	nan
293	repeat protein dssp4 H0.3D0.3 ps2 c1 1 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	PSIEDWLKKNHPGHPFKVARTPDGHYIA FDPKSGEGYFFDPDGHPGHPFKVARTPD GHYIAFDPKSGEGYFFDPDGHPGHPFKV ARTPDGHYIAFDPKSGEGYFFDPDGHPG HPFKVARTPDGHYIAFDPKSGEGYFFDP DGHPGHPFKVARTPDGHYIAFDPKSGEG YFFDPDGKVQKLDPDHL	0.888	nan

294	repeat protein dssp4 H0.3D0.3 ps3 c1 3 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	IDSLLAELKDHGADDVTIHHHDDDLTIK VRFPDGREHHIHIDDHGADDVTIHHHDD DLTIKVRFPDGREHHIHIDDHGADDVTI HHHDDDLTIKVRFPDGREHHIHIDDHGA DDVTIHHHDDDLTIKVRFPDGREHHIHI DDHGADDVTIHHHDDDLTIKVRFPDGRE HHIHIDDLDAQRDDLEH	0.816	nan
295	repeat protein dssp5 H0.2D0.2 ps1 c1 2 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	DPEDLKKRLEANGYSLTVNGDDVTLDGN KLSASGGGASVVITDANGYSLTVNGDDV TLDGNKLSASGGGASVVITDANGYSLTV NGDDVTLDGNKLSASGGGASVVITDANG YSLTVNGDDVTLDGNKLSASGGGASVVI TDANGYSLTVNGDDVTLDGNKLSASGGG ASVVITDDKARDLAKKL	0.922	nan
296	repeat protein dssp5 H0.2D0.2 ps1 c1 3 sym5 cap10 000003 model 4 ptm seed 0 unrelaxed A	GKDKLKAVLKSLGIDSIDMKPGDKITIS DGTLEISGGAKVTIKSLGIDSIDMKPGD KITISDGTLEISGGAKVTIKSLGIDSID MKPGDKITISDGTLEISGGAKVTIKSLG IDSIDMKPGDKITISDGTLEISGGAKVT IKSLGIDSIDMKPGDKITISDGTLEISG GAKVTIKDDIKKALKDL	0.946	nan
297	repeat protein dssp5 H0.2D0.2 ps3 c1 3 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	HHRKHAIDKIDAAVAAGAKKVHIHLDDP GLPREDLQDLGDEMIDAAVAAGAKKVHI HLDDPGLPREDLQDLGDEMIDAAVAAGA KKVHIHLDDPGLPREDLQDLGDEMIDAA VAAGAKKVHIHLDDPGLPREDLQDLGDE MIDAAVAAGAKKVHIHLDDPGLPREDLQ DLGDEMIKKADDHLQAK	0.859	nan
298	repeat protein dssp5 H0.2 ps1 c1 7 sym5 cap10 000009 model 4 ptm seed 0 unrelaxed A	HPLKKHLKKLGFKGHIEGVNEVSVENGE ITLTNVKKLELEHKHGFKGHIEGVNEVS VENGEITLTNVKKLELEHKHGFKGHIEG VNEVSVENGEITLTNVKKLELEHKHGFK GHIEGVNEVSVENGEITLTNVKKLELEH KHGFKGHIEGVNEVSVENGEITLTNVKK LELEHKHPEHLKHLKEE	0.925	nan
299	repeat protein dssp5 H0.3D0.3E0.3 ps1 c1 7 sym5 cap10 000006 model 4 ptm seed 0 unrelaxed A	EEDILAAVKEGKAELSGDTLTLTGEPTI SNLSYPGFKITNLKGGKAELSGDTLTLT GEPTISNLSYPGFKITNLKGGKAELSGD TLTLTGEPTISNLSYPGFKITNLKGGKA ELSGDTLTLTGEPTISNLSYPGFKITNL KGGKAELSGDTLTLTGEPTISNLSYPGF KITNLKGEEEAIKALQT	0.923	nan
300	repeat protein dssp5 H0.3D0.3E0.3 ps1 c1 7 sym5 cap10 000008 model 4 ptm seed 0 unrelaxed A	EEEKIEEIARKEGFEVVKVGEAPEGDRL FELKDPKGKKFSIELKEGFEVVKVGEAP EGDRLFELKDPKGKKFSIELKEGFEVVK VGEAPEGDRLFELKDPKGKKFSIELKEG FEVVKVGEAPEGDRLFELKDPKGKKFSI ELKEGFEVVKVGEAPEGDRLFELKDPKG KKFSIELEKANEILKEA	0.882	nan
301	repeat protein dssp5 H0.3D0.3E0.3 ps1 c1 9 sym5 cap10 000008 model 4 ptm seed 0 unrelaxed A	GEELEELARKAGFGEVKISKEGDKLVLT LKDPKTNETFTLELPAGFGEVKISKEGD KLVLTLKDPKTNETFTLELPAGFGEVKI SKEGDKLVLTLKDPKTNETFTLELPAGF GEVKISKEGDKLVLTLKDPKTNETFTLE LPAGFGEVKISKEGDKLVLTLKDPKTNE TFTLELPLEELREAEEA	0.904	nan

302	repeat protein dssp5 H0.3D0.3E0.3 ps1 c2 2 sym5 cap10 000000 model 4 ptm seed 0 unrelaxed A	DEELVAELEAAGWELVETLGSFRLFRNP DNTFVIIDEEGKLVGAGWELVETLGSFR LFRNPDNTFVIIDEEGKLVGAGWELVET LGSFRLFRNPDNTFVIIDEEGKLVGAGW ELVETLGSFRLFRNPDNTFVIIDEEGKL VGAGWELVETLGSFRLFRNPDNTFVIID EEGKLVGEEELKREFAE	0.92	nan
303	repeat protein dssp5 H0.3 ps1 c1 0 sym5 cap10 000007 model 4 ptm seed 0 unrelaxed A	SLEQHKAIASKLGLTVSASGPGATVSVS GNTVIVSGAHHATASKLGLTVSASGPGA TVSVSGNTVIVSGAHHATASKLGLTVSA SGPGATVSVSGNTVIVSGAHHATASKLG LTVSASGPGATVSVSGNTVIVSGAHHAT ASKLGLTVSASGPGATVSVSGNTVIVSG AHHATASITEEELKALI	0.929	nan
304	repeat protein dssp5 H0.3 ps2 c1 4 sym5 cap10 000001 model 4 ptm seed 0 unrelaxed A	REEARAHLRHHGFHFHPGVEIHHHPHGK IHARIRPGGRVHHKKHGFHFHPGVEIHH HPHGKIHARIRPGGRVHHKKHGFHFHPG VEIHHHPHGKIHARIRPGGRVHHKKHGF HFHPGVEIHHHPHGKIHARIRPGGRVHH KKHGFHFHPGVEIHHHPHGKIHARIRPG GRVHHKKPEERRRLAKH	0.95	nan
305	230211 W0.3 000069	GSIKDWEEWIEELEWWREKGWTKDEFIE WLEKQIEEWRKRGDRWAIERIEWIINKI KEGKTFDEIIKEWREWLKER	0.976	87.146
306	230211 W0.3 000284	KLTPEEIEEWLKWIKEWLEENPDWSWEE WRERIEREIEEWWAEHGIDDEEREWLER KIEEWLRELEEWRKRWKG	0.938	86.451
307	230211 W0.4 000445	GLSEAERWREEFERLWDEWREWIENGNW EEIREAWERLREIFERLREEGWFSPEEI ERWEEEWERWWEEAEERWRRWEG	0.958	90.207
308	230211 W0.3 000290	GAISKEEFERWKAEIEELIKWFKRNGWE DEIKKLKEWWEKFQEAWEQGDWDKIRRI WQEIKERWERWEKWIREG	0.984	89.279
309	230211 W0.5 000462	GLSPRERFEEEWEEWLWLWEEGRLSPEE FWEELRRWWEEWPGGLSEEERREWRERL EEWERWWERWEEEREAG	0.968	86.681
310	230211 W0.4 000495	GWEEWWERWERAGPREAWELWREWWERW REAGPSPEERREWLEWLERWAERLRERG DPEWVEETERLLARWRAWLEG	0.973	87.106
311	230211 W0.3 000088	KPLPNISEEEWEEFKRWFEKAWEEWRRW AEKGEEEAKRWREEWEEELREWCERWGI SDEEWEILWREFEEWFERRG	0.964	88.429
312	230211 W0.4 000225	GWWREWLEWIIERLRELGWEELVERIRR LWEEGRITWEELWEWIEEWWERNWSEEE RAEFRRWWERRWEEWRR	0.979	88.606
313	230211 W0.4 000275	PSREELWEWLERWRGASPEERERLWEEL DRWLRRMSPEEIRELFEWLKERWPENRE WIEELWRRWAWWLEWERG	0.901	86.569
314	230211 W0.4 000088	WSPEERRERIERAFEEAQRWWERWRRSG DEELWERCEEAWARIEAWWDQWREEGWD WADWRERWAEFRAEFERWRRG	0.983	90.3
315	230211 W0.3 000293	NWREWRARIEEWIEAGNIEEAWRWLREL REWWRERGEISPEDRRWFEEIWEWLRER FAGNREEWERLWKEW	0.957	88.504
316	230211 W0.5 000214	GAREELLAWWERWGWSREELREWWAEAR EWARREGWSPEEWLEWLAEWLRARGLWS PEEARRLRELLEAWWEAERG	0.934	85.246

317	230211 W0.3 000454	GKWEEIFRWFKREIERLLEELGGNKEWI ERIREWWEWIWEEWEKWRKRGWTEEEFE RFLERSFEELEKWLKEG	0.923	86.95
318	230211 W0.3 000466	PSREERLEELIEWLERNEGLSWQERIER LREWLERWDWWSREDREELLRWCEERGW PPELLEWLREWWRS	0.978	87.648
319	230211 W0.4 000083	GWSRWERLRRLIEWLRRAWGENWEEIRW IFEELWELLREDEEWREKFWEWLDEWLR KIDEEERRELWREWWEWRERRG	0.962	87.571
320	230211 W0.4 000144	MSIEERLNEWLERIEAGKKISDEELEEL LNWWWEEWDKLSEEERIEWWRRILEILK WSGWEEWIERIEEWIEKWE	0.992	85.848
321	230211 W0.4 000068	ASPEWRRWWERWREIWEEWEELKKEGWS REEIIERLAERWGFSKEEIRELWEWWER TGDWEEAFREFWEKRG	0.954	86.894
322	230211 W0.6 000334	GADEWRERVRERWEELEEWLERWGWGPD WREELERWREWWEGAGSEDEWREIWRQI EEWQERIEEEWERREGG	0.981	88.278
323	230211 W0.4 000116	KSREELIEEKRKELREWWERIKEWCRRN GIDWEEFEEWFEWIWEEFEEWWKGWSWE EIEEWWERIKEDFEERFEWWKKR	0.971	88.575
324	230126 uncond batch6 358	KKPLTEEDFREIRENLIRKIEEELKKRG IEISEELKRELEEKLEELDKLRPELENL TDEERREIIEKFISELIEEL	0.94	79.886
325	230126 uncond batch2 312	GIIPPFTEEELEEFVEEILALGGGITRE EIIERIRKLIGKTEEEVIEELDELLGDK KAIEKLRERLRRFRG	0.93	79.264
326	230126 uncond batch9 317	NLEEKFEEFIERRGGHITPEELKEFFEE LGFKEITPEDRQLIIELIKKGKIVLTPE DLKEIPLTDELREDIERIIESG	0.951	76.522
327	230126 uncond batch9 411	DFKEALDLIEKFLPNLNPEERDEVEKLI EEILKANQEELLERLEEIEKFFEELGSP EEKKLIEKIKESISSL	0.971	77.847
328	230126 uncond batch5 272	KSRLREKIKEIFEKFRQIFEELVAEGKL SEEERERIKKIYKEIEKIKKLAERGISE EDFEILKEIERLLKELEEL	0.946	78.78
329	230126 uncond batch2 134	KLPITEEELKRLGFTREEFREFIKEFEE KIKKEGMTLDEARREFIEELREKHKKIT PEEIKELPRHFKEELKER	0.913	76.775
330	230126 uncond batch9 401	GPAELEEERRRLREEARARGLEVEELPP PPGSTRDRARRLIAGGRVALPPPGITPE ERRELLRLIEEAIRERG	0.915	77.827
331	230126 uncond batch8 317	GERERQLKRIREELKELGISDREIEELL RLIRENPETFERLFADPTKERIEELFKE FRKNRKEREARLRRG	0.924	77.785
332	230126 uncond batch7 403	GDLEELEEEIVAHLLRRGITEEEARRLA REILEVIRAGISAEELRRFIEEELKGIS PEELRRRIEELRRERRG	0.933	76.738
333	230126 uncond batch8 328	GLTEEEIREEIEKLIEELLKLGIDKEEI LKYFDLKNITKEELKEFIEKIKKKLDEK EIKKLIEELEKLRKLIEKL	0.91	68.974
334	230126 uncond batch6 202	EIERLIREIERGEKETFRKRIEDFFKRT GFKITEEERENIERLIEEIESGKITPEE IERLERELREILKNG	0.908	68.006
335	230126 uncond batch7 105	GKAPLLEELVKKLEELLAEGATPEEIEA IIEELAKKYGITREELLEFIEKKYGIKP EEIVKKGKELIEEILAAIRKSR	0.921	78.6

336	230126	uncond	batch5	421	GEKREEILEELRKLIEAGKITLEEIREL IKNFSLSELRESLERLIKELREKGFSEE KIERIRELIEKLREKLG	0.939	69.903
337	230126	uncond	batch3	16	SPKILKELGITPEQFEELIEEIKNLKRE GISEEELKKRIIEHFKKEFNLTPEQLKR IEEFIDELLKEFEKS	0.954	79.374
338	230126	uncond	batch5	236	NPKLRELIESLKGLTLEELIARLKREYP DITPEEIEELIEELFKSGKLTPEEIREI AALYGITPEEIRKLLEAIEELNR	0.923	64.577
339	230126	uncond	batch2	217	KISPDEFLRELEKIKDKIREGLTNEELK EFEKEIEKLREEIKEELKKFDDEKREKL LRELEELEERLKKKLG	0.949	68.415
340	230126	uncond	batch5	226	NEKKREELEQLLEKLEEFRKRLEELIKK GEITLEEIQEEFENLLEELEKYGDEELI NKLREEIEEFRKRLK	0.95	79.79
341	230126	uncond	batch1	118	GTPDELFLRELKARGITLEEFFERFGFS EEERREIEEKFKAGEPFDREELIRFLER ILERGIKDPELREEIERLLEEIKG	0.958	78.947
342	230126	uncond	batch7	156	NEREKIRKLLERIRKELEEIRKKNGGIS PEELREFIRKLLEKEGIKLTPEERERIL EEIEEELEERRERLKEKKEG	0.919	75.632
343	230126	uncond	batch7	94	PSLRELLEELKGLTPEELREFIERFLEK HPLTEEEIEELREFFRERGIKIPLPKSL TEEERKEFIEEIRRLLKG	0.925	70.843
344	230126	uncond	batch5	349	SMTKEDFERFLKELIKKESPDEIIKYLI GKFTEEEFIELFTPKDLRELLKEFPLSD EEKETLEKILDKVEKLREIKG	0.953	76.987
345	230126	uncond	batch6	332	MTREEIEEILKEIRDEILELLKEIPDRE ELKERIEEILERLAKERNIDPEELRKLI TEDLEELLKEILKLRN	0.939	78.479
346	230126	uncond	batch9	94	GISQRDLREIIEKLKENGFKFSEEELIK ILEEKTLDEFLKKLEEKGSPKERKLIEE LEKFRRRKDEERKKAERE	0.935	76.133
347	230126	uncond	batch7	241	DPARELEELIRAARKDPEEFRKEAEKFR EYLLRSGISKEQFKQFVEEFKERFKPLS PEQKAKLEELLQEEFKAIEAKKSD	0.91	77.023
348	230126	uncond	batch6	51	GPTPEKREEKLRELENEIEEFIKEGRIT PEEREKLREEIKKLIRAGFDFEKIKEEI RELFNEIEEKRRKRRKEL	0.937	78.088
349	230126	uncond	batch8	460	NSLEEKLEKILKKGKELFKKFLEKKNNK KDLREIKNLDPRELIRRLLEIFKEHLKY LKNEEEIKELRELLENLREILEEN	0.924	73.279
350	230126	uncond	batch7	100	KKVEETLKRLRELIKAGGGDFDELLKEL EEIRKLLENLSEEERKKIIEELEELRKE FEKAIKNGSELGRKLLEELLG	0.95	79.153
351	230126	uncond	batch8	170	DLERLRKIIEIFKKGKTPEEIRELLEEI KKEELLEILKIDKEELREFLKKLGITKE EIEKLLEEIEEIVKKG	0.926	69.596
352	230126	uncond	batch2	388	NPKELIKFLRELLRKGISEEERERIREF LEELGINPEEILERLKEMTPEEFLEFLR NDPEIREKLREFLKRKKG	0.932	69.358
353	230126	uncond	batch6	205	EEKRERIEELERILKEKNITPKEIRELI REYIRKGKISEEELREILEFLKRKGIKI DLSKDDIEELKKLLESRK	0.93	77.759
354	230126	uncond	batch1	81	LSPEEVRSELEEFFAALKAKGKKISDEE IQELIEEIAKKGITDEEIDELLEELEEY GIDLDEIKDLISNIKK	0.959	67.35

355	230126 uncond batch8 77	ELKEKLIKSQEEIKKLPDKKERLEVIIE FFEKLRKLHEEGKISKEDFKELEEFIKD ISPEDKRELREELEKKRN	0.968	78.346
356	230126 uncond batch5 233	KLTPEDIREELEEFFKKLTPEELREALE RLIERCIEAGFTPEELRRFIEENCLEDI TPEMREKFPELIEKIEEYRKG	0.934	63.674
357	230223 V0.15 batch1 000019	DFERRLKEIEEVIEKFDKSNVRVEVTFE ELPNGKFKVSVRIDNGGKIFTFEFDEED VNQLVEFLKKRGIPVEVVRS	0.93	86.627
358	230223 V0.15 batch1 000175	GKVIVRVEPGFSEDERNEIENEVEKLFK EGVSIEEIRERIREWLERNNLGKDIEVE VERLEDGRWVVTVKR	0.954	88.086
359	230223 V0.15 batch0 000449	GKYEIVVTVNGLSKEEAEEIAKRVEKKF NVSVQVTNGGTIIVEVDGDSVSEEEIAA IRTEIRELAKEYNIEVSIEILRFS	0.944	85.405
360	230223 V0.15 batch2 000410	EISPGDEITVKPGEKLEVITNDPDISEE EIRKLLGPDVEVEKVEKLPNGRVIVTFK GGKTVRISKEDLKRLGSVRVVP	0.894	85.168
361	230223 V0.15 batch2 000186	NVVIEVRVSSDEEIQKKLNKIREFLKKK GIEVRISFDDSSNKVTVTFSKGNPDEIV DKLRKLGIEVIDVTR	0.959	86.334
362	230223 V0.15 batch5 000088	VEIVPDEITPEEIKELIKEGIPVKISVD SDEDMERVKEKVEEIKEELNVEEVRFEN NRVTIIVNGKKVKIEILKRPK	0.975	85.644
363	230223 V0.15 batch2 000164	GKGRVSVEIKKENGRWVVRVKPKGATFD DVLDQVRELINELVEEGRDVEVEVKDFS DEERERIRINIDKLLADIRAN	0.939	85.301
364	230223 V0.15 batch1 000133	KVKVEIIDLKGERETFEVEPNVDEFRKK VDEFKELIEAGEEVEVIEVDKNVPEEIR ELVERLRDVAEEYEEKRKAR	0.957	87.49
365	230223 V0.15 batch5 000352	GKIYVVRVNGDIDEIKSILKELDVNGKV FELNGVIIVVFEGVTKDEIERLVKKLRA KGFNVEIFEIENKEEFEEILKNG	0.979	86.95
366	230223 V0.15 batch3 000427	GRTRFRVTFPPNFSDEERNQIIEWLEKR NVPVQIVVDENGRRVIEFSVDRDKAEEL VEELLKELNLPNVKVVPLDD	0.952	87.943
367	230223 V0.15 batch0 000277	GRKVITFDGPITRAQFDRLIEEIRKLLN AGKVDAVVIEVNPDFSEEEVEELVRRIR EKTGFRVEVFRSENGKWVIRVERP	0.979	91.194
368	230223 V0.15 batch0 000257	GFTAEAQNFIEQLKKNNIVGGRVTIKNG KVTVINPEGETFVFEVDFKNEEEFEDLR KRIEELGIDKVTVTG	0.965	86.551
369	230223 V0.15 batch4 000289	SRAEKLEEIRKWAVKNNIDVVIIDPGGL SREEIKERIREFCKGRDRILVVTDDREY LDVIREVCRENNVEWKVVRVG	0.984	87.839
370	230223 V0.15 batch1 000423	EVSKVVEFNITKEEIDKVLNLVEEFRKK NNVEDLSVEFTEENGKYIIKVTIKDENG ERVIDEIVDEIRNL	0.968	86.076
371	230223 V0.15 batch4 000015	GRITVVWEELDPGTPVRIVRDKDGTYIV VIGKDFTEEQIRALIEEAFENGEVRIVV HKSLSEEERAKVEEIVEEYIKKG	0.956	88.319
372	230223 V0.15 batch3 000386	TAEITVSVEGTGGISDARIEELISELKE KVEQVLAQGLDIKVSVEASESEVEAAKR IKSEIQAKYPNVEIV	0.975	88.308
373	230223 V0.15 batch1 000484	APVVPGIRVERLPDGRVRVELPKDFTPE EAIRRCLECVRQLRKDGWKDVEVEVDSQ EIRDRIRAEIEAEGLKDVTVTIR	0.968	86.379

374	230223 V0.15 batch2 000357	GSVKELVEEIRRSVEEDREKVDELIEEL KKRGWEIRRIKSREEVEEFLKEVKDLGD VFVIRLPDGEWVVIVRPK	0.902	86.985
375	230223 V0.15 batch3 000278	DRDKFVEEFLKKVRELVEKGGTVTVENE DWRIEITLENNKFEVEVVDKNGEVVTRT FEINSVEEFRERLEKIIDDHKKRL	0.894	85.178
376	230223 V0.15 batch1 000449	NVTITVKIDPDMSFEERIQKIVEAVEEA RKKAKELNVKDVRVIVEGNISREDKKRI VEEIRERFPGFKIEFSNGVIEVVG	0.974	89.197
377	230223 V0.15 batch1 000251	FELKNGEFRVVFKGDGITEEEIKNQFKE ILKIIEKNGFKRVVVVFKNVPFSEEVRQ ILEEFQEELRKRGIEVEIE	0.952	86.701
378	230223 V0.15 batch3 000275	PVRIELEVVDPNEFRIRLVDLSNEEREK VIRELEEIAKEHNFTVTIFKDENGEVVV RISSDNISRDDLREVFERLRAI	0.951	85.931
379	230214 C0.2 batch0 000225 af2pred mpnn	GVRECLKICEECLRECNPEDFERCIECI KRKSGSPECIRICEEFIEAFRGG	0.964	90.261
380	cys partial 230214 C0.2 batch2 000132 000253 af2pred mpnn	GLCDECREILERLEALGCKPETLRECRR CVEDCERKPSAEEGERCCRACRRLVEKC G	0.977	91.687
381	cys partial 230214 C0.2 batch2 000132 000069 af2pred mpnn	GPRRECRELLERLEERGCPPETIEECRR CLEECERLPPGEECEPCCERCRRLVEEC G	0.968	91.374
382	cys partial 230214 C0.2 batch2 000132 000231 af2pred mpnn	ELCELCRELLAQFEALGCSREEIAECRD CCERCERSPPSEELIRCFERCVRLVEAC G	0.992	90.094
383	cys partial 230214 C0.2 batch2 000132 000441 af2pred mpnn	GLVERARELLERCRALGCSEEALAELER CLAECERAGSEEERRRCFERCRALLEAA G	0.967	91.714
384	230214 C0.2 batch1 000184 af2pred mpnn	NAESQEFCEEICRLCEECGFDVSECLAL CRESKAECEKIKEFCEKCLALQSN	0.969	90.16
385	cys partial 230214 C0.2 batch2 000132 000419 af2pred mpnn	GLCEECRELVERLERLGCPPELIRECRR CCEECESLPPGEEGRRCCERCRRLVREC G	0.98	91.191
386	cys partial 230214 C0.2 batch1 000330 000048 af2pred mpnn	GLSEEELRALARELCERCGPEGRRCIRE AERALREGDPEAVRECIRECERCLEEG	0.955	93.642
387	cys partial 230214 C0.2 batch2 000132 000609 af2pred mpnn	GLRRECRELVERCEEAGCSEEECRELER CVREAERSRDPETARRLCERCRRLLETC G	0.985	92.881
388	cys partial 230214 C0.2 batch2 000132 000212 af2pred mpnn	GLCERCRELVERLRERGCSREDLERCRR LVEECERSGSPELCERCCQRCEELVAEC G	0.988	93.524
389	cys partial 230214 C0.2 batch2 000132 000554 af2pred mpnn	SRDDECRRMRERLEEQGCSPETIQDCCR EIERCARLPRGEAVECVEELSRRCVERC G	0.984	91.218
390	cys partial 230214 C0.2 batch2 000132 000315 af2pred mpnn	GLLERCRRLFERAVELGCSEEVIRECLR CLEEVERLPPGEEGERCCERCRRLVERC G	0.965	92.376
391	230214 C0.2 batch1 000232 af2pred mpnn	GLTPEEFQALCDECLAEARACGVEDCVQ EVEELCRQGELTPEECREILADCERLAG	0.986	92.336
392	230214 C0.2 batch2 000132 af2pred mpnn	GLCEECRELVERLEARGCSPETIRECRR CVEDCERSPPGEEGERCCERCRRLVEQC G	0.979	90.722
393	cys partial 230214 C0.2 batch2 000132 000219 af2pred mpnn	GCLEEARRALEELERRGCPPELIRRCRE LIEELERLKSEEECRRAVEECRRLVEEC G	0.969	91.305

394	cys partial 230214 C0.2 batch2 000132 000757 af2pred mpnn	GPVEVCREALEQLRERGCPDEDIRQCQE CIERLERSGDPEEAIQCCERCQELVERC G	0.984	90.199
395	cys partial 230214 C0.2 batch2 000132 000812 af2pred mpnn	SPIEECRELVERLEELGCDPEKIRECRR CLEDCERGGSDEEVERCCERCRRLVEEC G	0.988	91.928
396	cys partial 230214 C0.2 batch2 000132 000735 af2pred mpnn	SLCEECRELLERLEERGCPPEVLREARE AVERCERGASGDERERLCERARELLERC G	0.987	90.127
397	cys partial 230214 C0.2 batch1 000330 000631 af2pred mpnn	NLSEEELRELCREVCERLGVGGDDTIRK AIRLCREHLGADREECREICERVCEES	0.966	90.65
398	cys partial 230214 C0.2 batch2 000132 000559 af2pred mpnn	GLPEECRELIEECERRGCSPETLARCRR LCERARGCSPEECRELCEECERLCVEAC G	0.992	91.465
447	attempt 02 01 000409 af2pred A	SKERKERLERKLKELIQLIEEYLKNPTK EKKKELKKAAKELTELFNEMKAENLYFQ GLSDEEIEAFVEELKELERALSKGSPEE IKERLKKLREEIEKFLKSIKKYK	0.925	85.227
448	attempt 02 02 000069 af2pred A	SAQKLIAEIKKLEEEIRELIELNLDDKE TVKRKLNKIKREIEAIRNRLKKENLYFQ GFTKEEIEEELKELEQEIEEAEKTDPEK AKEKLEEIRKRVKDLLVKLEKAK	0.9	90.671
449	attempt 02 02 000100 af2pred A	VTEEEIKKLKAELKKISELVDEAINASS KEERTKLIKEARREIERLISYLENLYFQ GDETKREIIRKLEKLKAAIEKLETREEA EKIKQEIKAIRKEIERLAEEIKK	0.895	91.562
450	attempt 02 02 000134 af2pred A	GKLPAERRAKLEEAISELEQLRKEGKKL RKVLLIAGSYTPEEKEELKAYLENLYFQ GIKFEIFDATGYTREEIIEFAEKNNFDL VISIGNGGLSDKDIEEIEKRIKN	0.879	86.343
451	attempt 02 03 000317 af2pred A	SNRVKKILEELKEAASKGIKAIETGDKE AALQAAEEIIEALKRLIKALEEENLYFQ GLDLDRILTELKEKIEKAKKAGDLEKLK EAIEKAKELIEKLSAEIKELEAN	0.929	89.303
452	attempt 02 04 000021 af2pred A	MTKEEEELELIDKVLEELKKAEEYIEEA TREGIEKAKEELEEAIKLLETNENLYFQ GLYDKVRGELEAALSYIEAGDIEKAKKE IKEAKEEVKRARKLVKERIEKKK	0.956	91.279
453	attempt 02 05 000184 af2pred A	GSFKERIREAVEKAREALKKNPNRKLYL VIETSDLEKFKPGDIQKVRDLFENLYFQ GVEIEVEAIKAGDEESIRAAKEKAKENG KVVLTVIIGSGPDLINIIRDALK	0.889	86.117
454	attempt 02 05 000274 af2pred A	GQLREFIEKLEELEKEIDALVKKKIDKE NLKKLIEKIKKLVEELEKLAKRENLYFQ GELEKLIERLRKVVEDLEKIPEEDNSRK ILKIEELKKEVKEIIKELKKALN	0.882	87.794
455	attempt 03 01 000147 af2pred A	ANDEKRKKELSELIKEIDSILKKVEAAK DALEAGDKAKAKVLIKQLKVELENLYFQ GQALKDENLEKNIENVLELIIKAGEEYD IEVLDKLIKEIEELIEQLRKEAD	0.949	91.263
456	attempt 03 02 000134 af2pred A	GKREQLREEIQQLVAEAKEAKKTAKQAL KIAKELGDAEAKKEAEKLLRETENLYFQ GEKLFKELEKLKRGDIGRLVNELRELRK ASKQSIDELKDLLQKLEKEKEEE	0.937	91.92
457	attempt 03 02 000201 af2pred A	KAAEKRALREEIERMKSEAERIARKLEA IKKAAKAKGLNDPTIERKVAEAENLYFQ GRELIEEAEEKLEKASEADIKELEKLIQ AAKAELEAAKKTLEEIEKIIEKN	0.948	91.972

458	attempt 03 02 000329 af2pred A	EPPTKKEQKKIIEGLKESLERLREQGRL TDEQIQEFIEQFERELKEGATFENLYFQ GKSFLDKIGLSSEEIKRLIEELEKQGYL TPDLVDELRELDAALSEQKAQLQ	0.903	84.614
459	attempt 03 03 000045 af2pred A	TQEEEREERVKKYNELLEKIENLVKLLV ETLKDKGKKVPEAELEEIKKSVENLYFQ GKKGEISNEELEELEALIRIIETIHPDV DPEQIKELRELLEEIKKLLDKRK	0.906	84.642
460	attempt 03 03 000231 af2pred A	DKVKKEIDKLKEVLEALEKAIIQAKEVI EKLAELGEKKEIEKAKEELKKLENLYFQ GKELIKKLLKEEEREAIKEIDQIKEQIE QEIKEFERSLADFRKKILEEKKQ	0.879	87.521
461	attempt 03 03 000389 af2pred A	KQQSEIIAKLKAQLEQLKKLREKLKDMQ REAKELGEDKTAKKLEEAIKVVENLYFQ GLEAIKAVEEKDAVELTKAEAKELKKAI EDLDRKVKEAKAVLAEIKAKLKS	0.909	86.304
462	attempt 03 03 000440 af2pred A	NKFAEELIKRLKEEIKQLKKIAEKIGDK KLKEIIQKTLEKLKNIPDDASLENLYFQ GEELLQALNEAIEADKKNKEIDEKLIEA LKDEAEKVIRNLKETLEEKRKEL	0.822	75.36
463	test 10 000378 A	TEPPKFFERIEQALEEAEKALQTFMEKA KAGGKIPTAEEIEAFERRTKTAIEQAIK LLEEDPDAEYTPAEKRSIIRLLEEVLKL VPLIIVEYASGELSPAEIAKLIREGEEF IALIKQLLAEYPDKLSEDEKALFESLAD TFKKVLDQLEGGASKGRRKRGIGAVLKV LTTGLPALISWIKRKRQQ	0.915	86.004
464	test 17 000294 A	SKPGALQQLIKEAIELAKEVEALLRKLL GVPGLSAEQKKEIEELLEEVEKIEARLK KLAEVPDAQLTEEQIKELSEILLELIVL IKKLYKLAIENGFTPEELKEIKSLFEKS IKIAVAIYKALPPGEVSEKEIALLRKLE KEIEELEARIEKLSDGRRKRGIGAVLKV LTTGLPALISWIKRKRQQ	0.889	85.279
465	test 18 000052 A	RKPDVKRIRDILREARIIAEKVVKMIKE VLEGKPRTKEAKKLIELAEEIERQLEEI ESLADEGDLPNLEERIVELAEECRNLIS ELELLLREGYKLTDEEQEKVAKEIISLA EQILEIAKAVFDVDDLTPEQKKALEEFK RAIEEAKKFARDFEEGRRKRGIGAVLKV LTTGLPALISWIKRKRQQ	0.896	89.177
466	test 20 000093 A	TKKPAKVQEVLEKILKELKELEKLIRRI LKGELSLEEAEEKLERAALKIEQSIESL EELRKRGEGLTEEELKALKAELDRFEKE LLALKKKPPEKKLSDEERAEIEEAIEAV VAAKKELDELLEGRRKRGIGAVLKVLTT GLPALISWIKRKRQQ	0.872	86.841
467	test 20 000160 A	KFQDKLEELILRLEEEIAEAKALLAEFE ASGPKTFTKEEIERLMEALEKIEEVILE IRRFGHRKPSRKELKEILEELEKAEEEI EKLEERLEKKGAPESLRRELDKFKESLR KARALLIKYMKAGRRKRGIGAVLKVLTT GLPALISWIKRKRQQ	0.886	88.686
468	test 20 000217 A	EDEDEARRLELLRTLEEEITVALDLLKE FLSAGGKSDREKEIANAFEQILLSLREL IRALAPGDKEDKKRERLRRELDKIKAAL EELKAALEEDPDVSDKEKEELRNLEEFI EAINRDMQAYLEGRRKRGIGAVLKVLTT GLPALISWIKRKRQQ	0.856	87.81

469	test	20	000501	A	PKLSKEEIRRRVREALAGLRSALRAAIN MEDSGLTPEELEKELQAVQAKMEKLLAA LEKEPLRPPLSKEEAEIFRKEIEEILKL IEELDEVYAERGFTDEDIIELAEFRAAL EEALKLVEELAEGRRKRGIGAVLKVLTT GLPALISWIKRKRQQ	0.914	85.575
470	test	20	000536	A	GRITKEEKEQLIELAEKIRELVKEAKEL AEAQGLTPEEKKKLSEISEELIQKAKEL EKLLENEGEPSEEEAIKKIISLVEELRR LILELQKLLEEGLSREERAALSRFLDEL KEAEAQAEDLLNGRRKRGIGAVLKVLTT GLPALISWIKRKRQQ	0.907	88.083
471	test	20	000562	A	GLKESLIAKAKELKEKVEKVQSEAKEAA KAVGNDEIREVVEEVEKLQDAVERFIKA LEKGVKIDTEEIKKEIRELGEAAAKLAK AIKSALQADPTASRAQIKKAKKIINQLR EARKEALQVAEGGRRKRGIGAVLKVLTT GLPALISWIKRKRQQ	0.92	85.262
472	test	21	000167	A	SRASKRKIKKAKKIAEDVIKEIEELKKA KRAPSGEEIKRLIERLKELREMIKEILK KGTDLTPEEAIELAELILECAVRAAEYF IENGIKSEEEVKFLLLLLEGTLEALGLV KKQGSGIDQEAKEKLARRALELQRLLRK LASRTGRRKRGIGAVLKVLTTGLPALIS WIKRKRQQ	0.863	85.225
473	test	23	000365	A	APTPEEKLKSAIKKMTALIEEVEKLLES GKKIPDEEERLIKRLIKKLKRLLASLKK REDAEPEEVLQLIEELKKEIEELKAELL SKPSITEEEEKLLKELSQVEKAIQQSMN LFKEGRRKRGIGAVLKVLTTGLPALISW IKRKRQQ	0.913	85.799
474	test	23	000422	A	STITREELKKKIAAAEEEFERLEELYKA GKELSFSELEALIKKVEELLKEIIVALE AGKRLSEEEIRKVLKLSERVLELIKKEA ERKGLITEAEKEELERMKAAIKEAREAI ERGNGRRKRGIGAVLKVLTTGLPALISW IKRKRQQ	0.898	85.906
475	test	23	000642	A	AKPTVDELRKRIKSIKETIERLRKLLKQ GEAVSPAELEKRLIASFEKLVALLRSLA AGYEPTEEEIEEIILYLETLIELVESLL DEKGITEAERKELEELKRALEEARQLLE AVASGRRKRGIGAVLKVLTTGLPALISW IKRKRQQ	0.806	86.358
476	test	23	000687	A	EISREMKRLIERAEAEVEEAIDLIEELP IGKDEKERAALKELIEAEKELIKAAIEA LKSNDPEAFEELVESIEKTLEAIEEFIK VLPDESDLELLFRESIRQMVEGFERARK EISLGRRKRGIGAVLKVLTTGLPALISW IKRKRQQ	0.94	90.171