

# **Modular repeat protein sculpting using rigid helical junctions**

## **Authors**

TJ Brunette<sup>1,2</sup>, Matthew J Bick<sup>1,2</sup>, Jesse M. Hansen<sup>1,3</sup>, Cameron M. Chow<sup>2</sup>, Justin M. Kollman<sup>1</sup>, and David Baker<sup>1,2,4</sup>

## **Affiliation**

1 Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

2 Institute for Protein Design, University of Washington, Seattle, WA 98195, USA

3 Graduate Program in Biological Physics, Structure, and Design (BPSD), University of Washington, Seattle, WA, USA

4 Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

## Contents

**Figure S1 | Computational protocol**

**Discussion S1 | Computational protocol**

**Figure S2 | Machine learning forward folding**

**Discussion S2 | Machine learning forward folding**

**Figure S3 | Rosetta fragment assembly sampling improvements**

**Discussion S3 | Crystal structure determination analysis**

**Table S1 | Crystallographic data collection and refinement statistics**

**Figure S4 | Structural validation by SAXS**

**Table S2 | Summary of SAXS data**

**Discussion S4 | Small Angle X-ray Scattering (SAXS) analysis**

**Figure S5 | Filtering of junction library**

**Discussion S5 | Filtering and coverage of junction library**

**Figure S6 | Joinability of DHR**

**Figure S7 | Connections between junctions**

**Table S4 | Summary of sculpt data**

**Discussion S6 | Protein sculpt analysis**

**Figure S8 | Ankyrin junction EM image**

**Table S5 | Sequences of junctions and experimental data link**

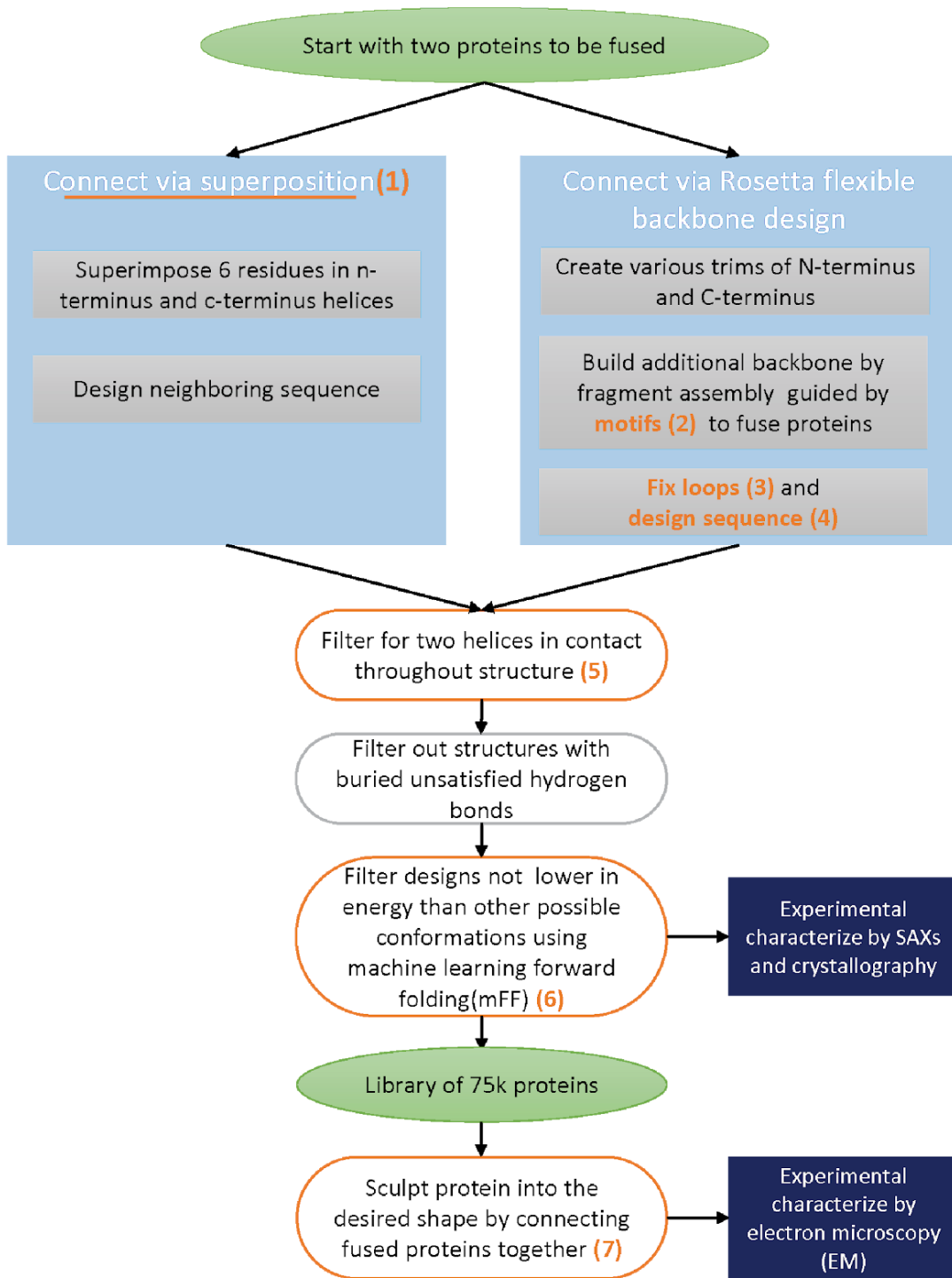
**Table S6 | Sequences of sculpts and experimental data link**

**Discussion S9 | Methods for expression, crystallization, SAXs and negative stain electron microscopy**

**Discussion S10 | Acknowledgements**

**Supplementary References**

**Figure S1 | Computational protocol**



Flowchart of the design protocol. The green boxes indicate input or output structures, light blue boxes show major components of the method and navy is for experimental techniques used to verify the designs. Orange is used to highlight new or improved algorithms developed for this paper.

## Discussion S1 | Computational protocol

### Overview

We developed two methods to rigidly fuse proteins together and used them to connect 44 designed helical repeat proteins (DHRs) into a building block library of 75k junctions, each with a unique shape. Proteins from this library were then used to sculpt larger, nanometer length proteins. To generate a library at this scale, we needed to significantly improve the speed and efficiency of our design software, Rosetta, and automate the design process. The new algorithms responsible for these speed improvements are implemented inside Rosetta and the code for each step is provided in Supplementary Materials `rosetta_examples` and is compatible with Rosetta post git version 74cba0b67de. The examples can be downloaded from here:

[http://files.ipd.uw.edu/brunette/helix\\_fusion\\_files/protein\\_fusion\\_scripts.tar.gz](http://files.ipd.uw.edu/brunette/helix_fusion_files/protein_fusion_scripts.tar.gz)

### A1. The superposition algorithm

In our approach to fuse two DHRs along a shared helix, six-residue helical segments from a first DHR were superimposed onto six-residue helical segments from a second DHR. A single repeat from each DHR was scanned. For overlaps less than 0.3Å RMSD that did not clash, the sequence was redesigned for positions within 6Å of the new DHR-DHR interface. Repack of side chains occurred for residues within 8Å. Residues on the terminal DHR repeat were not redesigned. During design, surface residues were restricted to hydrophilic and core residues to hydrophobic by a Rosetta layer design task operator. After design, the structures were filtered according to step B.

### A2. The Rosetta fragment assembly algorithm

A second way to make a rigid connection was to create additional residues between the two proteins using Rosetta fragment assembly. This proceeded in six steps:

#### **1. Create various DHR trims**

To explore a wide spectrum of possible junction geometries, the terminal helices were trimmed by one to four residues, which is enough to span one turn of an alpha-helix. For DHR

combinations that were unable to be joined due to the filters applied in step B, additional interface geometries were explored such as trimming one helix out of the two-helix repeat; to keep these additional geometries compatible with the building block library, two terminal repeats were maintained.

## **2. Backbone design using Rosetta fragment assembly guided by motifs**

For each DHR pair, additional amino acid residues were added using Rosetta fragment assembly between the two domains consisting of either a loop, a helix (with two loops), or two helices (with three loops). The lengths of the helices ranged from one less than the shortest helices of the DHRs being joined to one residue longer than the longest residue, and loops ranged from two to four residues. For structures with two helices, the helix length was restricted to be within one residue of the lengths of the DHR helices. All secondary structure possibilities consistent with these rules were exhaustively generated. Backbone coordinates were built up through 3,200 Monte Carlo fragment assembly steps with fragments harvested from a non-redundant set of structures from the PDB (1) starting from a structure with ideal helices and extended loops. Following each fragment insertion, the rigid body transform was propagated to the downstream repeat protein domain and the backbone in the flanking terminal repeat of the DHRs were kept rigid. The score that guided fragment assembly considers Van der Waal interactions, packing, backbone dihedrals angles and, for the first time, Residue-Pair-Transform(RPX) motifs (2). RPX motifs indicate when a portion of the backbone will pack together with hydrophobic residues in full-atom prior to assigning side chains (centroid representation). In this way, RPX motifs increase the accuracy of the centroid energy function.

## **3. Filter backbones to reduce flexibility**

To reduce flexibility across the junction, we require that at least two helices from each DHR and/or junction make contact across the new interface. We found that if a helix interacts with three or fewer other helices that structure had flexible point made up of a single helix. To determine which helices were in contact the Residue Pair Motifs (RPX) (2) was used. Structures

with three helices in contact at the centroid stage can become four helices during the subsequent full-atom relax; as such, structures with  $< 3$  helices in contact were filtered.

#### **4. Filter backbones with structural features dissimilar to those in solved protein structures**

The validation step most likely to reject a design is Rosetta *ab initio* structure prediction. Since sequence design and filtering are computationally expensive steps, it is important to quickly triage structures that would fail *ab initio*. Designs are more likely to fail structure prediction when parts of the design do not resemble natural proteins. To explore the foldability of designs, nine residue fragments from the design were compared to all nine-residue fragments in the PDB. Proteins were more likely to pass Rosetta *ab initio* if the loops are within 0.4Å RMSD and helices are within 0.14Å RMSD to a structure in the PDB. A helix that is above 0.14Å relative to all helices in the PDB appeared bent or kinked. All structures analyzed were helical with short (2-4 residue) loops so different values may be required when applying this filter to proteins with longer loops or sheets.

The algorithm to identify the most similar fragment took approximately one second to search through the four million fragments in the VALL PDB database (3). To achieve this speed, only fragments with the same secondary structure were compared, and RMSD was calculated using the Quaternion Characteristic Polynomial method (QCP kernel) (4, 5).

#### **5. Fix loops so they are structurally similar to those in the PDB**

A loop dissimilar to all loops in the PDB can often be repaired by swapping the designed loop with one from the PDB that better superimposes onto the end points of helices being bridged. To identify the loop that best matches onto the helix endpoints the two helical residues on either side of all short loops from the VALL pdb database were superimposed onto two stub residues at the end of the bridged helices. The four residue match with the lowest RMSD was considered the best match. To address small deviations in the overlapped residues the loop backbone was minimized after being placed by superposition. To explore a wide possibility of helical end point geometries the helices were extended and shrunk by three residues. The final loop RMSD was

measured using the algorithm from step 4. Structures with loops  $>0.4\text{\AA}$  RMSD after fixing were filtered.

## **6. Sequence design**

Rosetta design was used to design the amino acid sequence of residues in the junction and residues in the repeat that neighbors the junction. Surface residues were restricted to hydrophilic and core residues to hydrophobic by a Rosetta layer design task operator. Sequence was further optimized to satisfy buried hydrogen bonds, match secondary structure predicted from sequence (psipred), and bias the sequence toward protein fragments with similar structure. The unsatisfied hydrogen bonds (6) and PSIPRED (7) sequence match were optimized using the generic simulated annealing mover in Rosetta which applies a Monte Carlo search over sequence design.

Sequence composition was biased toward native protein fragments with similar local structure using a structure profile. The structural profile used the fragment lookback approach described in step 4 to identify the most structurally similar nine residue fragments where the RMSD to the design was lower than  $0.4\text{\AA}$ . Previously, structure profile generation would take 10-20 minutes and require a script outside of Rosetta (8). Using the fragment lookback approach the structural profile now takes seconds to build.

### ***B. Filter***

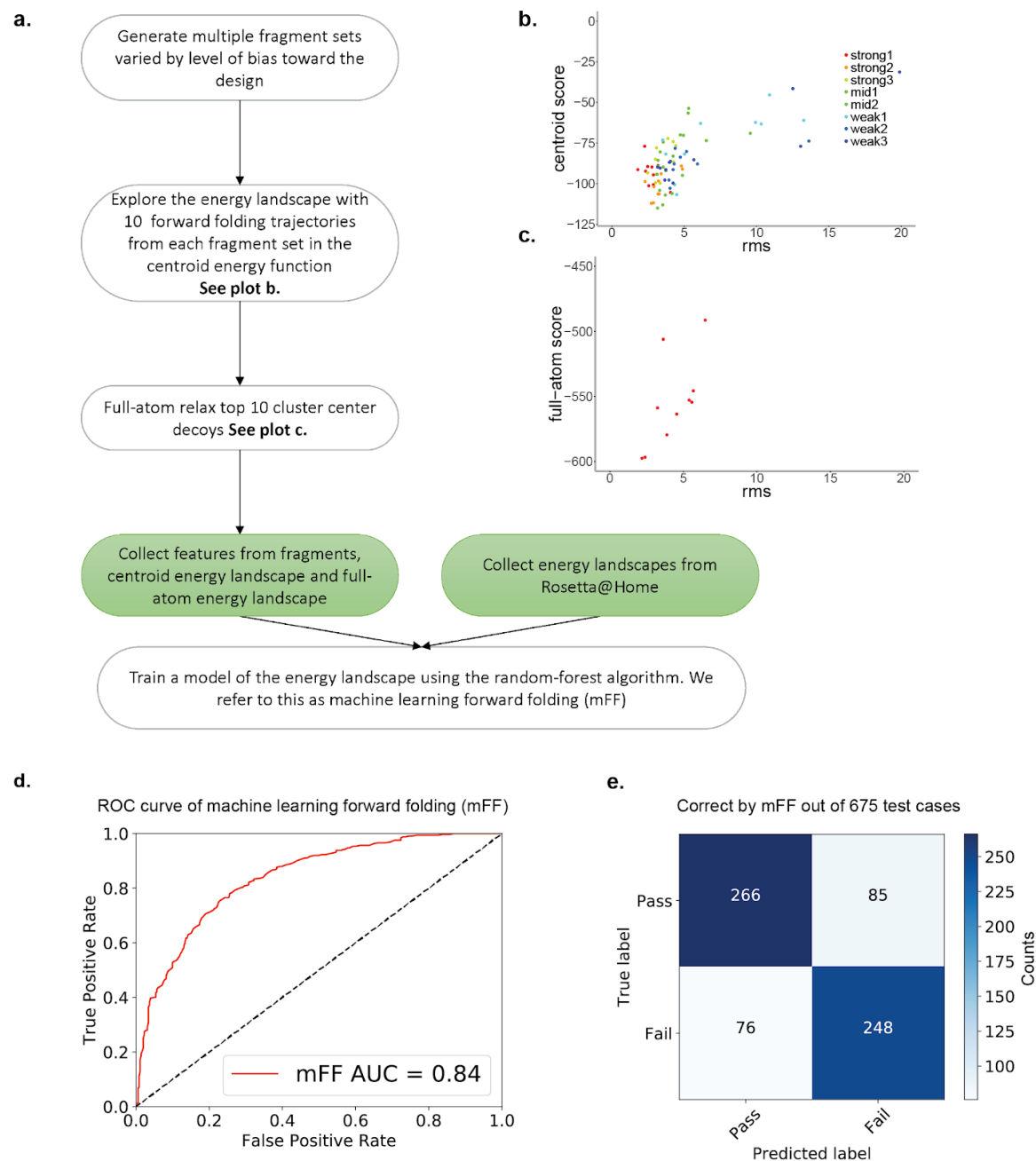
The junction library generated in the previous steps was filtered to ensure all proteins were of high quality and can be used to sculpt larger proteins. The proteins were filtered for uniqueness to  $1.0\text{\AA}$  RMSD, lack of unsatisfied hydrogen bonds, a large and broad hydrophobic interface across multiple helices, and to have the lowest energy compared to other potential folds as measured by Rosetta *ab initio*. Most of these filter steps can be run on millions of proteins, but evaluating if the designed protein was in a lower energy state than alternative conformations can take several days on hundreds of CPUs using Rosetta *ab initio*. To speed up Rosetta *ab initio*, machine learning was used to simulate *ab initio* on a single CPU in 3-4 hours with high accuracy. The Rosetta *ab initio* step is described in more detail in SI Appendix Discussion 2.

### **C. Sculpt**

For protein sculpting, possible junction combinations containing one or two junctions were enumerated. The junction combinations were stored in a blueprint file that contains the information necessary for Rosetta to build protein sculpts. Due to the huge number of possible junction combinations, only a small and random subset of the possibilities were made. Ordering was done by visual inspection and designs that clash were discarded. For symmetric designs, symmetry was applied after the monomer construction.

Large proteins composed of numerous repetitive amino acid stretches require genes that are difficult to synthesize. To alleviate this problem the surface residues of all helices not part of the symmetric interface were redesigned using Rosetta.

## Figure S2 | Machine learning forward folding



**a.** Flowchart of the machine learning forward folding algorithm (mFF). 2250 Rosetta@Home simulations were used to train the model with 70% used for training and 30% set aside for testing. The Rosetta@Home simulations took two–three weeks to generate sufficient samples for training while each run of mFF took three–four hours on a single core. **b.** Exploration of the energy landscape by the different fragment sets in centroid. Fragment sets strongly biased toward the design focus exploration of the energy landscape on the region closest to the design, while weakly biased sets explore more broadly. **c.** To speed the algorithm, only a subset of centroid models are relaxed in the full-atom energy function. The decoys chosen for relaxing are the low energy cluster centers **d.** the roc curve, and **e.** the confusion matrix illustrates the accuracy of mFF as compared to the Rosetta@Home simulation.



## Discussion S2 | Machine learning forward folding (mFF)

In *ab initio* structure prediction (also called forward folding), the energy landscape is explored using short simulations starting from an initial extended structure (decoy). In each step of the simulation, a 9 or 3 residue fragment from a solved protein structure is swapped into the decoy and accepted using the Metropolis Monte Carlo criteria. Each simulation results in a decoy with an energy and distance from the design measured as root mean square deviation (RMSD). The design is validated if the distribution of decoys produces a funnel to the low energy and low RMSD designs. Thousands of decoys are required to suggest a design is lower in energy than alternative minima. To generate those decoys, Rosetta@Home (9) is used to distribute the job to hundreds of users. A Rosetta@Home *ab initio* simulation can take several days, with a max throughput of 500-1000 simulations per week.

*Ab initio* validation contains more information than *ab initio* structure prediction, because structural prediction lacks the structural design data. Using information from the design can be used to bias exploration toward the design or not used so exploration broadly explores the entire energy landscape. To control this bias, 8 fragment sets were created that are subsets of the 200 fragments normally used in Rosetta *ab initio*. The 8 fragment sets used are listed with decreasing bias: top 3 by RMSD to design, top 15 by RMSD, from the first 25 fragments select the top 3, the top 3 plus a random 10 from 200, top 15 plus random 10, top 3 plus random 15, top 3 plus random 25, from the first 25 select a random 15. The top 200 fragments are ranked during fragment picking so fragments in the top 25 are more likely to be correct.

Using these 8 fragment sets ranging from strongly to weakly biased 10 centroid *ab initio* simulations were run. These 80 decoys were clustered and the low-energy cluster center is relaxed into the Rosetta full-atom energy function. It has been previously established that compute time can be saved by running full-atom Rosetta only on cluster centers (10).

Each of these eight centroids and one full atom simulations produces features that indicate if a protein would pass Rosetta *ab initio* structure prediction. These features are used to train a random forest that can predict if the protein design would pass *ab initio* structure prediction. The features used are the lowest rms structure, the score range between structures, the standard deviation in RMSD between structures and average RMSD to the design. Additional features are extracted from the fragment sets including the percentage of fragments lower than 0.5, 1 and 1.5 Å RMSD and the average fragment quality for the top 3 and top 15 fragments sets.

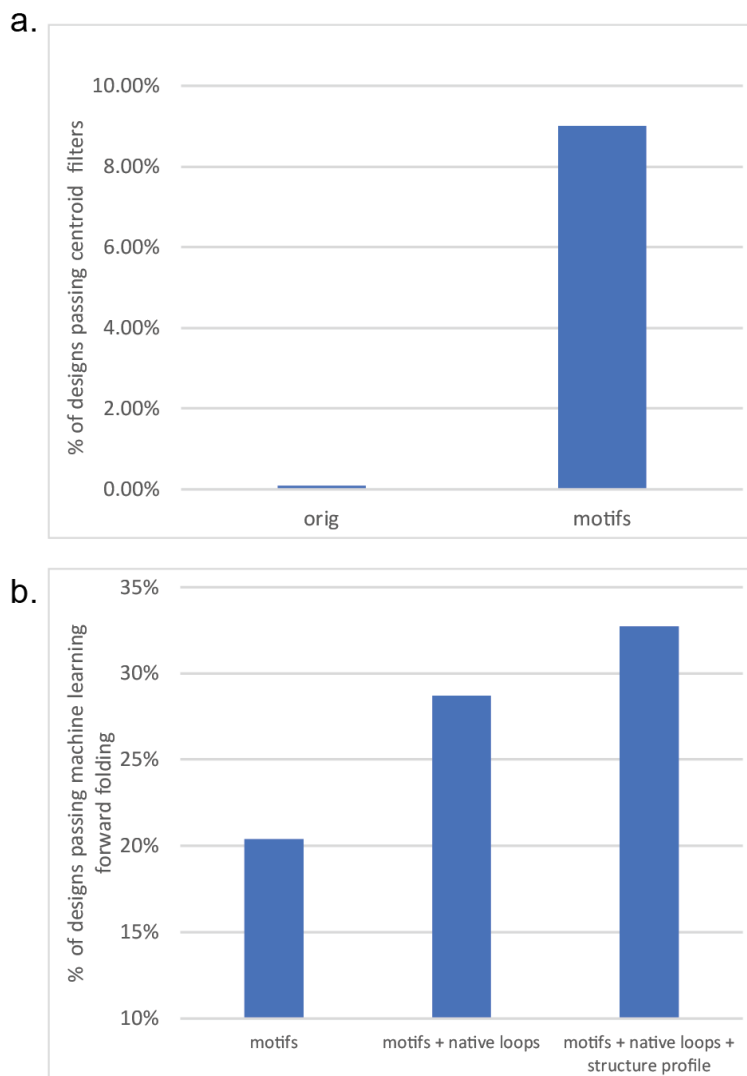
To train the model we collected 2250 *ab initio* simulations on Rosetta@Home split evenly between cases that pass *ab initio* and those that did not. The simulations were labeled as passing

*ab initio* if the `ff_metric` value is  $<25$ . `FF_metric` is an algorithm that uses the sum of RMSD in the lowest energy points to evaluate the funnel (11).

30% of the Rosetta@home simulations were set aside for testing and 70% used to train the model. The resulting random forest model had an AUC of 0.84 with error split between false positives, and false negatives. The top three features in the model are the low RMSD structure generated from the top 3, top 15 and top 3 plus 25 fragment sets.

Machine learning forward folding (mFF) takes about 3-4 hours on a single core as compared to several days on hundreds of user computers. This dramatic speed improvement allows us to simulate thousands of de novo protein designs when previously we could only simulate hundreds. It also allows us to screen designs before submitting to Rosetta@Home.

**Figure S3 | Rosetta fragment assembly sampling improvements**



To evaluate Rosetta flexible backbone sampling improvements we designed approximately 2000 Denovo Helical Repeat(DHR) proteins with the sampling strategies described in this paper. *Orig* is the method from (8). *RPX Motifs* is a centroid score term that indicates when the backbone packs together with hydrophobic residues. *Native loops* replace fragment sampled loops with their closest natural loop. *Structure profile* biases the sequence design toward sequences of naturally occurring proteins. **a.** RPX motifs made a 116x improvement in sampling efficiency. Only 0.08% of designs made with the original method pass the centroid filtering while 9% pass with RPX motifs. **b.** After centroid sampling, full-atom design occurs. Designs are evaluated by what percent pass machine learning forward folding. We see a 1.6x improvement between the original Rosetta design procedure (motifs) and those designs generated with native loops and the structure profile.

### Discussion S3 | Crystal structure determination analysis

Junction 19 is between DHR54 and DHR79 and had an RMSD of 1.14Å to the crystal structure. The main deviation between the design and crystal is observed in the c-terminal helix, the likely result of a crystal-packing artifact. The n-terminal repeat and the core rotamers are in their designed positions.

Junction 23 is between DHR14 and DHR18 and had an RMSD of 1.58Å to chain A of the crystal structure. We observed a slight deviation in the n-terminal repeat structure relative to the design. It appears that the n-terminal repeat twist does not occur in the junction itself but in the second repeat past the junction. There is a second chain resolved in the crystal structure, with an RMSD deviation of 1.5Å relative to the design. The N-terminal helix is not resolved in the structure and is presumed to be disordered.

Junction 24 is between DHR14 and DHR18 had an RMSD of 0.93 Å relative to the crystal structure. A 5-residue stretch in the c-terminal portion of the protein is disordered. Disorder of the c-terminal helix previously occurred to several of the DHR proteins (8).

The design of junction 34 between DH53 and DHR4 had an RMSD of 1.51Å to the crystal structure. There appears to be a slight twist in the junction.

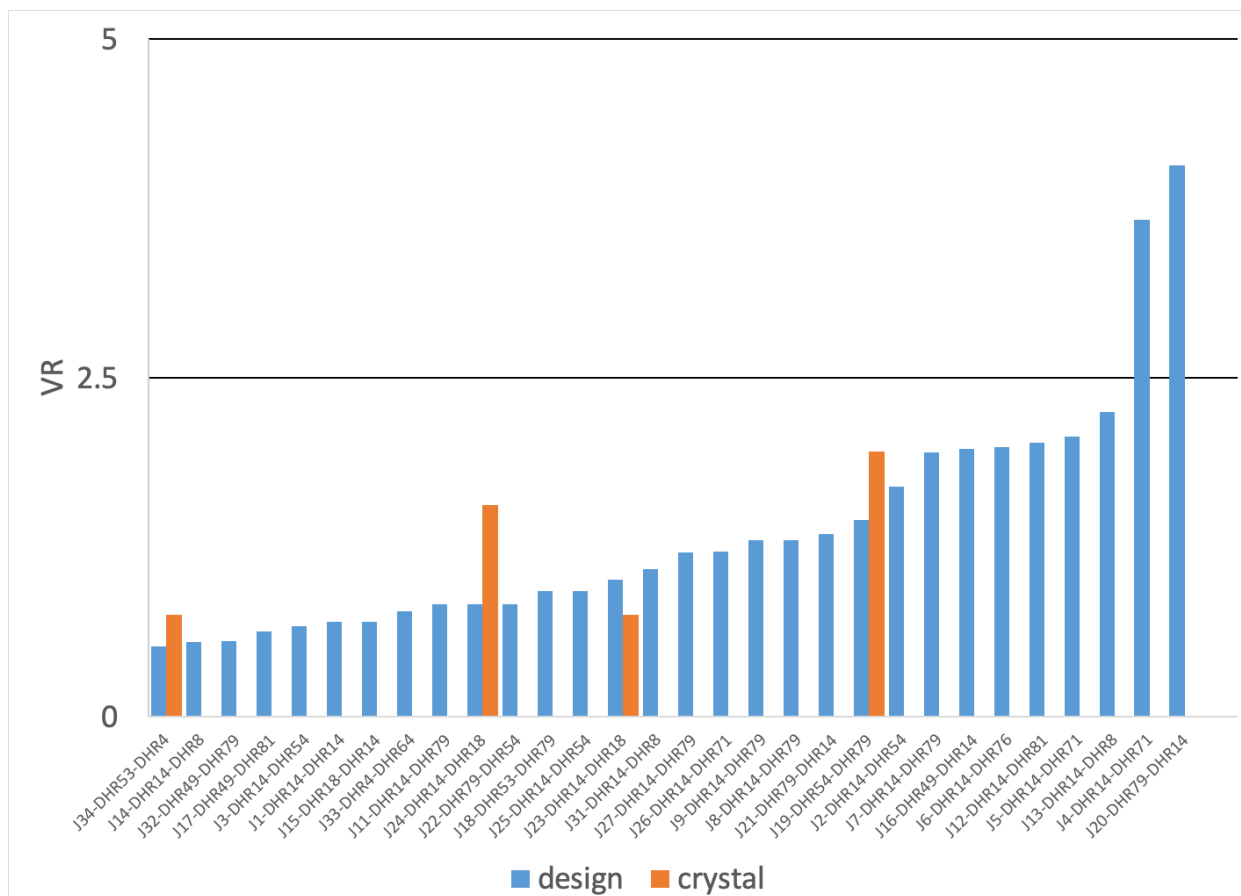
**Table S1: Crystallographic data collection and refinement statistics**

	<b>Junction 19</b> DHR54-DHR79	<b>Junction 23</b> DHR14-DHR18	<b>Junction 24</b> DHR14-DHR18	<b>Junction 34</b> DHR53-DHR4
<b>Protein Data Bank accession numbers</b>	6W2R	6W2V	6W2W	6W2Q
<b>Wavelength</b>	0.9999	0.9791	1	1
<b>Resolution range</b>	45.63 - 2.35 (2.43 - 2.35)	33.89 - 2.40 (2.49 - 2.40)	43.71 - 2.21 (2.29 - 2.21)	37.46 - 1.8 (1.86 - 1.80)
<b>Space group</b>	P 1 21 1	P 1 21 1	P 21 21 21	P 21 21 21
<b>Unit cell</b>	53.7 109.7 81.0 90 107.5 90	62.0 41.1 94.0 90 104.9 90	49.1 49.7 92.0 90 90 90	43.9 57.6 71.7 90 90 90
<b>Total reflections</b>	219869 (22021)	198525 (8047)	69268 (6908)	117427 (11054)
<b>Unique reflections</b>	37152 (3256)	17440 (1307)	11787 (1147)	17375 (1692)
<b>Multiplicity</b>	5.9 (6.0)	11.4 (6.0)	5.9 (6.0)	6.8 (6.5)
<b>Completeness (%)</b>	87.01 (63.89)	88.97 (62.88)	95.32 (85.93)	94.30 (80.32)
<b>Mean I/sigma(I)</b>	15.26 (1.32)	11.23 (2.39)	18.59 (2.05)	13.06 (1.06)
<b>Wilson B-factor</b>	48.6	69.6	43.6	30.8
<b>R-merge</b>	0.08 (1.76)	0.11 (0.57)	0.06 (0.94)	0.07 (1.31)
<b>R-meas</b>	0.09 (1.93)	0.13 (0.61)	0.07 (1.03)	0.07 (1.42)
<b>R-pim</b>	0.0367 (0.78)	0.0332 (0.22)	0.028 (0.42)	0.028 (0.55)
<b>CC1/2</b>	1 (0.57)	0.997 (0.87)	0.999 (0.7)	0.999 (0.54)
<b>CC*</b>	1 (0.85)	0.999 (0.97)	1 (0.91)	1 (0.84)
<b>Reflections used in refinement</b>	32655 (2385)	16292 (1113)	11244 (989)	16462 (1371)
<b>Reflections used for R-free</b>	1794 (133)	1643 (99)	1110 (101)	1648 (131)
<b>R-work</b>	0.24 (0.38)	0.24 (0.33)	0.23 (0.32)	0.20 (0.30)
<b>R-free</b>	0.27 (0.39)	0.27 (0.36)	0.25 (0.35)	0.23 (0.33)
<b>CC(work)</b>	0.97 (0.71)	0.95 (0.80)	0.97 (0.75)	0.96 (0.74)
<b>CC(free)</b>	0.95 (0.62)	0.95 (0.71)	0.96 (0.66)	0.95 (0.66)
<b>Number of non-hydrogen atoms</b>	5715	2763	1537	1533
<b>macromolecules</b>	5701	2762	1517	1435

<b>ligands</b>				1
<b>solvent</b>	14	1	20	97
<b>Protein residues</b>	863	435	224	196
<b>RMS(bonds)</b>	0.004	0.005	0.004	0.003
<b>RMS(angles)</b>	0.82	0.87	0.87	0.50
<b>Ramachandran favored (%)</b>	99.2	98.4	99.1	100.0
<b>Ramachandran allowed (%)</b>	0.82	1.63	0.90	0.00
<b>Ramachandran outliers (%)</b>	0.00	0.00	0.00	0.00
<b>Rotamer outliers (%)</b>	0.24	2.30	0.84	0.00
<b>Clashscore</b>	1.47	0.39	1.72	1.05
<b>Average B-factor</b>	74.93	95.72	66.01	37.72
<b>macromolecules</b>	74.97	95.73	66.22	37.29
<b>ligands</b>				53.79
<b>solvent</b>	55.57	75.15	50.62	43.91
<b>Number of TLS groups</b>	1	2	7	1

Statistics for the highest-resolution shell are shown in parentheses.

Figure S4 | Structural validation by SAXS



Vr values for the fit of SAXS profiles to design models, in blue, and crystal structures, in orange. The Vr cutoff value of 2.5 was calibrated using designs confirmed by crystallography in our previous paper on DHRs which have similar size and aspect ratio as junctions (8). 28 of 30 designs were validated.

**Table S2: Summary of SAXS analysis**

Junc # DHR <sub>1</sub> - DHR <sub>2</sub>	protein length	Porod	q range		Guinier				Real space p(r)			Model p(r)		Model Fit		Crystal p(r)		Crystal Fit		accepted	
			q min	q end	start range	end range	Rg	l0	dmax	Rg	l0	dmax	Rg	Vr 0.015<q q<0.25	chi	dmax	Rg	Vr 0.015<q q<0.25	chi		
1-14-14	199	4	0.01366	0.27898	0.01366	0.06662	19.4	843	61	19.7	808	65	17.6	0.70	1.18					Yes	
2-14-54	181	3.8	0.01254	0.42885	0.01254	0.05491	23.7	219	76	22.75	196	63	18.1	1.70	3.21					Yes	
3-14-54	188	3.9	0.12543	0.34027	0.01254	0.06383	20.4	1000	60	20.9	935	64	18.3	0.67	0.63					Yes	
4-14-71	208	3.9	0.01310	0.35643	0.01310	0.03930	33.3	677	110	34.8	671	79	19.5	3.67	26.2					No	
5-14-71	200	3.9	0.01756	0.31241	0.01756	0.05602	23.3	335	79	23.3	320	68	18.7	2.07	1.22					Yes	
6-14-76	208	4.0	0.01589	0.41270	0.01589	0.05324	24.3	1290	76	23.4	1170	62	18.0	1.99	4.86					Yes	
7-14-79	235	3.9	0.01533	0.41771	0.01533	0.05156	25.4	2140	78	24.0	1870	69	19.0	1.95	3.34					Yes	
8-14-79	237	4.0	0.01533	0.37704	0.01533	0.06216	20.8	850	63	20.9	861	62	18.5	1.30	0.84					Yes	
9-14-79	238	3.8	0.01366	0.40156	0.01366	0.05435	24.1	2270	77	23.4	2100	73	19.6	1.30	5.14					Yes	
11-14-79	235	3.6	0.01477	0.30071	0.01477	0.06327	20.6	291	63	21.8	309	74	20.3	0.83	2.46					Yes	
12-14-81	214	3.9	0.01310	0.40657	0.01310	0.05101	25.7	463	81	26.3	447	70	19.0	2.02	9.34					Yes*	
13-14-8	191	3.8	0.01477	0.37370	0.01477	0.05993	21.9	848	65	21.2	785	66	17.2	2.25	3.80					Yes	
14-14-8	196	4.0	0.01477	0.38484	0.01477	0.06662	19.4	1490	62	19.9	1540	64	17.5	0.55	1.07					Yes	
15-18-14	218	3.8	0.01366	0.37370	0.01366	0.06271	20.8	1680	64	20.9	1690	70	19.2	0.70	1.36					Yes	
16-49-14	179	4.0	0.01923	0.32133	0.01923	0.06327	20.7	8.78	63	19.9	7.91	59	17.0	1.98	1.89					Yes	
17-49-81	221	4.0	0.02202	0.36757	0.02202	0.06439	20.3	238	64	20.3	232	64	18.5	0.63	0.68					Yes	
18-53-79	248	3.9	0.01422	0.34529	0.01422	0.05937	22.1	602	70	22.3	605	77	20.3	0.93	0.51					Yes	
19-54-79	222	3.8	0.01923	0.31798	0.01923	0.06160	21.1	307	62	20.7	282	64	18.6	1.45	0.81	62	16.8	1.96	2.24		Yes
20-79-14	206	3.6	0.01589	0.34584	0.01589	0.04710	27.9	11.2	104	28.0	11.1	62	18.3	4.08	8.34					No	
21-79-14	210	4.0	0.02146	0.42885	0.02146	0.05937	21.9	144	62	21.1	128	64	18.4	1.35	1.33					Yes	
22-79-54	222	3.9	0.01366	0.45169	0.01366	0.06160	21.3	1440	66	20.8	1350	65	18.0	0.83	2.01					Yes	
23-14-18	224	3.9	0.01756	0.35086	0.01756	0.06327	20.6	32.2	67	20.6	31.1	63	18.7	1.01	2.31	65	19.1	0.75	3.35		Yes
24-14-18	230	3.9	0.01756	0.37036	0.01756	0.05602	23.3	17.3	88	23.0	15.8	75	19.4	0.83	1.83	64	18.2	1.56	3.74		Yes
25-14-54	187	4.0	0.01812	0.37872	0.01812	0.06550	19.9	21.7	62	19.8	20.5	66	17.3	0.93	2.04					Yes	
26-14-71	203	4.0	0.01868	0.40657	0.01868	0.05602	23.3	12.4	81	22.9	11.5	74	18.9	1.22	2.15					Yes	
27-14-79	220	4.0	0.01812	0.35086	0.01812	0.05936	21.9	8.2	71	21.5	7.8	62	18.8	1.21	0.90					Yes	
31-14-8	189	4.0	0.01979	0.34529	0.01979	0.07219	18.1	50.7	55	19.2	54.7	65	17.1	1.09	2.64					Yes	
32-49-79	210	3.8	0.01756	0.35643	0.01756	0.06439	20.2	286	66	20.6	279	68	19.2	0.56	0.78					Yes	
33-4-64	235	4.0	0.01812	0.37314	0.01812	0.06104	21.4	248	63	21.1	227	67	18.7	0.78	1.62					Yes	
34-53-4	208	3.9	0.01199	0.37314	0.01199	0.05658	22.8	449	71	21.8	402	75	19.4	0.52	2.27	70	18.5	0.75	6.05		Yes

Vr > 2.5  
 difference in real space Rg or Dmax between model and SAXS data > 30%  
 \* designs with dmax discrepancy > 30% and 2 < Vr < 2.5



#### **Discussion S4 | Small Angle X-ray Scattering (SAXS) analysis**

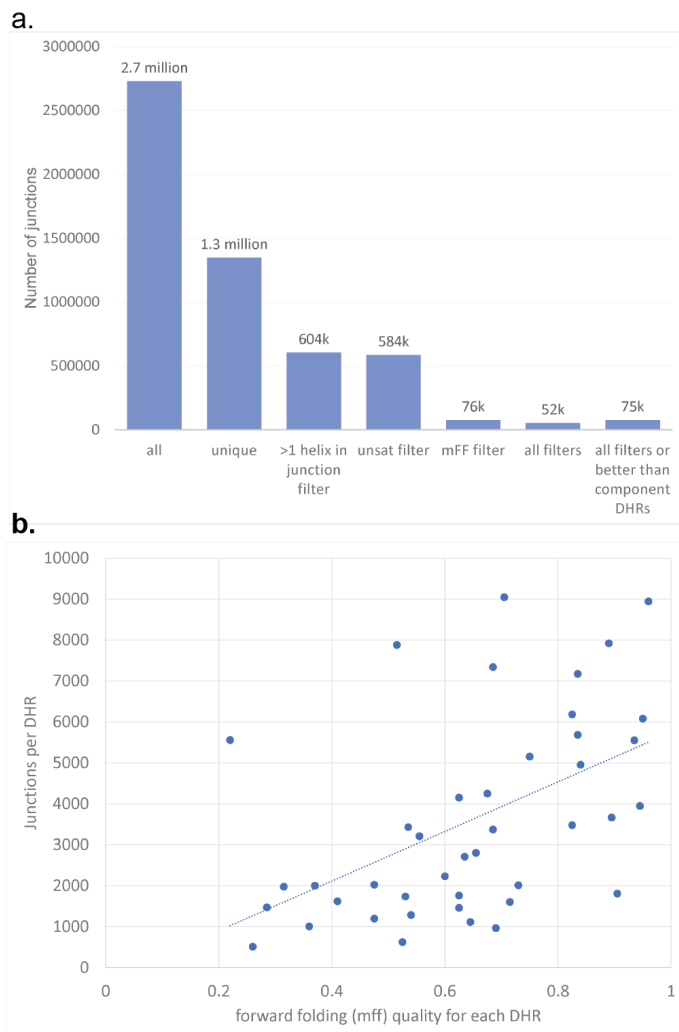
To characterize the structures of proteins we used Small Angle X-Ray Scattering (SAXS) analysis (12–15). with data collected at the SIBYLS beamline (16). The results are summarized in SI appendix Table 2: Summary of SAXS analysis. Data frames were merged using the SAXS Frameslice program. The Porod,  $q$  range, Guinier, realspace  $p(r)$ , model  $p(r)$  and crystal fit were solved using SCATTER 3.0g (14). The model fit measurements of the volatility of ratio ( $V_r$ ) and Chi were calculated using scripts from (15).

The protein designs and crystals were prepared for SAXS by adding missing residues and the n-terminal GWLEHHHHHH purification tag with Rosetta in the same way as in our previous paper (8). The tag was added using Rosetta *ab initio* structure prediction on Rosetta@Home. The lowest energy 100 decoy were then clustered.  $V_r$  and chi were calculated for the top 5 cluster centers and the lowest  $V_r$  was reported. Subsequent analysis within SCATTER was conducted using the design with the tag that produced the lowest  $V_r$ .

Data was collected on the 30 designs that were monomeric in SEC. The 28 designs with  $V_r < 2.5$  were considered successes. The 2.5  $V_r$  cutoff was the maximum  $V_r$  of a design that produced a crystal structure in our previous paper (8). Additionally, all 30 designs had a Porod of  $>3.8$  indicating a well-folded core. 27 of the designs had a  $V_r < 2.5$ , and real space radius of gyration ( $R_g$ ) and a maximum of distance distribution ( $d_{max}$ ) within 30% of the model. For 1 design, junction 12, the  $V_r$  was  $<2.5$  but the  $d_{max}$  was 38% of the model indicating there is likely aggregation.

The two failed proteins, Junction 4 and 20, had a  $V_r$  score greater than 2.5. These failed designs also had a  $d_{max}$  and  $R_g$  significantly higher than predicted indicating there was likely aggregation.

**Figure S5 | Filtering of junction library**



**a.** The number of designs left after each stage of filtering. Designs are filtered to 1.0 RMSD for uniqueness, 0 unsatisfied hydrogen bonds, 2 helices in connection throughout the structure, and lower energy than other conformation explored in the energy landscape (mFF). 52k designs pass all filters. Not all DHRs pass the filters so to enable all DHRs to be joined we also generated a second 75k database that includes junctions that were better than their component DHRs (See Discussion S5). **b.** The number of designs per junction correlates with the quality of the DHRs that make up the junction. Shown is the number of designs per DHR vs mFF quality of the component DHR. The counts in this graph are from the 75k library of junctions.

## **Discussion S5 | Filtering and coverage of junction library**

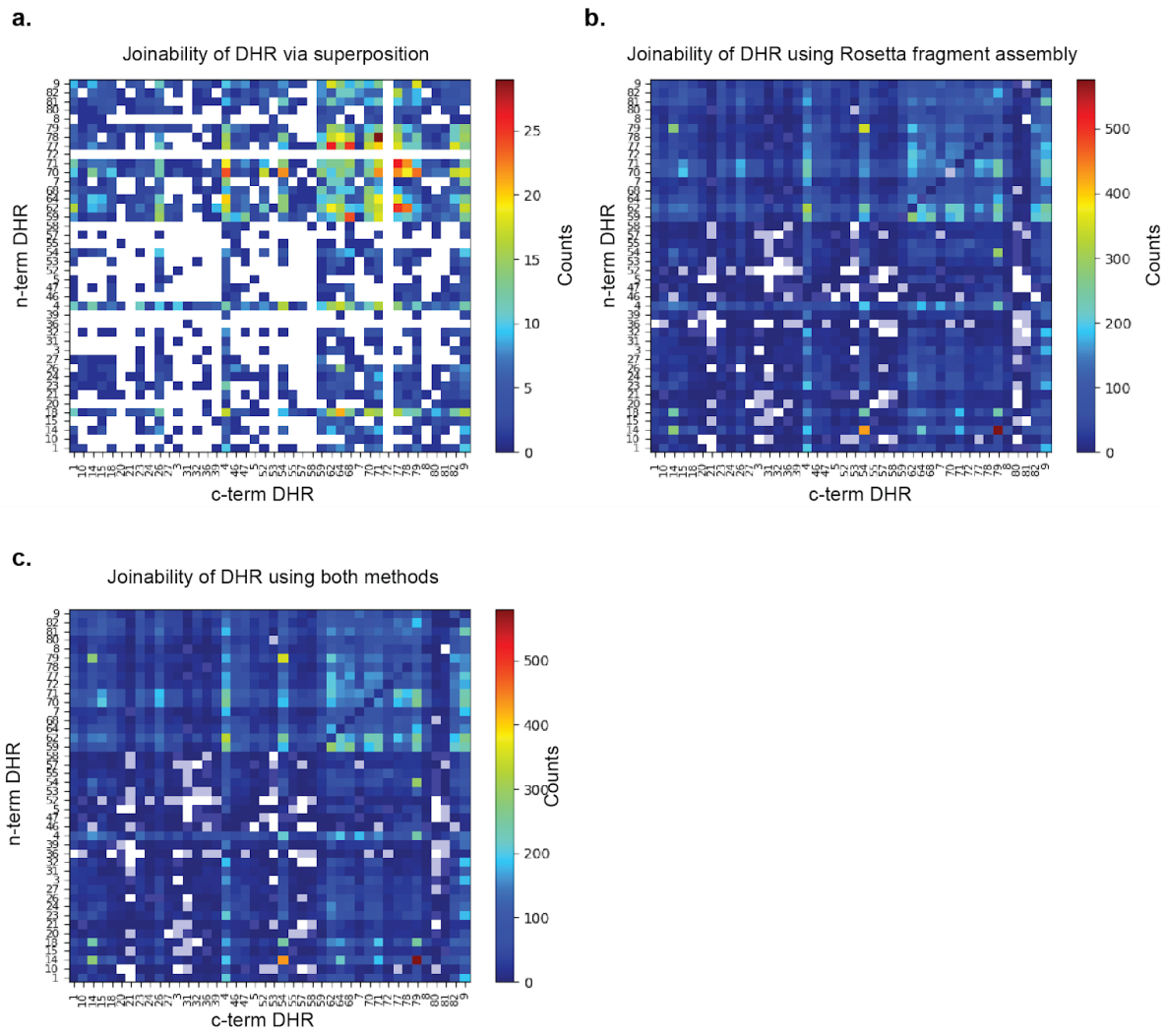
A key step in protein design is typically visual inspection to eliminate designs that appears good by Rosetta score metrics but poor by visual inspection. An example of this would be buried unsatisfied hydrogen bonds. The Rosetta metric for solvent accessible surface area (SASA) will evaluate a residue to be at the surface when the bond is close to a small pocket. While the protein designer may intuit that pocket is unlikely to exist so the hydrogen bond is unsatisfied in the core. The parameter to control pocket detection (SASA) could be tuned to match human intuition for that one case but in another case, a good design would be discarded.

For our filters, we attempted to identify thresholds that would allow all experimentally verified DHR to pass while filtering all designs that human intuition would discard. We were unable to identify a perfect filter threshold that would accomplish both goals. The filters we used are >1 helix in junction, no buried unsatisfied hydrogen bonds and that the design is the lowest point in the energy landscapes which was modeled with machine learning (mFF). For the filter thresholds that best matched human intuition 14 of the 44 experimentally verified DHRs would also be discarded; DHR53, 80 and 81 fail to have >1 helix in junction. DHR10, 52, 77, 78, 79 and 81 fail the unsatisfied hydrogen bond filter. And DHR1, 5, 10, 36, 46, 47, 53 and 59 fail mFF.

To allow DHRs to be joined where the DHR itself is below the filter cutoffs we relax the thresholds to require junctions be better than their component DHR. For >1 helix in a junction, the design must have more contact between neighboring helices than either component DHR. For unsatisfied hydrogen bonds, the junction must have fewer unsatisfied hydrogen bonds than the initial design. And for mFF, the junction must be more likely to fold than the average of the two parent DHRs. The resulting database of junctions contains 75k designs.

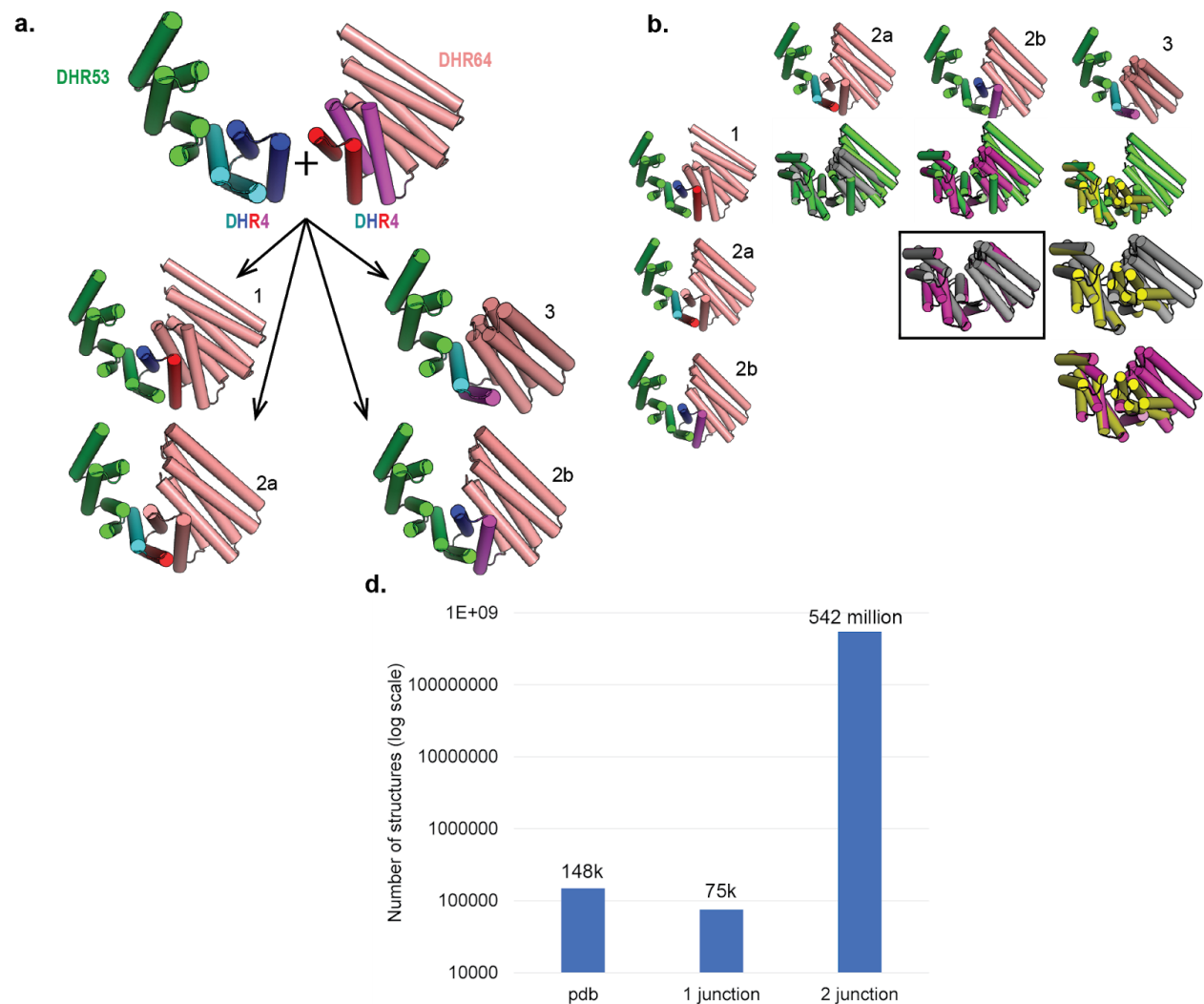
Fig S5 a shows the number of designs filtered at each stage. The joinability between DHR correlates with the quality of mFF of the parent DHR (Fig S5 b) .

**Figure S6 | Joinability of DHR**



Illustrations of the DHRs that can be connected together after filtering **a.** via superposition of helices. **b.** via Rosetta fragment assembly and **c.** both methods

## Figure S7 | Connections between junctions



**a.** From two junctions there are 4 possible structures, three of which are unique. The four ways to join junctions are by superimposing the outer two repeats (1), the inner two repeats (3) or one inner repeat to one outer repeat (2a and 2b). When one inner and one outer repeat is used the structure is identical independent of which of the junction provides the outer repeat. This is a byproduct of having 2 structurally identical and superimposable repeats at the end of each junction. Note: Sequence of connection type 1 are identical to the repeated sequence in the DHR. In case 2 and 3 each residue in the overlap derives its amino acid type for the residue from whichever building block has a residue closer. **b.** Superimposition of the 4 ways to join two junctions. The box highlights when the structure is identical. **c.** From the 75k designs in our databases, there are 542 million possible unique two junction combinations. If repeats protein extensions are counted the number of possibilities climbs into the billions.

**Table S4 | Summary of sculpt data**

**a.**

	Tested	Expressed and soluble	Correct oligomeric state by sec-mals	Correct Rg by SAXs	EM verified
Monomer sculpt	2	2	2	2	2
Oligomer sculpt	9	8	6	3	2,1 as monomer

**b.**

Oligomer Component	Tested	Expressed and soluble	Correct oligomeric state by sec-mals	Correct Rg by SAXs	EM verified	Note
C2-(tj18C2_V03)	2	2	2	1	1(as monomer)	dimer interface looks correct in SEC-mals and SAXs but not in negative stain EM
C3-(tj41C3_pm1v2)	2	2	1	0	0	
C4-(HR04C4_1)	2	2	2	2	2	
C5-(HR10C5_2)	3	3	2	0/1*	0	

**c.**

Sculpt	length	Sym.	Porod	q range		Guinier				Real space p(r)			Model p(r)		Model Fit		Sec-mals (kDA)		Real space	EM
				q min	q end	start range	end range	Rg	l0	dmax	Rg	l0	dmax	Rg	Vr	chi	Exp	Model	Exp	Valid
35	335	C5	4	0.0103	0.3035	0.0103	0.0215	60.8	29	154	52.0	24	151	45.8	7.72	3.46	102	113	No	No
36	357	C4	4	0.0103	0.2422	0.0101	0.0309	42.6	32	122	42.4	31	122	40.8	2.48	2.40	120	153	Yes	Yes
37	394	C5	3.5	0.0181	0.3364	0.0181	0.0410	55.7	144	162	59.8	148	176	53.0	7.68	1.32	215	219	Yes	No
38	337	C2	4.0	0.0109	0.2857	0.0109	0.0343	38.1	13	119	36.9	11.7	132	28.9	4.16	1.81	70	76	No	No
39	300	C4	3.9	0.0103	0.2689	0.0103	0.0309	41.9	24	115	41.4	22.4	109	36.9	6.59	3.01	133	140	Yes	Yes
40	549	C2	3.0	0.0109	0.2868	0.0109	0.0242	53.9	9.6	190	55.0	9.1	199	48.9	3.75	0.93	122	116	Yes	Yes*
40-mon																				
41	495	C3	3.9	0.0086	0.2645	0.0086	0.0209	64.8	55	160	57.1	46	141	42.7	2.69	23.37	158	170	No	No
42	452	C3	3.5	0.0103	0.3258	0.0103	0.2313	56.6	29	149	51.7	26	142	41.2	2.57	10.68	182	149	No	No
43	360	C5	4.0	0.0103	0.2868	0.0103	0.0304	42.9	27	125	42.8	27	186	53.6	8.87	12.24	118	204	No	No
44	968	-	3.8	0.0114	0.3425	0.0114	0.0332	57.2	13	178	59.8	12	177	51.8	5.02	9.83	-	-	Yes	Yes

Unable to set q\*Rg value in Guinier fitting < 1.3.

Highlights when Rg or Dmax differs >25% between model and data

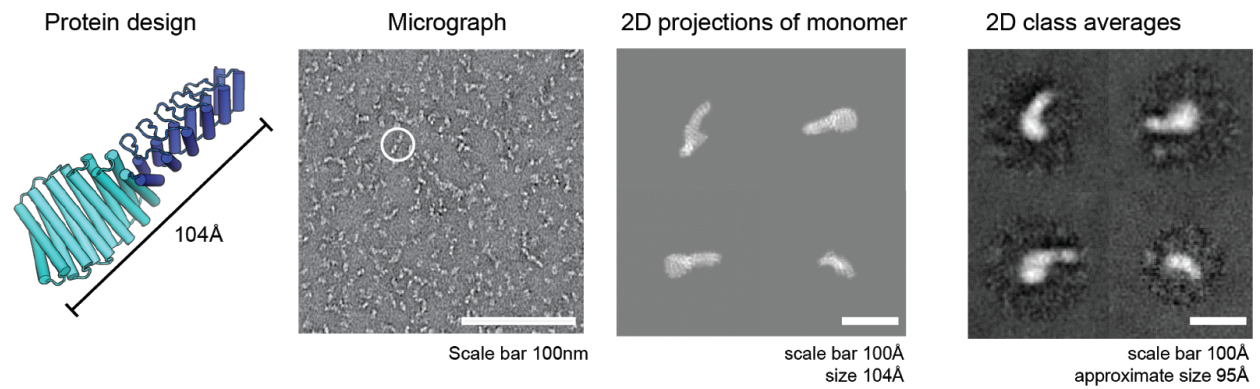
\* validated as a monomer by EM

**a.** All monomer sculpts were able to be validated by EM but only 22% of oligomers were correct. **b.** Oligomer success rate appears correlated to which oligomer is used. The C4 oligomer had a 100% success rate, while the C2,C3, and C5 oligomers fail more frequently. **c.** SAXs was run on all sculpts except Design 45 which expresses poorly and had been previously validated by EM.

## **Discussion S6 | Protein sculpt analysis**

100% of the monomer sculpts had the correct shape by electron microscopy (EM). While only 22% of the oligomer sculps were correct by EM. In most cases of EM failure, the SAXs Rg value does not match while the SEC mals size matches the correct oligomer. This suggests there may be re-arrangement happening at the interface or the interface is breaking. Also, all of the oligomer successes came from the same C4 building block. Future work will seek to identify the most stable oligomer building blocks or to design more robust building blocks. For details see Table S4.

**Figure S8 | Ankyrin junction EM image**



Characterization of DHR-ankyrin by negative stain EM. Column 1: design model with each junction in a different shade of blue. Column 2: raw negative stain micrograph. Column 3: 2D projections of the monomer design model. Column 4: 2D class averages of the design that appears to be structurally consistent with the 2D projection of monomer. Note the distinctive shape of the DHR component that is wider and shorter than the ankyrin component.



## Table S5 | Sequences

Experimental data such as gels, sec-mals, SAXs, and Rosetta energy landscapes can be found here:

[http://files.ipd.uw.edu/brunette/experimental\\_data\\_PNAS\\_2019.pdf](http://files.ipd.uw.edu/brunette/experimental_data_PNAS_2019.pdf)

Name	Sequence,
Junction 1 DHR14-DHR14	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPEVVKEIVTQLAQVAQESTNEELIREIIEVLKEL LKEAQTPEEQAFIAAAIAAAAAKSGNEEEVRQAIQKAAELASQTSEESVKELVRELAELAKKAKDPKAVEAIVQL LAELAKKSSDSELVNEIVKQLEEVAKKEATDKELVEHIEKILEELKKQSTDGWLEHHHHHHH
Junction 2 DHR14-DHR54	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPELVLEILKQLIEVLKKSQNEELQEEILEVLKELL QLGDLEVILRAAQLAAKKGDQEVVRAALEVAEKAIIKAAKRGNTDEVKALEVALKIAEDAGTEEAVRLALEVVK RVSDEAKKQGNEDAVKEAEVVRKKIEEESGTGWLEHHHHHHH
Junction 3 DHR14-DHR54	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPELVKEIVEQLLQVAQESTDEELLETILQVIKEL AKNAQSPEAALRAAEAILLAEKAGKLTETEEAKELLEIARAAIEAARSGNVEAVRKALELALQVAKSAGTEEAVR LALVVKRVSDEAKKQGNEDAVKEAEVVRKKIEEESGTGWLEHHHHHHH
Junction 4 DHR14-DHR71	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPEVVAEIVTQLLQVAKESTDVELILEIAEVLRL AEKAQSKELASKALSSAVEAVTYLAELLKEGPPNPEAALEAAEAALQAARLAAENGNEEAFKAAEAALQAAKI LVEVASESGDPELVEEAAKVAEEVVRKLAKKQGDEEVYEKARETAREVKEELKRVREEKGDGWLEHHHHHHH
Junction 5 DHR14-DHR71	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPEVISEILELLEEVARKSTDKELILEIVQVILQLA KRNHGSP LAVKAARIAAKLAADAGDAELALRAAEAVEIARTAVENGDDVEAKEAAEALEIAKKVVEAAASEKG DPELVEEAAKVAEEVVRKLAKKQGDEEVYEKARETAREVKEELKRVREEKGDGWLEHHHHHHH
Junction 6 DHR14-DHR76	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPEVISEILELLEEVARKSTDKELILEIVQVILQLA KRNHGSP LAVKAARIAAKLAADAGDAELALRAAEAVEIARTAVENGDDVEAKEAAEALEIAKKVVEAAASEKG DPELVEEAAKVAEEVVRKLAKKQGDEEVYEKARETAREVKEELKRVREEKGDGWLEHHHHHHH
Junction 7 DHR14-DHR79	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPEVVAKILQALAEVAQQSTDPELARRIIEVIAEL AKESGDEALLQAAEAKEAAQKGNTELLAVLQALLVAEVLIVAEQARENGNEELAEARELIRAVAEIATEAV QQGNPELVERVARLAKKAAELIKRAIRAKEG NRDERREALERVREVIERIEELVRQNGWLEHHHHHHH
Junction 8 DHR14-DHR79	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPEKAIQIEIAERLAEVAKESQDEELITLVLNLLS TSTDPEALEQIARAVLELARQNGDEELAQLAEEALRAVQTAKEAKEKGDDELQAALLIALAAAAAALIAAKQ TGDPEVRELAQKLVELAQTAATQVKQNPKDEEVNEALKKIVKAIQEAESLREAEESGDPEKREKARERVREA VERAEVQRDPSSGWLEHHHHHHH
Junction 9 DHR14-DHR79	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPTLISKIAERL TEVAEQGTNDELLVQIYVLLRIL QNGQTDDLKRVKNAIKVLQKVVSNRDAADLAAKAVRKAEDTLREHPDSSDVEKALKLVEEAQKAAERARE AADRTGTEDVQRLAQELIRLAIEAALQVSDPSSEEVNEALKKIVKAIQEAESLREAEESGDPEKREKARERV REAVERAEEVQRDPSSGWLEHHHHHHH
Junction 10 DHR14-DHR79	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPEVLEIVEQLAQVATEAQDPELVSRILEVLAR LAETLTNPEALSTVIQILTELAKELLEQGNLEAAEAIAIALEALAKTTGDEEVKRAAEALAKLALQAAQEATEAAQ RTGDPEVKKLAQKLAKLAATAALQILQNPDDVEEVNEALKKIVKAIQEAESLREAEESGDPEKREKARERVREA VERAEVQRDPSSGWLEHHHHHHH
Junction 11 DHR14-DHR79	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPELVKKVSVLLAEVAVESKNEELIQEIIIEVLKELI SSIQDPEQLKELAQELKEQLQEALKEGDYDAAKVLAEALAAAAKESGDEDLAEAAKLIKAAEAIKRAKEAADR TGDPEVQKLAEEELARLAEALQVLQDPKDEEVNEALKKIVKAIQEAESLREAEESGDPEKREKARERVREAV ERAEEVQRDPSSGWLEHHHHHHH
Junction 12 DHR14-DHR81	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPEVIRILELLKQVLKESTDPELQARILLVLARL ASQQGNLREAAARLAVRAAEATAKAGDQEALKEALEIARKALEEAQQARQAKNEGDLETAKALIAIALIAAAI VACTSGDKKEEAERAYEDARRVEEERKVKESAEQQGDSEVKRLAEAEQLAREARRHVQECRNGWLEHH HHHH
Junction 13 DHR14-DHR8	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPTLVAKILADLAEAALEAKDPELVQRRIIEILQELA KQATSEDLTLIAQLAISAAARAQNGDEAVAKVALALLQAVKLALENGNEVAATIIVAKKILEALKENPSDEMAK

	KMLELAKRVLDAAKNNDDETAREIARQAAEEVEADRENNSGWLEHHHHHH
Junction 14 DHR14-DHR8	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEKQSTDPPEAVKEVAIQLAAVAAQAQDPPELVKRIAQILEEI LQQFPDDEAAREALQIARAILVLEALHSSNSEEFKVKAKALLEAVLLALENGDPKVALEIARAAEAIIRALRENPS DEMAKKMLELAKRVLDAAKNNDDETAREIARQAAEEVEADRENNSGWLEHHHHHH
Junction 15 DHR18-DHR14	MDIEKLCCKAESEAREARSKAEELRQRHPDSQAARDAQKLASQAEEAVKLACELAQEHPNADRACACILLASA AAYAASKAVEDAQRHPDNQTARDKIKEASRIAEVLIQFCRAAQENNDQKALDVLEKLATVASESGNEHVLIIVE VLAILAQTTITNKDDVIQAVDIARKIAEESTNSELVNEIVKQLEEVAKAATDKELVEHIEKILEELKKQSTDGWLEHH HHHH
Junction 16 DHR49-DHR14	MDSEEEQERIRRIKKEARKSGTEESLRQAIEDVAQLAKKSQDPEVIAHAVHVIKIAQTSGSEEAKQQALRAVTE ILSNASEEEIIEALKEALETAQQEGDDEALKLLVAAAAAAAKNSKDPDAIKEIVQLLLEAAKNSTDSELVNEIVKQL EEVAKAATDKELVEHIEKILEELKKQSTDGWLEHHHHHH
Junction 17 DHR49-DHR81	MDSEEEQERIRRIKKEARKSGTEESLRQAIEDVAQLAKKSQDPEVLRRTAVEVIKIEAETSGSPEALYEAIQAVIEIA RSAQDEEALATAAIAAELADQLLQTASESGDEEALTEAAELAREILREARRVLEQAQRSGNLEVAAKALIAIALA ILVIAKVACQKGDKEEAERAYEDARRVEEARKVKESAEQGDSEVKRLAEEAEQLAREARRHVQECRGNW LEHHHHHH
Junction 18 DHR54-DHR79	MSNDEKEKLLKELKRAEELAKSPDPEDLKEAVRLAEVVRERPGSEDAKKALKIVIKAAAELAKAPNPEALKEAI EALQKVAEHSNSEEVKEAIEAIKSVLEAAREALESGDEEAAQELARLAYRAAQLLIKLEDSQDDEEKKALLAVQ ALAAAAQALQAASQTGDPEVIELAQKVELAETAATQVEQNPKEEVNEALKKIVKAIQEAESLREAEESGDP EKREKARERVREAVERAEEVQRDPSSGWLEHHHHHH
Junction 19 DHR54-DHR79	MTTEDERRELEKVARAKAIEAAREGNTDEVREQLQRALEIARESGBTAVKLALDVALRVAQEAARKGNKDAIDE AAEVVRIAEEESNSDALEQALRVLEEIYAKAVLKSEKTEDAKKAVKLVQEAYKAAQRAIEAAKRTGTPDVIKLAIK LAKLARAALAVIKRPKSEEVNEALKKIVKAIQEAESLREAEESGDPKEKREKARERVREAVERAEEVQRDPSS GWLEHHHHHH
Junction 20 DHR79-DHR14	MSSDEEEARELIERAKEAAERAQEAERTGDPRVRELARELKRLAQEAEEVKRDPSSRITLDILKAVIEAIEVA VRSLEKAYRNGNPEDVKKASKIVEEAVRLAEAAATKGNVQEINKAAREATKNNNEDLVRIAVKAAAAAAKETQTK DDVVKIVDELKRIAKNNTNSELVNEIVKQLEEVAKAATDKELVEHIEKILEELKKQSTDGWLEHHHHHH
Junction 21 DHR79-DHR14	MSSDEEEARELIERAKEAAERAQEAERTGDPRVRELARELKRLAQEAEEVKRDPSSKTTLIALKLIIEIIEAV RALEEAIKGNPEEVKATKIVEKAVRLAEIQHGNQKQIARAAADIKLAIESGNEDVARKVVVVAELAQTGT NKDVTVEIVKALEKIARQGTNSELVNEIVKQLEEVAKAATDKELVEHIEKILEELKKQSTDGWLEHHHHHH
Junction 22 DHR79-DHR54	MSSDEEEARELIERAKEAAERAQEAERTGDPRVRELARELKRLAQEAEEVKRDPSSKDTLRALSIIIIAIEAV IALEVAQKQGNPKVKERASQLVEEAVRAAEEVQNDPTDDAVYNAVHTLARAALDAVKNPDRDTRDVKKALEVV ARLAIARQGSTDAVRDALKVALKIARTAGNEEAVRLALEVVKRVSDAEEKQGNEDAVKEAEEVRKKIEEESGT GWLEHHHHHH
Junction 23 DHR14-DHR18	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEKQSTDPNLVAEVVRALEVAKTSTDTLIREIIVLLELA SKLRDPQAVLEALQAVAEARELAEKTGDPKAECAEAVSAAAEAVKKAADLLKRHPGSEAAQAALAKAAAAE AVLIACLLALDYPKSDIAKKCIKAAASEAAEEASKAAEEAQRHPDSQKARDEIKEASQKAAEEVKERCERAQEH AGWLEHHHHHH
Junction 24 DHR14-DHR18	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEKQSTDPNVVAEIVYQLAEVAEHSTDPPELIKEILQEALRL AEEQGDEELAEARLALKAARLLEEARQLLSKDPENEAKECLKAVRAALEAALLALLLAKHPGSQAAQDAV QLATAALRAVEAACQLAKQYPNSDIKCIKAAASEAAEEASKAAEEAQRHPDSQKARDEIKEASQKAAEEVKER CERAQEHNPAGWLEHHHHHH
Junction 25 DHR14-DHR54	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEKQSTDPKEVVKRIVELLTEVAKESTDVELIAEIIAVLIELAA HASSETLQEQANLIRELLHEAASGNKEAVQILLEAIAELAVKAARKGNVEAVKLALQAALVAESAGTEEA LEVVKRVSDAEEKQGNEDAVKEAEEVRKKIEEESGTGWLEHHHHHH
Junction 26 DHR14-DHR71	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEKQSTDEELVNRIVEALEEVAKESTDPQLIIEILLVALLA VESGGTEKADEALRRITEQAREAAQGGDAEAVLEAARAALQAAKAAAEKGDDEEVFKSAAEAALTIKELVEAA SEKGDPELVVEAAKVAEEVRKLAKKQGDDEEVYKARETAREVKEELKRVREEKGDGWLEHHHHHH
Junction 27 DHR14-DHR79	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEKQSTDPPEVVEIVEQLLQVAQEAQDPPELVKEIIRILKEL AKTAENEEAAATALLVAEALVLAELLARTTGDD SARQAELAKEAAEAAKRAQEAARTGDPEVKRLALELV

	RLAAEAAEEVTKNPDDEEVNEALKKIVKAIQEAVESLREAEEESGDPEKREKARERVREAVERAEEVQRDPSSG WLEHHHHHH
Junction 28 DHR14-DHR79	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDPTLVAKIALLAEVAAEAQDPELIKRIELRQLIK NAKSDEARKAAKALAEAVEVALKAAQQLKQNPEDESARQALELILEAVEAAARALKAALETGSPEVIELALKLAE LAIEAARQVLKNPDNEEVNEALKKIVKAIQEAVESLREAEEESGDPEKREKARERVREAVERAEEVQRDPSSGW LEHHHHHH
Junction 29 DHR14-DHR8	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDKEAIKDIVRALKEVLKHSQDDELREQILVLL AAQAGDVEEALERLAQEAKKEGDEEALKVLLKALAEAVRTAKENGNEVAATVAEAAAKIATALRENPSDEM AKKMLELAKRVLDAAKNNDDETAREIARQAAEEVEADRENNSGWLEHHHHHH
Junction 30 DHR14-DHR8	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDDEEAVKEVVRQLALVAATATDPELIAEILQVILQLA EQAGDEEVAAARQALEEIKQAQEQGSEAVLVAALAVAVLAAAANGNPEVARVVKHAARLIKEALEENPSDE MAKKMLELAKRVLDAAKNNDDETAREIARQAAEEVEADRENNSGWLEHHHHHH
Junction 31 DHR14-DHR8	MDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAEELAKQSTDKKLALQIVLLAEVLQEAQDPELAIKAEELAEII KEAGGSEDALQIVQEIATLRQGNEEVAVLAVLLIAVILALQNGNPEVAHEVARVAREILKALEENPTDEMAKK MLELAKRVLDAAKNNDDETAREIARQAAEEVEADRENNSGWLEHHHHHH
Junction 32 DHR49-DHR79	MDSEEEQERIRILKEARKSGTEESLRQAIEDVAQLAKKSQDEEVLREAVEVITQAARDSGSSEALQQAVRAVL EIAKSGKDVEAAAHAAKLLLEKNPEDESAREALELVERAVQAAQEAQEAANRTGDPEVQELAEKLLALAADAA AQVVKNPDDEEVNEALKKIVKAIQEAVESLREAEEESGDPEKREKARERVREAVERAEEVQRDPSSGWLEHHH HHH
Junction 33 DHR4-DHR64	MSYEDECEEKARRVAEKVERLKRSGTSEDEIAEEVAREISEVIRTLKESGSSDEEIIATCVLILAAAARALKESG VSDEQINRILATLIKEVLRALNQETNKSNEEILRELLQALIELASKSDSETALLAVQLVVLAKVALEVAQSEGSEE ALELALAEAAEAARLAKEVLRATENGNEVARRAVELVKRVAELLERIARESGSEEAKERAERVREEARELQE RVKELREREGDGWLEHHHHHH
Junction 34 DHR53-DHR4	MSNDEKEKELKELLKRAEELAKSPDPEDLKEAVRLAEEVVRERPGSEAAKKALEIIQEAELLKKSPPDPEIIAAA RALLKIAATTGDNEAAKQAEAAASKAAQLAEQRGDDELVCEALALLIAAQVLLKQGGTSDEEVAEHVARTISQL VQRLKRKGASYEVIKECVQRIVEIIVEALKRSGTSEDEINEIVRRVKSEVERTLKESGSSGWLEHHHHHH

**Table S6 | Sequences of sculpts**

name	Sequence
Sculpt 35 HR10C5_2 -(DHR10-DHR14) -DHR14 <sub>3</sub>	MSAEKMLMAKLIIVAENAKRKGDDTLIAIMAAKLAFEIVRIAEEAGIDSSEVLELAIRLIKEVVENAQRE GYDISIAALAAAMAFALVAIAAKRAGITSPEVLKLAIIKLVVLAQQLSGYDIEEAAKKAETFLRVAAEEARE KGIDPREVIARSIADAAEEAATLAVRKGDEESLKSIVRLAATAAKTAKNPEVITKIVNLLLEIAERATDNELV NEIVKQLAEVAKEATDKDLVIHVIRILAEAKHSTDSSELVNEIVKQLAEVAKRATDKELVIEVRILAEAKES TDSRLVEEIVRQLKEVAERATDKELVEEIEKILEELKKESTDGWLEHHHHHH
Sculpt 36 (DHR53-DHR4) -DHR4 <sub>2</sub> -HR04C4_1	MHHHHHHHGGSGENLYFQGGSGWGNEEEEKELLERAKELAKSPDPEDLKEAVRLAEEVVRERPG SEAAKKALEIIQEAELLKESPDPEAIIAARALLKIAATTGDEEAAKEAIEAAEKAARLAEERGDDDELVCE ALALLIARVLLLKQQGTSDEEVAETVARTISKLVKRLKKKGASEEVICCVARIVAEIVKALKRSGTSEEEI AEIVARVISEVIRTLKEESGSSYEVICCVARIVAEIVEALKRSGTSAVEIAKIVARVISEVIRTLKESGSSYEVI CECVARIVAEIVEALKRSGTSAIIALIVALVISEVIRTLKESGSSFEVILECVIRIVLEIIEALKRSGTSEQDV MLIVMAVLLVVLATLQLSGS
Sculpt 37 HR10C5_2 -(DHR10-DHR9) -DHR9 <sub>3</sub>	MSAEKMLMAKLIIVAENAKRKGDDTLIAIMAAKLAFEIVRIAEEAGIDSSEVLELAIRLIKEVVENAQRE GYDISIAALAAAMAFALVAIAAKRAGITSSEVLELAIRLIKVVENAQREGYDIEEAAAARAAAEAFERVAEAA KRAGITSSKAIKIAELIEVVVRAASRNHDISKAARKAAETIKTAADLAKKGNPDELAKHIAKTVEELKRN GVSEDEIARTVAIIAFVIQALKSSGSSSEDIATIVARIVAEIVRALKRSGTSEDEIAEIVAKVISEVIRTLKES GSSHEVIKIVARIVAEIVEALKDSGTSEEEIAKIVAHVISEVIRTLKESGSSSEVIHHIVKRIVHEIVKALKES GTSEDEIREIVKHVEHEVERTLHESGSSGWLEHHHHHH
Sculpt 38 (DHR14-DHR18) -tj18C2_V03	MHHHHHHHGGSGENLYFQGGSGWGSSEVNERVKQLAEKAKEATDKEEVIEVKELAEAKQSTDPNL VAEVVRALTEVAKTSTDTLIREIIVLLELASKLRDPQAVLEALQAVAEALAEKKTGDPIAKLCAIIVSL AAEAVKKAELLKRHPDSQAAQDALKLAKQAAEAVLLACLLALEHPNAAIILCVAAIAAAIAASMAAALA QRHPDSQAARDAIKLASQAAEAVKLACELAQEHPNAKIAVLCILAAALAAIAAALAALLAQLHPDSQAAR DAIKLASQAAEAVKLACELAQEHPNADIAEKICILLAILAALLAILAALLAMLHPDSQLARDLIDLASELAEV KERCER
Sculpt 39 DHR53 <sub>2</sub> -(DHR53-DHR4) -HR04C4_1	MHHHHHHHGGSGENLYFQGGSGWGNEEEEHLKELLKRAEELAKSPDPDDLREAVRLAEEVVRTRPG SELAKKALEIILRAAEELAKLPDPEALHEAVRAAEHVRSQPGSEAAKEALRIIQAELLKESPDPTAIIR AARALLKIARTTGDEEAAKEAIEAAKKAADLARERGDDELVCEALALLVAAQVELLKQQGTSVEIAKIVA RVISEVIRTLKEKGSSEYVICCVARIVAEIVEALKRSGTSAIIALIVALVISEVIRTLKESGSSFEVILECVIR IVLEIIEALKRSGTSEQDVMLIVMAVLLVVLATLQLSGS
Sculpt 40 (ank1-DHR18) -DHR18 <sub>2</sub> -tj18C2_V03	MHHHHHHHGGSGENLYFQGGSGWGSSELGKRLIEAAENGNKDRVKDLIENGADVNASDSDGRTPLHH AAENGHKEVVKLLISKGADVNAKDSGRTPLHHAENGHKEVVKLLISKGADVNAKDSGRTPLHHA ENGHKEVVKLLISKGADVNAKADRGMTPLHFAAWRGHKEVVKLLISKGADLNTSAKDGATPVLLALRR GDEEVVRLLEEAKKRGFDEFARCAEAAELAEALKLAELLRYPNDEAARLAHHLAKLAEVLAELACI LASEHPNADIAKLCIKAASEAAEAAASKAAELAQRHPSQAARDAIKLASQAAEAVKLACELAQEHPNADI AKLCIIAASLAAEAAASKAAELAQRHPSQAARDAIKLASQAAEAVKLACELAQEHPNAAIILCVAAIAAAI AASMAAALAQRHPDSQAARDAIKLASQAAEAVKLACELAQEHPNAAIILCVAAIAAAI HPDSQAARDAIKLASQAAEAVKLACELAQEHPNADIAEKICILLAILAALLAILAALLAMLHPDSQLARDLID LASELAEVVKERCER
Sculpt 41 (139_tj41C3_pm1v2_ DHR27) tj41C3_pm1v2 -(TJ41-DHR27)	MIEEVVAEMIDILAESSKKSIEELARAADNKTTEKAVAEIEEIIARLATAAIQLIEALAKNLASEEFMARISA IAELAKKAIEAIYRLADNHTTDTFMARAIAAIANLAVTAILAIALASNHTTEEFMARISAIAELAKKAIEAIY RLADNHTTDFMAAAIEAIALLATLAILAIALASNHTTEEFMAKAIASIAELAKKAIEAIYRLADNHTNEELI RHAIEIIEIAEIAARAIIIEAKRLKSEYALHALRAVLEIIEHALERIARKADKEEKKALELLIEVAREIYRLAE EAAKRAKDEEEAAKIAVIAAEAILLELRAQRKVTDNEVIEKLEVVKEIIRLAEEMKKMTDEEEAAKIAKE ALEAIKMLARAVEEVTDNEVIEKLEVVKEIIRLAEEMKKMTDEEEAAKIAKEALEAIKMLARAVEEVTDK ERIEQLLREVKEIIRRAEEESRKETDDEEAAKRAREALRRIRERAREVEEDKSGWLEHHHHHH
Sculpt 42 tj41C3_pm1v2 -(TJ41-DHR1)	MIEEVVAEMIDILAESSKKSIEELARAADNKTTEKAVAEIEEIIARLATAAIQLIEALAKNLASEEFMARISA IAELAKKAIEAIYRLADNHTTDTFMARAIAAIANLAVTAILAIALASNHTTEEFMARISAIAELAKKAIEAIY RLADNHTTDFMAAAIEAIALLATLAILAIALASNHTTEEFMAKAIASIAELAKKAIEAIYRLADNHTNEEI HEAAEAILRIAAEIRAIIEELVRRSKSEIEERAKKLEIEIARKAIEAALRLGSEEIARVAYILIEIIKRRHPGD KEEAAEAIARKIIEQIIRTLPGGCDCVAKAASSIIRAVIEKPNPYSEVVADVAAAIVKAIIEGNPNPDCDCVAKA ASSIIRAVIEKPNPYSEVVADVAAAIVKAIIEGNPNPGRDCVRAASSIIRAVQEKNPNYSEVVEDVKRAIEK AIKEGNPNPWWLEHHHHHH



## **Discussion S9 | Methods for expression, crystallization, SAXs and negative stain electron microscopy**

### **Protein expression and characterization:**

Genes were synthesized and cloned by IDT into pET29b. Genes were optimized for *E. coli* expression using DNAworks (17). For the 34 junction proteins, an additional c-terminal tag of GWLEHHHHHH was added; W was added for tracking protein concentration through absorbance at A280. For the protein “sculpt” the tag was changed to the n-terminal HHHHHHHGGS (His tag), GENLYFQG (TEV site), GSGWG (flexible region + W), except for cases where the n-terminal was part of the dimer interface. In those cases, the original c-terminal tag was used. The genes for the 800+ residue protein “L” and “V” sculpt were synthesized by Genscript.

Proteins were expressed in *E. coli* Lemo21s using 500  $\mu$ M isopropyl- $\beta$ -D-thiogalactopyranoside (IPTG) after 4 hours at 37 °C in Terrific Broth (TB) growth medium. Cells were harvested by centrifugation and lysed using a Microfluidizer (Microfluidics) and purified by metal ion affinity (IMAC) and size-exclusion chromatography (SEC). The lysis buffer was 20 mM Tris pH 8.0, 500mM NaCl, DNase, 0.25% CHAPS. The wash buffer was 20 mM Tris pH 8.0, 500mM NaCl, 30mM imidazole. The elution buffer was 20 mM Tris pH 8.0, 150mM NaCl and 250mM imidazole. Following the IMAC step, proteins were dialyzed in 20mM Tris 150mM NaCl pH 8.0. Protein concentrations were measured using a NanoDrop spectrophotometer (Thermo Scientific). Thermal denaturation and secondary structure content were monitored by circular dichroism (CD) using an AVIV 420 spectrometer (Aviv Biomedical). Oligomeric states were measured by analytical gel filtration (Superdex 75 or 200, GE Healthcare) coupled with multiple-angle light scattering (SEC-MALS). Molecular weights were confirmed by mass spectrometry on an LCQ Fleet Ion Trap Mass Spectrometer (Thermo Scientific).

### **Crystallization:**

All crystallization trials were carried out at 22 °C in 96-well format using the hanging-drop method. Crystal trays were set up using a Mosquito crystallization robot enclosed in a humidifying chamber (TTP labtech). Drop volumes ranged from 200 to 400 nl and contained protein to crystallization solution in ratios of 1:1, 2:1 and 1:2. All crystals were frozen in liquid nitrogen prior to shipment to the Advanced Light Source (ALS, Berkeley, CA) or the Advanced Photon Source (APS, Lemont, IL) for diffraction data collection. All datasets were integrated and scaled in HKL2000 (18). Diffraction data quality was assessed using Xtriage in the Phenix software suite (19). Phase information was obtained by molecular replacement in PHASER (20), using either the original Rosetta Design models or related low-energy variants as the search models. Initial models were automatically obtained using Phenix.autobuild (21). Final models

were produced after iterative rounds of manual building in Coot (22) and refinement with Phenix.refine (23). Final resolution cutoffs were determined by monitoring the refinement statistics in the context of the reflection data completeness and the CC  $\frac{1}{2}$  values of the original diffraction data (24). The geometric quality of the final models was assessed using Molprobit (25).

*Junction 19* – Crystals were grown in Qiagen JCSG+ condition E5 (0.1M CAPS pH 10.5, 40% MPD) and required no additional cryopreservation. Diffraction data was collected on ALS beamline 8.2.2., 280 images with 1 ° increments.

*Junction 23* – Crystals were grown in Qiagen MPD condition A9 (0.2 Ammonium chloride, 40% MPD) and required no additional cryopreservation. Diffraction data were collected on APS beam line NE-CAT 24-ID-C, 1200 images with 0.25 ° oscillations.

*Junction 24* – Crystals were grown in Qiagen JCSG+ suite condition D9 (0.19M Ammonium sulfate, 25.5% (w/v) PEG 4000, 15% (v/v) glycerol) and required no additional cryopreservation. Diffraction data was collected on ALS beamline 8.2.2., 150 images with 1 ° oscillations.

*Junction 34* – Crystals were grown in Qiagen JCSG Core III suite condition G5 (0.2M calcium chloride dihydrate, 20% (w/v) PEG 3500. Crystals were briefly soaked in crystallization condition supplemented with 25% (v/v) PEG 400 as a cryoprotectant. Diffraction data was collected on ALS beamline 8.2.2., 200 images with 1 ° oscillations.

Data collection and refinement statistics are given in SI appendix Table 1

#### **SAXS:**

SAXs data was collected at the SIBYLS 12.3.1 beamline at the advanced light source LBNL (13, 16, 26) using the same method as used in (8). Data was averaged and sliced using the SAXs Frameslice program and analyzed using SCATTER 3.0g program (14). An in-depth analysis of the SAXs method can be found in the supplementary information.

#### **Negative Stain Electron Microscopy**

Samples were applied to glow-discharged continuous carbon film EM grids and stained with 1% uranyl formate. Designs that failed with the uranyl formate stain were tried with nano-tungsten stain but these still failed. Screens were run on an FEI Morgagni 268 electron microscope operating at an accelerating voltage of 100 kV. Grids were then examined using a Tecnai Spirit G2 transmission electron microscope operating at an acceleration voltage of 120 kV.

Micrographs were acquired at a magnification of 67,000x and pixel size of 1.60Å with a Gatan Ultrascan 4000 CCD via Legikon software (27). Approximately 100 micrographs were collected per sample at a defocus range between 1-1.5µm. Image processing, including CTF estimation, particle picking, and 2D reference-free classification, was performed using the software package cisTEM (28). Multiple rounds of 2D classification were carried out to remove junk particles, and

selected representative final averages are shown. The 2D projection images in Fig. S8 were generated using the v4 projection tool in the Eman 1.9 software package (29).



## **Discussion S10 | Acknowledgements**

We thank Lauren Carter, Cameron M. Chow, Alex Young-Seug Kang and the rest of the IPD staff for help with protein production and setting up crystal screening trays. We thank Alexis Courbet and Rubul Mout for teaching TB how to create EM grids. We also thank Luki Goldschmidt for compute infrastructure support. The majority of the compute work was facilitated by the Hyak and Mox supercomputer system at the University of Washington. The electron microscopy work was carried out at the Arnold and Mabel Beckman Cryo-EM Center at the University of Washington.

The SAXs work was conducted at the Advanced Light Source (ALS), a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the Department of Energy, Office of Basic Energy Sciences, through the Integrated Diffraction Analysis Technologies (IDAT) program, supported by DOE Office of Biological and Environmental Research. Additional support comes from the National Institute of Health project ALS-ENABLE (P30 GM124169), a High-End Instrumentation Grant S10OD018483, the Open Philanthropy Project at the Institute for Protein Design, the Alzheimer's Disease Research Center (ADRC P50 AG005136) and the Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research. This work was also supported by NIGMS (R01GM12764 and R01GM118396 to J.M.K.).

## Supplementary References

1. C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
2. J. A. Fallas, *et al.*, Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem.* **9**, 353–360 (2017).
3. P. Bradley, K. M. S. Misura, D. Baker, Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).
4. D. L. Theobald, Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallogr. A* **61**, 478–480 (2005).
5. P. Liu, D. K. Agrafiotis, D. L. Theobald, Fast determination of the optimal rotational matrix for macromolecular superpositions. *J. Comput. Chem.* **31**, 1561–1563 (2010).
6. J. B. Maguire, S. E. Boyken, D. Baker, B. Kuhlman, Rapid Sampling of Hydrogen Bond Networks for Computational Protein Design. *J. Chem. Theory Comput.* **14**, 2751–2760 (2018).
7. D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
8. T. J. Brunette, *et al.*, Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).
9. R. Das, *et al.*, Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69 Suppl 8**, 118–128 (2007).
10. T. J. Brunette, O. Brock, Guiding conformation space search with an all-atom energy potential. *Proteins* **73**, 958–972 (2008).
11. G. J. Rocklin, *et al.*, Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
12. R. P. Rambo, J. A. Tainer, Super-resolution in solution X-ray scattering and its applications to structural systems biology. *Annu. Rev. Biophys.* **42**, 415–441 (2013).
13. G. L. Hura, *et al.*, Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat. Methods* **6**, 606–612 (2009).
14. R. P. Rambo, J. A. Tainer, Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **496**, 477–481 (2013).
15. G. L. Hura, *et al.*, Comprehensive macromolecular conformations mapped by quantitative

- SAXS analyses. *Nat. Methods* **10**, 453–454 (2013).
16. S. Classen, *et al.*, Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the Advanced Light Source. *J. Appl. Crystallogr.* **46**, 1–13 (2013).
  17. D. M. Hoover, J. Lubkowski, DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43 (2002).
  18. Z. Otwinowski, W. Minor, Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
  19. P. D. Adams, *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
  20. A. J. McCoy, *et al.*, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
  21. T. C. Terwilliger, *et al.*, Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).
  22. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
  23. P. V. Afonine, *et al.*, Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).
  24. P. A. Karplus, K. Diederichs, Linking crystallographic model and data quality. *Science* **336**, 1030–1033 (2012).
  25. V. B. Chen, *et al.*, MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
  26. S. Classen, *et al.*, Software for the high-throughput collection of SAXS data using an enhanced Blu-Ice/DCS control system. *J. Synchrotron Radiat.* **17**, 774–781 (2010).
  27. C. Suloway, *et al.*, Automated molecular microscopy: the new Legimon system. *J. Struct. Biol.* **151**, 41–60 (2005).
  28. T. Grant, A. Rohou, N. Grigorieff, cisTEM, user-friendly software for single-particle image processing. *Elife* **7** (2018).
  29. S. J. Ludtke, P. R. Baldwin, W. Chiu, EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).