nature portfolio

https://doi.org/10.1038/s41586-021-04383-5

Supplementary information

A backbone-centred energy function of neural networks for protein design

In the format provided by the authors and unedited

Supplementary Information for

A backbone-centred energy function of neural networks for protein design

Bin Huang[1,#], Yang Xu[1,#], Xiuhong Hu[1,#], Yongrui Liu[1], Shanhui Liao[1], Jiahai Zhang[1], Chengdong Huang[1,2], Jingjun Hong[1], Quan Chen[1,2,*], Haiyan Liu[1,2,3,*]

[1]MOE Key Laboratory for Membraneless Organelles and Cellular Dynamics, Hefei National Laboratory for Physical Sciences at the Microscale, School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230027, China.

[2]Biomedical Sciences and Health Laboratory of Anhui Province, University of Science and Technology of China, Hefei, Anhui 230027, China.

[3]School of Data Science, University of Science and Technology of China, Hefei, Anhui 230027, China.

[#]these authors contributed equally to the work.

*to whom correspondence should be addressed.

e-mails: chenquan@ustc.edu.cn, hyliu@ustc.edu.cn.

Table of Contents

Supplementary Methods

1. The SCUBA energy function

1.1. Energy components

The overall SCUBA effective energy as a function of atomic coordinates is defined as the sum of usual covalent and steric components and a statistically learned component,

$$E_{total}(\boldsymbol{r}) = E_{covalent}(\boldsymbol{r}) + E_{steric}(\boldsymbol{r}) + E_{statistical}(\boldsymbol{r}), \quad \text{(S1)}$$

in which $\mathbf{r}$ stands for coordinates of mainchain non-hydrogen atoms.

The covalent component $E_{covalent}(\mathbf{r})$ is a sum of harmonic functions depending on individual bond lengths, bond angles or improper torsional angles. For mainchain atoms, the steric component $E_{steric}(\mathbf{r})$ is a sum of simple atom-pairwise Lenard-Jones potentials removed of the attractive tails.

The statistical component $E_{statistical}(\mathbf{r})$ is a sum of different types of statistical energies derived using the neighbor counting-neural network (NC-NN) approach, in which the neural networks are simple fully-connected three-layer perceptrons. As shown previously in Ref 29, such networks can faithfully represent high-dimensional molecular energy surfaces, while its analytic form and computational efficiency is useful for applications involving extensive structure sampling and optimization.

Trained on a dataset of 12465 non-redundant natural protein structures determined at resolutions of 2.5 Å or above by X-ray crystallography and with sequence identity between any two chains below 50%[27,30], each type of NC-NN-learned energy terms describes the summed effects of different kinds of interactions on the correlated distributions of a given set of sequentially or spatially local structural variables, or formally,

$$E_{statistical}(\boldsymbol{r}) = \sum_{type} w^{type} E^{type}_{statistical}(\boldsymbol{r}), \quad \text{(S2)}$$

with

$$E^{type}_{statistical}(\boldsymbol{r}) = \sum_i e^{type}_{NC-NN}(\boldsymbol{Q}^{type}_i) \quad \text{(S3)}$$

here $\boldsymbol{Q}^{type}_i$ represent the *i*-th subset of structural variables on which the e^{type}_{NC-NN} term depends. The index *i* varies over all the backbone positions if the energy type depends on sequentially local variables, or over pairs of backbone positions if the energy type depends on spatially local variables.

The NC-NN-learned energy function types and the structural variables on which they depend are listed in Extended Data Fig. 2**a**. They have been chosen to comprehensively cover interactions that potentially determine backbone designability in sidechain type-independent or insensitive ways. We note that the $e_{NC-NN}^{local-HB}$ energy term carries essentially redundant information about local backbone conformations as compared with $e_{NC-NN}^{\varphi-\psi-1}$ and $e_{NC-NN}^{\varphi-\psi-5}$. However, the $e_{NC-NN}^{local-HB}$ term explicitly depends on inter-atomic distances and can thus improve the mainchain hydrogen-bond distance distributions in helices, which would be somewhat too broad if only the torsional angle-dependent $e_{NC-NN}^{\varphi-\psi-1}$ and $e_{NC-NN}^{\varphi-\psi-5}$ energy terms had been considered in SCUBA.

At the beginning, the overall SCUBA energy function had been defined without considering any explicit sidechains. The optimized backbone atom positions on such a backbone-only energy landscape were found to be too "blurred" for subsequent amino acid sequence design (alanine or glycine selected for too many positions on such backbones, see Extended Data Fig. 5**c**). Thus we included explicit sidechains that are described with, besides usual harmonic covalent energies, two additional energy components. One is a sum of the $e_{NC-NN}^{rotamer}$ terms which have been listed in Extended Data Fig. 2**a**. The other is a sidechain packing component, $E^{SC-packing}(\mathbf{r})$, which is a sum over sidechain-sidechain and sidechain-mainchain terms, each atom pairwise term being simply inter-atomic distance-dependent, composed of a Lennard-Jones form repulsive part and an inverted Gaussian form attractive part. The resulting single-well distance-dependent function has two adjustable parameters. The first is the minimum energy distance, which is related to optimum packing distances and has been treated as atom type specific and taken from a previous ABACUS sequence design model[27,28]. The second is the well depth, which, for simplicity, has been treated to be the same for all atom pairs and merged into the weight parameter $w^{SC\text{-}packing}$ for combining this sidechain packing component with other energy components.

In backbone sampling and optimization prior to sequence design, the explicit sidechain-dependent terms are either not considered, or computed with relatively featureless sidechain types (such as the LVG sequences described in the main text). These latter sidechains serve to averagely model the various sidechain-related factors that affect the physical plausibility or designability of backbone structures. These factors may include chiral covalent structures of sidechains, backbone-dependent

sidechain conformation preferences, and sidechain-excluded volumes in mainchain-sidechain and sidechain-sidechain packing.

1.2. Learning the energy terms by fitting kernel estimations to perceptron neural networks

1.2.1. The general idea

For a given set of structural variables and the training structural data, learning an energy term by NC-NN comprises two stages. In the first stage(*i.e.*, NC), single-point estimations of a statistical energy are obtained with simple non-parametric, kernel-based density-estimations in the high-dimensional structural variable space. In the second stage (*i.e.*, NN), a neural network (here a fully-connected three-layer perceptron) representing the statistical energy as an analytical function of the structural variables is trained on a large set of NC-estimated single-point energies.

1.2.2. Defining statistical energy as ratio between probability densities

To formally describe the NC-estimation of single-point energies, for a set of correlated structural variables $\boldsymbol{Q}$, the effective energy function to be learned from observed data is defined as

$$e(\boldsymbol{Q}) = -\,ln\left[\frac{\rho^{observed}(\boldsymbol{Q})}{\rho^{reference}(\boldsymbol{Q})}\right], \tag{S4}$$

in which $\rho^{observed}(\boldsymbol{Q})$ is the probability density of the observed data distributed in the $\boldsymbol{Q}$ space, and $\rho^{reference}(\boldsymbol{Q})$ is the probability density of the same set of variables but expected for an idealized reference system, in which the interactions described by $e(\boldsymbol{Q})$ is absent. Extended Data Fig. 2**a** gives the respective reference distributions of the NC-NN-learned energy functions considered in SCUBA.

1.2.3. Estimating single-point statistical energies by kernel density estimation (neighbor counting)

At any probing point $\boldsymbol{Q}^{probe}$ in the $\boldsymbol{Q}$ space, a single-point estimation of $\rho^{observed}(\boldsymbol{Q}^{probe})\,/\,\rho^{reference}(\boldsymbol{Q}^{probe})$ can be obtained from data by calculating $n_{neighbor}^{observed}(\boldsymbol{Q}^{probe})/n_{neighbor}^{reference}(\boldsymbol{Q}^{probe})$, in which $n_{neighbor}^{observed}(\boldsymbol{Q}^{probe})$ stands for the number of observed data points that fall within the neighborhood of $\boldsymbol{Q}^{probe}$, while $n_{neighbor}^{reference}(\boldsymbol{Q}^{probe})$ stands for the number of neighbors of $\boldsymbol{Q}^{probe}$ among a set of reference data points computationally drawn according to $\rho^{reference}(\boldsymbol{Q})$.

Practically, given the observed and reference data points, $n_{neighbor}^{observed}(\boldsymbol{Q}^{probe})/n_{neighbor}^{reference}(\boldsymbol{Q}^{probe})$ can be estimated with the help of a kernel function as

$$\frac{n_{neighbor}^{observed}(\boldsymbol{Q}^{probe})}{n_{neighbor}^{reference}(\boldsymbol{Q}^{probe})} \approx \frac{N_{total}^{reference}}{N_{total}^{observed}} \frac{\sum_{i\in \text{observed}} h(\boldsymbol{Q}^{probe},\boldsymbol{Q}_i)}{\sum_{i\in reference} h(\boldsymbol{Q}^{probe},\boldsymbol{Q}_i)}. \tag{S5}$$

Because the observed and reference neighbor numbers are estimated with the same kernel function $h(\boldsymbol{Q},\boldsymbol{Q}')$, the ratio $n_{neighbor}^{observed}(\boldsymbol{Q}^{probe})$ / $n_{neighbor}^{reference}(\boldsymbol{Q}^{probe})$ is insensitive to the exact choice of the kernel as long as the kernel has the general property that

$$h(\boldsymbol{Q},\boldsymbol{Q}') = \begin{cases} 1 & \textit{if } \boldsymbol{Q} \textit{ and } \boldsymbol{Q}' \textit{ are sufficiently similar,} \\ 0 & \textit{if } \boldsymbol{Q} \textit{ and } \boldsymbol{Q}' \textit{ are very different. Otherwise} \\ & \textit{switched between 1 to 0 according to similarity.} \end{cases} \tag{S6}$$

If necessary, the similarity criteria (or the radius of the kernel) in formula S6 can be adaptively chosen, being stricter for $\boldsymbol{Q}^{probe}$ in regions densely populated by training data, and being more relaxed for $\boldsymbol{Q}^{probe}$ in sparsely populated regions. Because a smaller radius of the kernel function leads to higher resolution of the resulting model (which is accompanied by increased statistical uncertainty, owing to fewer points being counted as neighbors), the adaptive kernel can balance this trade-off between resolution and statistical uncertainty according to the local distributions of the training data.

1.2.4. Training a fully-connected multi-layer perceptron network to represent the statistical energy function

Each perceptron neural network in SCUBA comprises one input layer, one hidden layer and one single-node output layer, nodes between neighboring layers fully connected. Nodes of the hidden layer use logistic functions for activation. Mathematically, each network performs the following transformation from input to output,

$$f(\mathbf{x}) = b^{2\to3} + \sum_{j=1}^{N_2} w_j^{2\to3} \left\{ 1+\exp\left[-\left(\sum_{i=1}^{N_1} w_{ij}^{1\to2} x_i + b_j^{1\to2} \right) \right] \right\}^{-1}, \tag{S7}$$

in which N_1 is the input dimension size or number of nodes in the input layer, N_2 the number of nodes in the hidden layer, $w_{ji}^{1\to2}$ and $w_j^{2\to3}$ the weights for the connections from the first to the second layers and from the second to the third layers,

respectively, $b_j^{1\to 2}$ and $b^{2\to 3}$ the respective biases of the second and the third layer nodes.

Generally, the input vector $\mathbf{x}$ encodes the structural variables contained in $\boldsymbol{Q}$ on which the effective energy depends, in the following way: each angle variable θ is encoded with a series of triangular function values, for example, $(\sin k\theta, \cos k\theta, k = 1,2,4)$; each inter-atomic distance variable d is transformed into a series of Gaussian function ($\exp\left[-\frac{(d-c_i)^2}{\sigma_i^2}\right]$) centered at c_i and with standard deviations σ_i [29].

As in many machine learning practices, the schemes and parameters for encoding the input and the number of nodes in the neural network have been manually chosen with trials and errors, while the weight and bias parameters have been learned, here from a large number of aforementioned NC-estimated single point energy values.

For every final NC-NN-trained energy function, we have verified not only that the fully-connected three-layer perceptron network can satisfactorily reproduce the NC-estimated energies, but also that the distribution of computationally sampled data points according to $\tilde{\rho}(\boldsymbol{Q}) \propto exp\left[-e_{NC-NN}^{type}(\boldsymbol{Q})\right]$ can closely mimic the corresponding distribution of observed data points in the $\boldsymbol{Q}$ space.

1.2.5. Details of learning the three-layer perceptron neural networks from data

For each statistical energy term, the loss was the sum of squares for error. The optimization algorithm used was the momentum optimizer as implemented in the TensorFlow machine learning package[35]. The training/test accuracy were measured by the squares of error averaged over the training/test data points.

More details for each of the energy terms are given below.

For $e_{NC-NN}^{\varphi\text{-}\psi\text{-}1}$:

a) Observed data points for NC: 1,252,924 backbone positions in the training proteins that are not in helices and strands.
b) Reference data points for NC: not needed for a uniform distribution.
c) Probing points for training NN: 32,400 uniformly distributed points on the 2D plane.
d) NN attributes: 2 input dimensions encoded by 12 input-layer nodes, 16 middle-layer nodes, and 225 parameters. The mean square fitting error is 0.03.

For $e_{NC-NN}^{\varphi\text{-}\psi\text{-}5}$:

a) Observed data points for NC: 1,722,251 five-residue segments in the training proteins.

b) Reference data points for NC: independent combinations of separately drawn points from one 1-D distributions for ψ, three 2-D distributions for (φ, ψ), and one 1-D distributions for φ. For each distribution 50,000 points were drawn.

c) Probing points for training NN: 165,000 observed points from training data and 165,000 points drawn from the reference distribution.

d) NN attributes: 8 input dimensions encoded by 64 input-layer nodes, 72 middle-layer nodes, and 4,753 parameters. The mean square fitting error is 0.52.

For $e_{NC-NN}^{site-pair}$:

a) Observed data points for NC: 27,590,092 backbone pairs with inter Cα distance below 12.1 Å in the training proteins.

b) Reference data points for NC: independent combinations of separately drawn points from one 6-D distribution for the inter-site translations and rotations, and two 4-D distributions, each for the four (φ, ψ) angles surrounding one site. The number of reference points covering the 6-D space is the same as the observed data points. 2,805,520 observed, four-residue segments served as reference points for the 4-D distributions.

c) Probing points for training NN: 95,000 observed points from training data and 176,000 points drawn from the reference distribution.

d) NN attributes: 14 input dimensions encoded by 191 input-layer nodes (the 6-D translation and rotations are first surrogated by inter-atomic distances and then encoded), 64 middle-layer nodes, and 12,353 parameters. The mean square fitting error is 0.60.

For $e_{NC-NN}^{local\text{-}HB}$:

a) Observed data points for NC: 767,524 pairs of three-atom units from the training proteins, with the two units separated by two to four residues along the primary sequence. and O-N distance blow 5.5 Å.

b) Reference data points for NC: 7,675,240 pairs of three-atom units with

uniformly distributed relative translations and rotations, and with O-N distance below 5.5 Å.

c) Probing points for training NN: 760,000 observed points from training data and 76,000 points drawn from the reference distribution. Here the number of probing points drawn from the reference distributions has been intentionally chosen to be one order of magnitude less than the probing points taken from the observed data points, so that the training could focus on fitting the lower energy regions (populated by the probing points taken from the observed data).

d) NN attributes: 5 input dimensions encoded by 35 input-layer nodes, 32 middle-layer nodes, and 1,185 parameters. The mean square fitting error is 1.13 (the variation range of the energy is about 23 in arbitrary unit).

For $e_{NC-NN}^{rotamer}$

a) Observed data points for NC: 4×10^4 to 2.8×10^5 points from training structures for various sidechain types.

b) Reference data points for NC: 2×10^5 points with backbone φ-ψ angles randomly drawn from observed data combined independently with sidechain torsional angle(s) randomly drawn from uniform distributions.

c) Probing points for training NN: 4×10^4 to 2×10^5 observed points from training data, and the same numbers of points drawn from respective reference distributions.

d) NN attributes: the number of input dimensions is (2 + *number_of_flexible_sidechain_torsional_angles*), each torsional angle is encoded with 6 input-layer nodes. The number of middle-layer nodes is 24. The number of parameters is 1 + 24 × (6 × *number_of_input_dimensions* + 2). The mean square fitting errors are around 0.1 or smaller for various sidechain types.

We note that the NN fitting errors listed above are the training errors. Estimated using data points that had not been used for training, the test errors were similar to the training errors if the test data were composed of the same fractions of observed versus computationally-drawn probing points as the training data. The test errors were smaller than the training errors if the test data were composed solely of observed probing data points. This is understandable because the NC-estimated training energies themselves

are noisy. Compared with the computationally drawn probing points, the observed data as probing points are more enriched in regions of higher probability densities, and their NC energies suffer less from statistical uncertainty. Thus, in regions where the observed data points are located, the perceptron models are statistically more accurate than the levels indicated by the overall fitting errors. This is a desired property in applications, because it is the accuracy in these regions that determines the quality of the optimized backbones.

1.3. Calibrating the weights of the energy terms in SCUBA

To sample or optimize protein structures on the SCUBA energy landscape, we applied stochastic dynamics (SD) simulations in which the Langevin equations of motion are integrated to obtain the time trajectories of atoms moving in a frictional medium at a given temperature[26]. SD simulations of a test set of natural proteins starting from their native X-ray crystal structures have been employed to calibrate the few undetermined energy weight parameters in the overall SCUBA energy function *(i.e.*, w^{type} in formula S2). The objective for calibration was to stabilize the native conformational states relative to conformations deviating from the native states. In addition, we would like the overall interaction strength to be as weak as possible, so long as the native structures can retain their stability at a reduced temperature of $T_r = 1$.

We carried out the calibration using an approach that is conceptually similar to force field refinements using thermodynamics cycles in conformational space[36]. Briefly, to refine a set of trial parameters, two set of simulations using the trial parameters were carried out on the test proteins. One set comprised unrestrained simulations in which the structures were allowed to freely deviate from the starting native structures according to the uncalibrated energy function. The other set comprised restrained simulations in which sampling in the conformational space was restricted to regions near the native states (here we restrained the secondary structure mainchain atomic RMSD from native structures to be below 2.5 Å). Then the various energy components averaged over conformations sampled in the two sets of simulations were compared. If an energy component is systematically lower in conformations with larger RMSDs from the native structures, its weight would be reduced in the next round of test simulations. Otherwise, its weight could be kept the same or tentatively increased.

We estimated an initial set of weights in this way by first considering only two small globular test proteins, one all α (PDB ID 3l32) and another α+β (PDB ID 1a6j, chain A). The resulting weights were then further refined by considering 33 manually selected test proteins (see Extended Data Fig. 2**b**), which are relatively small, globular, soluble, with X-ray structures of relatively high-resolutions, containing no disulfide bond, and belonging to different fold classes.

After obtaining a calibrated set of energy weights ($w^{\varphi-\psi-1}=2$, $w^{\varphi-\psi-5}=0.5$, $w^{site-pair}=0.32$, $w^{local-HB}=0.6$, $w^{rotamer}=2.4$ and $w^{sc-packing}=3.1$), we carried out restraint-free SD simulations at $T_r=1$ with each of the weights separately and systematically scaled by a value between 0.2 and 2.0 with the other weights fixed, or simulations with the total energy scaled by a value between 0.25 to 2.5. The results confirmed that the native structures of a majority (26 out of 33) of the tested proteins located close to minima on the SCUBA energy landscape (the average backbone RMSD of structures sampled in the SCUBA simulations from respective native structures are below 2 Å), with the chosen weights provided sufficiently strong interactions to maintain the stability of the native structures at $T_r=1$.

After calibrating the weights on the proteins with their explicit natural sidechains, we carried out simulations in which the sidechain types were changed according to the LVG sequences (i.e., leucine on α helices, valine on β strands and residues in loops removed of sidechains) to check if the native backbones remained stable with non-native sidechains according to SCUBA.

1.4. Computational costs

We note that the time needed for learning the NC-NN models are irrelevant to most applications because it has already been done once and for all. That said, the most time-consuming part for NC-NN learning was the NC step, which took a few days on a multi-CPU workstation for all the SCUBA energy terms. The subsequent training of the NNs were inexpensive and took hours on common desktops. For costs of the SCUBA-driven SD simulations, to simulate a backbone of approximately 100 residues for 1 *ps* takes 10 to 15 seconds using 4 cores of an Intel Xeon E5-2680 v4 CPU. Usually 300 to 400 *ps* of simulations are used to finish the overall optimization stages of a single backbone (see below).

2. *De novo* backbone design

2.1. Constructing initial structures according to sketches

The basic strategy of designing a *de novo* backbone with SCUBA is to carry out stochastic dynamics simulations with simulated annealing (SASD) starting from initial structures that are built according to a set of user-defined specifications or a sketch. By using an initial structure conforming to the sketch, only the part of the conformational space that are relevant to the given sketch are searched for physically plausible backbones.

In the current work, sketches have been composed following the "periodic table" abstraction of protein structures described by Taylor et al[32]. In that model, a well-folded protein was abstracted into secondary structure elements organized into roughly parallel layers, each layer comprising either a β-sheet of multiple strands or approximately parallel or antiparallel helices. For each sketch, we predefined the number of secondary structure segments, their approximate sizes, and their orders and directions in different abstractive layers. According to a sketch, we first generated peptide segments in local conformations of helices or strands at approximate relative positions, and then built loops to connect them.

The secondary structure segments were placed in one of two ways. The first way (see Extended Data Fig. 3) was to place the starting or ending atom of each segment at a cross point of a 2-D planar grid and then to grow the entire segments in perpendicular to the grid plane (which is perpendicular to the planes of the secondary structure layers). The end points of segments in the same secondary structure layer fell onto the same straight line, the straight lines corresponding to different layers were parallel and separated by 10 Å intervals. The distances between neighboring helices in the same layer were approximately 10 Å, and those between neighboring strands were approximately 8 Å. To grow each segment, the internal backbone torsional angles were randomly sampled according to the designated secondary structure type of the segment. The second way of placing the secondary structure segments was to use a computer graphic system to interactively place idealized helical or strand segments at approximate positions. Initial structures of the EXTD and the H4 sketches have been generated in the second way, while initial structures of the other sketches (H2E4 and those in Extended Data Fig. 4) have been generated in the first way.

To generate a loop to connect two pre-placed secondary structure segments, we first generated an unclosed loop using backbone (φ, ψ) torsional angles randomly selected according to the Ramachandran backbone angle distribution of protein coils,

and then attempted to close the loop with the kinematic loop closure algorithm[37]. The loop length was first set at 3 and then gradually increased, until a kinematic loop closure solution could be found within 1000 loop generation attempts for the given loop length.

We note that the lengths of the α-helix and β-strand segments defined in the sketches are only approximate, because extension or shrinking of the predefined secondary structure segments can take place spontaneously in SCUBA-driven SASD.

2.2. Optimizing the backbone structures by SASD simulations

After building the initial backbones, we optimized them with SCUBA-driven SASD simulations in two substages. The backbone-only model was used in substage 1 while the LVG sequences was used in substage 2.

In substage 1, the local conformations of the designated helix or β-strand segments were restrained, and the steric interactions involving loop atoms were multiplied by a factor of 0.01. These treatments simplified the overall energy landscape and reduced the search space, so that approximate SCUBA minima conforming to the user-defined sketches could be efficiently located. The simulations in substage 1 included an initial 10 *ps* simulation at a low temperature and with a large frictional coefficient (the reduced temperature $T_r = 0.1$ and the frictional coefficient $\gamma = 5\ ps^{-1}$). This simulation served to remove any possible tense strains in the initial backbones. The resulting backbones were then optimized by a 60 *ps* SASD simulation, in which the temperature has been varied between 2.0 and 0.5 in 6 cycles, each cycle comprising a 4 *ps* SD at $T_r = 2.0$, followed by a 1 *ps* SD in which T_r gradually decreased from 2.0 to 0.5, and then by a 5 *ps* SD at $T_r = 0.5$. To compensate for the thermal expansion effects of the higher temperatures in simulated annealing, the overall radius of gyration has been restrained (see also the legend of Extended Data Fig. 2**b**). With the above protocol, a majority of the simulations starting from varied initial structures constructed for the same sketch could produce backbone structures meeting the design specifications defined by the sketch. In addition, the lowest SCUBA energies usually converged in the last several simulated annealing cycles.

In substage 2 of backbone optimization, LVG sidechains were added, local conformation restraints were removed, and interactions involving loop residues were recovered to full strength. SD simulations of this substage comprised a 4 *ps* relaxation with $\gamma = 5\ ps^{-1}$ and $T_r = 0.2$, followed by a 120 *ps* SASD run with the temperature changed between 2.0 and 0.5 in 10 cycles, with the radius of gyration restraint applied.

The resulting structures were finally refined by another 120 *ps* SASD run with the temperature cycled between 0.5 and 0.2. In majority of the final refinement simulations, both the structures and the total SCUBA energies fluctuated with relatively small amplitudes, indicating that stable minima on the SCUBA energy surface were reached.

2.3. Resampling and optimization of loops

The final round of experimental tests on the H2E4 and the H4 designs have been carried out with designed backbones generated after further extensive resampling and optimization of loops. The loop resampling and optimization were performed on the backbones optimized by the above two-substage SASD simulations.

In the loop resampling process, the regular secondary structure regions have been fixed. The starting/ending positions of each loop were systematically varied within the last/first three backbone positions of the loop's flanking secondary structure elements. For the H2E4 backbones, the loop lengths were gradually increased from 3 up to the loop lengths before resampling. For the H4 backbones, we considered the minimum loop lengths for which loop closure solutions could be found, as well as loop lengths one-residue longer than the respective minimum lengths. For each loop length, 1000 randomly generated and closed starting loop configurations were separately optimized with SCUBA-driven SD simulations ($T_r = 0.5$) until the energy fluctuations became small.

From the loop sampling results, a set of candidate backbone structures were selected for each loop. For the H2E4 designs, this set comprised the 10 lowest energy non-redundant structures of the optimum loop length (meaning the lowest per-residue SCUBA energy estimated for conformations of this loop length was lower than the corresponding energies estimated for any other loop lengths). The set was augmented with all (non-redundant) sampled loop structures that were one-residue shorter but with per-residue SCUBA energies lower than the 10^{th} lowest energy of the optimum loop length.

To obtain complete backbone structures including several loops, combinations of candidate structures of different loops were sampled, and the final total SCUBA energies were compared between different combinations. For each backbone considered for loop resampling and optimization, 10 lowest energy final backbone structures were considered for subsequent sequence selection. This led to 500 final H2E4 backbone structures used for subsequent sequence selection (50 backbones

optimized from different initial structures × 10 alternative structures with re-optimized loops).

For the H4 designs, 116 backbone structures optimized from 30 initial backbones of different topologies (see topologies in Figs. 3**c-f)** were subjected to loop resampling and optimization. For each H4 backbone which contains 3 loops, systematic combination of 10 candidate structures for each loop led to 1000 loop-optimized backbone structures. From the resulting total 116000 backbone structures, 3382 structures of lower through-space interaction energies for loops (according to the total $e_{NC-NN}^{site-pair}$ involving loop residues) were considered for subsequent sequence selection.

3. Sequence selection with ABACUS2 and filtering with Rosetta biased forward folding

3.1. Backbone relaxation-sequence selection iterations

The backbones optimized with the LVG sequences (and after loop resampling for the finally experimentally tested H2E4 and H4 proteins) were used for ABACUS2 sequence selection in the first iteration. In later iterations, the backbones were relaxed by SCUBA-driven SD with the explicit sidechains of the ABACUS2 designed sequences.

When repeatedly applying this backbone relaxation-sequence selection iteration, we found that excluding the sequences selected in the first iteration, the sequences selected for backbones relaxed in different later iterations were of more than 50% amino acid sequence identity and of similar ABACUS2 energies. On the other hand, in the selected sequences, the frequencies of alanine and glycine gradually increased with the number of iterations. This phenomenon is likely to be caused by that a small residue introduced in an earlier iteration could lead the room for sidechain placement at the corresponding position to shrink in subsequent backbone relaxation. Then larger residues could no longer be selected for that position in later iterations.

Because of this caveat of introducing an increasing number of small residues with the increasing number of iterations, and also because of the relatively high-similarity of the sequences selected in different iterations, we limited the number of backbone relaxation-sequence selection iterations to two or three.

More specifically, the protocol applied to generate the H2E4 sequences was the following: given an optimized target backbone, ABACUS2 sequence design was carried out with the sidechain atom radius parameters multiplied by 0.9 to account for

the fact that the backbones had been optimized with not the actual sidechains. Then the backbones were relaxed with SCUBA-driven SASD simulations (4 *ps* constant temperature SD with $\gamma = 5\ ps^{-1}$ and $T_r = 0.2$ followed by 20 *ps* simulated annealing SD with $\gamma = 0.5\ ps^{-1}$ and T_r decreased from 1.0 to 0.2). Five different sets of random initial velocities were used for the SD simulations to produce 5 different relaxed backbone structures. Each structure was separately subjected to two iterations of ABACUS2 sequence selection (no downscaling of the atomic radius parameters) followed by backbone relaxations with the same SASD protocol. The five resulting backbones were used for final ABACUS2 sequence selection, 20 sequences selected for each backbone and the lowest energy one retained. With the standard set of ABACUS2 parameters, somewhat too high proportions of solvent-exposed positions on β-strands were selected to be threonine and those on α helices selected to be alanine. This could potentially increase the tendency of the designed proteins to aggregate. To investigate into this, we designed additional sequences with the ABACUS2 residue type-specific reference energy parameters[28] adjusted in three different ways to increase the probability of selecting other polar residues at such positions.

More specifically, the XM1* proteins of the H2E4 architecture have been designed with the default ABACUS2 reference energies. The XM2* proteins have been designed with the ABACUS2 reference energies for polar residues at non-buried helix and strand positions (except for threonine at strand positions) changed by -0.8 (arbitrary energy unit, the same below). This adjustment increased the usage of polar residues at non-buried positions on secondary structures. In addition, reference energies for isoleucine and valine at non-buried strand positions were increased by 1.0 and 1.5, respectively, to reduce the use of these nonpolar residue types at such positions. The AM1* proteins have been designed with the ABACUS2 reference energy for alanine at non-buried helix positions increased by 0.5. This adjustment reduced the usage of alanine at helix positions. The AM2* proteins have also been designed with the ABACUS2 reference energy for alanine at non-buried helix positions increased by 0.5. In addition, the reference energy for threonine at strand positions have been increased by 1.2, and the reference energies for isoleucine and valine at non-buried strand positions have been increased by 0.9. The latter adjustments increased the usage of other residue types (especially those polar types) at non-buried strand positions. The different adjustments of the reference energies did not affect the usage of residue types for loops or for buried positions.

Because the H2E4 sequences designed with different ABACUS2 reference energies differed only at solvent-exposed positions of regular secondary structure elements, we considered the differently selected sequences as equally plausible in subsequent computational filtering and experimental characterization.

The specific protocol to generate the EXTD and the H4 sequences was the following: given an optimized backbone, 10 sequences were designed using ABACUS2 (standard parameters) and the lowest energy one was kept. Then the backbone was relaxed by one cycle of 4 *ps* simulated annealing with the reduced temperature dropped from 1.0 to 0.2. Then the backbone relaxation-sequence selection iteration was repeated three more times with the same SASD protocol. For each designed backbone, the sequence of the lowest ABACUS2 energy among all iterations (not necessarily from the last iteration) was retained.

3.2. Computationally filtering the designed sequences

The sequences designed for the loop re-optimized H2E4 and H4 backbones were computationally filtered by Rosetta biased forward folding[10,15] before experimental characterization. In these folding simulations, fragment conformations pre-predicted based on sequences were assembled into overall structures. To predict if a given sequence was likely to fold into a particular target structure, the pre-predicted conformations for each fragment included only three members that were of the lowest RMSDs from the corresponding fragment in the target among all conformations predicted based on the sequence[10]. With this restricted selection of possible fragment conformations, only a small number of overall structures need to be generated and the lowest RMSD of these structures from the target can be used as a judging criterion. For every ABACUS2-selected sequence for the H2E4 sketch (H4 sketches), 50 (200) biased forward folding models were generated. The lowest RMSD from biased forward folding (and the Rosetta energy score of the lowest RMSD model) was considered as filtering criteria for final experimental tests. For the H2E4 sketch, 2000 selected sequences (4 sequences selected for each of the 500 backbone structures by using different ABACUS2 reference energies) were computationally filtered with biased forward folding. From each group of sequences designed with different ABACUS2 reference energies, 6 or 9 sequences associated with lowest RMSD Rosetta predictions were selected for experimental characterization. From the 3382 ABACUS2-selected sequences covering the four different H4 sketches, 8 sequences associated with the

lowest RMSD of Rosetta biased forward predictions (2 sequences for each sketch) were selected for experimental characterization.

3.3. Designing helical proteins of novel architectures

First, 10,000 initial structures of randomly arranged helical segments were generated in the following way. Six peptide backbone segments of 10 to 25 residues in helical conformations were randomly generated by sampling the Ramachandran torsional angles in the helical region. They were placed at uniformly-distributed random positions and in uniformly-distributed random orientations in the same spatial region, their centers of geometries fallen within a sphere of a radius of 15 Å.

Then each of the initial configurations were optimized by two-stage SCUBA-driven simulated annealing to obtain a diverse set of plausibly packed helices, with the following protocol. Stage 1: 50 *ps* SASD simulation of the backbone-only model with the temperature changed between 1.0 and 0.5. At this stage, the steric term in SCUBA was not considered because the initial configurations contained numerous steric clashes. Stage 2: 50 *ps* SASD simulation with all sidechains changed into leucine and the temperature changed between 2.0 and 0.5.

The 6-segments configurations thus obtained were used as the basis for building designable backbones of novel helical architectures. For a given configuration, we consider every two helix segments in that configuration in turn and try to find ways of connecting them with a loop of 3 or 4 residues, allowing possible truncation of either segments in the connected chain. This is done by considering all possible combinations of a N-side connecting (truncating) point on one helix paired with a C-side connecting (truncating) point on the other helix.

The plausibility of a pair of connecting (truncating) points was empirically scored in the following way. We extracted all helix-loop-helix motifs of 3 to 4 loop residues from the non-redundant PDB structures and collected the following set of six atomic coordinates: if i is the residue to start the loop on helix A and j is the residue to end the loop on helix B, then we collected coordinates of the six Cα atoms of residues i, i-2, and i-4 from helix A and of residues j, j+2 and j+4 from helix B. We applied the kernel neighbor counting approach (see also section 1.2.3) to estimate the local density in the space of the relative geometries between the six atoms. The local density for the natural loops were estimated, ranked, and then used as reference values for scoring the connecting points.

By retaining all paired connecting points that led to local density higher than 50% of the natural loops, we built a pool of plausible paired connecting points between the helical segments in the SCUBA-optimized configurations. Then we enumerated all possible schemes of ordering the segments from the N to the C terminus so that as many as possible segments can be sequentially connected through the retained connecting point pairs. For each ordering and connecting scheme, we built closed loops of 3 to 4 residues and optimized loop backbones by SCUBA, and obtained a configuration of a single connected chain. The configuration was further relaxed by 50 *ps* SCUBA SASD simulated annealing with the temperature changed between 1.0 and 0.5, with leucine sidechains on all helix residues and no sidechains on loop residues. Then ABACUS2 sequence selection was applied to the backbone, followed by two iterations of SCUBA backbone relaxation followed by ABACUS2 sequence reselection. The sequence of the lowest ABACUS2 energy was taken as the final design result.

During the above process, the following filters have been applied at various stages: the connected single chain should comprise at least five (truncated) helices, all the connected helices must form a single compact structure, the Dali[38] Z-score of the backbone with the most similar PDB structure must be below 6.0 for structure novelty, the average per-residue ABACUS2 energy must be below -0.5, and the Rosetta biased forward folding result should agree with the design model (RMSD below 1.5 Å).

Finally, we selected 13 design results for experimental characterization based on combined considerations of structure novelty, ABACUS2 sequence energies, and Rosetta biased forward folding results. The model backbone structures of these designs are shown in Extended Data Fig. 7**c**.

4. Experimental methods

4.1. Preparation of recombinant proteins

DNA sequences encoding the designed proteins were synthesized and cloned into the *Nde*I and *Xho*I sites of pET-22b(+) by TsingKe Biotech and General Biotech. D12, D22 and D53 were subcloned into MBP expression vectors V28E2 or V28E4[39] with tiny modification of N-terminal residues (see sequences in Supplementary Table 1) for crystallization. The plasmids were transformed into *E. coli* BL21(DE3) and induced at OD_{600}~0.8 with 0.5 *mM* IPTG for 20 hours at 16℃. For ^{1}H-^{15}N HSQC NMR study, uniformly ^{15}N-labeled proteins were prepared by growing the bacteria in inorganic medium (24 *g/L* KH_2PO_4, 5 *g/L* NaOH, 0.5 *g/L* $^{15}NH_4Cl$, 2.2 *mM* $MgSO_4$, 0.1 *mM* $CaCl_2$ and 2.5 *g/L* glucose) using $^{15}NH_4Cl$ as isotope source. Cells were harvested and

sonicated in buffer containing 20 *mM* Tris and 500 *mM* NaCl at pH 7.8. The level of expression and solubility of the designed proteins were evaluated using SDS-PAGE. The soluble supernatant was purified by Ni^{2+} affinity chromatography using 500 *mM* NaCl, 20 *mM* Tris-HCl and 350 *mM* imidazole. Eluted proteins were concentrated in buffer containing 20 *mM* Tris, 300 *mM* NaCl and 1 *mM* EDTA, pH 7.8 and subjected to gel filtration in a Superdex 75 column with the ÄKTA purifier system (GE Healthcare). The monomeric fractions were collected for structure characterization.

4.2. NMR data acquisition

All NMR data were acquired at 298 *K* on a Bruker DMRX500 spectrometer equipped with triple resonances, self-shielded z-axis gradient probes. The NMR samples typically contained 0.35 *mM* ^{15}N-labeled proteins, 20 *mM* NaH_2PO4, 50 *mM* NaCl, 1 *mM* EDTA (pH 6.2) and 10% (v/v) D_2O. Data were processed using the programs NMRDraw/NMRPipe[40] and SPARKY[41].

4.3. Crystallization and X-ray diffraction analysis

Crystallization screening was conducted at 289 *K* applying the sitting-drop vapor diffusion method and commercial screens. Purified proteins at 10-20 *mg/ml* were added to the respective crystallization buffer. The crystals used for data collection were grown in 24 hours in 2.0 *M* sodium chloride and 10% w/v PEG 6000 for XM2H, 3.5 *M* sodium formate and 0.1 *M* sodium acetate trihydrate, pH 4.6 for AM2M, 20% w/v PEG 10000 and 0.1 *M* Hepes, pH 7.5 for H4A2S, 1.1 *M* ammonium tartrate dibasic and 0.1 *M* sodium acetate trihydrate, pH 4.6 for H4C2R, 20% w/v PEG 3350, 0.2 *M* potassium chloride, pH 7.0 for EXTD-3, 25% w/v PEG 3350, 0.1 *M* sodium Acetate for D12, 0.1 *M* sodium HEPES, 50% v/v PEG500 MME, pH 7.24 for D22 and 2.2 *M* sodium malonate, pH7.0 for D53. Crystals of H4A1R appeared in 7 days in buffer containing 2.0 *M* potassium/sodium phosphate, pH 7.0. All crystals were shortly soaked in reservoir solution supplemented with 15% glycerol (v/v) except 25% glycerol (v/v) for D53 and then flash frozen in liquid nitrogen. The diffraction data of XM2H, AM2M, H4A1R and D53 were collected on Rigaku XtaLAB PRO 007 HF at 1.5406 Å and processed using the CrysAlisPro software suite (CrysAlisPro Software System, Version 1.171.39.35c, Rigaku). Data sets of EXTD-3, H4A2S, H4C2R, D12 and D22 were collected on BL18U1, BL19U1[42] or BL02U1 at 0.9785 Å to 0.9791 Å and processed using XDS[43] at Shanghai Synchrotron Facility (SSRF). The designed structures served as search model for molecular replacement with PHENIX[44]. Model rebuilding was performed in Coot and the final structures were refined by PHENIX. The statistics for

data collection and structural refinement are summarized in Extended Data Fig. 8**a**. Figures of protein structures were made using the PyMOL[45] program.

4.4. Circular dichroism

Circular dichroism (CD) data were collected on a AppliedPhotophysics Chirascan™ V100 spectrophotometer using a 1 *mm* path length quartz cuvette. Protein samples were prepared in 20 *mM* Na_2HPO_4 buffer (pH 8.0) and protein concentrations were adjusted to 0.6-0.8 *mg/ml* before measurement. The CD spectra were scanned in the far-UV range (200-260 nm) and the thermal denaturation CD data were obtained by measuring ellipticity every 5 degree of temperature increasing from 20 ℃ to 95 ℃ at λ=218 *nm* for XM2H or λ=222 *nm* for H4A1R. The CD spectra data were processed and the secondary structure contents were estimated with built-in software suite Pro-Data Viewer v4.5 and Deconvolution v2.1 respectively.

Supplementary References

35 Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *OSDI'16: Proc. 12th USENIX Conf. Operating Systems Design and Implementation* (chairs Keeton, K. & Roscoe, T.) 265–283 (USENIX Association, 2016).

36 Cao, Z. & Liu, H. Using free energy perturbation to predict effects of changing force field parameters on computed conformational equilibriums of peptides. *The Journal of Chemical Physics* **129**, 015101, doi:10.1063/1.2944248 (2008).

37 Coutsias, E. A., Seok, C., Jacobson, M. P. & Dill, K. A. A kinematic view of loop closure. *Journal of Computational Chemistry* **25**, 510-528, doi:10.1002/jcc.10416 (2004).

38 Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Research* **44**, W351-W355, doi:10.1093/nar/gkw357 (2016).

39 Jin, T. et al. Design of an expression system to enhance MBP-mediated crystallization. *Scientific Reports* **7**, 40991, doi:10.1038/srep40991 (2017).

40 Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR* **6**, 277-293 (1995).

41 Lee, W., Westler, W. M., Bahrami, A., Eghbalnia, H. R. & Markley, J. L. PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy. *Bioinformatics* **25**, 2085-2087, doi:10.1093/bioinformatics/btp345 (2009).

42 Zhang, W. et al. The protein complex crystallography beamline (BL19U1) at the Shanghai Synchrotron Radiation Facility. *Nuclear Science and Techniques* **30**, 170, doi:10.1007/s41365-019-0683-2 (2019).

43 Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallographica Section D* **66**, 133-144, doi:10.1107/S0907444909047374 (2010).

44 Adams, P. D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallographica Section D* **58**, 1948-1954, doi:10.1107/S0907444902016657 (2002).

45 The PyMOL Molecular Graphics System, Version 1.5 Schrödinger, LLC.

Supplementary Tables

Supplementary Table 1. Protein sequences of designs selected for experimental testing.

Topo	Protein	Amino acid sequence
EXTD	EXTD-1	AMTLSDCQQELSLVQTVTRGSRAFLSREEAQHFVKEC GLLNCEAVLELLICHLHWGMQVIKILKEFRARMRAAT TREALRQAVQELYDMMKEHAQNGSMLQILYEKIKKIL NDNFEELSLEEIAARVMAAVKRVLEMTECVQMRALG AVASLGCTDLLPQEHILLLTRPRLQELSAGSPGPVTNK ATKILRHFEASC
	EXTD-2	AMTLSDCQQELSLVQTVTRGSRAFLSREEAQHFVKEC GLLNCEAVLELLICHLRSPGGQLILKALQYAIEVVKME NVSVREAAQQLKERVKALGFQEGVGEILRRILDIVIKL AEKGDVEEMKETIKRLVARARAWIECVQMRALGAVA SLGCTDLLPQEHILLLTRPRLQELSAGSPGPVTNKATKI LRHFEASC
	EXTD-3	AMTLSDCQQELSLVQTVTRGSRAFLSREEAQHFVKEC GLLNCEAVLELLICHLRLGMEIMKLGRQLREAVRAND VDAMLKIAKEIIKVIGETGLDEVYRQLLKAAKEFLERR AENFSHEEAVAFAQQIIQLIKQVECVQMRALGAVASL GCTDLLPQEHILLLTRPRLQELSAGSPGPVTNKATKILR HFEASC
	EXTD-4	AMTLSDCQQELSLVQTVTRGSRAFLSREEAQHFVKEC GLLNCEAVLELLICHLRRAMRNFEATQQLIDMIRKAG GNQDNVREAIKELLERARAGEHNITEVQIQILKALYQF LQQHDWSLEKAIKFVKETAKRLFKEYECVQMRALGA VASLGCTDLLPQEHILLLTRPRLQELSAGSPGPVTNKA TKILRHFEASC
	EXTD-5	AMTLSDCQQELSLVQTVTRGSRAFLSREEAQHFVKEC GLLNCEAVLELLICHLQRMNVEAVRELIAALQRLKAG SDVLETIKRVIEIIKKIAYEITEELQKILDYLRAAYRRGQ VSDEVIAAVKKAVEFLRKWVECVQMRALGAVASLGC TDLLPQEHILLLTRPRLQELSAGSPGPVTNKATKILRHF EASC
	EXTD-6	AMTLSDCQQELSLVQTVTRGSRAFLSREEAQHFVKEC GLLNCEAVLELLICHLTRMNLAAVREAIRAVERARAG EDFREIIKQLIEIIKKIAFSITEELRKLLEQLKKYYKTGK WSDEVLAVMQKAVDFLRAFVECVQMRALGAVASLG CTDLLPQEHILLLTRPRLQELSAGSPGPVTNKATKILRH FEASC

	EXTD-7	AMTLSDCQQELSLVQTVTRGSRAFLSREEAQHFVKEC GLLNCEAVLELLICHLRRKNFEQALERLRKLIREAAAG KYTQEDILKFIVQVAIQLLYMSEEEIRAAVRRIVQQLW SSIEEMDGLFKIAKVVVQFAKKYMECVQMRALGAVA SLGCTDLLPQEHILLLTRPRLQELSAGSPGPVTNKATKI LRHFEASC
	EXTD-8	AMTLSDCQQELSLVQTVTRGSRAFLSREEAQHFVKEC GLLNCEAVLELLICHLHGNTAVQRALERLRKLYQDTS MSVRERIEKAIAILTALAPPSGDSWLRKTLEKVRRIAQ QGNFSEETAKKILETAIEMVKKFIECVQMRALGAVASL GCTDLLPQEHILLLTRPRLQELSAGSPGPVTNKATKILR HFEASC
H2E4	XM1A	LTWTATVDSASPEAAAAAAARRLAERVREAGITSPATV TATANGITFTYTVPVSLSEEGLAALVAAAVRALLEAH RATNGQSATLTAG
	XM1B	LAATATVDSASEEAVRAAAAAVVARLVAAGVEGKVT VTATANGVTFTYTADLRRSEEAAAAVAAALAAAGLA AARATGGGTVTLTAV
	XM1C	RAATATVDSRSEEALRAAARALVAAILARFTSPHEVVI TVTANGRVTFTYTVEVDGSEEALAAALAAAIAALRAA VEATGSSATVTAV
	XM1D	ERFTVTVDSASLAALRAAAAKAAALIKAANISGKITTT VTANGVTYTFTVTGPLSEETIAAVAAAILRAARAAVA AGGTSLTITAG
	XM1E	RSYSATVDSNSKEAILAAARRLAERAVAAGVRTSVTV TATADGTKFTFTAPLDESEEGIAAVAAAALRAGLAAV RATGGRSVTLTVV
	XM1F	LPVTATIDASATREQVRAAARAAARRLRELGVSGTVV VTATAGDVSFTYTADLDGSEEALAAVAAAAFAAGLA ALKAAGGTKTVTLTAF
	XM2G	GMKYSAEVSSSMSLEEIRQVIAELIKQAVRAAGGRKG KVRVRVETSDGYVFEFEAEISSEEKIRELAERAAQAILE AGGKSARVEAEV
	XM2H	HSWSATVDSRSEEAVRAAARRLAERLLAAGISGKIKIE VEANGIKYEYEVEGPATEEVAKKIVEYAVAAALRAIA AGATSVTITVG
	XM2J	VAEYEYEVPAGNIEEVFKIVEEALIAAYEAAGHSARVR IELEANGVRMEYEFEISSEEDAREIAQRVAEAFKKYPSE RATVKVSV
	XM2K	WYATATVDSRDREAAREAARRLAQALLEANVTGKIY VRVEAGGHEYEFEAEGPPSEEVAQAIVEAAVRAVLEGI AQGAKSATVEVG

	XM2M	LKVRVRLENPTLEQIREAARRAVEALVRLGVEQSKVR VRVRGGEEYEYEAELTRTEESIKRVAEALYKAGKEAL EHTDTSIELEAV
	XM2N	NWFEVEIDSRSEEVAREGAREAARRVVAAGITGKKVR AEAEGNGVKYSFEAEMNASDEESIQRVIEAAAEAVRR AAAAGADKVRLRVG
	AM1A	LPVTVTAPTADLEALRAAARAAVAFLVAAGVEGPVV VTATAGDVSYTYTAEARADDEEGIKRVYDALVAAAL RAAEAGYGTVTLTAY
	AM1B	LSFTVTVPPDSPLEAIRAAAEALAEGLRRAGVKGEVVV TATAGPYTATYTATLDGTEESLKKVIEAAVAAGKELL ELTGRTYPITLSAV
	AM1C	NTFTATVNSASEEDIREAAKRLAERLRAAGVEGKLTV TATANGVTFTYTVTGPVSEEVLEKVIEFMVEAALAAA AAGAKSITLTAG
	AM1D	LWFTATVDSAAMEDIEEAARRLAERVVAAGISNGEVT VTATANGVTFTVTVPADATSPEGIAAVLAAALAAAKA AAAAGGTSITLEVA
	AM1E	VIVMTVTVETSSTEEAAREAAKWAIEYIIKYPASTITVT ITVGGVTVTVTASGSIEEAIAAVLAAAQAAIEAPATGA ATATVVTG
	AM1F	LSWTATVDSRSEEVIREAARRLAELAVKAGISNQKLTV TATANGITYTFTADLGRDDEEGIAAVVAALVAAALAA AAAGATTVTLTVV
	AM1G	RTYTATVDPDATEEQIRAAARRLAERAVAAGVRGPVT VTVTTGDVTFTFTADLDGTEEGIAAVVAAVVRGALAA LRAAGGTRPVTLTVG
	AM1H	NTWTATAPAGNLEALKAAAKRLVERVVAAGIRNKTV TVTVTGGDFTYTFEAPVKAGDPEGRAALLAAILAGAR AAAAAGAQTITITVG
	AM1J	RSATATADASDEEELRAAAQRLVEWLKEAGAVGEVV VTAVAGDYTYTATAPASADDEEHLKAVLEAALRAAR YAAALGVGKVTLTAG
	AM2K	LSAEAEVDADDLEGAREAAERAVEILKKLGVSGKITV TAKAGDFEYTFTFEASADDEEALKQVKELALEALRQA MAAGVGKVTLTVG
	AM2M	ASAEAEVKPDATIEEIRAAARRLAEALRKAGVSGPVTV TAEAGDVSFSYTADLDGTEEGLKRVVEAIVRAAIAAL KATGGTKPVLLSAV
	AM2N	LAFTVTVDSASLEALREGAERAAEYIKKHNIVGKIVVK VTAGGVTYTATAEGPVSEETLERVIEAIVRAARRAVEA GGTSLTITVV

	AM2P	LLFTATADSRSPEAALAAAERLAEFIKKANATGTITVTI EANGVTYTFEVTGPLSEETIKRVKEAIAEGILAGAAAG ADRVTATAG
	AM2Q	LRFEAEAPASDEEGLKEAARRLAERVAAAGVAGLVV VEARAGDVEYRAEVPAAADDEEGLKAIVEAALRAAR RAAEAGAGKVTITVV
	AM2R	HAFTATVDSRSLEAIRAAAQRLADWIREANLSGEIVVE AEANGVRYTFTVSGPLSEETLKRVFEAAVEAAKAAVA AGGTKVTITAG
	AM2T	ISWTATVDSASEEVIREAARRLAELLVKAGIENKEVRV TAEGNGQRYTFTAPMRASDEEGIAAVVAALVRAALA AARAGATSITLTVS
	AM2V	LAWEATVDTASLEEIRAAARRLAERVLAAGIRSDVTV TAEANGTVFEVTVPVTETEEGLKALVEAAVAAALEAV KKTNGGSVLLRVV
	AM2X	LSWSATAPGGNLEALRAAARRLAERLIAAGVRNQKVT VTATGGDHEFEFTVEASAGDPESFAAVVAAALRAARA ALAAGATEVTLTVG
H4	H4A1R	DEYKKYYQQAIQLIQQLKKALEGNPEMKKLADKVLA LLKQAYAAFKAGRSPEEIRALLRKAIEAAKKLAKLGA SLGGFDLAKRIIELLKKMYELG
	H4A1S	VDAALALARAAAAAMRALLKRAPPGSETSQAIKQLY QLMLEMANATTVDEMIAAAKKAIEVAKQLIAQGNPQI IQAAQLTIDFAKKVIDALR
	H4C1R	LDELKKKLQEFIQKAIDMIKAHAGDPEGIRAVLRATLQ RAKELLKKHGASDDFIKKIIDFAQKMLDYIKQQNLGPD GVIKTIKAIAQMVLDFG
	H4C1S	LDELLRKAVEYIKKAIAAARAGDIDAAIAYAKKAIEIA QKIVKIAPPGSKISQIARKIIEAAKELIEALREGDDEKIK KAIEKLKRAAEEALKA
	H4A2R	WDDLAKKLIELLKKAIELLKQHNMDPEFLKLLKEVAQ ALIRAIKAGRLSPEVIKLAIEAAKLAIQAARHGDDELRA QAAALLRQVLELVKKLL
	H4A2S	EDYLKLLEEALKIAREVLENYPLTPVMRAAARAIIEAV KMAKKYGDEELIKLVVEAARLLRQAAKQGDLELARQ ALAAARQALAFARRVA
	H4C2P	DEAFKRAEELVRQAAEAAANMTEGGLEKVLALLRAA AAALRAAGFSLDDIRAMARKAVELAKRLGATDEQLA AAQAAAQRAVDG
	H4C2R	HPEIVAAAVAFVRQIWEYARQGMSLDEMIAWAVKYA KKIFDLVKKMGASDEVLKKVMDAVLAAAQAYAQQL NDEAAQRLLVAAQVIVQVLQQL
	D12	MDKVAVMAAMARAALEMSLEEAAQYAVELGAGPET LKRIRQATSVREVAILIAISWYPENEELAKKVVDRVL

Novel helical protein	D19	GWAAARVAAAGELPIEGVLAAVKALGLSDEFIAAGLA MARLVVGALRGETSPEIRAAVAAMRAGEPGKEQAAQ FLFDAAG
	D22	IPEVQAALQAALNSDDDAVVAWVQAVMDLIKNGKVS EEEAKLMLAAAAAYAGVLSPELMKALGLTPEQIARAL ARLRQLLG
	D25	GYQELVQQILAVLRRNPGNAVETAIVAAKYAGLSDEFI EAIRRWGQLYASGVESPAVQAIIDAAKAMYQGRPGLV EYAKIIIAAVRRLYERMG
	D46	DEIQKMIEEFAKALGATSIDQLLAFLRALVRRHPDNDL VRAALQAGLSPAIAFKLLAKLAGGDLEKALRIALEG
	D47	GDAAAARAALRAGKTAEEIVKLLKQLNINPGAIAVAK AVVRLGVGAEGVVLAVLLALRDKSNLDKVIQQVLAW HG
	D53	GDPIDILIKIAKALGISDRAAKILAEAAKAAGITSAEGA LRLANGEYPEAWKYAIELAKKRGDDAALAALRRAFG
	D55	GHPVVVQALMDAMGPEGAAQIQQWLKKAGGDTLEV MRFTARMNPEAARRAAARLRRLGVSPELIAALLAAAN G
	D60	SLEEKIKKAIEVARRAAALGAVSPEAAKIAIELLKSGLS VQQAIDFLKKNNIGQFAILALSLDSTEKVVEKVKKG
	D74	DPLVAMAAALRAAIKRAGGDAAAAVQALYKMFPGD GVARAMVAGEASPELIRELVKAALEGDVEKIKKVMK EILDFARSG
	D78	SLVEKAKKLLGVSVEGVRAAARYAEAFNGDARRVAR ALVEAGLVSPEGRAFLLRALAGGPGLDVVRKGVQVIL EYLRR
	D80	GEVLETLRRIVELGLPGIGIYQAYAQGLTVAQIAAALR ARGFTPEEVKKAAEAALRTGSPATAAIIREIVKALG
	D81	GDIRAALQRLIEAAKRFMAAPMSEEQRLAALIAIFIAKL LKSGVSVEEAARFAIELGSIDLIAYAAALARALGASPV AVALVRAVAAAAG
Novel helical protein with fused MBP-tag	MBP-D12	KIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTV EHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSG LLAEITPAAAFQDKLYPFTWDAVRYNGKLIAYPIAVEA LSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFN LQEPYFTWPLIAADGGYAFKYAAGKYDIKDVGVDNA GAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETA MTINGPWAWSNIDTSAVNYGVTVLPTFKGQPSKPFVG VLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDK PLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDAALAAAQTNAAA ADKVAVMAAMARAALEMSLEEAAQYAVELGAGPET LKRIRQATSVREVAILIAISWYPENEELAKKVVDRVL

	MBP-D22	KIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTV EHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSG LLAEITPAAAFQDKLYPFTWDAVRYNGKLIAYPIAVEA LSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFN LQEPYFTWPLIAADGGYAFKYAAGKYDIKDVGVDNA GAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETA MTINGPWAWSNIDTSAVNYGVTVLPTFKGQPSKPFVG VLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDK PLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDAALAAAQTNAAR AAAIAEVQAALQAALNSDDDAVVAWVQAVMDLIKN GKVSEEEAKLMLAAAAAYAGVLSPELMKALGLTPEQI ARALARLRQLLG
	MBP-D53	KIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTV EHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSG LLAEITPAAAFQDKLYPFTWDAVRYNGKLIAYPIAVEA LSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFN LQEPYFTWPLIAADGGYAFKYAAGKYDIKDVGVDNA GAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETA MTINGPWAWSNIDTSAVNYGVTVLPTFKGQPSKPFVG VLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDK PLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDAALAAAQTNAAR AAAIDILIKIAKALGISDRAAKILAEAAKAAGITSAEGA LRLANGEYPEAWKYAIELAKKRGDDAALAALRRAF

Supplementary Table 2. DNA sequences of designs selected for experimental testing.

Protein	DNA sequence
EXTD-1	ATGGCTATGACCCTGAGCGATTGTCAGCAGGAACTG AGCCTGGTTCAGACCGTTACCCGTGGCAGCCGTGCC TTTCTGAGTCGTGAAGAAGCCCAGCATTTTGTGAAA GAATGCGGTCTGCTGAATTGCGAAGCCGTTCTGGAA CTGCTGATTTGCCATCTGCATTGGGGTATGCAGGTT ATTAAGATTCTGAAAGAATTCCGCGCCCGTATGCGC GCCGCCACCACAAGAGAAGCACTGCGCCAGGCAGT GCAGGAACTGTATGATATGATGAAAGAACATGCCC AGAATGGCAGTATGCTGCAGATTCTGTATGAAAAAA TTAAGAAGATCCTGAACGACAATTTCGAAGAACTGA GCCTGGAAGAAATTGCCGCCCGCGTTATGGCAGCAG TGAAACGTGTTCTGGAAATGACCGAATGCGTGCAGA TGCGTGCACTGGGTGCCGTTGCCAGTCTGGGCTGCA CCGATCTGCTGCCGCAGGAACATATTCTGCTGCTGA CCCGCCCGCGCCTGCAGGAACTGAGTGCAGGCAGTC CGGGTCCGGTTACCAATAAGGCCACCAAAATTCTGC GCCATTTTGAAGCCAGCTGTCTCGAGCACCACCACC ACCACCAC
EXTD-2	ATGGCTATGACCCTGAGCGATTGTCAGCAGGAACTG AGCCTGGTTCAGACCGTTACCCGTGGCAGCCGTGCC TTTCTGAGTCGTGAAGAAGCCCAGCATTTTGTGAAA GAATGCGGTCTGCTGAATTGCGAAGCCGTTCTGGAA CTGCTGATTTGCCATCTGCGTAGTCCGGGCGGTCAG CTGATTCTGAAAGCACTGCAGTATGCAATTGAAGTG GTGAAAATGGAAAATGTTAGTGTTCGTGAAGCCGCC CAGCAGCTGAAAGAACGCGTGAAAGCACTGGGTTT TCAGGAAGGTGTGGGTGAAATTCTGCGCCGCATTCT GGATATTGTTATTAAGCTGGCAGAAAAAGGTGACGT GGAAGAAATGAAAGAAACCATTAAGCGTCTGGTGG CCCGTGCACGCGCATGGATTGAATGCGTGCAGATGC GTGCACTGGGTGCCGTTGCCAGTCTGGGCTGCACCG ATCTGCTGCCGCAGGAACATATTCTGCTGCTGACCC GCCCGCGCCTGCAGGAACTGAGTGCAGGCAGTCCG GGTCCGGTTACCAATAAGGCCACCAAAATTCTGCGC CATTTTGAAGCCAGCTGTCTCGAGCACCACCACCAC CACCAC
EXTD-3	ATGGCTATGACCCTGAGCGATTGTCAGCAGGAACTG AGCCTGGTTCAGACCGTTACCCGTGGCAGCCGTGCC TTTCTGAGTCGTGAAGAAGCCCAGCATTTTGTGAAA GAATGCGGTCTGCTGAATTGCGAAGCCGTTCTGGAA CTGCTGATTTGCCATCTGCGTCTGGGTATGGAAATT

	ATGAAACTGGGCCGTCAGCTGCGCGAAGCAGTTCGT GCCAATGATGTTGATGCCATGCTGAAAATTGCCAAA GAAATTATTAAGGTTATTGGCGAAACCGGCCTGGAT GAAGTGTATCGTCAGCTGCTGAAAGCAGCAAAAGA ATTTCTGGAACGTCGCGCCGAAAATTTTAGTCATGA AGAAGCAGTGGCATTTGCACAGCAGATTATTCAGCT GATTAAGCAGGTTGAATGCGTGCAGATGCGTGCACT GGGTGCCGTTGCCAGTCTGGGCTGCACCGATCTGCT GCCGCAGGAACATATTCTGCTGCTGACCCGCCCGCG CCTGCAGGAACTGAGTGCAGGCAGTCCGGGTCCGGT TACCAATAAGGCCACCAAAATTCTGCGCCATTTTGA AGCCAGCTGTCTCGAGCACCACCACCACCACCAC
EXTD-4	ATGGCTATGACCCTGAGCGATTGTCAGCAGGAACTG AGCCTGGTTCAGACCGTTACCCGTGGCAGCCGTGCC TTTCTGAGTCGTGAAGAAGCCCAGCATTTTGTGAAA GAATGCGGTCTGCTGAATTGCGAAGCCGTTCTGGAA CTGCTGATTTGCCATCTGCGTCGTGCAATGCGCAAT TTTGAAGCCACCCAGCAGCTGATTGATATGATTCGC AAAGCCGGCGGCAATCAGGATAATGTTCGTGAAGC CATTAAGGAACTGCTGGAACGTGCACGCGCAGGCG AACATAATATTACCGAAGTTCAGATTCAGATTCTGA AAGCCCTGTATCAGTTTCTGCAGCAGCATGATTGGA GCCTGGAAAAAGCCATTAAGTTTGTGAAAGAAACC GCAAAACGCCTGTTTAAAGAATATGAATGCGTGCAG ATGCGTGCACTGGGTGCCGTTGCCAGTCTGGGCTGC ACCGATCTGCTGCCGCAGGAACATATTCTGCTGCTG ACCCGCCCGCGCCTGCAGGAACTGAGTGCAGGCAG TCCGGGTCCGGTTACCAATAAGGCCACCAAAATTCT GCGCCATTTTGAAGCCAGCTGTCTCGAGCACCACCA CCACCACCAC
EXTD-5	ATGGCTATGACCCTGAGCGATTGTCAGCAGGAACTG AGCCTGGTTCAGACCGTTACCCGTGGCAGCCGTGCC TTTCTGAGTCGTGAAGAAGCCCAGCATTTTGTGAAA GAATGCGGTCTGCTGAATTGCGAAGCCGTTCTGGAA CTGCTGATTTGCCATCTGCAGCGTATGAATGTGGAA GCCGTGCGTGAACTGATTGCAGCCCTGCAGCGTCTG AAAGCCGGTAGTGATGTTCTGGAAACCATTAAGCGT GTTATTGAAATTATCAAGAAGATCGCATACGAAATT ACCGAAGAACTGCAGAAAATTCTGGATTATCTGCGT GCCGCATATCGCCGTGGTCAGGTGAGTGATGAAGTG ATTGCAGCCGTGAAAAAAGCCGTGGAATTTCTGCGC AAATGGGTTGAATGCGTGCAGATGCGTGCACTGGGT GCCGTTGCCAGTCTGGGCTGCACCGATCTGCTGCCG CAGGAACATATTCTGCTGCTGACCCGCCCGCGCCTG

	CAGGAACTGAGTGCAGGCAGTCCGGGTCCGGTTACC AATAAGGCCACCAAAATTCTGCGCCATTTTGAAGCC AGCTGTCTCGAGCACCACCACCACCACCAC
EXTD-6	ATGGCTATGACCCTGAGCGATTGTCAGCAGGAACTG AGCCTGGTTCAGACCGTTACCCGTGGCAGCCGTGCC TTTCTGAGTCGTGAAGAAGCCCAGCATTTTGTGAAA GAATGCGGTCTGCTGAATTGCGAAGCCGTTCTGGAA CTGCTGATTTGCCATCTGACCCGTATGAATCTGGCA GCCGTGCGCGAAGCAATTCGTGCAGTTGAACGCGCA CGCGCAGGTGAAGATTTTCGCGAAATTATTAAGCAG CTGATTGAAATTATCAAGAAGATTGCCTTTAGCATT ACCGAAGAACTGCGCAAACTGCTGGAACAGCTGAA AAAATATTATAAAACCGGTAAATGGAGCGATGAAG TGCTGGCAGTGATGCAGAAAGCCGTGGATTTTCTGC GCGCATTTGTTGAATGCGTGCAGATGCGTGCACTGG GTGCCGTTGCCAGTCTGGGCTGCACCGATCTGCTGC CGCAGGAACATATTCTGCTGCTGACCCGCCCGCGCC TGCAGGAACTGAGTGCAGGCAGTCCGGGTCCGGTTA CCAATAAGGCCACCAAAATTCTGCGCCATTTTGAAG CCAGCTGTCTCGAGCACCACCACCACCACCAC
EXTD-7	ATGGCTATGACCCTGAGCGATTGTCAGCAGGAACTG AGCCTGGTTCAGACCGTTACCCGTGGCAGCCGTGCC TTTCTGAGTCGTGAAGAAGCCCAGCATTTTGTGAAA GAATGCGGTCTGCTGAATTGCGAAGCCGTTCTGGAA CTGCTGATTTGCCATCTGCGTCGTAAAAATTTCGAA CAGGCCCTGGAACGTCTGCGTAAACTGATTCGCGAA GCCGCCGCAGGCAAATATACCCAGGAAGATATTCTG AAATTCATTGTGCAGGTGGCCATTCAGCTGCTGTAT ATGAGTGAAGAAGAAATTCGTGCAGCCGTGCGCCG TATTGTTCAGCAGCTGTGGAGTAGCATTGAAGAAAT GGATGGTCTGTTTAAAATCGCCAAAGTTGTTGTGCA GTTTGCCAAAAAATATATGGAATGCGTGCAGATGCG TGCACTGGGTGCCGTTGCCAGTCTGGGCTGCACCGA TCTGCTGCCGCAGGAACATATTCTGCTGCTGACCCG CCCGCGCCTGCAGGAACTGAGTGCAGGCAGTCCGG GTCCGGTTACCAATAAGGCCACCAAAATTCTGCGCC ATTTTGAAGCCAGCTGTCTCGAGCACCACCACCACC ACCAC
EXTD-8	ATGGCTATGACCCTGAGCGATTGTCAGCAGGAACTG AGCCTGGTTCAGACCGTTACCCGTGGCAGCCGTGCC TTTCTGAGTCGTGAAGAAGCCCAGCATTTTGTGAAA GAATGCGGTCTGCTGAATTGCGAAGCCGTTCTGGAA CTGCTGATTTGCCATCTGCATGGTAATACCGCCGTG CAGCGCGCACTGGAACGTCTGCGTAAACTGTATCAG

	GATACCAGTATGAGTGTGCGTGAACGCATTGAAAA AGCAATTGCAATTCTGACCGCACTGGCCCCGCCGAG CGGTGACTCATGGCTGCGTAAAACCCTGGAAAAAGT TCGCCGTATTGCCCAGCAGGGTAATTTTAGTGAAGA AACCGCAAAAAAGATCCTGGAAACCGCCATTGAAA TGGTGAAAAAATTCATTGAATGCGTGCAGATGCGTG CACTGGGTGCCGTTGCCAGTCTGGGCTGCACCGATC TGCTGCCGCAGGAACATATTCTGCTGCTGACCCGCC CGCGCCTGCAGGAACTGAGTGCAGGCAGTCCGGGT CCGGTTACCAATAAGGCCACCAAAATTCTGCGCCAT TTTGAAGCCAGCTGTCTCGAGCACCACCACCACCAC CAC
XM1A	CTGACCTGGACCGCAACCGTGGATAGCGCCAGCCCG GAAGCCGCAGCAGCAGCAGCTCGCCGTCTGGCAGA ACGTGTGCGCGAAGCAGGCATTACCAGTCCGGCCAC CGTTACCGCAACCGCCAATGGCATTACCTTCACCTA TACCGTGCCGGTGAGCCTGAGTGAAGAAGGTCTGGC AGCACTGGTTGCCGCAGCAGTTCGTGCCCTGCTGGA AGCACATCGTGCAACCAATGGCCAGAGCGCCACCCT GACCGCCGGC
XM1B	CTGGCAGCTACCGCAACCGTTGATAGCGCAAGTGAA GAAGCAGTGCGTGCCGCAGCCGCAGCAGTTGTGGC CCGTCTGGTTGCCGCAGGTGTGGAAGGTAAAGTTAC CGTTACCGCAACCGCCAATGGCGTGACCTTCACCTA TACCGCCGATCTGCGTCGTAGTGAAGAAGCCGCAGC AGCAGTGGCCGCCGCACTGGCAGCTGCTGGTCTGGC AGCAGCCCGCGCTACCGGTGGTGGTACCGTTACCCT GACCGCAGTG
XM1C	CGTGCAGCTACCGCAACCGTTGATAGCCGTAGCGAA GAAGCACTGCGTGCAGCCGCCCGTGCACTGGTGGCT GCTATTCTGGCACGCTTCACCAGCCCGCATGAAGTG GTTATTACCGTTACCGCAAATGGTCGCGTGACCTTC ACCTATACCGTGGAAGTTGATGGTAGCGAAGAAGC CCTGGCAGCCGCCCTGGCCGCTGCAATTGCAGCCCT GCGCGCAGCCGTGGAAGCAACCGGTAGTAGCGCAA CCGTTACCGCCGTG
XM1D	GAACGCTTCACCGTTACCGTGGATAGCGCCAGCCTG GCCGCCCTGCGTGCTGCAGCAGCTAAAGCCGCCGCC CTGATTAAAGCAGCCAATATTAGTGGCAAAATTACC ACCACCGTGACCGCAAATGGCGTGACCTATACCTTC ACCGTGACCGGTCCGCTGAGTGAAGAAACCATTGCC GCCGTGGCAGCCGCCATTCTGCGCGCAGCACGTGCC GCTGTTGCCGCTGGTGGTACCAGCCTGACCATTACC GCCGGT

XM1E	CGTAGCTATAGTGCAACCGTTGATAGTAATAGCAAA GAAGCCATTCTGGCCGCAGCACGTCGTCTGGCCGAA CGCGCAGTTGCAGCAGGCGTTCGTACCAGTGTGACC GTTACCGCCACCGCCGATGGCACCAAATTCACCTTC ACCGCACCGCTGGATGAAAGTGAAGAAGGTATTGC CGCAGTTGCCGCAGCCGCCCTGCGTGCAGGTCTGGC AGCTGTTCGCGCCACCGGTGGTCGCAGTGTGACCCT GACCGTTGTT
XM1F	CTGCCTGTTACCGCCACCATTGATGCCAGCGCAACC CGTGAACAGGTGCGTGCCGCAGCACGTGCCGCAGCT CGTCGTCTGCGTGAACTGGGTGTGAGTGGCACCGTT GTGGTTACCGCCACAGCAGGCGATGTGAGCTTCACC TATACCGCAGATCTGGATGGCAGCGAAGAAGCACT GGCCGCCGTGGCCGCAGCAGCATTCGCAGCTGGTCT GGCAGCCCTGAAAGCAGCCGGTGGCACCAAAACCG TGACCCTGACCGCCTTC
XM2G	GGTATGAAATACAGTGCAGAAGTTAGCAGCAGCAT GAGTCTGGAAGAAATTCGTCAGGTGATTGCAGAACT GATTAAACAGGCAGTTCGTGCAGCAGGGGGGCGGA AAGGAAAAGTTAGAGTGAGAGTTGAAACCAGCGAT GGATATGTTTTTGAGTTTGAGGCAGAGATTAGCAGT GAGGAAAAAATACGGGAATTAGCAGAAAGAGCAGC ACAGGCAATACTGGAAGCAGGTGGAAAAAGCGCAA GAGTTGAAGCAGAAGTG
XM2H	CATAGCTGGAGCGCAACCGTTGATAGCCGGAGCGA AGAAGCAGTTCGGGCAGCAGCACGTCGTCTGGCAG AACGTCTGTTAGCAGCAGGTATTTCAGGTAAAATTA AAATTGAGGTGGAGGCAAATGGAATTAAATATGAG TATGAAGTGGAGGGTCCGGCAACCGAAGAAGTTGC AAAAAAAATCGTGGAATATGCAGTTGCAGCAGCCC TGCGTGCAATCGCAGCAGGAGCAACCAGTGTTACCA TTACAGTTGGT
XM2J	GTTGCAGAGTATGAGTATGAGGTTCCGGCGGGTAAT ATTGAGGAAGTTTTTAAAATTGTGGAGGAAGCACTG ATTGCGGCGTATGAAGCAGCAGGTCATAGTGCGCG GGTTCGTATTGAGTTAGAAGCAAATGGGGTGCGTAT GGAATATGAATTTGAAATAAGTAGCGAGGAAGATG CGAGAGAAATTGCACAGCGGGTTGCAGAAGCATTT AAAAAGTATCCGAGTGAACGGGCAACGGTGAAAGT GAGCGTT
XM2K	TGGTATGCGACGGCCACGGTTGATAGTCGTGATCGT GAAGCGGCCCGTGAGGCGGCCCGTAGACTGGCACA GGCGCTGCTGGAAGCGAATGTTACGGGTAAAATTTA TGTTCGGGTTGAGGCCGGTGGCCATGAATATGAATT

	TGAAGCAGAAGGACCGCCGAGCGAAGAAGTTGCAC AAGCAATTGTCGAAGCCGCTGTTAGAGCCGTTCTGG AAGGGATTGCACAGGGTGCGAAAAGTGCAACAGTC GAAGTTGGT
XM2M	CTGAAAGTTCGTGTTCGTCTGGAAAATCCTACCCTG GAACAGATTCGTGAAGCAGCACGTCGTGCAGTTGA AGCACTGGTTCGTTTAGGTGTTGAACAGTCTAAAGT TCGTGTGCGTGTTCGCGGTGGTGAAGAATATGAATA TGAAGCAGAACTGACCCGCACCGAGGAAAGTATAA AACGTGTGGCAGAAGCACTGTATAAGGCGGGTAAA GAAGCACTGGAGCATACGGATACAAGCATTGAACT GGAAGCAGTG
XM2N	AACTGGTTTGAAGTGGAAATTGATAGCCGCAGCGA AGAAGTGGCACGCGAAGGCGCACGCGAAGCGGCAA GACGCGTGGTTGCAGCAGGAATCACCGGCAAAAAA GTGCGGGCGGAAGCAGAAGGTAATGGGGTGAAGTA TAGCTTTGAAGCAGAAATGAATGCAAGCGATGAAG AAAGCATACAGAGAGTTATTGAGGCAGCAGCAGAA GCAGTTCGGCGGGCGGCAGCAGCAGGAGCAGATAA AGTGCGGCTGCGGGTGGGT
AM1A	CTGCCTGTTACCGTTACCGCCCCGACCGCAGATCTG GAAGCCCTGCGTGCCGCAGCCCGTGCTGCTGTTGCA TTTCTGGTTGCAGCAGGCGTTGAAGGTCCGGTTGTG GTGACCGCAACCGCCGGCGATGTTAGTTATACCTAT ACCGCAGAAGCACGTGCAGATGATGAAGAAGGCAT TAAGCGCGTGTATGATGCCCTGGTTGCAGCGGCCCT GCGTGCGGCAGAAGCAGGTTATGGCACCGTTACCCT GACCGCATAT
AM1B	CTGAGTTTTACCGTTACCGTGCCGCCGGATAGCCCG CTGGAAGCCATTCGTGCAGCAGCCGAAGCACTGGC AGAAGGTCTGCGCCGCGCAGGCGTGAAAGGTGAAG TTGTGGTTACCGCAACCGCAGGTCCGTATACCGCCA CCTATACCGCCACACTGGATGGTACCGAAGAAAGTC TGAAAAAAGTGATTGAAGCAGCCGTTGCAGCCGGC AAAGAACTGCTGGAACTGACCGGTCGCACCTATCCG ATTACCCTGAGTGCCGTT
AM1C	AACACCTTTACCGCCACCGTGAATAGCGCAAGTGAA GAAGATATTCGTGAAGCCGCCAAACGCCTGGCCGA ACGCCTGCGTGCAGCCGGTGTGGAAGGTAAACTGA CCGTTACCGCAACCGCCAATGGTGTGACCTTTACCT ATACCGTGACCGGCCCGGTTAGTGAAGAAGTGCTGG AAAAAGTGATTGAATTCATGGTGGAAGCAGCCCTG GCCGCCGCAGCCGCAGGTGCAAAAAGCATTACCCT GACCGCCGGT

AM1D	CTGTGGTTTACCGCAACCGTGGATAGTGCCGCAATG GAAGATATTGAAGAAGCCGCACGCCGTCTGGCAGA ACGCGTTGTTGCAGCCGGTATTAGCAATGGCGAAGT TACCGTGACCGCAACCGCCAATGGTGTTACCTTTAC CGTTACCGTGCCGGCAGATGCCACCAGTCCGGAAGG TATTGCCGCAGTTCTGGCAGCAGCCCTGGCAGCAGC AAAAGCAGCAGCCGCAGCAGGTGGCACCAGTATTA CCCTGGAAGTGGCA
AM1E	GTTATCGTGATGACCGTGACCGTTGAAACCAGCAGC ACCGAAGAAGCCGCACGCGAAGCAGCAAAATGGGC AATTGAATATATTATTAAGTACCCGGCCAGCACCAT TACCGTTACCATTACCGTGGGCGGCGTTACCGTTAC CGTGACCGCAAGCGGCAGTATTGAAGAAGCAATTG CCGCCGTGCTGGCCGCCGCCCAAGCTGCAATTGAAG CCCCGGCCACCGGTGCAGCCACCGCAACCGTTGTGA CCGGT
AM1F	CTGAGCTGGACCGCCACCGTTGATAGTCGCAGTGAA GAAGTTATTCGCGAAGCAGCACGTCGCCTGGCAGA ACTGGCCGTTAAAGCAGGTATTAGTAATCAGAAACT GACCGTGACCGCCACCGCCAATGGCATTACCTATAC CTTTACCGCCGATCTGGGCCGTGATGATGAAGAAGG CATTGCAGCAGTTGTGGCCGCCCTGGTGGCAGCCGC TCTGGCAGCAGCAGCAGCGGGTGCAACCACCGTTAC CCTGACCGTTGTG
AM1G	CGTACCTATACCGCAACCGTGGATCCGGATGCAACC GAAGAACAGATTCGTGCCGCAGCCCGCCGTCTGGCC GAAAGAGCTGTGGCCGCCGGTGTTCGCGGTCCGGTT ACCGTGACCGTTACCACCGGCGATGTGACCTTTACC TTTACCGCCGATCTGGATGGTACCGAAGAAGGCATT GCCGCCGTTGTGGCAGCAGTTGTGCGCGGCGCACTG GCCGCACTGCGTGCAGCAGGTGGCACCCGTCCGGTT ACCCTGACCGTTGGT
AM1H	AACACCTGGACCGCAACCGCCCCGGCAGGTAATCTG GAAGCACTGAAAGCAGCCGCAAAACGTCTGGTTGA ACGCGTGGTTGCAGCAGGCATTCGTAATAAGACCGT GACCGTGACCGTTACCGGCGGTGACTTTACCTATAC CTTTGAAGCCCCGGTGAAAGCAGGCGATCCGGAAG GTCGTGCCGCACTGCTGGCCGCAATTCTGGCCGGCG CCCGTGCAGCAGCCGCTGCAGGTGCACAGACCATTA CCATTACCGTGGGC
AM1J	CGTAGCGCTACCGCAACCGCCGATGCAAGTGATGA AGAAGAACTGCGCGCAGCCGCACAGCGTCTGGTGG AATGGCTGAAAGAAGCAGGTGCCGTGGGTGAAGTG GTGGTGACCGCAGTTGCAGGTGACTATACCTATACC

	GCCACCGCCCCGGCCAGCGCTGATGATGAAGAACA TCTGAAAGCCGTGCTGGAAGCAGCCCTGCGCGCCGC ACGTTATGCCGCAGCACTGGGTGTGGGCAAAGTTAC CCTGACCGCCGGT
AM2K	CTGAGCGCTGAAGCAGAAGTTGATGCAGATGATCTG GAAGGTGCCCGTGAAGCAGCAGAACGTGCAGTTGA AATCCTGAAAAAGCTGGGTGTTAGCGGTAAAATTAC CGTTACCGCAAAAGCAGGTGATTTTGAATATACCTT TACCTTTGAAGCAAGTGCAGATGATGAAGAAGCACT GAAACAGGTGAAAGAACTGGCACTGGAAGCACTGC GTCAGGCCATGGCAGCAGGTGTGGGTAAAGTTACCC TGACCGTTGGT
AM2M	GCAAGTGCAGAAGCAGAAGTAAAACCGGATGCAAC AATTGAAGAAATTCGTGCAGCAGCACGTCGTCTGGC AGAAGCACTGCGTAAAGCAGGTGTGAGCGGTCCGG TTACAGTGACCGCAGAAGCAGGTGATGTGAGTTTTA GCTATACCGCAGATCTGGATGGTACCGAAGAAGGTC TGAAACGTGTTGTTGAAGCCATTGTGCGTGCGGCAA TTGCAGCACTGAAAGCAACCGGTGGTACCAAACCG GTGCTGCTGAGCGCGGTT
AM2N	CTGGCATTTACTGTTACTGTTGATTCTGCTTCTCTGG AAGCTCTGCGTGAAGGTGCTGAACGTGCTGCCGAAT ATATTAAAAAGCATAATATCGTGGGCAAAATCGTTG TTAAAGTTACTGCAGGTGGTGTTACCTATACTGCTA CCGCCGAAGGTCCGGTTTCTGAAGAAACACTGGAAC GTGTGATTGAAGCCATCGTGCGTGCAGCACGTCGTG CAGTTGAAGCAGGCGGTACCAGCCTGACAATTACCG TTGTT
AM2P	CTGCTGTTTACCGCCACCGCAGATTCTCGTTCCCCGG AAGCCGCACTGGCCGCCGCAGAACGTCTGGCTGAAT TTATTAAAAAGGCCAATGCCACCGGTACCATTACCG TTACTATTGAAGCCAATGGCGTTACCTATACCTTTG AAGTCACCGGTCCGCTGAGCGAAGAAACCATTAAA CGTGTTAAAGAAGCAATCGCAGAAGGTATTCTGGCA GGCGCTGCAGCAGGCGCGGATCGTGTTACAGCAAC CGCCGGT
AM2Q	CTGCGTTTTGAAGCCGAAGCCCCTGCCTCCGATGAA GAAGGTCTGAAAGAAGCCGCCCGTCGTCTGGCCGA ACGTGTTGCCGCCGCCGGTGTTGCCGGTCTGGTTGT TGTCGAAGCCCGTGCAGGTGATGTAGAATATCGTGC AGAAGTTCCGGCAGCAGCAGATGATGAAGAAGGCC TGAAAGCAATTGTTGAAGCAGCACTGCGTGCGGCAC GTCGTGCAGCAGAAGCAGGTGCAGGTAAAGTTACC ATTACCGTGGTT

AM2R	CATGCATTTACCGCCACCGTTGATTCACGTTCACTG GAAGCAATTCGTGCAGCCGCACAGCGTCTGGCCGAT TGGATTCGTGAAGCCAACCTGTCCGGTGAAATTGTT GTTGAAGCCGAAGCAAATGGTGTTCGTTATACCTTT ACCGTTTCAGGTCCGCTGAGCGAAGAAACCCTGAAA CGTGTTTTTGAAGCAGCAGTGGAAGCAGCAAAAGC GGCAGTTGCAGCAGGTGGTACCAAAGTGACGATTA CCGCAGGC
AM2T	ATCAGCTGGACCGCAACCGTTGATAGCGCCTCAGAA GAAGTCATTCGCGAAGCCGCCCGTCGCCTGGCAGAA CTGCTGGTTAAAGCAGGCATTGAAAACAAAGAAGT TCGCGTTACGGCAGAAGGTAATGGTCAGCGTTATAC CTTTACCGCACCGATGCGTGCAAGCGATGAAGAAG GTATTGCAGCAGTTGTCGCAGCACTGGTTCGTGCAG CACTGGCAGCAGCACGTGCAGGTGCAACAAGCATT ACACTGACCGTTTCC
AM2V	CTGGCATGGGAAGCTACCGTTGATACCGCAAGCCTG GAAGAAATTCGTGCAGCCGCACGTCGTCTGGCAGA ACGTGTTCTGGCTGCTGGTATTCGTTCTGATGTTACT GTTACTGCTGAAGCAAATGGTACCGTTTTTGAAGTT ACGGTTCCTGTTACCGAAACCGAAGAAGGTCTGAAA GCACTGGTGGAAGCAGCTGTTGCAGCAGCACTGGA AGCGGTTAAAAAGACAAATGGTGGTAGCGTTCTGCT GCGTGTTGTT
AM2X	CTGAGCTGGAGCGCCACCGCACCTGGTGGCAATCTG GAAGCACTGCGTGCCGCAGCACGTCGTCTGGCAGA ACGTCTGATTGCCGCAGGCGTTCGTAATCAGAAAGT TACCGTTACCGCAACCGGTGGTGATCATGAATTTGA ATTTACCGTTGAAGCATCTGCTGGTGATCCGGAAAG CTTTGCAGCAGTTGTTGCAGCAGCACTGCGTGCTGC ACGCGCAGCACTGGCAGCAGGTGCAACAGAAGTGA CCCTGACCGTGGGT
H4A1R	CATATGGGCGATGAATACAAAAAATACTACCAGCA GGCCATCCAGCTGATCCAGCAGCTGAAAAAAGCCCT GGAAGGCAATCCGGAAATGAAGAAGCTGGCCGATA AAGTTCTGGCCCTGCTGAAACAGGCCTACGCCGCCT TCAAAGCCGGCCGCAGCCCGGAAGAAATCCGCGCC CTGCTGCGCAAAGCCATCGAAGCCGCCAAGAAACT GGCCAAACTGGGCGCCAGCCTGGGCGGCTTCGATCT GGCCAAACGCATCATCGAACTGCTGAAAAAAATGT ACGAACTGGGCGGCCTCGAG
H4A1S	CATATGGGTGTTGACGCTGCATTGGCACTAGCCCGT GCGGCCGCGGCAGCAATGAGAGCCCTACTAAAGCG CGCCCCGCCGGGCAGCGAAACCAGCCAGGCCATCA

	AACAGCTGTACCAGCTGATGCTGGAAATGGCCAATG CCACCACCGTTGATGAAATGATCGCCGCCGCCAAAA AAGCCATCGAAGTTGCCAAACAGCTGATCGCCCAG GGCAATCCGCAGATCATCCAGGCCGCCCAGCTGACC ATCGATTTCGCCAAAAAAGTTATCGATGCCCTGCGC GGCCTCGAG
H4C1R	CATATGGGCCTGGATGAACTGAAAAAAAAACTGCA GGAATTCATCCAGAAAGCCATCGATATGATCAAAGC CCATGCCGGCGATCCGGAAGGCATCCGCGCCGTTCT GCGCGCCACCCTGCAGCGCGCCAAAGAACTGCTGA AAAAACATGGCGCCAGCGATGATTTCATCAAAAAA ATCATCGATTTCGCCCAGAAAATGCTGGATTACATC AAACAGCAGAATCTGGGCCCGGATGGCGTTATCAA AACCATCAAAGCCATCGCCCAGATGGTTCTGGATTT CGGCGGCCTCGAG
H4C1S	CATATGGGCCTGGATGAACTGCTGCGCAAAGCCGTT GAATACATCAAGAAGGCCATCGCCGCCGCCCGCGC CGGCGATATCGATGCCGCCATCGCCTACGCCAAAAA GGCCATCGAAATAGCGCAGAAAATCGTTAAAATCG CCCCGCCGGGCAGCAAAATCAGCCAGATCGCCCGC AAAATCATCGAAGCCGCCAAAGAACTGATCGAAGC CCTGCGCGAAGGCGATGATGAAAAAATCAAAAAGG CGATCGAGAAACTCAAACGCGCCGCCGAAGAAGCC CTGAAAGCCGGCCTCGAG
H4A2R	CATATGGGCTGGGATGATCTGGCCAAAAAACTGATT GAGCTGCTGAAAAAAGCGATCGAACTGTTAAAACA GCATAATATGGATCCGGAATTCCTGAAACTGCTGAA AGAAGTTGCCCAGGCCCTGATCCGCGCCATCAAAGC CGGCCGCCTGAGCCCGGAAGTTATCAAACTGGCCAT CGAAGCCGCCAAACTAGCAATACAGGCCGCGCGCC ACGGAGACGACGAACTAAGGGCTCAGGCAGCCGCC CTGCTGCGCCAGGTTCTGGAACTGGTTAAAAAACTG CTGGGCCTCGAG
H4A2S	CATATGGGCGAAGATTACCTGAAACTGCTGGAAGA AGCCCTGAAAATCGCCCGCGAAGTTCTGGAAAATTA CCCGCTGACCCCGGTTATGCGCGCCGCCGCCCGCGC CATCATCGAAGCCGTTAAAATGGCCAAAAAATACG GCGATGAAGAACTGATCAAACTGGTTGTTGAAGCCG CCCGCCTGCTTCGGCAGGCCGCCAAACAAGGGGATC TGGAACTTGCAAGGCAAGCCCTGGCCGCGGCTCGCC AAGCCCTGGCATTCGCCCGCCGCGTTGCCGGCCTCG AG
H4C2P	CATATGGGCGATGAAGCCTTCAAACGCGCCGAAGA ACTGGTTCGCCAGGCCGCCGAAGCCGCCGCCAATAT

	GACCGAAGGAGGCCTCGAAAAGGTGCTGGCCCTGC TCCGAGCAGCCGCCGCCGCTTTAAGGGCCGCCGGCT TCAGCCTGGATGATATCCGCGCCATGGCCCGCAAAG CCGTTGAACTGGCCAAACGCCTGGGCGCCACCGATG AACAGCTGGCCGCAGCCCAGGCCGCCGCCCAGCGC GCCGTTGATGGCGGCCTCGAG
H4C2R	CATATGGGCCATCCGGAAATCGTTGCCGCCGCCGTT GCCTTCGTTCGCCAGATCTGGGAATACGCCCGCCAG GGCATGAGCCTGGATGAAATGATCGCCTGGGCCGTT AAATACGCCAAAAAAATCTTCGATCTGGTTAAAAAA ATGGGCGCCAGCGATGAAGTTCTGAAAAAAGTTAT GGATGCCGTTCTGGCCGCCGCCCAGGCCTACGCCCA GCAGCTGAATGATGAAGCCGCCCAGCGCCTGCTGGT TGCCGCCCAGGTTATCGTTCAGGTTCTGCAGCAGCT GGGCCTCGAG
D12	GATAAGGTGGCAGTTATGGCAGCAATGGCCCGCGC CGCACTGGAAATGAGCCTGGAAGAAGCCGCCCAGT ATGCAGTTGAACTGGGCGCCGGTCCGGAAACCCTGA AACGCATTCGTCAGGCAACCAGTGTTCGCGAAGTGG CAATTCTGATTGCAATTAGTTGGTATCCGGAAAATG AAGAACTGGCAAAAAAGGTTGTTGATCGCGTGCTG
D19	GGTTGGGCAGCCGCTCGTGTTGCAGCCGCAGGTGAA CTGCCGATTGAAGGTGTGCTGGCCGCCGTTAAAGCA CTGGGTCTGAGTGATGAATTCATTGCAGCAGGTCTG GCAATGGCACGCCTGGTGGTTGGCGCCCTGCGTGGT GAAACCAGCCCGGAAATTCGTGCAGCAGTTGCCGCC ATGCGTGCAGGCGAACCGGGCAAAGAACAGGCAGC CCAGTTTCTGTTTGATGCAGCCGGT
D22	ATCCCGGAAGTTCAGGCAGCACTGCAGGCAGCCCTG AATAGTGATGATGATGCAGTTGTGGCATGGGTTCAG GCAGTTATGGATCTGATTAAGAATGGTAAAGTGAGT GAAGAAGAAGCAAAACTGATGCTGGCCGCCGCCGC AGCCTATGCAGGCGTTCTGAGTCCGGAACTGATGAA AGCACTGGGTCTGACCCCGGAACAGATTGCCCGTGC CCTGGCACGCCTGCGTCAGTTACTGGGC
D25	GGTTATCAGGAACTGGTTCAGCAGATTCTGGCCGTT CTGCGTCGTAATCCGGGTAATGCCGTGGAAACCGCA ATTGTTGCAGCAAAATATGCCGGTCTGAGCGATGAA TTCATTGAAGCAATTCGCCGCTGGGGTCAGCTGTAT GCAAGTGGCGTTGAAAGTCCGGCAGTTCAGGCAATT ATTGATGCAGCCAAAGCAATGTATCAGGGCCGCCCG GGTCTGGTGGAATATGCAAAAATTATTATTGCCGCA GTGCGCCGCCTGTATGAACGTATGGGT

D46	GATGAAATCCAGAAAATGATCGAAGAATTTGCCAA AGCCCTGGGTGCAACCAGCATTGATCAGCTGCTGGC ATTTCTGCGCGCCCTGGTGCGTCGCCATCCGGATAA TGATCTGGTTCGTGCAGCACTGCAGGCCGGCCTGAG CCCTGCAATTGCCTTTAAACTGCTGGCCAAACTGGC AGGTGGTGACCTGGAAAAAGCACTGCGTATTGCCCT GGAAGGT
D47	GGTGACGCTGCCGCTGCTCGTGCCGCTCTGCGTGCA GGTAAAACCGCCGAAGAAATTGTTAAACTGCTGAA ACAGCTGAATATTAATCCGGGCGCAATTGCCGTGGC CAAAGCAGTTGTGCGTCTGGGCGTGGGCGCCGAAG GTGTTGTTCTGGCAGTGCTGCTGGCACTGCGCGATA AAAGCAATCTGGATAAAGTGATTCAGCAGGTGCTG GCCTGGCATGGT
D53	GGTGACCCTATTGATATTCTGATTAAGATTGCCAAA GCCCTGGGTATTAGCGATCGTGCAGCCAAAATTCTG GCCGAAGCCGCAAAAGCAGCAGGTATTACCAGTGC AGAAGGCGCACTGCGTCTGGCCAATGGCGAATATCC GGAAGCATGGAAATATGCCATTGAACTGGCAAAAA AGCGTGGTGACGATGCAGCCCTGGCAGCACTGCGTC GCGCCTTTGGC
D55	GGTCATCCGGTTGTGGTGCAGGCCCTGATGGATGCA ATGGGTCCGGAAGGCGCAGCACAGATTCAGCAGTG GCTGAAAAAAGCCGGTGGCGATACCCTGGAAGTTA TGCGCTTTACCGCACGCATGAATCCGGAAGCAGCAC GTCGTGCAGCCGCACGTCTGCGCCGTCTGGGTGTTA GTCCGGAACTGATTGCCGCCCTGCTGGCCGCCGCAA ATGGT
D60	AGTCTGGAAGAAAAAATCAAGAAGGCAATTGAAGT TGCCCGCCGCGCCGCAGCACTGGGTGCAGTGAGTCC GGAAGCCGCCAAAATTGCAATTGAACTGCTGAAAA GCGGCCTGAGCGTGCAGCAGGCCATTGATTTTCTGA AAAAGAATAATATCGGTCAGTTCGCCATTCTGGCCC TGAGCCTGGATAGTACCGAAAAAGTGGTTGAAAAA GTTAAAAAGGGC
D74	GATCCGCTGGTGGCAATGGCAGCAGCACTGCGCGCC GCCATTAAGCGTGCCGGCGGTGACGCCGCCGCCGCA GTGCAGGCATTATATAAAATGTTTCCGGGTGACGGC GTGGCACGCGCAATGGTGGCCGGTGAAGCAAGCCC GGAACTGATTCGTGAACTGGTTAAAGCCGCCCTGGA AGGTGACGTGGAAAAAATTAAGAAAGTGATGAAAG AGATCCTGGATTTTGCACGTAGCGGC
D78	AGCCTGGTGGAAAAAGCAAAAAAGCTGCTGGGCGT TAGCGTTGAAGGTGTTCGTGCAGCAGCACGCTATGC

	CGAAGCATTCAATGGCGATGCCCGTCGCGTGGCCCG TGCATTAGTTGAAGCAGGTCTGGTGAGTCCGGAAGG TCGCGCATTTCTGCTGCGTGCCCTGGCAGGCGGCCC GGGTTTAGATGTGGTTCGTAAAGGTGTGCAGGTTAT TCTGGAATATCTGCGTCGC
D80	GGTGAAGTTCTGGAAACCCTGCGCCGCATTGTTGAA CTGGGTCTGCCGGGCATTGGCATCTATCAGGCATAC GCTCAGGGCCTGACCGTTGCCCAGATTGCAGCAGCA CTGCGCGCACGCGGCTTTACCCCGGAAGAAGTTAAA AAAGCCGCCGAAGCCGCCCTGCGCACCGGTAGCCCT GCTACAGCCGCAATTATTCGTGAAATTGTTAAAGCC CTGGGC
D81	GGTGACATTCGTGCAGCCCTGCAGCGTCTGATTGAA GCAGCAAAACGCTTTATGGCAGCCCCGATGAGCGA AGAACAGCGCCTGGCCGCACTGATTGCCATTTTTAT TGCAAAACTGCTGAAAAGTGGCGTTAGCGTGGAAG AAGCCGCCCGTTTTGCAATTGAACTGGGTAGTATTG ATCTGATTGCCTATGCCGCCGCACTGGCCCGTGCCT TAGGTGCTAGCCCGGTTGCCGTGGCCCTGGTGCGTG CTGTTGCCGCAGCAGCAGGC
MBP- D12	AAAATCGAAGAAGGTAAACTGGTAATCTGGATTAA CGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGG TAAGAAATTCGAGAAAGATACCGGAATTAAAGTCA CCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCC CACAGGTTGCGGCAACTGGCGATGGCCCTGACATTA TCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTC AATCTGGCCTGTTGGCTGAAATCACCCCGGCCGCAG CGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATG CCGTACGTTACAACGGCAAGCTGATTGCTTACCCGA TCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAG ATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAG ATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGG TAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTA CTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTA TGCGTTCAAGTATGCAGCCGGCAAGTACGACATTAA AGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGG GTCTGACCTTCCTGGTTGACCTGATTAAAAACAAAC ACATGAATGCAGACACCGATTACTCCATCGCAGAAG CTGCCTTTAATAAAGGCGAAACAGCGATGACCATCA ACGGCCCGTGGGCATGGTCCAACATCGACACCAGC GCAGTGAATTATGGTGTAACGGTACTGCCGACCTTC AAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTG AGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGA GCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGAC

	TGATGAAGGTCTGGAAGCGGTTAATAAAGACAAAC CGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAG AGTTGGCGAAAGATCCACGTATTGCCGCCACTATGG AAAACGCCCAGAAAGGTGAAATCATGCCGAACATC CCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACT GCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTC GATGCAGCCCTGGCAGCCGCGCAGACTAATGCAGC GGCCGCAGATAAGGTGGCAGTTATGGCAGCAATGG CCCGCGCCGCACTGGAAATGAGCCTGGAAGAAGCC GCCCAGTATGCAGTTGAACTGGGCGCCGGTCCGGAA ACCCTGAAACGCATTCGTCAGGCAACCAGTGTTCGC GAAGTGGCAATTCTGATTGCAATTAGTTGGTATCCG GAAAATGAAGAACTGGCAAAAAAGGTTGTTGATCG CGTGCTG
MBP-D22	AAAATCGAAGAAGGTAAACTGGTAATCTGGATTAA CGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGG TAAGAAATTCGAGAAAGATACCGGAATTAAAGTCA CCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCC CACAGGTTGCGGCAACTGGCGATGGCCCTGACATTA TCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTC AATCTGGCCTGTTGGCTGAAATCACCCCGGCCGCAG CGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATG CCGTACGTTACAACGGCAAGCTGATTGCTTACCCGA TCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAG ATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAG ATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGG TAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTA CTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTA TGCGTTCAAGTATGCAGCCGGCAAGTACGACATTAA AGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGG GTCTGACCTTCCTGGTTGACCTGATTAAAAACAAAC ACATGAATGCAGACACCGATTACTCCATCGCAGAAG CTGCCTTTAATAAAGGCGAAACAGCGATGACCATCA ACGGCCCGTGGGCATGGTCCAACATCGACACCAGC GCAGTGAATTATGGTGTAACGGTACTGCCGACCTTC AAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTG AGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGA GCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGAC TGATGAAGGTCTGGAAGCGGTTAATAAAGACAAAC CGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAG AGTTGGCGAAAGATCCACGTATTGCCGCCACTATGG AAAACGCCCAGAAAGGTGAAATCATGCCGAACATC CCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACT GCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTC

	GATGCAGCCCTGGCAGCCGCGCAGACTAATGCAGCT CGTGCGGCCGCAATCGCGGAAGTTCAGGCAGCACT GCAGGCAGCCCTGAATAGTGATGATGATGCAGTTGT GGCATGGGTTCAGGCAGTTATGGATCTGATTAAGAA TGGTAAAGTGAGTGAAGAAGAAGCAAAACTGATGC TGGCCGCCGCCGCAGCCTATGCAGGCGTTCTGAGTC CGGAACTGATGAAAGCACTGGGTCTGACCCCGGAA CAGATTGCCCGTGCCCTGGCACGCCTGCGTCAGTTA CTGGGC
MBP-D53	AAAATCGAAGAAGGTAAACTGGTAATCTGGATTAA CGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGG TAAGAAATTCGAGAAAGATACCGGAATTAAAGTCA CCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCC CACAGGTTGCGGCAACTGGCGATGGCCCTGACATTA TCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTC AATCTGGCCTGTTGGCTGAAATCACCCCGGCCGCAG CGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATG CCGTACGTTACAACGGCAAGCTGATTGCTTACCCGA TCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAG ATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAG ATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGG TAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTA CTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTA TGCGTTCAAGTATGCAGCCGGCAAGTACGACATTAA AGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGG GTCTGACCTTCCTGGTTGACCTGATTAAAAACAAAC ACATGAATGCAGACACCGATTACTCCATCGCAGAAG CTGCCTTTAATAAAGGCGAAACAGCGATGACCATCA ACGGCCCGTGGGCATGGTCCAACATCGACACCAGC GCAGTGAATTATGGTGTAACGGTACTGCCGACCTTC AAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTG AGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGA GCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGAC TGATGAAGGTCTGGAAGCGGTTAATAAAGACAAAC CGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAG AGTTGGCGAAAGATCCACGTATTGCCGCCACTATGG AAAACGCCCAGAAAGGTGAAATCATGCCGAACATC CCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACT GCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTC GATGCAGCCCTGGCAGCCGCGCAGACTAATGCAGC GCGTGCAGCCGCAATTGATATTCTGATTAAGATTGC CAAAGCCCTGGGTATTAGCGATCGTGCAGCCAAAAT TCTGGCCGAAGCCGCAAAAGCAGCAGGTATTACCA GTGCAGAAGGCGCACTGCGTCTGGCCAATGGCGAA

	TATCCGGAAGCATGGAAATATGCCATTGAACTGGCA AAAAAGCGTGGTGACGATGCAGCCCTGGCAGCACT GCGTCGCGCCTTTGGC