The Emergence of Novel Versus Known Three-Dimensional Structures from Random Sequences.

Rose Yang[a,‡ *], Hyunjun Yang[b,‡ *], Anton Davydenko[b], Zack Mawaldi[a], Rian Kormos[a], Dru Myerscough[a], Yibing Wu[a,*], William F. DeGrado[a,*]

[a]Department of Pharmaceutical Chemistry, Cardiovascular Research Institute, University of California, San Francisco, CA 94158
[b]Department of Biochemistry, Brandeis University, Waltham, MA 02453

[‡]R. Y. and H. Y. contributed equally.

* To whom correspondence may be addressed. Email: yibing.wu@ucsf.edu, yanghj@brandeis.edu, and william.degrado@ucsf.edu.

Abstract: It has been hypothesized that while random sequences are unlikely to fold into proteins of the length of globular proteins, repeated random sequences are more likely to adopt stably folded structures, with implications for molecular evolution. We used structure prediction methods to determine the foldability of approximately 120-residue sequences composed of 5- to 60-residue random repeats. With repeats of less than 30-residues, sequences were frequently discovered (1–12%) that fold with high confidence. For less than 60-residue repeats, we frequently observe β-solenoids, similar to those seen in natural proteins. We observe solenoids stabilized by apolar packing as well as ones stabilized by polar interactions with Ca^{2+} in the core of the structure as in natural RTX domains. Helical bundles were observed with high frequency when insertions or deletions (INDELs) were included between blocks of repeating sequences. We also observed a new super-secondary structure consisting of a tightly wound α-helical screw, and experimentally confirmed its stability and structure by CD spectroscopy and X-ray crystallography. Thus, structure predictors can discover structures that are well out of the distribution of the data upon which they were trained. Beyond 40-residue repeat lengths very few sequences were predicted to fold. The small number of structures we observed were representative of well-established major classes of tertiary structures; greater sampling would be needed to discover novel structures from a random distribution. These studies illuminate dark matter regions of protein structure space and support previous predictions that proteins evolved through the assortment of shorter peptide sequences.

Significance statement. The availability of powerful and accurate programs for predicting protein three-dimensional structures enables one to ask fundamental questions concerning the origin of folded functional proteins during evolution. We show that 120-residue proteins composed of random sequences repeated in tandem are predicted to be much more likely to fold than fully random proteins. These studies validate previous predictions that proteins evolved through the repetition and assortment of short peptide sequences. Also, some of the predicted structures represent novel conformations, which were confirmed experimentally. These findings advance our understanding of molecular evolution and have implications for design of novel proteins.

Introduction. Proteins are molecular machines that rely on well-defined tertiary structures to carry out many of their remarkable functions. The molecular mechanisms by which proteins evolved their modern-day sequences and structures have been a puzzle for decades. Early studies focused on the astronomical number of possible sequences[1,2]; for even a 100-residue protein there are 20^{100} possible permutations for a sequence of this length composed of the 20 naturally occurring amino acids. However, more than a single sequence can fold into a single tertiary structure, and there are many possible tertiary structures into which proteins can fold. Moreover, in pioneering work, Margaret Dayhoff identified approximate symmetry in modern-day protein sequences and structures, and suggested that proteins evolved from peptides that assembled into protein-like structures.[3,4] Gene duplication resulted in repeated sequences in which each individual repeat could evolve independently and asymmetrically under selective pressure for function.[5,6,7,8,9,10,11] In the resulting structures, the original repeating units are arranged into structures with approximate rotational or dihedral symmetry.[12] Other repeating sequences are arranged into elongated solenoidal structures with repeat units arrayed in pseudo-translational or screw symmetries. Such repeat proteins are found in enzymes, binding proteins and structural proteins.[9,13,14]

While there have been extensive investigations into design of repeat proteins[15,16], in each case the design process was biased towards the practical objective of obtaining a predetermined outcome, such as helical repeat proteins[17,18,19] trefoils[20,21], or solenoids[22,23,24]. Recent studies have focused on the design of proteins starting from random sequences, using mutations and a guiding function to reach a folded structure.[25,26] Here, we instead seek to determine the extent to which fully random repeating sequences can fold into native structures without rounds of sequence improvement. What fraction of random sequences are “in-distribution” and hence predicted to fold into tertiary structures already known to occur in the Protein Data Bank (PDB)? We also probe how the probability of folding varies with respect to the length of the sequence repeat, and the presence of insertions and deletions between the repeats. Next, we ask whether there are structures that are very far from the distribution of known structures, but that can nonetheless be identified using programs that are trained on natural proteins. We identify and experimentally characterize a new super-secondary structure.

Historically, designable sequences are defined as ones with an energetically favorable tertiary structure accompanied by a large energy gap between this structure and the next most stable member of the structural ensemble.[27,28,29] Designable structures can accommodate a large number of sequences; the more sequences that can fold into the same structure, the more designable it is defined to be. Here, we determine how repeat length and the presence of INDELs influence the foldability of random sequences.

Results

Foldability of random repeat proteins. To estimate foldability, we generated 120-residue proteins composed of repeating units ranging from 5 to 120 amino acids. Two amino acid alphabets were tested: (i) the full set of 20 canonical residues and (ii) a reduced set of 10 “primitive” residues thought to be the ones available early in molecular evolution (G, A, S, T, D,

E, I, L, P, V)[30]. Similar trends were observed for both alphabets up to 40-residue repeats; therefore, unless otherwise noted, we focus on results from the 20-residue alphabet.

For each repeat length, approximately one million structures were predicted using RaptorX[31], chosen for its speed and accuracy with small proteins.[32] Models were clustered using a greedy algorithm (see methods) with a 1.0 Å RMSD cutoff, which provided sufficient resolution to separate folds with qualitatively different motifs and φ/ψ angle distributions. Long monomeric helices were frequently predicted but excluded from further analysis. Representative cluster centroid sequences were subsequently re-predicted with AlphaFold3 (AF3)[33]; the RaptorX and AF3 models were generally consistent within 1.0 Å RMSD.

The fraction of sequences predicted to fold ("foldable sequences") depended strongly on repeat length (**Fig. 1A–F**). We required a value of pLDDT > 90 as a stringent test of foldability. The foldability was 3.6% for pentamers, rose to a maximum of 17.2% for 10-residue repeats, and declined sharply thereafter to 0.01% for proteins composed of two 60-residue repeats and 0.001% ppm for fully random sequences (**Fig. 1G**).

The variation in designability with respect to repeat length (**Fig. 1H**) suggested an interplay between the structural diversity available at each repeat length and the designability of the corresponding folds. To quantify the *structural diversity* available to a given repeat length, we count the number of distinct clusters required to achieve 99% coverage of the total number of foldable sequences for that repeat length. The number of structural clusters increases with the repeat length until a length of 30 residues is reached and declines rapidly thereafter (**Fig. 1I**). We define the *designability* of a given structural cluster as the number of sequences in our sample that are predicted to fold into that structure. We also defined a metric to compare the designability of the full ensemble of structures associated with each repeat length ($\#seq_{ave}$, **Fig. 1H**) as the mean number of sequences in each cluster for that repeat length. For short repeat lengths of less than 13-residues, a small number of highly designable structures dominated the foldable space, resulting in high values of $\#seq_{ave}$.

Together, these data show that there are relatively few designable conformations for structures with short repeat lengths (**Fig. 1H**), but this number increases as the number of mainchain torsional degrees of freedom increases. However, at repeat lengths longer than 30 residues, foldable sequences occur with such low frequencies that we observe far fewer within our sample size of one million sequences/repeat length.

Sequence and structural features of highly designable structural repeats.

Pentameric repeats. For pentameric repeats, the sequence space is small enough (approximately 20^5 / 5 = 640,000 unique sequences) to be exhaustively sampled (**Fig. 2**). Of these, 3.6% were predicted to fold into one of five β-solenoid–like structures that together account for ~98% of foldable sequences (pLDDT > 90, **Fig. 2A**). The two dominant clusters (cluster_0 and cluster_1, **Fig. 2B, D**) are structurally very similar, and together comprised ~89% of all models. They both form right-handed, square β-solenoids (4.00 ± 0.05 repeats/turn and a pitch of 4.6 Å/turn) formed

by repeating β-strands connected by sharp turns. The trajectory of the chain at the vertices of the squares resembles that of a classic turn (**Fig. 2F, G**): a mix of type I and type II β-turn conformations are found in cluster_0 and predominantly type I β-turns in cluster_1.

Database searches confirmed that the centroid structures closely matched known β-solenoids in natural proteins (**Fig. 2C, E**). The geometry of the 5-mer clusters 0 and 1 were observed 444 and 452 times, respectively, indicating that these motifs occur frequently in nature. The sequence logos for both natural and random pentameric repeats showed similar hydrophobic enrichment at positions 1 and 3, consistent with packing of side chains into the solenoid core (**Fig 2H, I**). [14,23,34,35,36] Finally, predictions restricted to the 10 "primitive" amino acids yielded the same four structures, although triangular solenoids were relatively more common than square structures in this reduced alphabet (**Fig. S1**).

Hexameric repeats. For repeats longer than 5 residues, exhaustive enumeration of sequence space was no longer feasible with our computing resources, so we analyzed approximately one million random sequences. 5.0% of these sequences were folded into one of 10 different solenoidal topologies (**Fig. 3**). The three largest clusters are triangular and account for **83.5**% of the random sequences analyzed; square and oval β-solenoids were much less common (**Fig 3A**). The two most frequent clusters are right-handed solenoids with a triangular cross-section (3.0 residues/turn) and an axial rise of approximately 4.5 Å/turn (**Fig. 3B, D)**. The β-strands are one residue longer in the 6-mers than in the above 5-mers. The corners are formed by type II and type I turns in clusters 0 and 1, respectively, which were enriched in glycine (**Fig. 3F, G**). The structures were qualitatively similar to known structures of repeating hexamers and database searches confirmed structural similarity to natural proteins, returning 435 and 75 hits for clusters 0 and 1, respectively (**Fig. 3C, E)**. In both natural and the randomly generated sequences, hydrophobic residues are strongly enriched at positions 1 and 3 of the β-strand segments, packing into a central apolar core **(Fig. 3H, I)**.

7–10 residue repeats. Repeat lengths in the 7–10 residue range marked a transition from compact uniform, β-solenoids toward more diverse, elongated and structurally adaptive solenoids. Heptameric and nonameric repeats produced relatively high foldability fractions (3.6% and 13.2%, respectively). Compared to the 5- and 6-mers, their folds were more numerous and varied: In 7-mers, a triangular arrangement with three repeats per turn was occasionally observed, but sidechain packing was looser than in hexameric solenoids. In both 7- and 9-mers, β-arcade geometries were common, in which two repeat units formed paired β-arches.[14,37,38] The cores were stabilized by apolar sidechains, though in some clusters internal Asn and Gln residues contributed hydrogen-bonded zippers.[39,40,41]

Even-numbered repeats tended to form symmetrical rectangular assemblies whose long edges are regular β-strands. Eight-residue repeats (**Fig 4A, B)** produced two major rectangular clusters (~45% of models) with tight arches forming the short edges. Consistent with their rectangular symmetry, they have 2.0 turns/repeat and an axial rise of 4.8 Å. Ten-residue repeats (**Fig 5A, B)** tended to have a rectangular cross-sectional β-solenoid motif (2.0 repeats/turn), which is elongated or widened relative to the 8-mer.

While the cores of the 9-mers and 10-mers were generally well packed with hydrophobic (Ala, Val, Ile, Leu, Met, Phe) or Ser/Thr, we also observed Asn or Gln residues with their sidechains forming extensive sidechain-sidechain and sidechain-mainchain, as seen in natural amyloids and solenoids.[39,40,41] We also were initially surprised to see that members of 9-mer and 10-mer clusters had Asp residues pointing inward proximal to several mainchain carbonyl groups. AlphaFold is known to predict metal-binding sites, even without the metal ions included, and the oxygen-rich environment appeared ideal for binding Ca(II) ions. Indeed, a search of the PDB identified repeating RTX Ca(II)-binding motifs and enzymes (e.g., 1RU4) with closely related structures in 7788 proteins. RTX motifs play an important role in bacterial secretion. Thus, our workflow identifies metal ion-binding motifs as well as structural motifs.

Together, these findings highlight a progression in which the number of repeats required to complete a turn of the solenoid decreases with increasing repeat length: 5-mer repeats form squares with four repeats/turn, 6-mers form triangles with three repeats/turn, 8-mers to 12-mers form increasingly variable shapes with two repeats/turn; 15–20-mers have a single repeat/turn, and 40 residue repeats have 0.5 repeats/turn (**Fig. 5C)**. As a result, the number of residues per turn of the solenoid stays roughly constant between 15 and 30 residues. This behavior appears to be dictated by the need to maintain efficient packing of interior sidechains. The emergence of greater geometric diversity at longer repeat lengths reflects increased conformational freedom as the number of backbone torsional angles increases. We also note that although both right-handed and left-handed β-solenoids are seen in known structures, with 70% frequency representing right-handed structures[14], we found almost exclusively right-handed solenoids in our models. This likely reflects the fact that the great majority of monomeric solenoids are right-handed, while solenoids that associate to higher order structure are often left-handed. Additionally, it has been noted that AlphaFold2 tends to predict right- rather than left-handed solenoids for self-associating β-solenoids known to be left-handed from independent structural studies.

Discovery of a novel structural unit comprising a tightly wrapped α-helical screw. Interestingly, we also observed a novel super-secondary structure, which we define as an α-helical screw, in which a series of interconnected helices are tightly wrapped around a central axis. Helical screws were observed most frequently in 8-mer repeats (133 occurrences); they were also present in 7-mer (30 occurrences) and to a lesser extent in other repeat lengths (**Fig. 6A, B**). We observe variations on this basic motif, which differ in how tightly the superhelix is wound (**Fig. 6D**). We identify and name a given screw conformation based on its repeat length and the number of repeats per turn (e.g., an 8-mer$_{2.5}$ would have 2.5 8-mer units per superhelical turn). We observe a range of 2.5 to 4.0 units per turn of the helical screw, and all were found to wind in a clockwise manner. A TM-align[42] search of the PDB revealed only a single example of an 8-mer helical screw (**Fig. 6C**), and no hits were found for other helical screw conformations. Thus, helical screws represent a “dark-matter” portion of protein conformational space.[43,44]

The 8-residue helical screws are all stabilized by hydrophobic clustering of apolar sidechains in the core of the superhelical screw, as seen in the predicted structures and sequence logos of the motif (**Fig. 6E, F**). Moreover, Pro occurs frequently at the N1 position of the helix, and short polar

residues Ser, Thr, Asp, Asn occupy the N-cap positions where they form hydrogen bonds to the amides exposed at the N-terminus of the helices.[45]

Given the novelty of this structural type, we sought to experimentally characterize an 8-mer helical screw, which we term HeliScrew. To reduce sequence redundancy and ensure a water-soluble exterior, we used Chroma[46] to design six helical screw sequences (HeliScrew1–6) derived from the original cluster centroid structure (HeliScrew7). All seven proteins expressed robustly in *E. coli* and were soluble in aqueous buffer at pH 7.5 by SDS-PAGE and SEC FPLC (**Fig. S22–S25**). All constructs eluted as apparent monomers except HeliScrew6, which showed a shifted higher–molecular weight peak by SEC–FPLC, indicative of higher-order assembly.

All seven proteins were subjected to X-ray crystallization screening. HeliScrew4 crystallized in 1.8 M ammonium citrate tribasic, pH 7.0, yielding rod-shaped crystals suitable for X-ray diffraction. Data were collected at Argonne National Laboratory (APS beamline 24-ID-E). The structure was solved by molecular replacement and refined in space group $P12_11$ to 2.20 Å resolution. The asymmetric unit contains two crystallographically independent HeliScrew4 molecules, one of which is shown in **Fig. 6G**. Comparison of the experimentally determined HeliScrew4 structure (magenta) with the Chroma design model (green) shows close agreement (Cα backbone r.m.s.d. = 1.05 Å; **Fig. 6G**). Comparison of the HeliScrew4 crystal structure with the cluster centroid HeliScrew7 also shows close agreement (Cα backbone r.m.s.d. = 1.38 Å). Circular dichroism spectroscopy of HeliScrew4 revealed a typical α-helical signature with negative minima at 208 and 222 nm (**Fig. S26**), suggesting that the protein adopts the designed α-helical structure in solution.

The high degree of “expressibility” and water solubility of these helical screw sequences as well as the excellent correspondence between the predicted and the crystallographically observed α-helical screw structures shows that they are indeed highly designable examples of structures from the “dark matter” region of conformational space.

Globular folds emerge in 60-mer and non-repeating structures. Beyond a repeat length of 40 residues, there is an abrupt change from repeating solenoidal structures to globular structures –all with folds that are well represented in the PDB. For 60-mers we observed a mixture of all-α folds, all-β, and α/β structures (**Fig. 7A**). All represented common topologies, and we found that their tertiary structures were covered to an extent of 80% or greater using two CATH domains[47] using CATHcover (**Fig. 7B**). No sequences were predicted to fold with a pLDDT > 90 for the fully random 120-mer, although 11 structures were found to have values of pLDDT > 80. Their structures included 3 helical bundles, 1 all-β structure and 7 α/β structures (**Fig. S27**), and all were covered at 80.0% percent or more by 2 or 3 CATH domains.

Insertions and deletions encourage formation of a-helical bundles. We observed helical structures with much lower frequency than for β-solenoids. Helical hairpins, with two helices running anti-parallel to one another were seen 393 times for 7-mer repeats (1.1% of the folded sequences) and 286 times (0.3%) for 11-mer repeats. Three- and four-helix bundles were even

more rare, being observed only in the 40-mer repeats (6 times, 1.4%), and the 30-residue repeats (25 times, 1.3%), respectively.

We were initially surprised by the relative paucity of conventional helical bundles and coiled coils in our random proteins, given that they occur widely in natural proteins. Instead, sequences that might have been expected to form coiled coils were predicted to form extended, monomeric α-helices. However, when we assessed the *intermolecular* folding of sequences composed of random 7-mer repeats as non-covalent homodimers, we found 15.4% of the sequences had a value of pLDDT > 90. The corresponding values for 3-, 4-, 5- and 6-stranded structures were 3.5%, 6.8%, 1.2%, and 0.3%, respectively.

These calculations suggested that intramolecularly folded helical bundles might be formed with high frequency only if one allowed occasional breaks between blocks of heptad repeats to allow the protein helices to fold back on one another. To test this idea, we systematically introduced INDELs, consisting of either 1- or 2-residue deletions or 1 to 4-residue Gly insertions between blocks of four repeated random heptads (to create app. 28-residue helices). This pattern was repeated four times to create potential 4-helix bundles. The presence of deletions indeed increased the fraction of sequences that folded with high confidence by 20-fold, from 0.002% to approximately 0.04%. More dramatic, 300 to 500-fold increases were seen with 1 to 4-residue Gly insertions, reaching 0.93% for a two-residue insertion (**Fig. 8A**). The large majority (39.7%) of the predicted structures represented classical left or right-turning helical bundles (**Fig. 8B, C**). Similar results were seen for sequences intended to form 5- or 6-helix bundles, respectively, although they were found to fold with lower frequency, as expected from the fact that they occur less frequently than 2, 3, or 4-helix bundles in nature (**Fig. S34, S35**).

Conclusions

Our results show that the inclusion of repeats enriches the fraction of otherwise random sequences that can adopt folded conformations. With repeat lengths of 5 to 20 residues, over 1% of the sequences are predicted to fold, mostly into repeating solenoids or β-hairpins, although we also observe helical bundles and tightly wound helical screws that represent a new family of super-secondary structures. With the inclusion of INDELs between blocks of repeats, we also observe the formation of helical bundles at a frequency of 0.5 to 1%.

On the other hand, beyond a repeat length of 30 residues, the frequency of observing predicted folded structures decreased sharply, reaching 0.001% for sequences lacking repeats. Also, these structures were all globular. Evidently, without the imposition of a sequence repeat, the probability of adopting a repeating conformational form becomes very low, and only more asymmetric arrangements typical of the domains seen in globular proteins are observed. The resulting structures are representative of ones seen in globular proteins, and include the α-helical, α/β, and all-β classes of proteins.

These observations answer our first questions concerning how frequently random sequences are likely to fold into structures similar to those found in nature. However, this does not necessarily

guarantee that all predicted proteins would fold as monomeric species in aqueous solution. In nature, solenoids and repeating β-hairpins are capped on at least one end to prevent end-to-end polymerization. In the absence of such repeat-breaking features we expect that proteins discovered from precisely repeating sequences would be prone to forming polymeric structures similar to amyloid fibrils. Indeed, short, amyloid-forming peptides have been hypothesized to have been important players in the molecular evolution of proteins.[48,49,50] They are known to have catalytic functions[41,51,52,53] and they can form membrane-like tubes and spheres as well as phase-segregated states[,54,55,56].[57] A solenoid combines many copies of monomeric amyloidogenic peptides into a single chain, allowing each monomer to mutate independently to evolve function. Indeed, solenoids form the frameworks of many enzymes across multiple catalytic classes, and they serve as structural domains in multi-domain proteins, anti-freeze proteins, cell-surface proteins, and viruses.[35]

We also noted that up to half of the predicted structures do not appear to have sufficiently polar exteriors to maintain water-solubility. Some of the helical bundle proteins had fully hydrophobic exteriors, which would make them good candidates for insertion into membranes. On the other hand, we are unaware of any transmembrane solenoids and so we expect that a large fraction of the predicted structures would be insoluble if expressed or chemically synthesized. Thus, the predictions shown here should be considered upper limits of the fraction of proteins that would be folded and soluble in membranes or aqueous solution.

We, like previous investigators[58], were also initially concerned about the presence of polar residues in the cores of the predicted structures of solenoids. While these features were previously considered to be indicative of false predictions, closer inspection indicated that the polar residues could impart function. For example, one prominent feature was a series of interior-facing Asp residues, that comprise the Ca(II)-binding repeats in natural RTX motifs[54,59,60] Other polar residues in the interior include Asn residues, which are common in zippers that stabilize parallel β-sheet structures and buried salt bridges, which are common in both water-soluble solenoids and amyloids. We also observe clusters of His, Glu/Asp and Cys residues that could bind transition metals known to catalyze hydrolytic and oxidative reactions. Indeed, metallo-amyloids that contain His-Xxx-His and related motifs form highly active esterases, carbonic anhydrases, and oxidases.[41,53,61]

Our next question was whether structure prediction programs would be able to extrapolate from the distribution found in the PDB to identify new folds. Previously, folding predictors have been used during the design of proteins that have structures with arrangements of secondary structures not frequently or ever seen in the PDB.[43,44] Here, we discovered an entirely new family of tightly wound helical screws. We observed different conformers for different helical screw repeat lengths, forming distinct sub-families with characteristic superhelical parameters. Only a single member of one of the helical screw family had previously been seen in nature, which prompted us to express and determine the structure of a *de novo* helical screw. The structure is in close (< 1.38 Å backbone RMSD; < 1.58 Å all heavy atom RMSD) agreement with the predicted structure, indicating that the motif is designable. Why then, do we not see it frequently in natural proteins? The sequence requirements required to adopt this structure appear to be restrictive: the core is

tightly packed with sterically restricted apolar side chains, and the helix is generally initiated by strong helix-promoting sequence, including a helix-initiating Pro near the N-terminus. It is possible that the sequence requirements for folding are so restrictive that there is little ability for variation in the sequence to generate function.

Our remaining question is whether there are other stably folded structures that are not predicted by current prediction methods. Any structure that is highly extended would have been removed by our radius of gyration filter, which we found necessary to eliminate very long helices or extended structures. Beyond this limitation, the question hinges on whether the set of structural fragments and the interactions required for assembling them into higher order structure represent a covered set. Grigoryan and coworkers have argued that a complete set of fragments and instructions to assemble them is largely covered in the PDB.[29,44,62] Indeed, at the level of sequentially contiguous backbone structures, it is likely that all energetically viable sequentially local structures (such as α-helices, hairpins, omega loops, polyproline II helices…), and the way that these fragments can interact is limited by simple physical forces that are captured in folding predictors. Thus, it is likely that—within our sampling size limit—we have identified most if not all possible designable folds up to lengths of approximately 30 residues. Beyond this length, our sampling is too low to identify rare folds, and the few folds we find are well precedented in the PDB. Given our discovery of a completely novel super-secondary structure in the regime where we do have good sampling, it is reasonable to expect that there might be other folds that would be discovered in a larger sample. Such hypothetical structures would not be part of the modern repertoire for a variety of reasons: 1) they might have not been sampled during evolution, 2) they might be refractory to experimental methods of structure determination, 3) they might have been sampled during evolution but failed to confer a functional advantage, or 4) they might possess an inherent liability, such as toxicity or a tendency to misfolding. New tertiary structures might, however, confer advantages for applications or environments not currently found in nature.

Methods

RaptorX random repeat generation. Each n-residue repeat was repeated in tandem to produce a 120-residue chain, with truncation or extension applied when the 120/n was not an integer. Repeat lengths of 5 - 13, 15, 17, 20, 24, 30, 40, 60, and 120 residues were analyzed. All sequences were submitted to RaptorX on UCSF's Wynton high-performance compute cluster, and models with mean pLDDT ≥ 90 were retained. For the primitive-alphabet analysis approximately 250,000 sequences per repeat length were generated.

Clustering method. A single model was selected at random as the initial seed, and all other structures were compared to it using TM-align (v2020). For every comparison, the full 120-amino acid sequence was used. Models with RMSD ≤ 1.2 Å and TM-score ≥ 0.7 relative to the seed were assigned to the same cluster. A new unassigned model was then chosen as the next seed until all structures were classified. Cluster sequences and their predicted structures are available as **supplementary data**.

Structure analysis. Aligned sequences were reconstructed from TM-align output logs. Consensus sequences were generated by applying a column-wise majority rule to the aligned sequences. Cluster centroid structures were analyzed using Bio.PDB (Biopython v1.83) to calculate φ and ψ dihedral angles across repeat segments. Torsion angles were projected onto APBLE Ramachandran map assignments to classify local backbone preferences, including α-helical, β-sheet, left-handed helical, and extended regions.[63] Position frequency matrices were converted to sequence logos, with amino acids colored according to a fixed scheme described in supplementary methods. For helical repeat motifs, residues were annotated using heptad notation (A–G).

Protein design. Initial structural models for protein design were derived from the cluster centroid of helical screws from the largest 8-mer cluster. Each starting backbone (74 total centroids) was first perturbed by adding noise (t = 0.5) using partial diffusion module in Chroma[46], followed by denoising and subsequent sequence design steps. For each seed, three designed sequences were generated, yielding 296 sequences in total (74 original + 222 designed). All sequences were evaluated by ESMFold[64] to obtain predicted structures. The designed models were chosen from the top 25th percentile of ESMFold pLDDT scores and top 25th percentile of backbone agreement between design and prediction (Cα RMSD). Only models passing both thresholds were chosen, preserving a pool of 74 top-scoring sequences per iteration. This iterative diffusion–design–prediction loop was repeated for 20 cycles, yielding 74 structures that converged toward high-confidence models with high pLDDT and low RMSD. The seven top ranking designs (HeliScrew1–6) and the original cluster centroid (HeliScrew7) were selected for expression and experimental characterization. Sequences for all designed models are provided in the supplementary data (**Fig. S22)**.

Protein expression and purification. Genes encoding the seven design proteins and the original cluster centroid protein (Twist Bioscience) were cloned into a pET-16b expression vector (Amp marker; BamH1 and Not1) containing an N-terminal His-tag-SUMO fusion tag and a cleavage site for human SENP2 protease. Gene sequences are provided in supplementary data (**Fig. S23**). Constructs were transformed into BL21(DE3) *E coli* cells, which were streaked onto

LB agar plates. A single colony from each transformation was inoculated into LB medium, and protein expression was induced with 0.5 mM IPTG at an OD of ~0.5. Cultures were harvested after 4 h of induction at 37 °C. Cell pellets were collected by centrifugation and resuspended in lysis buffer (50 mM NaPhos, 150 mM NaCl, pH 7.5), followed by sonication. Lysates were clarified by centrifugation and loaded onto nickel-nitrilotriacetic acid (Ni-NTA) agarose bead affinity columns for immobilized metal affinity chromatography (IMAC). Bound proteins were eluted using 250 mM imidazole containing lysis buffer. Eluted proteins were dialyzed overnight to remove imidazole and subsequently cleaved with recombinant human SENP2 protease to remove the His-SUMO tag, leaving a single additional serine at the N-terminus. Cleaved proteins were separated by a second round of IMAC purification. Final purification was performed using size-exclusion chromatography (SEC) FPLC on a Superdex 200 Increase 10/300 GL column equilibrated in 50 mM sodium phosphate, 150 mM NaCl, pH 7.5. Peak fractions corresponding to monomeric species were pooled and concentrated for downstream biophysical characterization, including CD spectroscopy and X-ray crystallography.

Protein crystallization. Crystallization trials were set up for all seven helical screw proteins, testing 384 conditions (Hampton screen Index, Crystal, PEG, JCSG). Heliscrew4 formed in 1.8 M ammonium citrate, pH 7.0; crystallization by hanging drops gave X-ray diffraction-grade crystals. Diffraction data were collected at the Advanced Photon Source, Beamline 24-ID-E (NE-CAT; 12.66 keV; 0.2 ° oscillation over 360° (**Table S2**, pdb 9ZEM). The structure of the helical screw was solved by molecular replacement using the designed model in Phenix.phaser[65]. Iterative model building and refinement was carried out in Phenix.autobuild[66], Coot[67], and Phenix.refine[68].

Circular Dichroism (CD). Far UV CD spectra were recorded using a Jasco J-810 instrument for Heliscrew4 in H_2O) in a 1 mm pathlength cell. Spectra were collected in 0.5 nm increments from 260–190 nm at a rate of 50 nm/min; 3 scans were averaged.

Data availability:

All sequences and computed structures for the proteins are available here. (Dataset DOI: 10.5061/dryad.vmcvdnd4m). Statistics concerning the size of clusters and pdb codes and chain/sequence identifiers are provided in the **supplementary data.**

Code availability

Code to cluster the predicted structures are available here. (https://github.com/degrado-lab/repeating-sequences.git)

Supplementary data files:

PDB files for all members of all clusters and all cluster centroids organized by repeat length.

Sequences of all centroids.

Excel files with designability metrics for each cluster.

Figures S1–35 and Table S1–S2.

Acknowledgements. W.F.D. acknowledges support from the National Science Foundation (NSF, CHE-2528384 and MCB-2306190) and the National Institutes of Health (R35GM122603). H.Y. acknowledges support from the National Institutes of Health (R00AG084926). X-ray diffraction studies were performed under APS beam time award (DOI: https://doi.org/10.46936/APS-191095/60014853) from the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science user facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. H.Y. thanks Dr. Kay Perry for beamline support. Z.M. acknowledges support from the NIH (NIH T32GM149436). D.M. acknowledges support from the NSF Graduate Research Fellowships Program (GRFP).

Author contributions

R.Y., H.Y., Y.W., and W.F.D designed the research; R.Y., H.Y., R.K. A.D. Z.M., D.M. and Y.W. generated the structures. R.Y., H.Y., Y.W., and W.F.D analyzed the data; and R.Y., H.Y., Y.W., and W.F.D wrote the paper. Y.W. and W.F.D. supervised the study. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

References

1 Levinthal C (1969) How to fold graciously. In Mössbauer Spectroscopy in Biological Systems, Proceedings of a meeting held at Allerton House, Monticello, IL, eds. Debrunner P, Tsibris JCM, Münck E. (University of Illinois Press, Urbana), pp. 22–24.

2 Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. Proc Natl Acad Sci USA 89(1):20–22.

3 Eck RV, Dayhoff MO (1966) Evolution of the structure of ferredoxin based on living relics of primitive organisms. Science 152:363–366.

4 Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 118:e2016239118.

5 Söding J, Lupas AN (2003) More than the sum of their parts: On the evolution of proteins from peptides. Bioessays 25:837–846.

6 Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. Nat Rev Genet 11:572–582.

7 Alva V, Söding J, Lupas AN (2015) A vocabulary of ancient peptides at the origin of folded proteins. eLife 4:e09410.

8 Tokuriki N, Tawfik DS (2009) Protein dynamism and evolvability. Science 324(5924):203–207.

9 Deryusheva EI, Machulin AV, Galzitskaya OV (2023) Diversity and features of proteins with structural repeats. Biophys Rev 15:1159–1169.

10 Kocher CD, Dill KA (2024) Origins of life: The protein folding problem all over again? Proc Natl Acad Sci USA 121:e2315000121.

11 Muthahari YA, Magnus L, Laurino P (2025) From duplication to fusion: Expanding Dayhoff's model of protein evolution. Protein Sci 34:e70054.

12 Goodsell DS, Olson AJ (2000) Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29:105–153.

13 Binz HK, Plückthun A (2005) Engineered proteins as specific binding reagents. Curr Opin Biotechnol 16:459–469.

14 Kajava AV, Steven AC (2006) β-rolls, β-helices, and other β-solenoid proteins. Adv Protein Chem 73:55–96.

15 Petersen M, Barrick D (2021) Analysis of tandem repeat protein folding using nearest-neighbor models. *Annu Rev Biophys* 50:245–265.

16 Ferreiro DU, Komives EA, Wolynes PG (2021) Frustration, function, and folding. Annu Rev Biophys 50:321–345.

17 Regan L, DeGrado WF (1988) Characterization of a helical protein designed from first principles. Science 241(4868):976–978.

18 Brunette TJ, Parmeggiani F, Huang P-S, Bhabha G, Ekiert DC, Tsutakawa SE, Hura GL, Tainer JA, Baker D (2015) Exploring the repeat protein universe through computational protein design. Nature 528(7583):580–584.

19 Parmeggiani F, Randles LG, Burrone OR, Plückthun A (2015) Designed armadillo-repeat proteins as modular peptide-binding scaffolds. Proc Natl Acad Sci USA 112:1129–1138.

20 Terada D, Voet ARD, Noguchi H, Kamata K, Ohki M, Addy C, Fujii Y, Yamamoto D, Ozeki Y, Tame JRH, Zhang KYJ (2017) Computational design of a symmetrical β-trefoil lectin with cancer cell binding activity. Sci Rep 7:5943.

21 Tenorio CA, Parker JB, Blaber M (2022) Functionalization of a symmetric protein scaffold: Redundant folding nuclei and alternative oligomeric folding pathways. Protein Sci 31:e4301.

22 Peralta MDR, Karsai A, Ngo A, Sierra C, Fong KT, Hayre NR, Mirzaee N, Ravikumar KM, Kluber AJ, Chen X, Liu G-Y, Toney MD, Singh RR, Cox DL (2015) Engineering amyloid fibrils from β-solenoid proteins for biomaterials applications. ACS Nano 9(1):449–463.

23 MacDonald JT, Kabasakal BV, Godding D, Kraatz S, Henderson L, Barber J, Freemont PS, Murray JW (2016) Synthetic β-solenoid proteins with the fragment-free computational design of a β-hairpin extension. Proc Natl Acad Sci USA 113(37):10346–10351.

24 Pretorius D, Nikov GI, Washio K, Florent S-W, Taunt HN, Ovchinnikov S, Murray JW (2025) Designing novel solenoid proteins with in silico evolution. bioRxiv:2025.04.23.646631.

25 Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Baker D (2021) De novo protein design by deep network hallucination. Nature 600:547–552.

26 Sahakyan H, Babajanyan SG, Wolf YI, Koonin EV (2025) In silico evolution of globular protein folds from random sequences. Proc Natl Acad Sci USA 122(27):e2509015122.

27 Szczepaniak K, Lach G, Bujnicki JM, Dunin-Horkawicz S (2014) Designability landscape reveals sequence features that define axial helix rotation in four-helical homo-oligomeric antiparallel coiled-coil structures. J Struct Biol 188(2):123–133.

28 Shakhnovich EI (1998) Protein design: a perspective from simple tractable models. Fold Des 3(3):R45–R58.

29 Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci USA 103(8):2605–2610.

30. Makarov M, Sanchez Rocha AC, Krystufek R, Cherepashuk I, Dzmitruk V, Charnavets T, Faustino AM, Lebl M, Fujishima K, Fried SD, Hlouchova K (2023) Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. J Am Chem Soc 145:5320–5329.

31 Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7:1511–1522.

32 Agarwal S, McShan AC (2024) The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. Nat Chem Biol 20:950–959.

33 Jumper J, Evans R, Pritzel A, Green T, Tunyasuvunakool K, Žídek A, Bridgland A, Meyer C, Kohl SAA, Potapenko A, Ballard A, Cowie A, Figurnov M, Kohli P, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Hassabis D (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 631:349–356.

34 Vetting MW, Hegde SS, Fajardo JE, Fiser A, Roderick SL, Takiff HE, Blanchard JS (2006) The pentapeptide repeat proteins. Biochemistry 45:1–10.

35 Kajava AV (2012) Tandem repeats in proteins: from sequence to structure. J Struct Biol 179(3):279–288.

36 Zhang R, Kennedy MA (2021) Current understanding of the structure and function of pentapeptide repeat proteins. Biomolecules 11:638.

37 Yoder MD, Keen NT, Jurnak F (1993) Unusual structural features in the parallel β-helix in pectate lyases. Structure 1:241–251.

38 Hennetin J, Jullian B, Steven AC, Kajava AV (2006) Standard conformations of β-arches in β-solenoid proteins. J Mol Biol 358(4):1094–1105.

39 Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJW, McFarlane HT, Madsen AØ, Riekel C, Eisenberg DS (2007) Atomic structures of amyloid cross-β zippers. Nature 447:453–457.

40 Wu H, Saltzberg DJ, Kratochvil HT, Jo H, Sali A, DeGrado WF (2019) Glutamine side chain $^{13}C=^{18}O$ as a non-perturbative IR probe of amyloid fibril hydration and assembly. J Am Chem Soc 141:7320–7326.

41 Rufo CM, Moroz YS, Moroz OV, Stöhr J, Smith TA, Hu X, DeGrado WF, Korendovych IV (2014) Short peptides self-assemble to produce catalytic amyloids. Nat Chem 6:303–309.

42 Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309.

43 Harteveld Z, Van Hall-Beauvais A, Morozova I, Southern J, Goverde C, Georgeon S, Rosset S, Defferrard M, Loukas A, Vandergheynst P, Bronstein MM, Correia BE (2024) Exploring “dark-matter” protein folds using deep learning. Cell Syst 15:898–910.e5.

44 Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D (2012) Principles for designing ideal protein structures. Nature 491:222–227.

45 Richardson JS, Richardson DC (1988) Amino acid preferences for specific locations at the ends of α-helices. Science 240:1648–1652.

46 Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, Frappier V, Lord DM, Ng-Thow-Hing C, Van Vlack ER, Tie S, Xue V, Cowles SC, Leung A, Rodrigues JV, Morales-Perez CL, Ayoub AM, Green R, Puentes K, Oplinger F, Panwar NV, Obermeyer F, Root AR, Beam AL, Grigoryan G (2023) Illuminating protein space with a programmable generative model. Nature 623:1070–1078.

47 Waman VP, Bordin N, Lau A, Kandathil S, Wells J, Miller D, Velankar S, Jones DT, Sillitoe I, Orengo C (2025) CATH v4.4: major expansion of CATH by experimental and predicted structural data. Nucleic Acids Res 53(D1):D348–D355.

48 Reches M, Gazit E (2004) Formation of closed-cage nanostructures by self-assembly of aromatic dipeptides. Nano Lett 4:581–585.

49 Carny O, Gazit E (2005) A model for the role of short self-assembled peptides in the very early stages of the origin of life. FASEB J 19:1051–1055.

50 Maury CPJ (2018) Amyloid and the origin of life: self-replicating catalytic amyloids as prebiotic informational and protometabolic entities. Cell Mol Life Sci 75:1499–1507.

51 Hirst AR, Roy S, Arora M, Das AK, Hodson N, Murray P, Marshall S, Javid N, Sefcik J, Boekhoven J, van Esch JH, Santabarbara S, Hunt NT, Ulijn RV (2010) Biocatalytic induction of supramolecular order. Nat Chem 2(12):1089–1094.

52 Zozulia O, Dolan MA, Korendovych IV (2018) Catalytic peptide assemblies. Chem Soc Rev 47(10):3621–3639.

53 Arad E, Jelinek R (2022) Catalytic amyloids. Trends Chem 4:907–917.

54 Shur O, Dooley K, Blenner M, Baltimore M, Banta S (2013) A designed, phase changing RTX-based peptide for efficient bioseparations. BioTechniques 54(4):197–206.

55 Frederix PWJM, Scott GG, Abul-Haija YM, Kalafatovic D, Pappas CG, Javid N, Hunt NT, Ulijn RV, Tuttle T (2015) Exploring the sequence space for (tri-)peptide self-assembly to design and discover new hydrogels. Nat Chem 7:30–37.

56 Peralta MDR, Karsai A, Ngo A, Sierra C, Fong KT, Hayre NR, Mirzaee N, Ravikumar KM, Kluber AJ, Chen X, Liu G-Y, Toney MD, Singh RR, Cox DL (2015) Engineering amyloid fibrils from β-solenoid proteins for biomaterials applications. ACS Nano 9(1):449–463.

57 Mandal D, Shirazi AN, Parang K (2014) Self-assembly of peptides to nanostructures. Org Biomol Chem 12:3544–3561.

58 Pratt OS, Elliott LG, Haon M, Mesdaghi S, Price RM, Simpkin AJ, Rigden DJ (2025) AlphaFold 2, but not AlphaFold 3, predicts confident but unrealistic β-solenoid structures for repeat proteins. Comput Struct Biotechnol J 27:467–477.

59 Linhartová I, Bumba L, Mašín J, Basler M, Osička R, Kamanová J, Procházková K, Adkins I, Hejnová-Holubová J, Sadílková L, Morová J, Šebo P (2010) RTX proteins: a highly diverse family secreted by a common mechanism. FEMS Microbiol Rev 34(6):1076–1112.

60 Mohd Rozi MFA, Raja Abd Rahman RNZ, Leow ATC, Mohamad Ali MS (2022) Ancestral sequence reconstruction of ancient lipase from family I.3 bacterial lipolytic enzymes. Mol Phylogenet Evol 168:107381.

61 Friedmann MP, Torbeev V, Zelenay V, Sobol A, Greenwald J, Riek R (2015) Towards prebiotic catalytic amyloids using high-throughput screening. PLoS One 10:e0143948.

62 Grigoryan G, DeGrado WF (2011) Probing designability via a generalized model of helical bundle geometry. J Mol Biol 405(4):1079–1100.

63 Fiser A, Do RK, Šali A (2000) Modeling of loops in protein structures. Protein Sci 9(9):1753–1773.

64 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379(6637):1123–1130.

65 McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Crystallogr 40(4):658–674.

66 Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Zwart PH, Hung LW, Read RJ, Adams PD (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr D Biol Crystallogr 64:61–69.

67 Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66:486–501.

68 Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D Biol Crystallogr 68:352–367.

Figure Legends

Figure 1. Foldability of random repeat proteins as a function of repeat length. (A-F) Distributions of mean pLDDT for proteins 5, 8, 10, 24, 40 and 120-residue repeats. The red dashed line marks a pLDDT score of 90. (G) Fraction of sequences predicted to fold (pLDDT ≥ 90) per repeat length. (H) Number of structural clusters to achieve 99% coverage of foldable sequences per repeat length. (I) Mean designability of structural clusters ($\#seq_{ave}$) per repeat length.

Figure 2. Features of 5mer repeat proteins. (A) Histogram depicting frequency and structure of each cluster. (B, D) Cross-sectional models of the two dominant pentameric clusters. (F, G) Plots of φ/ψ angles of centroids highlighting β-strand geometries (positions 1–3) and turn conformations (positions 4–5). (C, E) Closely matching natural β-solenoids from the PDB. (H, I) Sequence logos for the clusters.

Figure 3. Features of 6mer repeat proteins. (A) Histogram depicting frequency and structure of each cluster. (B, D) Cross-sectional structures of the two dominant pentameric clusters. (F, G) Centroid Ramachandran plots highlighting conserved β-strand φ/ψ geometries (positions 1–3). (C, E) Closely matching natural β-solenoids from the PDB. (H, I) Sequence logos for the clusters.

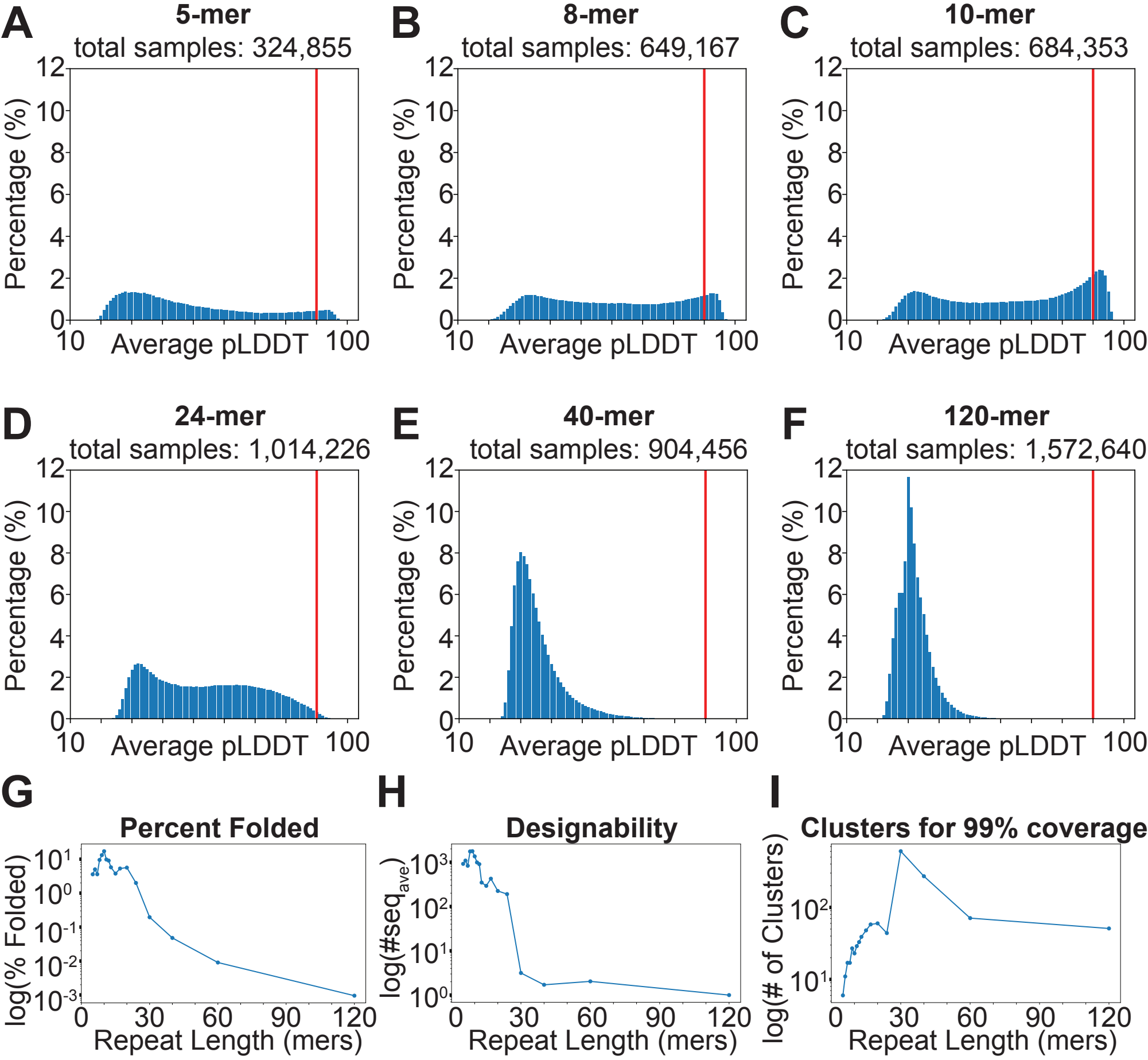
Figure 4. Features of 8mer repeat proteins. (A) Sequence logos and secondary structure assignments for the three largest clusters. (B) Axial views of structures and comparisons with natural proteins. (C). Members of 9mer and 10mer repeat clusters that resemble natural RTX calcium-binding motifs, with AlphaFold3 predictions likewise positioning Ca^{2+} at acidic residues.

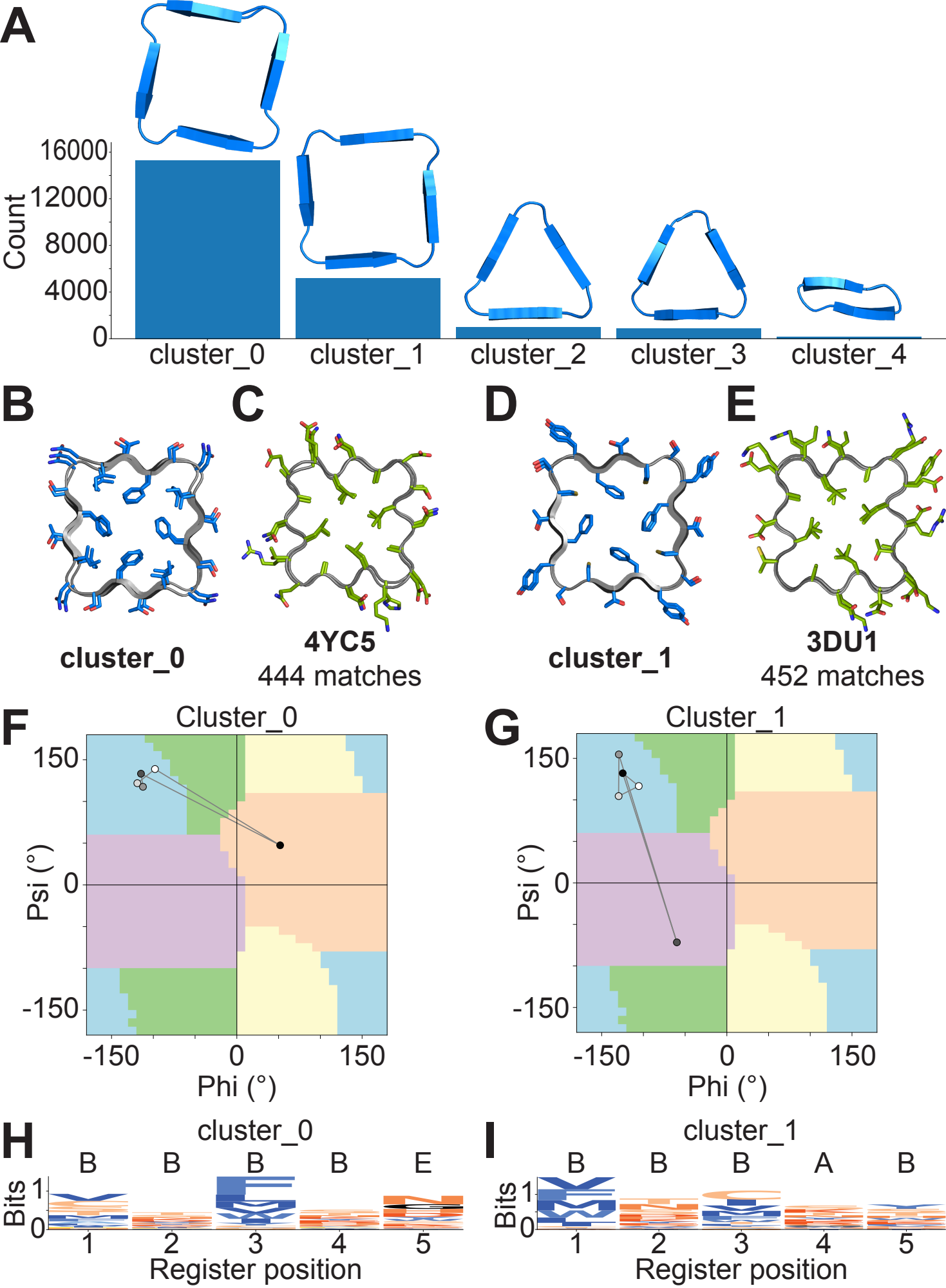
Figure 5. Features of 10mer repeat proteins. (A, B) Sequence logos and cross-sectional structures for the three major clusters. (C) Centroid structures for repeats ranging 5 – 60 residues. As repeat length increases, the number of repeats per turn (red) decreases.

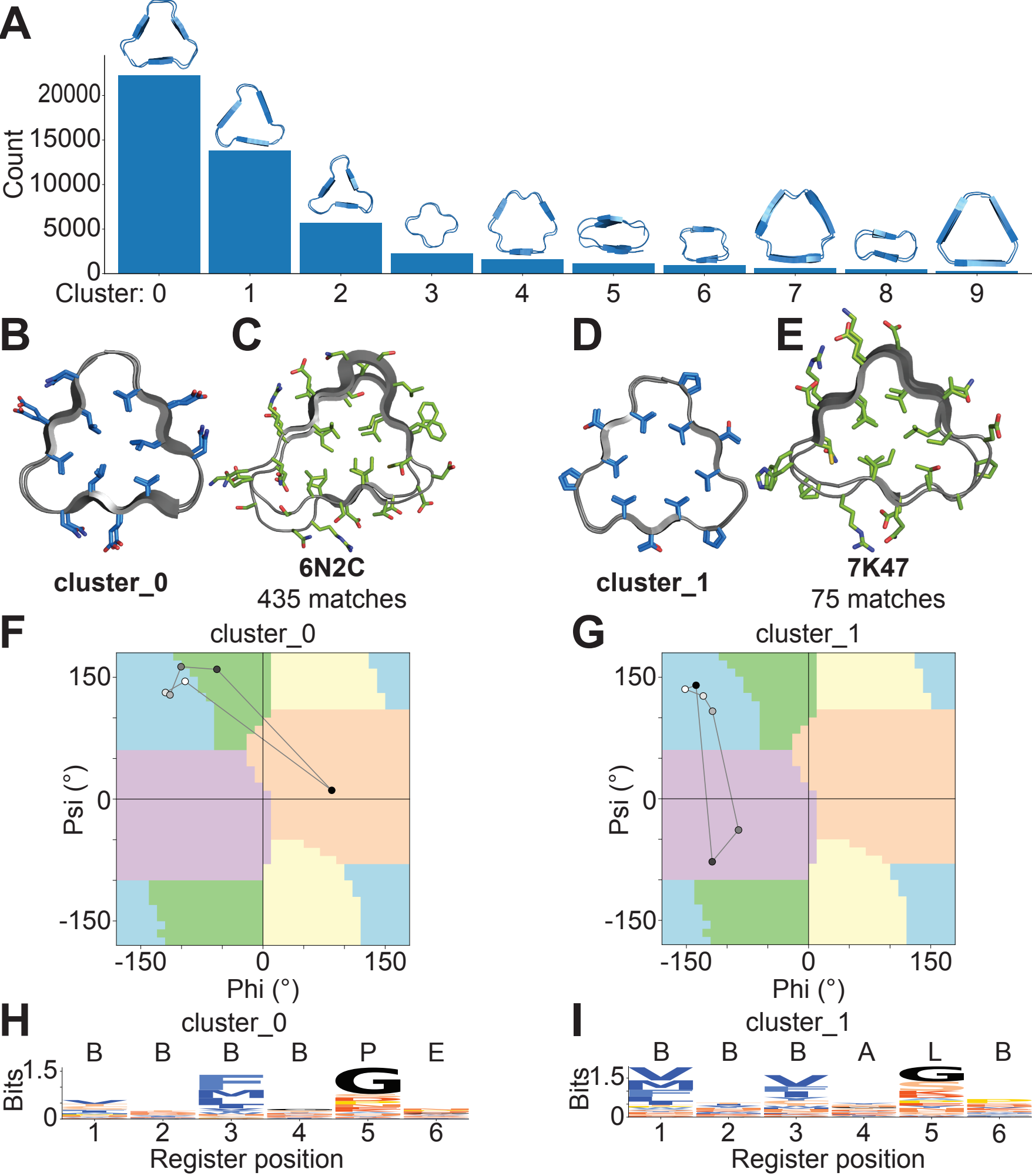
Figure 6. (A) Frequency of occurrence of α-helical screws as a function of repeat length. (B) Centroid structures from the largest clusters for 7mer, 8mer, and 12mer repeats. (C) Structure of the natural protein 7QLR containing a similar helical screw. (D) Examples of three distinct α-helical screw geometries observed for 8mer repeats. (E) Cross-sectional structure of a member of the largest 8-mer cluster. Hydrophobic core-packing residues (green) and a helix-initiating Pro (yellow) are shown. (F) Sequence logo representation of the 8mer cluster with APBLE annotation[63]. (G) Comparison of experimentally determined structure HeliScrew4 (magenta) with the design model HeliScrew7 (green), showing close agreement (b.b. r.m.s.d. = 1.38 Å).

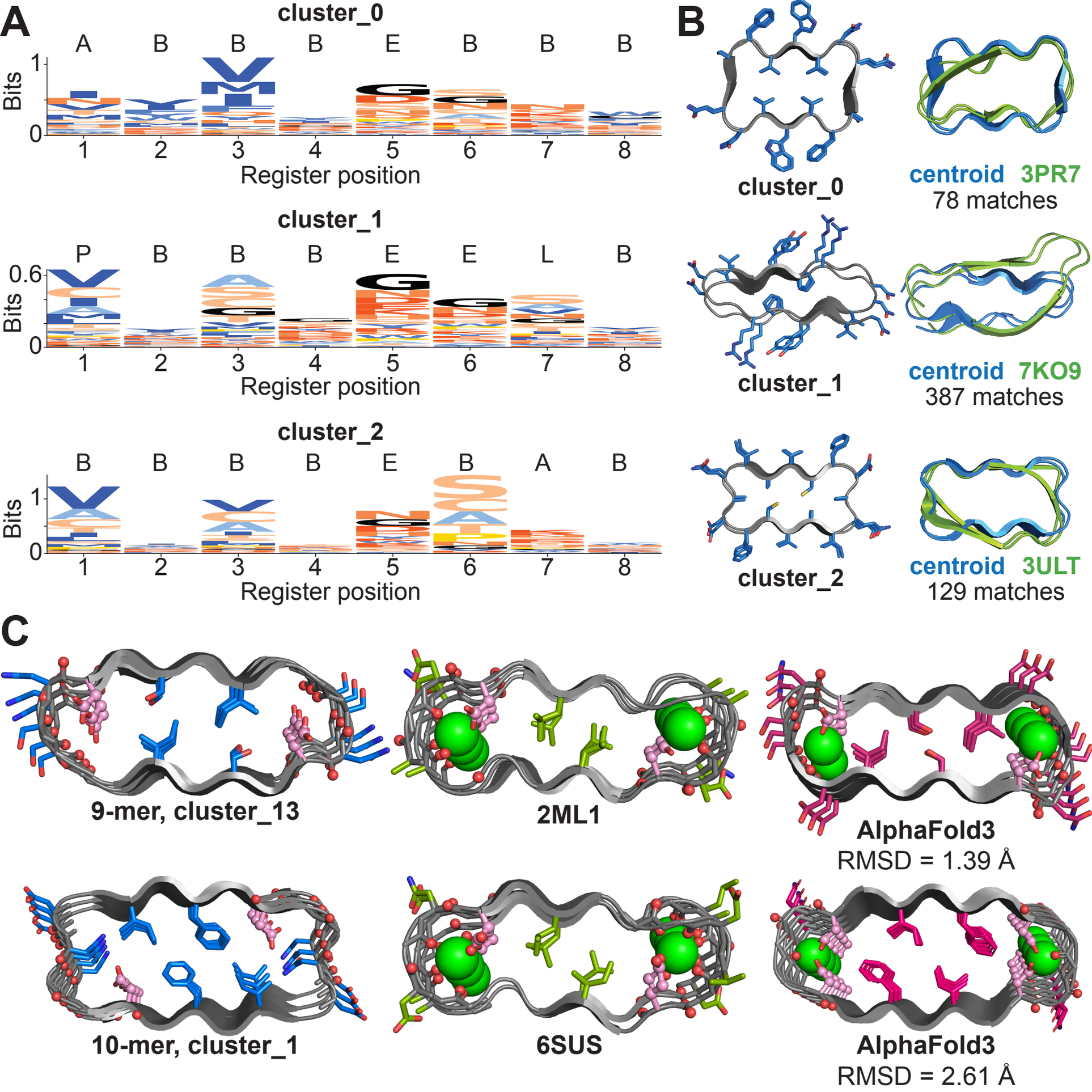
Figure 7. (A) 60mer repeat proteins with α, β, and mixed α/β architectures, and the number of structures observed in each cluster. (B) Comparisons between designed repeat proteins and their closest natural CATH domains. Centroids for 60mer_cluster_3 and 120mer_cluster_0 align closely with natural folds, achieving ~80% coverage with two CATH domains. (C) Overlays with AF3 predictions.

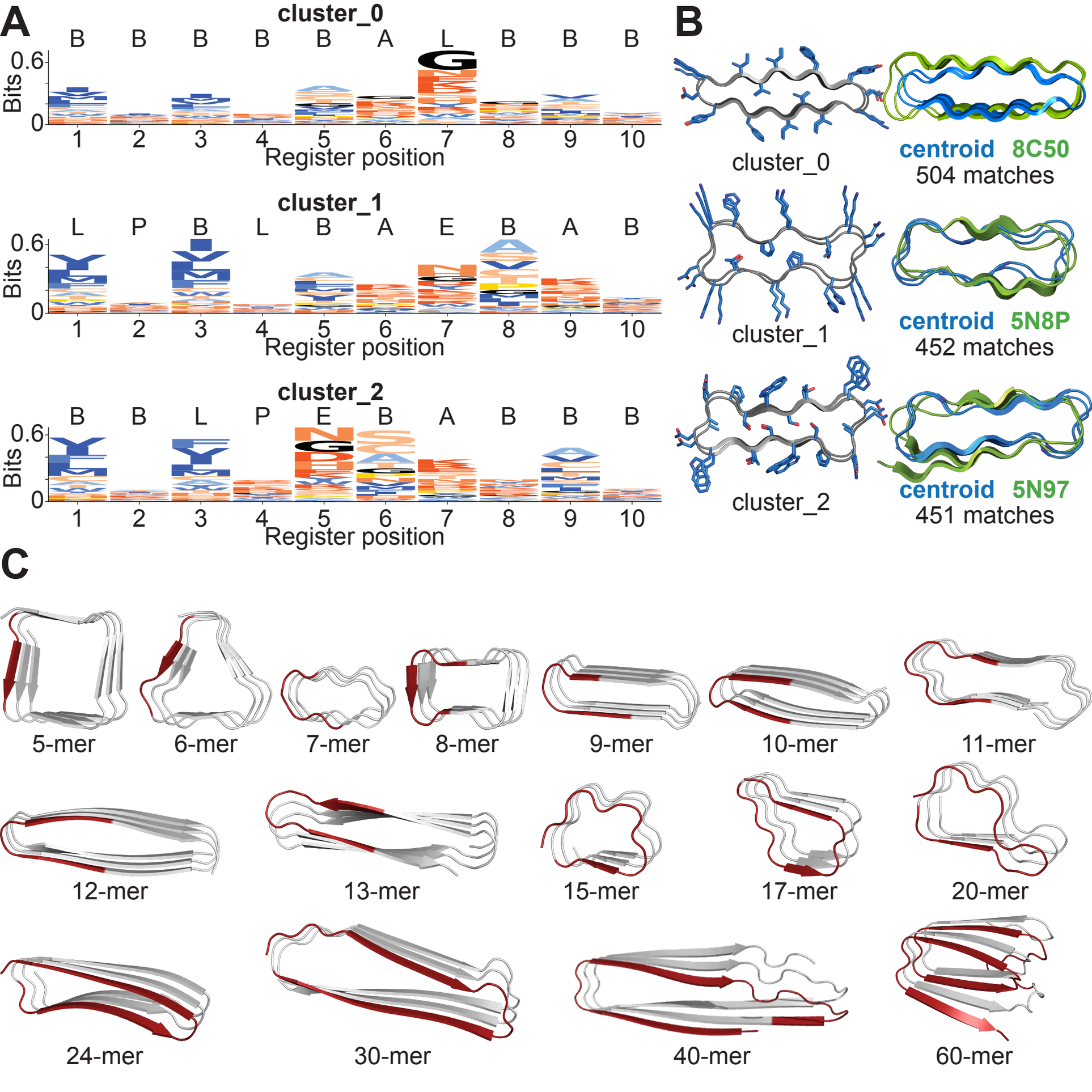
Figure 8. Effect of INDELs on the monomolecular folding of sequences with blocks of four, 7-residue repeats. (A) The percentage of sequences that folded into helical bundles as a function of the indel type. (B) Frequency and structure of each cluster. (C) Sequence logos for the largest cluster with 2-Gly insertions. Clusters 0, 1, 3 and 7 are left- and right-turning 4-helix bundles, while the remaining structures contain helical hairpin units. Hydrophobic residues at positions 1 and 4 project tend to project towards the interior of each 4-helix bundle.

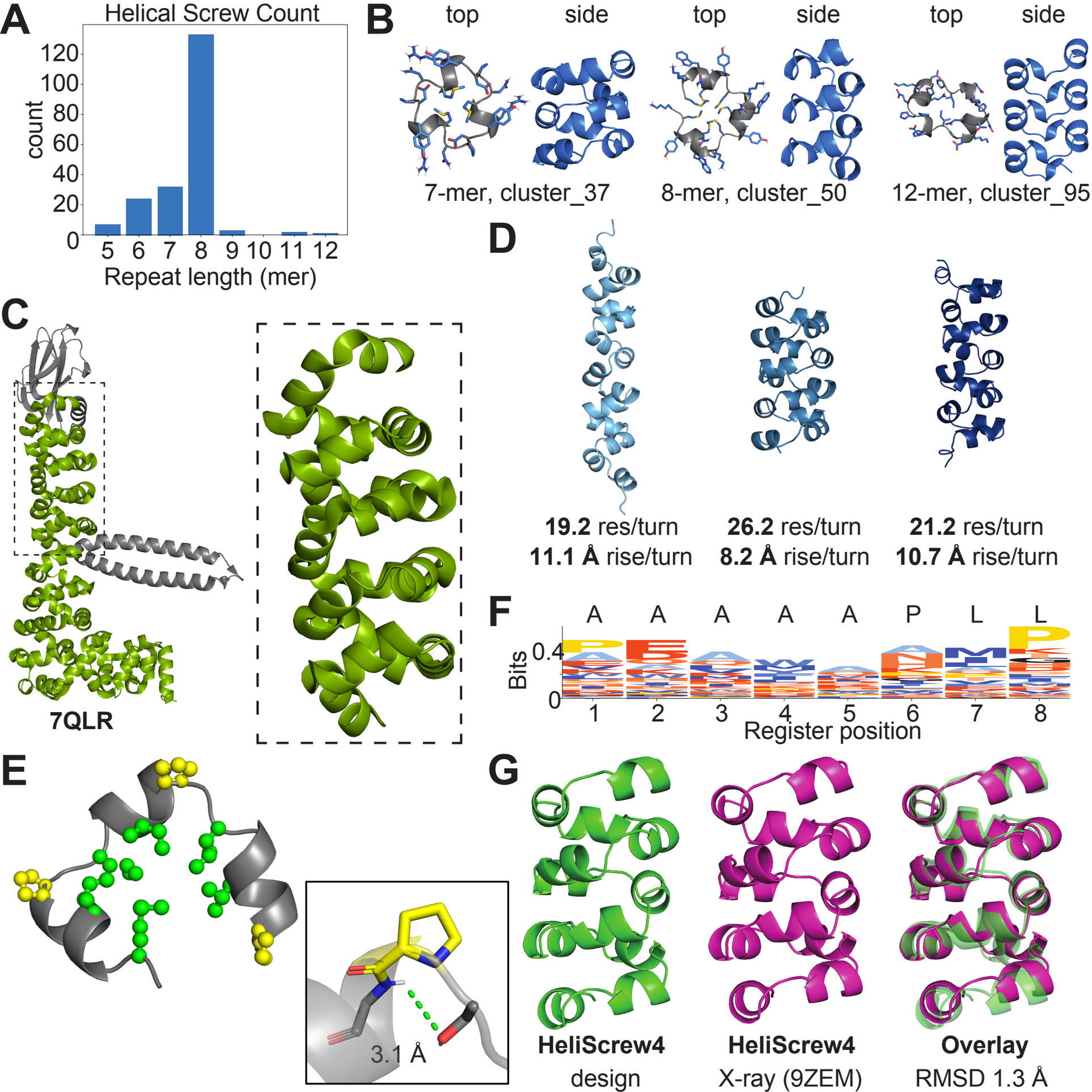

A
5-mer
total samples: 324,855
B
8-mer
total samples: 649,167
C
10-mer
total samples: 684,353
D
24-mer
total samples: 1,014,226
E
40-mer
total samples: 904,456
F
120-mer
total samples: 1,572,640
Percentage (%)
Average pLDDT
G
Percent Folded
log(% Folded)
H
Designability
log(#seq_ave)
I
Clusters for 99% coverage
log(# of Clusters)
Repeat Length (mers)

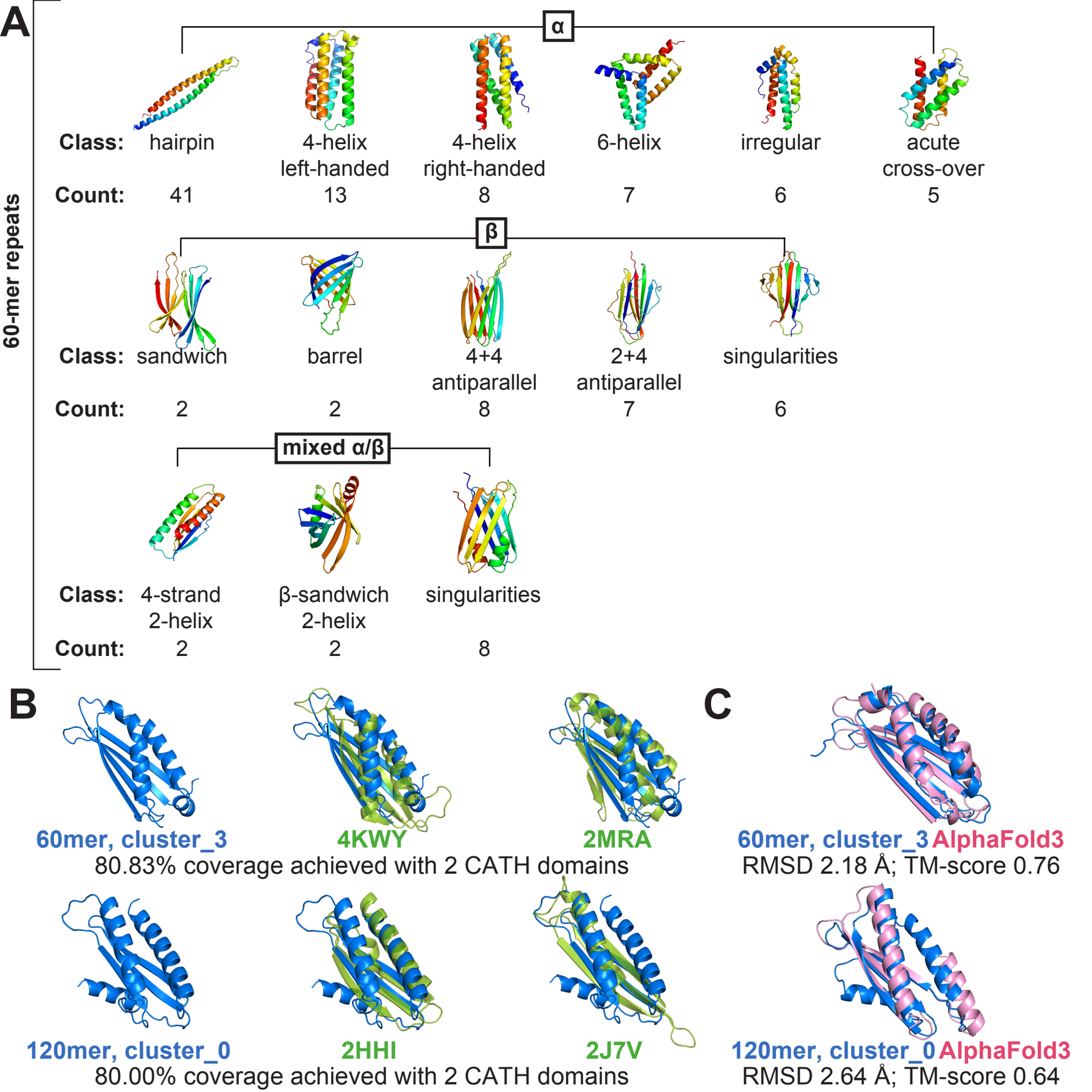

A
Count
16000
12000
8000
4000
0
cluster_0
cluster_1
cluster_2
cluster_3
cluster_4
B
cluster_0
C
4YC5
444 matches
D
cluster_1
E
3DU1
452 matches
F
Cluster_0
Psi (°)
Phi (°)
150
0
-150
G
Cluster_1
Psi (°)
Phi (°)
150
0
-150
H
cluster_0
B B B B E
Bits
1
0
1 2 3 4 5
Register position
I
cluster_1
B B B A B
Bits
1
0
1 2 3 4 5
Register position

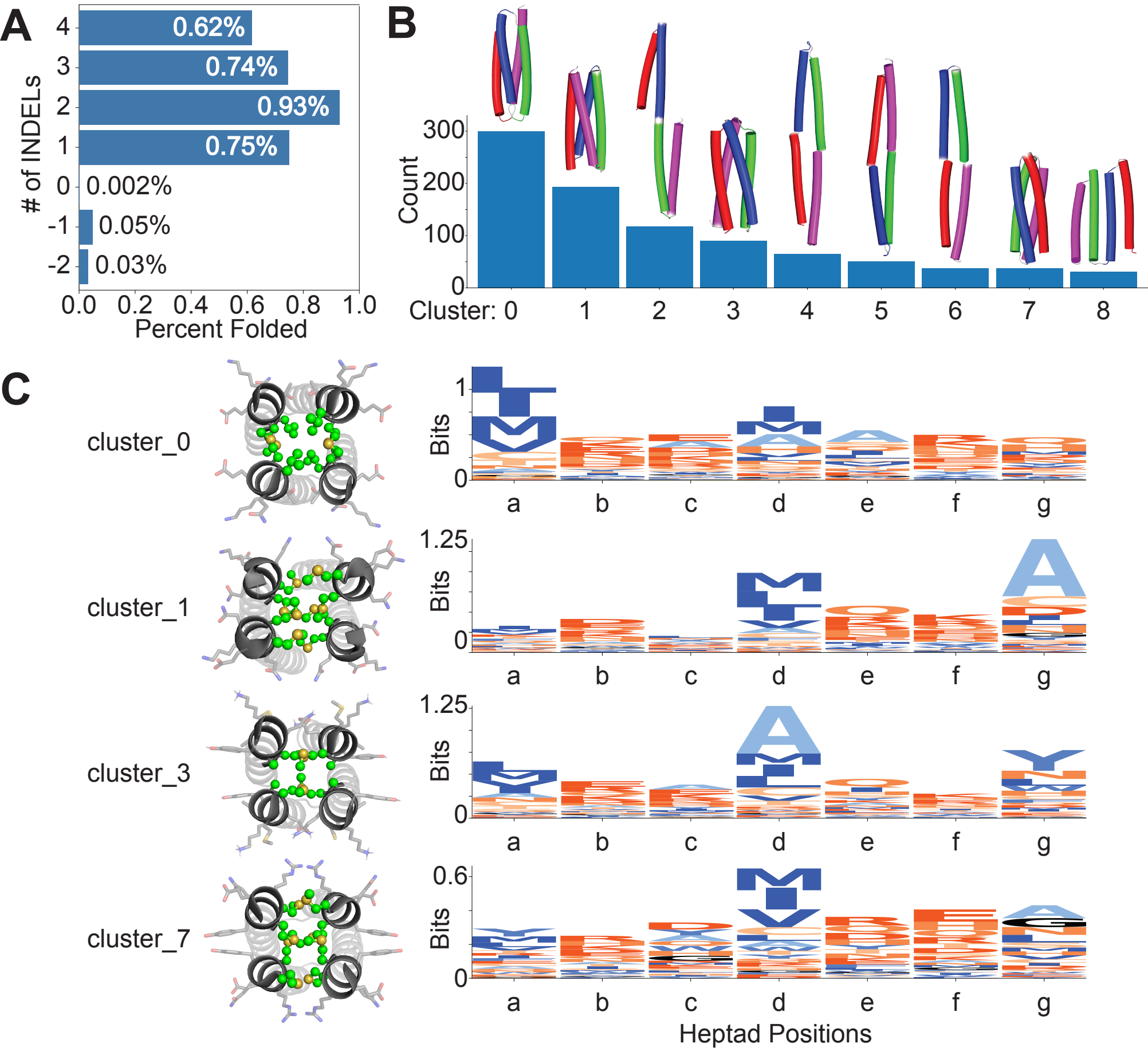

A
Count
20000
15000
10000
5000
0
Cluster: 0
1
2
3
4
5
6
7
8
9
B
C
D
E
cluster_0
6N2C
435 matches
cluster_1
7K47
75 matches
F
cluster_0
G
cluster_1
150
0
-150
Psi (°)
Phi (°)
H
cluster_0
I
cluster_1
B B B B P E
B B B A L B
Bits
1.5
0
Register position


A
cluster_0
A B B B E B B B
Bits
1
0
1 2 3 4 5 6 7 8
Register position
cluster_1
P B B B E E L B
Bits
0.6
0
1 2 3 4 5 6 7 8
Register position
cluster_2
B B B B E B A B
Bits
1
0
1 2 3 4 5 6 7 8
Register position
B
cluster_0
centroid 3PR7
78 matches
cluster_1
centroid 7KO9
387 matches
cluster_2
centroid 3ULT
129 matches
C
9-mer, cluster_13
2ML1
AlphaFold3
RMSD = 1.39 Å
10-mer, cluster_1
6SUS
AlphaFold3
RMSD = 2.61 Å


A
cluster_0
B B B B B A L B B B
Bits
0.6
0
1 2 3 4 5 6 7 8 9 10
Register position
cluster_1
L P B L B A E B A B
Bits
0.6
0
1 2 3 4 5 6 7 8 9 10
Register position
cluster_2
B B L P E B A B B B
Bits
0.6
0
1 2 3 4 5 6 7 8 9 10
Register position
B
cluster_0
centroid 8C50
504 matches
cluster_1
centroid 5N8P
452 matches
cluster_2
centroid 5N97
451 matches
C
5-mer
6-mer
7-mer
8-mer
9-mer
10-mer
11-mer
12-mer
13-mer
15-mer
17-mer
20-mer
24-mer
30-mer
40-mer
60-mer


A
Helical Screw Count
count
0
20
40
60
80
100
120
5 6 7 8 9 10 11 12
Repeat length (mer)
B
top
side
top
side
top
side
7-mer, cluster_37
8-mer, cluster_50
12-mer, cluster_95
C
7QLR
D
19.2 res/turn
11.1 Å rise/turn
26.2 res/turn
8.2 Å rise/turn
21.2 res/turn
10.7 Å rise/turn
F
A A A A A P L L
Bits
0.4
0
1 2 3 4 5 6 7 8
Register position
E
3.1 Å
G
HeliScrew4
design
HeliScrew4
X-ray (9ZEM)
Overlay
RMSD 1.3 Å


A
60-mer repeats
α
Class: hairpin
Count: 41
4-helix left-handed
13
4-helix right-handed
8
6-helix
7
irregular
6
acute cross-over
5
β
Class: sandwich
Count: 2
barrel
2
4+4 antiparallel
8
2+4 antiparallel
7
singularities
6
mixed α/β
Class: 4-strand 2-helix
Count: 2
β-sandwich 2-helix
2
singularities
8
B
60mer, cluster_3
4KWY
2MRA
80.83% coverage achieved with 2 CATH domains
120mer, cluster_0
2HHI
2J7V
80.00% coverage achieved with 2 CATH domains
C
60mer, cluster_3 AlphaFold3
RMSD 2.18 Å; TM-score 0.76
120mer, cluster_0 AlphaFold3
RMSD 2.64 Å; TM-score 0.64


A
# of INDELs
4
3
2
1
0
-1
-2
0.62%
0.74%
0.93%
0.75%
0.002%
0.05%
0.03%
0.0
0.2
0.4
0.6
0.8
1.0
Percent Folded
B
Count
300
200
100
0
Cluster: 0
1
2
3
4
5
6
7
8
C
cluster_0
cluster_1
cluster_3
cluster_7
Bits
1
1.25
0.6
0
a
b
c
d
e
f
g
Heptad Positions