

## Control of repeat protein curvature by computational protein design

Keunwan Park<sup>1,4</sup>, Betty W. Shen<sup>2</sup>, Fabio Parmeggiani<sup>1,4</sup>, Po-Ssu Huang<sup>1,4</sup>, Barry L. Stoddard<sup>2</sup>, and David Baker<sup>1,3,4,\*</sup>

<sup>1</sup> Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

<sup>2</sup> Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

<sup>3</sup> Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

<sup>4</sup> Institute for Protein Design, University of Washington, Seattle, WA 98195, USA

### Abstract

Shape complementarity is an important component of molecular recognition, and the ability to precisely adjust the shape of a binding scaffold to match a target of interest would greatly facilitate the creation of high affinity protein reagents and therapeutics. Here we describe a general approach to control the shape of the binding surface on repeat protein scaffolds, and apply it to leucine rich repeat proteins. First, a set of self-compatible building block modules are designed that when polymerized each generate surfaces with unique but constant curvatures. Second, a set of junction modules that connect the different building blocks are designed. Finally, new proteins with custom designed shapes are generated by appropriately combining building block and junction modules. Crystal structures of the designs illustrate the power of the approach in controlling repeat protein curvature.

Repeat protein scaffolds have attracted much attention as alternative binding scaffolds to antibodies<sup>1-4</sup> and also as building blocks of protein nanomaterials<sup>5-7</sup> because of their intrinsic modularity and high stability. The leucine rich repeat (LRR) is a repeat protein scaffold with a horseshoe-like global structure in which the concave surface is often a binding interface<sup>8</sup>. While LRRs share a common structural motif (LxxLxLxxN/C), different LRR modules generate proteins with distinct global curvatures when the repeat modules are packed on themselves<sup>9</sup>. Irregular LRR modules are frequently observed interspersed within arrays of canonical repeat modules; their presence contributes to the curvature diversity within the family. For example, Toll-like receptor 4 (TLR4) contains three distinct regions

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:[http://www.nature.com/authors/editorial\\_policies/license.html#terms](http://www.nature.com/authors/editorial_policies/license.html#terms)

\*Correspondence: [dabaker@u.washington.edu](mailto:dabaker@u.washington.edu).

#### Author contributions

K.P. and D.B. conceived the project; K.P. performed the computational design with assistance from F.P. and P.S.H.; K.P. expressed, purified, and characterized the designs with assistance from F.P.; B.W.S. crystallized the designs and determined the structures; K.P. and D.B. drafted the manuscript with input from all authors; B.L.S. and D.B. supervised research.

#### Accession Codes

Atomic coordinates and structure factors have been deposited in the Protein Data Bank with the accession codes 4R58 (DLRR\_A), 4R5C (DLRR\_E), 4R5D (DLRR\_G3), 4R6J (DLRR\_H2), 4R6F (DLRR\_I), and 4R6G (DLRR\_K).

of LRR repeats, each having different curvatures which collectively generate a surface with high shape complementarity to the target surface of the MD2 protein<sup>10</sup>. Current engineering approaches have focused on changing residues at the binding surfaces of an already existing or consensus repeat protein<sup>11-16</sup>, varying the numbers of repeat modules<sup>17-19</sup>, and fusing naturally occurring repeat proteins<sup>10,20,21</sup>. While powerful, these strategies do not allow customization of repeat protein curvature for a specific application.

Here, we describe a general computational design strategy to create new repeat proteins with custom-specified curvature. We demonstrate the power of the approach by designing twelve novel proteins with different curvatures. Crystal structures show that the method allows control of repeat protein curvature with atomic-level accuracy.

## Results

### Strategy for curvature-tunable scaffold design

Our design strategy has three steps (**Fig. 1a**). The first step is the design of a set of idealized self-compatible building block modules ( $BB_1, BB_2, \dots, BB_n$ ) from which a series of proteins of variable length  $BB_i^n$  can be created directly by varying the number of building block repeats without any further engineering. These “homo-building block” proteins will have a constant curvature defined by the base building block module. The second step is the design of a set of junction modules ( $JN_{BB_i \rightarrow BB_j}$ ) that connect building block module  $i$  to building block module  $j$ . A critical feature of the design at step one and two is that the interfaces between individual building blocks, as well as those between building blocks in junction modules, have sufficiently low energy that the orientation between all units depends only upon the identity of adjacent repeats and is independent of the longer-range context. This enables the third and final step -- general module assembly -- the combination of building blocks and junction modules to generate a protein with a desired overall curvature. While the overall strategy is applicable to any repeat protein, in this paper we focus on LRRs. We describe the computational design and experimental characterization for each step in the following sections.

### Step 1: Building block module selection and design

Nature provides a diverse set of LRR modules, with lengths from 20 to 30 amino acids<sup>8</sup>, but only a few possess high self-compatibility such that repeated stacking of the same module generates a well-folded protein structure. We generated a Markov transition model for naturally occurring LRR proteins to investigate the overall patterns of module organization in LRR structures. In the model, nodes correspond to individual modules (represented by the module length: L22 indicates an LRR module with 22 residues, etc) and edges to transitions between modules with strength proportional to the transition frequency observed between the modules in the Protein Data Bank (PDB) (see Methods section). The resulting transition network (**Fig. 1b**) has strong self-edges corresponding to packing of identical modules for L22 and L24, and strong mutual transitions between L28 and L29. Accordingly, we selected these LRR types to design the idealized building blocks (**Fig. 1c-d**).

A recently developed Rosetta repeat protein idealization method<sup>22</sup> was used to design ideal versions of each unit. Different instances of the naturally occurring repeat units have somewhat variable sequences; the idealization process generates a single low energy repeat unit (both sequence and structure) guided by the available information for the family. Briefly, an idealized poly-valine backbone structure with identical repeats was generated using RosettaRemodel<sup>23</sup> with LRR family-specific constraints. Rosetta sequence design guided by a family-specific sequence profile was then carried out, while constraining the sequences to be identical for each repeat. The idealization of the L24 module (DLRR\_B) is described in Parmeggiani et al.<sup>22</sup> We applied the idealization procedure to the L22 module (DLRR\_A) and the two-unit {L28→L29} module (DLRR\_C), and obtained the sequences and models in **Figure 1d**.

Genes were synthesized for proteins containing 5 to 7 idealized building block modules. The N-terminal capping domain of internalin B was fused to DLRR\_A and DLRR\_B to enhance protein solubility and expression<sup>12,20</sup> whereas DLRR\_C was expressed without a capping motif; instead the sequences of the N and C terminal repeats were redesigned to eliminate exposed hydrophobic residues. The idealized repeat designs were expressed in *E.coli* and found to be soluble and to have high thermal stability (**Fig. 2c**).

We solved the crystal structures of DLRR\_A (L22<sup>6</sup>) and DLRR\_B (L24<sup>7</sup>) (**Table 1**) and found that they closely match the design models (DLRR\_A at C $\alpha$  root mean square deviation (RMSD) 1.4 Å; DLRR\_B at C $\alpha$  RMSD 1.7 Å, **Fig. 3a-b**). The crystal structures contain water-mediated networks localized to the convex side of the repeats; it may be possible to incorporate these in future design calculations (**Supplementary Fig. 1a**). Each of the idealized building block repeats has the expected overall curvature: repeats of the L22 and L24 building blocks generate solenoid-like structures, whereas repeats of the {L28→L29} building block are almost circular and have a more curved concave surface. Parametric descriptions of the global shapes generated by each building block repeat are provided in **Supplementary Figure 1b** and **Supplementary Table 1**.

## Step2: Design of junction modules

We devised a computational protocol for junction module design which takes advantage of the conserved motif (LxxLxLxxN/C) in the idealized LRR building blocks: the core residues are kept constant to maintain a stable hydrophobic core, while the evolutionarily variable positions, primarily located on the convex side, are optimized to create a low energy interface between adjacent modules. To generate a junction module  $JN_{BB_i \rightarrow BB_j}$  connecting building block *i* and building block *j*, we start from a two-unit  $BB_i^2$  module and a one-unit  $BB_j$  module (**Fig. 2a**). The second unit in  $BB_i^2$  is superimposed on  $BB_j$  by aligning the core motif residues. RosettaCM<sup>24</sup> is then used to generate a hybrid structure  $BB_i \rightarrow BB_j$  with coordinates based on those of the first unit in  $BB_i^2$  before the core motif and those of  $BB_j$  after the motif. The residues at the fusion interface are optimized using RosettaDesign<sup>25</sup>. This redesigned hybrid two-unit structure  $BB_i \rightarrow BB_j$  is the junction module  $JN_{BB_i \rightarrow BB_j}$  between building block *i* and building block *j* (**Fig. 2b**).

A special case of a junction module is a three-unit module  $JN_{BB_i \rightarrow BB_w \rightarrow BB_i}$  that connects two identical copies of the same building block but has a structure different than that of the

building block (**Fig. 2b**). We call such junction modules between two identical building blocks ‘wedge’ modules. Like other junction modules, wedge modules produce a local change in the protein curvature. We designed and characterized five junction modules connecting the building block modules described in the previous section.

**L22→L24 and L24→L22.** A junction module for L22→L24 was generated previously without hydrophobic core design<sup>12</sup>, and hence direct fusion constructs between L22 and L24 were made in both directions (i.e. L22→L24 and L24→L22) to test compatibility between the two idealized modules. The hybrid model structures showed high structural compatibility without further design. Hence the junction modules in these cases are simply the fusion of the two building blocks.

Two fusion proteins for L22→L24 (DLRR\_D) and L24→L22 (DLRR\_E) were expressed in *E. coli*, and found to be soluble and monomeric in Size-Exclusion Chromatography coupled to Multiple Angle Light Scattering (SEC–MALS) experiments (**Supplementary Fig. 1e**). Far-UV Circular Dichroism (CD) spectra and thermal denaturation profiles suggested well-packed structures with the expected secondary structure content (**Fig. 2d**). The fusion proteins had similar or higher stability than the original L22 (DLRR\_A) or L24 (DLRR\_B) designs (**Table 2**); L22 and L24 evidently have high compatibility despite the rare occurrence of fusions between them in nature (**Fig. 1b**). The crystal structure for L24→L22 (DLRR\_E) was determined at 1.9 Å resolution and showed high consistency with the design model (**Fig. 3c**) and the original L22 or L24 structures (**Supplementary Fig. 1c**).

**L22→L28 and L24→L28.** Designs of junction modules for L22→L28 and L24→L28 are challenging because of substantial differences in their structures, including their module length (22 or 24 vs. 28), their secondary structure in the variable region (3<sub>10</sub>-helix or loop vs. α-helix), their curvature on the concave surface (moderately-curved vs. highly-curved) and their global shape (super-helical vs. circular). The initial fusion models generated by RosettaCM<sup>24</sup> (before redesign) contained side chain clashes and cavities at the interface between the modules (**Supplementary Fig. 1d**). Residues at the fusion interface were therefore redesigned to improve the all-atom Rosetta energy and packing as assessed by RosettaHoles<sup>26</sup>. The junction designs were based solely on building block models generated by Rosetta<sup>25</sup> because the crystal structures of the building blocks were not determined. Six designs for L22→L28 (DLRR\_F) and six designs for L24→L28 (DLRR\_G) were experimentally characterized (**Table 2**).

All designs were expressed in *E. coli* and found to be highly soluble and monomeric in SEC–MALS experiments (**Supplementary Fig. 2 and 3**). They displayed well-defined far-UV CD spectra with minimum near 218 nm, similar to those of previously characterized LRRs with primarily beta sheet secondary structure. Thermal denaturation experiments showed cooperative unfolding for all fusion designs (**Fig. 2d**), suggesting a well-packed hydrophobic core. Fusion of more stable LRR modules to less stable LRR modules via a well-designed junction appears to increase overall stability: the stability of all the junction module containing designs was greater than that of the original {L28→L29}<sup>5</sup> design (DLRR\_C).

We determined the crystal structure of the L24→L28 fusion (DLRR\_G3) to evaluate the accuracy of the design. The crystal structure, determined at 2.5 Å resolution, shows the atomic details of the junction module as well as the structures of L24 and {L28→L29} modules (**Fig. 3d**). The assumption underlying our approach that curvature can be locally controlled is supported by the similarity of the L24 modules (C $\alpha$  RMSD 0.3 Å) in DLRR\_G3 structure to those in the all L24 DLRR\_B structure, and by the similarity of the {L28→L29} modules (C $\alpha$  RMSD 1.3 Å) to the {L28→L29} modules in DLRR\_C model. The key core side chain interactions in the junction module are very similar in the design model and crystal structure (C $\alpha$  RMSD 0.9 Å, **Fig. 3d**).

**L24→L32→L24.** In addition to the junction modules linking the different building block modules, we designed a wedge module inserted between L24 modules. In native LRR proteins, inserting an ‘irregular’ module between the regular modules is a common way to generate structural diversity by altering the overall curvature or forming a binding interface other than the concave surface (for example, the diverse LRR module organization and irregular binding surfaces in TLR family<sup>27</sup> and plant LRR proteins<sup>28</sup>). The idealized L24 repeat structure (DLRR\_B) was chosen as a base scaffold because it had the highest stability among the three idealized LRRs.

For the wedge module design, L24→[any-length-of-LRR]→L24 triples were retrieved from the LRRML database<sup>29</sup> to identify irregular modules flanked by the L24 modules. A total of 21 unique irregular modules were identified. We selected the 32-length LRR unit (L32) found in the Toll-like receptor 3 structure<sup>30</sup> (PDB ID: 2A0Z, 532–563) as a starting point. L32 has a relatively rigid and structured loop located on its convex surface which could be useful in future binding pocket designs. The junction module design process was applied to the two fusion interfaces (L24→L32 and L32→L24) which resulted in the wedge module JN<sub>L24→L32→L24</sub> (DLRR\_H). Four designs for L24→L32→L24 were selected and experimentally characterized (**Table 2**).

All designs were expressed in *E.coli* and found to be soluble. Two designs were monomeric in SEC-MALS experiment. Thermal denaturation experiments showed that insertion of the wedge module generally decreased stability of the base scaffold, but unfolding was still cooperative (**Fig. 2d**; **Supplementary Fig. 4**). The crystal structure of DLRR\_H2 determined at 2.9 Å resolution was consistent to the design model (C $\alpha$  RMSD 0.9 Å), confirming the accuracy of the junction module design protocol (**Fig. 3e**).

### Step 3: Curvature specification by general module assembly

The crystal structures described thus far demonstrate that the building block modules (L22, L24, L28, and L29), junction modules (L22→L24, L24→L22, L24→L28, L28→L29, and L29→L28) and wedge modules (L24→L32→L24) all have structures that are very similar to the design models regardless of overall protein context. In principle, this enables design of combinations of modules to achieve a desired curvature. We represent the space of possible LRR structures as a network consisting of building block modules (nodes) connected by junction modules (edges) as in Figure 1b (**Fig. 4a**). Any sequence of modules generated by following the edges in the network corresponds to an LRR structure with unique curvature. For example, all the 18,786 possible fusion structures consisting of 12

modules are depicted in **Figure 4b** as lines connecting the center of masses for each repeat module in the structure. The curvature diversity is orders of magnitude greater than that of the original LRRs containing the same number of building block modules.

We chose to use models of the individual building block and junction modules extracted from the crystal structures described thus far in the general module assembly process rather than the original design models of these units. While the building block modules are similar to previously described structures, the designed junction modules have sequences (**Supplementary Fig. 5a**) and structures (**Supplementary Fig. 5b**) quite different from previously described LRRs. Because of the imperfect state of computational protein design, we consider the crystal structures (which differ from the design models by C $\alpha$  RMSD 0.2–1.0 Å) to be more accurate representations of the structures these modules are likely to adopt in new designs (**Supplementary Fig. 5c**).

### General module assembly and experimental characterization

As a proof-of-concept for general module assembly, we designed four multiple fusion constructs (DLRR\_I, DLRR\_J, DLRR\_K, and DLRR\_L in **Fig. 4c**). The designs contain more than two fusion interfaces which result in large super-helical structures comparable in size to TLR4<sup>31</sup> (PDB ID: 3FXI, 626 residues) and plant steroid receptor BRI1<sup>32</sup> (PDB ID: 3RIZ, 743 residues). The module organization and module origins for each design are shown in **Table 2**, **Figure 4c**, and **Supplementary Table 2**.

Experimental characterization showed that the general module assembly protocol is quite robust. All of the multiple fusion designs were expressed in *E. coli* and found to be soluble, monomeric with well-defined CD spectra, cooperative unfolding transitions, and high thermal stability (**Fig. 4c**; **Supplementary Fig. 1e**). This is notable as all are quite large and complex proteins. We succeeded in solving the crystal structures of two of the designs as described in the following two paragraphs.

Design DLRR\_I contains two successive L32 wedge modules with multiple flanking N and C-terminal L24 modules. We solved the crystal structure of DLRR\_I at 1.7 Å resolution (**Fig. 4d**). Consistent with the assumption of context independent structure of the individual modules, the two L32 wedge modules in DLRR\_I and the single L32 wedge module in DLRR\_H2 are nearly identical over the backbone and core side chains (C $\alpha$  RMSD 0.3–0.5 Å). Over the full 10 repeat unit structure, the crystal structure is closer to the model (C $\alpha$  RMSD 0.5 Å, **Supplementary Fig. 5c**) assembled from the crystal structures of the individual building block and junction modules extracted from DLRR\_B and DLRR\_H than to the model (C $\alpha$  RMSD 1.7 Å, **Fig. 4d**) assembled from the design models of the individual modules (**Table 2**), supporting our decision to use the crystal structures of the building blocks rather than the original design models in the general module assembly calculations.

Design DLRR\_K consists of two L24 modules followed by the L32 module, three additional L24 modules, the L24→L28 junction module, and three {L28→L29} modules—a total of 15 repeat units. Such complexity of module organization is rarely if ever observed in naturally occurring LRRs. The protein is monomeric and stable, with a T<sub>m</sub> of 75 °C. The crystal structure of DLRR\_K at 2.8 Å resolution is very close to the general module assembly

model (built from crystal structures of the individual modules from previous structures), with a C $\alpha$  RMSD of 1.1 Å (**Fig. 4e**).

Taken together, these data suggest that general module assembly based on designed building block and junction modules can produce new structures with predefined shapes with high robustness and accuracy.

## Discussion

We have described a general approach to creating repeat proteins with custom-designed shape through combination of designed building block and junction modules. The generation of scaffolds with defined curvatures using our computational approach is very likely simpler than that which occurred during the complex evolution of naturally occurring LRRs, and is considerably more controlled than what can be achieved in library selection approaches. The strategy allows the ready programming of a rich diversity of scaffolds with distinct curvatures: over 18,000 distinct 12 repeat unit structures can in principle be generated with our current set of building block and junction modules (**Fig. 4b; Supplementary Table 3**). The stable and well-expressed DLRR\_L design (**Fig. 4c**) has a complex organization with five different types of modules (19 repeat units) in total; for this length there are over 5,000,000 distinct possibilities with our current module set, and increasing the repertoire of idealized building block and junction module designs would enrich the curvature diversity still further.

Our approach integrates protein structural analysis with energy driven design calculations to arrive at the idealized building block and junction modules, and computation and experiment to achieve high accuracy models of the complex repeat proteins generated by the module assembly process. While a completely energy driven approach would be preferable on aesthetic grounds, making use of information extracted from naturally occurring LRRs and from the crystal structures of idealized LRRs described in this study allows the generation of large families of LRR proteins with tunable curvatures to address current challenges. The critical role of computation in the overall process is illustrated by the junction modules: both the sequences (**Supplementary Fig. 5a**) and the structures (**Supplementary Fig. 5b**) of the designed junction modules differ considerably from their closest counterparts in naturally occurring LRRs and hence could not have been obtained without energy driven design calculations. These calculations are not perfect, however, and since the small differences between the design models and the corresponding crystal structures are amplified through lever arm effects when many modules are combined, we use crystal structures of the designed building block and junction modules in the general module assembly calculations rather than the original design models.

The ability to custom design repeat proteins with well-defined shapes and curvatures has immediate application to the design of a next generation of high affinity binding proteins. Studies of native protein-protein interactions have shown that shape complementarity is a major determinant of protein binding affinity<sup>33-36</sup>. In particular, naturally occurring LRR-based binding proteins often achieve high affinity and specificity by having shapes closely conforming to the surfaces of the target proteins. The importance of this shape tuning for

LRR protein molecular recognition is illustrated in **Supplementary Figure 5d** for the naturally occurring LRR proteins internalin A (InlA) and ribonuclease inhibitor (RI). Each protein has a curvature adapted to its target (E-cadherin and ribonuclease A, respectively) resulting in well-packed complementary protein-protein interfaces with hot-spot clusters at both the N and C termini. In contrast, swapping the respective target for each of the LRR proteins, namely InlA:ribonuclease and RI:E-cadherin complexes, results in significant clashes and large gaps.

With the capability provided by the approach described in this paper, it is now possible to design novel proteins with high backbone shape complementarity to essentially any macromolecular target of interest. Coupled with protein interface design methodology previously used to create new binding proteins based on already existing scaffolds<sup>37,38</sup>, this should allow the design of high affinity and specificity binding proteins. Such an approach complements directed evolution methods<sup>13,39,40</sup> for obtaining high affinity binding proteins based on a single stable protein backbone which although powerful still require considerable effort. For creating a high affinity binding protein to a target of interest in the near future, a combination of our shape complementary scaffold design approach, protein-protein interface design for chemical complementarity, and limited directed evolution to optimize interactions not accurately described by computational design may prove particularly effective.

## Methods

### Markov transition model for natural LRR modules

To construct a Markov transition model for natural LRR modules, all set of two consecutive LRR modules were collected from the LRRML database<sup>29</sup> and labeled based on the module length. From these data, we computed the transition probability  $P_{a \rightarrow b} = N_{a \rightarrow b} / \sum_i N_{a \rightarrow i}$ , where  $N_{a \rightarrow i}$  represents the frequency of transitions from module length  $a$  to  $b$  in the PDB. In the network model in **Figure 1b**, the size of a node was scaled by frequency of a module length in the PDB, and the thickness of an edge by the transition probability.

### Computational design of junction modules

The initial fusion models were generated by RosettaCM<sup>24</sup> from the motif-aligned scaffolds as described in the main text (**Fig. 2a**), and refined with Rosetta relax protocol with coordinate constraints<sup>41</sup> to reduce perturbation of the structure. The fusion interface between the two heterogeneous building blocks was redesigned to improve structural compatibility using Rosetta FastRelax protocol. The protocol runs four cycles of repack, design and minimization, and during each cycle the weight for the repulsive energy term gradually increases to obtain well-packed and low-energy structure. During the design procedure, residue type constraints were added in order to favor original residue identities. After generating 1,000 design sequences, the top 10% of design sequences both by Rosetta energy and packing were retrieved and manually inspected to select the final sequences.

### LRR structure modeling by iterative module assembly

Building block and junction module structures were extracted from the crystal structures of the designed LRR proteins containing one or two building block module types. Specifically,



two-unit or three-unit module structures of Ncap-L22 (DLRR\_A), L22-L22 (DLRR\_A), Ncap-L24 (DLRR\_B), L24-L24 (DLRR\_B), L22→L24 (DLRR\_B), L24→L22 (DLRR\_E), L24→L28→L29 (DLRR\_G3), L28→L29 (DLRR\_G3), L29→L28 (DLRR\_G3), and L24→L32→L24 (DLRR\_H2) were used to elongate a LRR structure through module assembly mediated by the common flanking module. For example, module assembly of L22→L24 and L24→L22 through the common L24 unit generates the three-unit structure L22→L24→L22. The module assembly was then iteratively applied to elongate the overall structure one by one module, which resulted in the mature form of a general LRR structure. Finally, energy minimization with Rosetta was performed to eliminate potential structural defects. The crystal structure of L22→L24 was obtained from DLRR\_B that has the fusion of L22-containing N-terminal capping domain (of internalin B) to L24. The L22→L28 was not used in the general module assembly due to the lack of the crystal structure.

### Gene cloning, protein expression and purification

Genes encoding building block LRRs were synthesized and cloned into pET21\_NESG (DLRR\_A) or pET15\_NESG (DLRR\_B and DLRR\_C) expression vectors by GeneScript. The gene fragments for each junction module were separately prepared by PCR assembly of six to eight 50–60 nucleotide oligos or by gene synthesis from Integrated DNA Technologies. Another gene fragment for the building block module to be fused was also obtained by PCR. The two gene fragments were then inserted into the plasmid of the appropriate building block protein by Gibson cloning<sup>42</sup>. The C-terminal 6xHis tag was added to all design sequences with Gly-Ser or Gly-Ser-Trp linkers where Trp was for measuring protein concentration easily.

The proteins were expressed in *E. coli* BL21 Star (DE3) cells at 37°C for 4 hours after induction with 0.1 mM IPTG. The cell pellets were resuspended in 20 ml of lysis buffer containing 20 mM Tris, 500mM NaCl, 30 mM imidazole, and 5% v/v glycerol (pH 8.0). Roche complete EDTA-free protease inhibitor tablet, lysozyme (1mg/ml), and DNase (1mg/ml) were also added to the lysis buffer. After sonication, the proteins were purified with a Ni-NTA column and eluted with 20 mM Tris, 500 mM NaCl, 250 mM imidazole (pH 8.0). The proteins were further purified with Superdex 200 column (GE Healthcare) equilibrated in 20 mM Tris and 50 mM NaCl at pH 8.0. The soluble expression and purity were also tested with SDS-PAGE and mass spectrometry (LCQ Fleet Ion Trap Mass Spectrometer, Thermo Scientific).

### Biophysical characterization

Circular Dichroism (CD) using AVIV 62S DA spectrometer was used to investigate secondary structure contents and thermal stability. Far-UV CD spectra from 200 nm to 260 nm were measured for the protein samples in 20 mM Tris and 50 mM NaCl (pH 8.0). Thermal denaturation experiment was also performed by following the minimum at 218 nm and increasing temperature from 25°C to 90°C. Size-Exclusion Chromatography coupled to Multiple Angle Light Scattering (SEC-MALS) was performed to access oligomeric state of protein samples. Superdex 200 10/300 GL column (GE Healthcare) was equilibrated in phosphate buffered saline (PBS) buffer and used on HPLC system (LC 1200 Series, Agilent

Technologies) connected to miniDAWN TREOS static light scattering detector (Wyatt Technologies). The collected data were analyzed by ASTRA software (Wyatt Technology).

### X-ray crystallography

Crystals of designed LRR repeat proteins were grown by standard vapour phase diffusion methods using a TTP labtech 'Mosquito' crystallization robot with 50 nanoliter drops of protein at concentrations ranging from 15 mg/mL to 40 mg/mL equilibrated against 100 volumes of microliter individual reservoir solutions. The reservoir compositions that produced each crystal are provided in **Supplementary Table 4**. Crystals were then flash-cooled by rapid emersion into artificial mother liquors corresponding to the crystallization reservoir solutions supplemented with either ethylene glycol (to 25% v/v) or with PEG3350 (to 35% w/v). Diffraction data were collected on cryocooled crystals using either an in-house CCD area detector with a rotating anode x-ray generator (DLRR\_A, DLRR\_G3, DLRR\_H2, DLRR\_K) or with a CCD area detector at the Advanced Light Source X-ray synchrotron facility (DLRR\_E, DLRR\_I). All data were processed and scaled using program suite HKL2000<sup>43</sup>. Molecular replacement was performed using program PHASER<sup>44</sup> with computational coordinates of the individual designs produced by Rosetta as search models. Model building was performed using COOT<sup>45</sup> and refinement was performed using program REFMAC<sup>46</sup>.

### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

### Acknowledgements

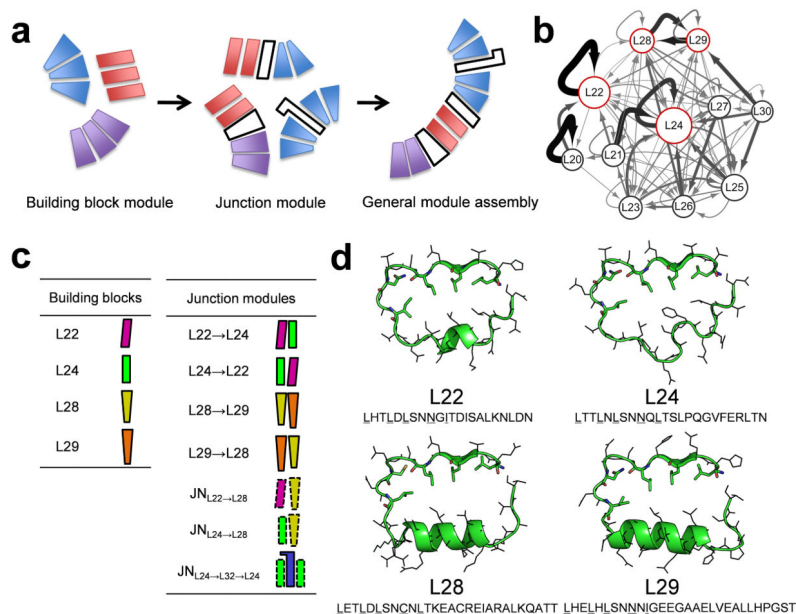
We thank J. Bolduc for data collection for DLRR\_A and the members of the protein production facility at the Institute for Protein Design (Seattle, WA) for protein production. This work was supported by grants from the Defense Threat Reduction Agency (HDTRA1-11-1-0041) (to D.B.) and National Institutes of Health (R01 GM49857) (to B.L.S.).

### References

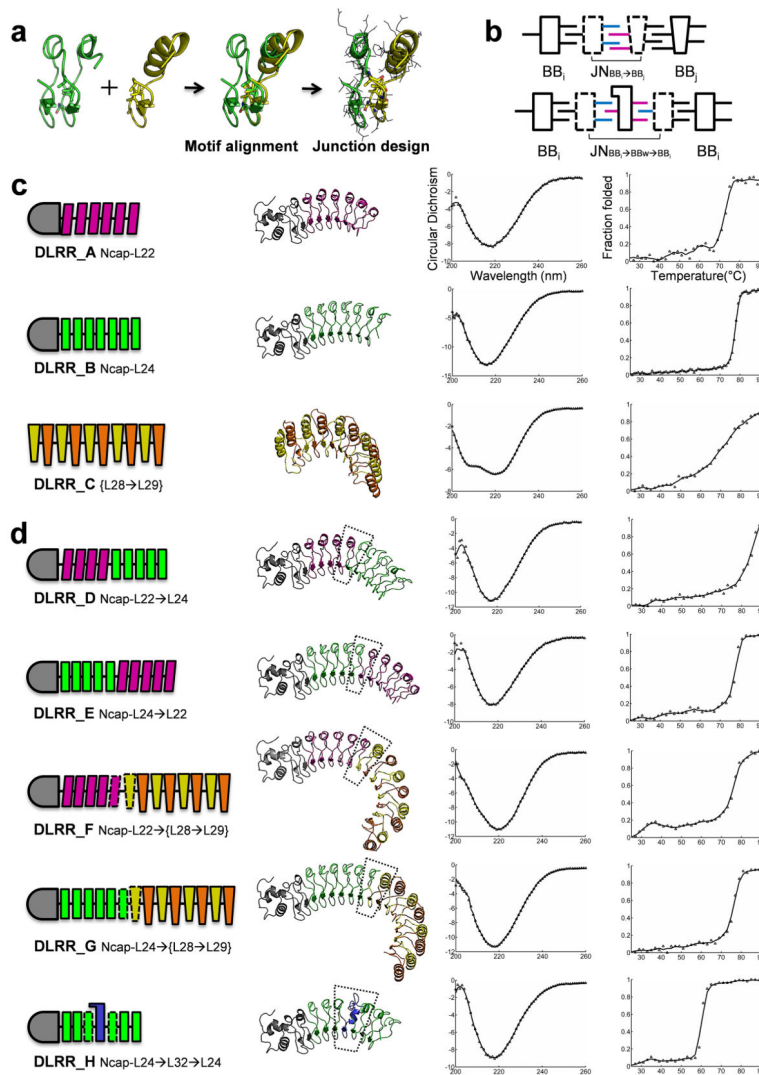
1. Binz HK, Amstutz P, Pluckthun A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotechnol.* 2005; 23:1257–68. [PubMed: 16211069]
2. Skerra A. Alternative non-antibody scaffolds for molecular recognition. *Curr Opin Biotechnol.* 2007; 18:295–304. [PubMed: 17643280]
3. Gebauer M, Skerra A. Engineered protein scaffolds as next-generation antibody therapeutics. *Curr Opin Chem Biol.* 2009; 13:245–55. [PubMed: 19501012]
4. Javadi Y, Itzhaki LS. Tandem-repeat proteins: regularity plus modularity equals design-ability. *Curr Opin Struct Biol.* 2013; 23:622–31. [PubMed: 23831287]
5. Grove TZ, Regan L, Cortajarena AL. Nanostructured functional films from engineered repeat proteins. *J R Soc Interface.* 2013; 10:20130051. [PubMed: 23594813]
6. Phillips JJ, Millership C, Main ER. Fibrous nanostructures from the self-assembly of designed repeat protein modules. *Angew Chem Int Ed Engl.* 2012; 51:13132–5. [PubMed: 23150207]
7. Han SH, Lee MK, Lim YB. Bioinspired self-assembled peptide nanofibers with thermostable multivalent alpha-helices. *Biomacromolecules.* 2013; 14:1594–9. [PubMed: 23550841]
8. Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol.* 2001; 11:725–32. [PubMed: 11751054]

9. Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, Matsushima N. Structural principles of leucine-rich repeat (LRR) proteins. *Proteins*. 2004; 54:394–403. [PubMed: 14747988]
10. Kim HM, et al. Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. *Cell*. 2007; 130:906–17. [PubMed: 17803912]
11. Parker R, Mercedes-Camacho A, Grove TZ. Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Sci*. 2014; 23:790–800. [PubMed: 24659515]
12. MauLee SC, et al. Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proc Natl Acad Sci U S A*. 2012; 109:3299–304. [PubMed: 22328160]
13. Binz HK, et al. High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat Biotechnol*. 2004; 22:575–82. [PubMed: 15097997]
14. Forrer P, Stumpp MT, Binz HK, Pluckthun A. A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett*. 2003; 539:2–6. [PubMed: 12650916]
15. Grove TZ, Cortajarena AL, Regan L. Ligand binding by repeat proteins: natural and designed. *Curr Opin Struct Biol*. 2008; 18:507–15. [PubMed: 18602006]
16. Main ER, Xiong Y, Cocco MJ, D'Andrea L, Regan L. Design of stable alpha-helical arrays from an idealized TPR motif. *Structure*. 2003; 11:497–508. [PubMed: 12737816]
17. Filipovska A, Rackham O. Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Mol Biosyst*. 2012; 8:699–708. [PubMed: 22234420]
18. Reichen C, Hansen S, Pluckthun A. Modular peptide binding: from a comparison of natural binders to designed armadillo repeat proteins. *J Struct Biol*. 2014; 185:147–62. [PubMed: 23916513]
19. Mak AN, Bradley P, Cernadas RA, Bogdanove AJ, Stoddard BL. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science*. 2012; 335:716–9. [PubMed: 22223736]
20. Ryou JH, Park K, Lee JJ, Kim D, Kim HS. Soluble expression of human glycoprotein Ibalpha in *Escherichia coli* through replacement of the N-terminal capping domain. *Protein Expr Purif*. 2014; 101C:21–27. [PubMed: 24925643]
21. Jung K, et al. Toll-like receptor 4 decoy, TOY, attenuates gram-negative bacterial sepsis. *PLoS One*. 2009; 4:e7403. [PubMed: 19816595]
22. Parmeggiani F, et al. A general computational approach for repeat protein design. *Journal of Molecular Biology*. in press.
23. Huang PS, et al. RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One*. 2011; 6:e24109. [PubMed: 21909381]
24. Song Y, et al. High-resolution comparative modeling with RosettaCM. *Structure*. 2013; 21:1735–42. [PubMed: 24035711]
25. Leaver-Fay A, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*. 2011; 487:545–74. [PubMed: 21187238]
26. Sheffler W, Baker D. RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci*. 2009; 18:229–39. [PubMed: 19177366]
27. Park BS, Lee JO. Recognition of lipopolysaccharide pattern by TLR4 complexes. *Exp Mol Med*. 2013; 45:e66. [PubMed: 24310172]
28. Matsushima N, Miyashita H. Leucine-Rich Repeat (LRR) Domains Containing Intervening Motifs in Plants. *Biomolecules*. 2012; 2:288–311. [PubMed: 24970139]
29. Wei T, et al. LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs). *BMC Struct Biol*. 2008; 8:47. [PubMed: 18986514]
30. Bell JK, et al. The molecular structure of the Toll-like receptor 3 ligand-binding domain. *Proc Natl Acad Sci U S A*. 2005; 102:10976–80. [PubMed: 16043704]
31. Park BS, et al. The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. *Nature*. 2009; 458:1191–5. [PubMed: 19252480]
32. Hothorn M, et al. Structural basis of steroid hormone perception by the receptor kinase BRI1. *Nature*. 2011; 474:467–71. [PubMed: 21666665]

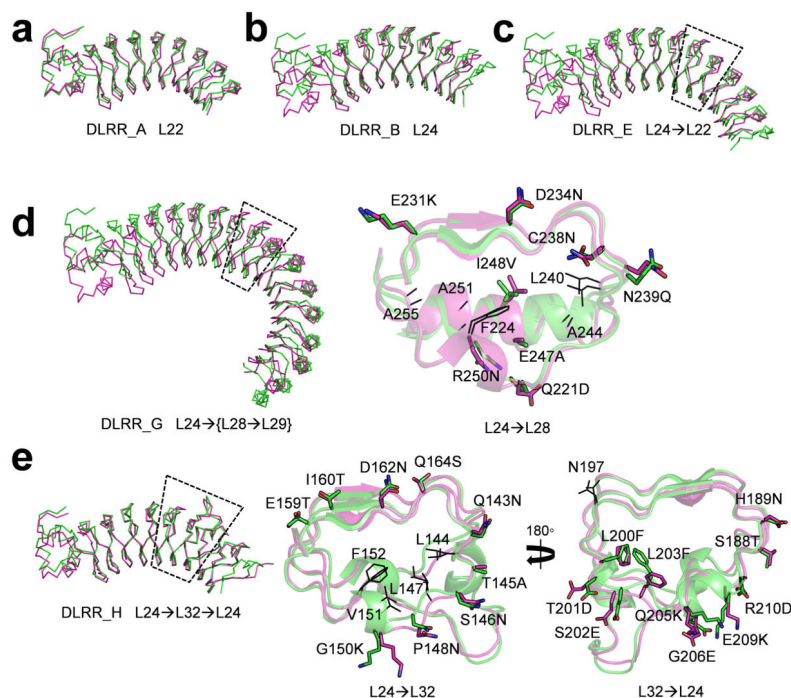
33. Chen R, Weng Z. A novel shape complementarity scoring function for protein-protein docking. *Proteins*. 2003; 51:397–408. [PubMed: 12696051]
34. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*. 1997; 272:106–20. [PubMed: 9299341]
35. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. *J Mol Biol*. 1993; 234:946–50. [PubMed: 8263940]
36. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*. 1996; 93:13–20. [PubMed: 8552589]
37. Fleishman SJ, et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*. 2011; 332:816–21. [PubMed: 21566186]
38. Procko E, et al. Computational design of a protein-based enzyme inhibitor. *J Mol Biol*. 2013; 425:3563–75. [PubMed: 23827138]
39. Epa VC, et al. Structural model for the interaction of a designed Ankyrin Repeat Protein with the human epidermal growth factor receptor 2. *PLoS One*. 2013; 8:e59163. [PubMed: 23527120]
40. Lee JJ, et al. A high-affinity protein binder that blocks the IL-6/STAT3 signaling pathway effectively suppresses non-small cell lung cancer. *Mol Ther*. 2014; 22:1254–65. [PubMed: 24682171]
41. Nivon, LG, Moretti R, Baker D. A Pareto-optimal refinement method for protein design scaffolds. *PLoS One*. 2013; 8:e59004. [PubMed: 23565140]
42. Gibson DG, et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009; 6:343–5. [PubMed: 19363495]
43. Zbyszek Otwinowski WM. Processing of X-ray diffraction data collected in oscillation mode. *Methods in Enzymology*. 1997; 276:20.
44. McCoy AJ, et al. Phaser crystallographic software. *J Appl Crystallogr*. 2007; 40:658–674. [PubMed: 19461840]
45. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*. 2004; 60:2126–32. [PubMed: 15572765]
46. Winn, MD, Murshudov GN, Papiz MZ. Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods Enzymol*. 2003; 374:300–21. [PubMed: 14696379]

**Figure 1.**

(a) Overview of curvature-tunable scaffold design: idealized building block module design, junction module design, and general module assembly. (b) Module organization of natural LRR modules is represented by a network where nodes represent modules and edges transitions between modules. The size of nodes and the thickness of edges are proportional to the frequencies observed in the PDB. (c) Graphical representation of building block and junction modules. (d) Idealized building block module structures and sequences. The highly conserved residues are shown in sticks.

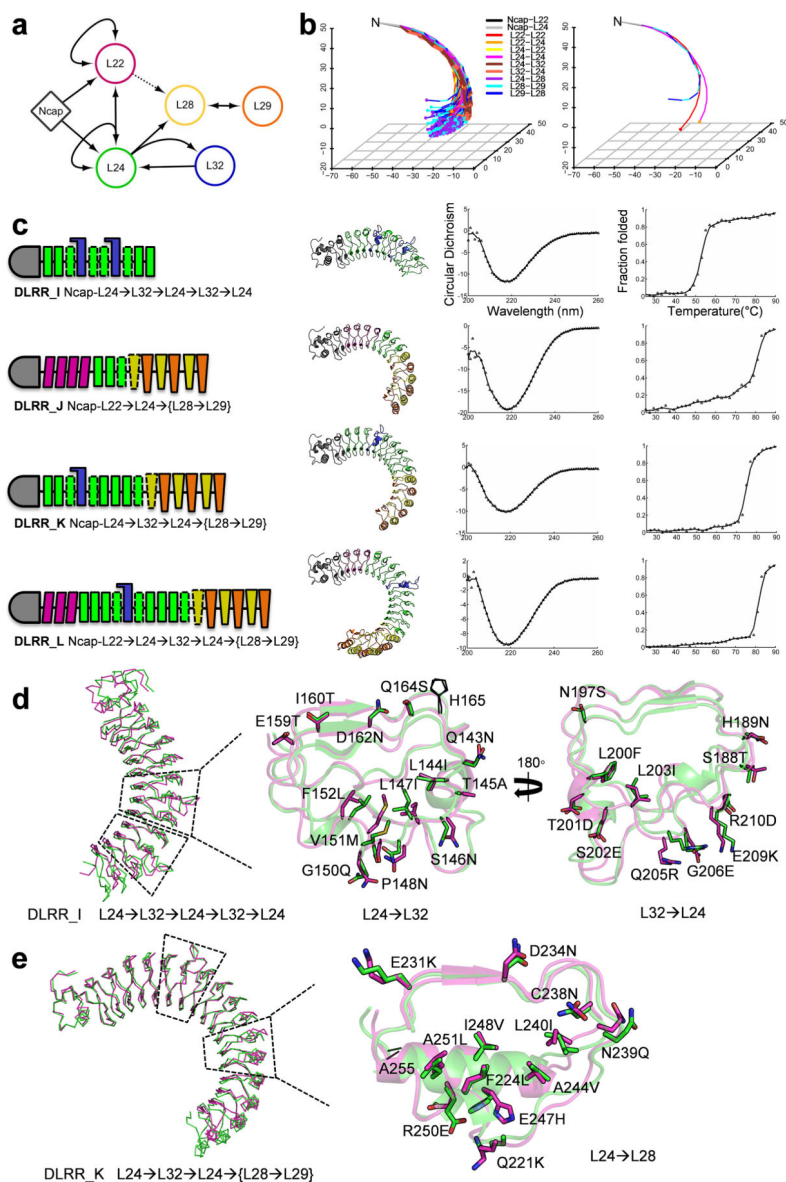


**Figure 2.** Building block and junction module design. (a) Junction module design illustrated by the L24→L28 fusion design. (b) Schematic representing a two-unit junction module  $JN_{BB_i} \rightarrow BB_j$  and a three-unit wedge module  $JN_{BB_i} \rightarrow BB_w \rightarrow BB_j$ . (c) The idealized building block modules for L22, L24, and {L28→L29}. (d) The junction modules for L22→L24, L24→L22, L22→L28 (DLRR\_F6), L24→L28 (DLRR\_G2), and L24→L32→L24 (DLRR\_H2). For each design, module organization, model structure, far-UV CD spectra (mean residue ellipticity,  $[\theta] \times 10^{-3} \text{ deg cm}^2 \text{ dmol}^{-1}$ ), and thermal denaturation profile following the wavelength 218 nm are displayed from left to right. When multiple designs exist, data of the most stable one are shown. The backbone modules are indicated in different colors: magenta-L22, green-L24, yellow-L28, and orange-L29 (Fig. 1c). The designed junction modules are represented by dotted lines, and the N-terminal capping domain connected to L22 or L24 is shown in gray.



**Figure 3.**

Crystal structures of the idealized building block and junction module designs. The design models (green) for (a) L22, (b) L24, (c) L24→L22, (d) L24→L28 (DLRR\_G3), and (e) L24→L32→L24 (DLRR\_H2) superimposed on the crystal structures (magenta). Close-up views for the designed junction modules (dashed region) are shown for (d) and (e). Residues mutated from the original building block sequences are annotated and shown as sticks. Additional residues that vary within the designs are shown as lines. Except for that of DLRR\_A, the crystal structures have missing electron density at the C-terminus (10–20 amino acids). PyMOL (<http://www.pymol.org/>) are used in all structural visualization.

**Figure 4.**

Control of curvature by general module assembly (a) Network representing the generation of a new LRR structure using the idealized building block modules (nodes) and the designed junction modules (edges). The edge for L22→L28 is represented by a dotted line due to the lack of crystal structure, and not used in the module assembly. (b) Curvature diversity of general 12-unit LRR structures (left) is compared to that of the three building block structures (right). The LRR structures are generated by arbitrary module assembly based on the network (a), and the lines connecting the center of masses for each module are only shown for clarity. The structures are aligned based on the N-cap domain, and the C-termini are represented by dots colored by the relevant LRR modules. (c) Module organization, model structure, far-UV CD spectra, and thermal denaturation results for the four multiple fusion designs. (d,e) Comparison between the design model (green) and the crystal structure



(magenta) for DLRR\_I (d) and DLRR\_K (e) with a close-up view of the fusion interfaces. Since the two L32 modules in DLRR\_I are nearly identical with a C $\alpha$  RMSD 0.2 Å, only the first L32 is shown.

Table 1

Data collection and refinement statistics (molecular replacement)

Crystal	DLRR_A	DLRR_E	DLRR_G3	DLRR_H2	DLRR_I	DLRR_K
<b>Data collection</b>						
Space group	P2 <sub>1</sub>	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	F222	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	C2	P2 <sub>1</sub> 2 <sub>1</sub>
Cell dimensions						
<i>a, b, c</i> (Å)	57.66, 245.07, 57.73	32.12, 77.71, 101.89	91.13, 136.38, 161.74	89.78, 96.50, 136.36	109.49, 42.71, 67.82	36.87, 93.37, 126.24
<i>α, β, γ</i> (°)	90, 115.36, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 102.4, 90	90, 90, 90
Resolution (Å)	50(2.36)*	42.6(1.93)	23.5(2.53)	50(2.9)	50(1.73)	50(2.8)
<i>R</i> <sub>sym</sub>	0.081(0.183)	0.063(0.171)	0.067(0.153)	0.092(0.529)	0.076(0.252)	0.192(0.742)
<i>I</i> / <i>σI</i>	24.0(6.4)	17.7(8.0)	15.5(4.1)	17.2(3.85)	33.7(4.5)	8.9(2.3)
Completeness (%)	96.7(83.7)	99.8(96.0)	98.1(85.5)	99.8(99.6)	96.3(83.7)	99.7(99.2)
Redundancy	5.7(3.0)	6.4(5.0)	4.5(1.9)	7.2(7.1)	10.3(2.3)	6.2(5.8)
<b>Refinement</b>						
Resolution (Å)	50(2.36)	42.6(1.93)	23.5(2.53)	50(2.9)	50(1.73)	50(2.8)
No. reflections	34180	19993	17061	25484	31150	10729
<i>R</i> <sub>work</sub> (%)	18.9(22.3)	15.86(17.50)	18.47(23.4)	21.16(32.8)	17.07(21.50)	20.75(28.4)
<i>R</i> <sub>free</sub> (%)	24.2(27.7)	22.38(23.7)	24.65(36.1)	25.15(48.5)	21.99(31.70)	28.53(36.0)
No. atoms						
Protein	6771	2388	3456	7841	2577	3582
Ligand/ion	8	12	29	20	---	1
Water	230	106	96	1	199	18
<i>B</i> -factors						
Protein	12.13	14.65	11.73	69.14	10.84	16.96
Ligand/ion	35.53	39.67	54.64	85.26	---	42.89
Water	18.0	35.39	21.91	50.44	25.43	18.60
r.m.s. deviations						
Bond length (Å)	0.0137	0.0181	0.0138	0.0126	0.0194	0.0136
Bond angles (°)	1.661	1.810	1.629	1.651	2.052	1.475

\* Values in parentheses are for highest-resolution shell.

Table 2

Summary of fusion designs and experimental results

Design Name	Module organization <sup>§</sup>	Modules <sup>¶</sup> (repeat units)	designs tested	Soluble	Folded (CD)	Mono meric	X- ray	T <sub>m</sub> (°C)	RMSD <sup>1</sup>	RMSD <sup>2</sup>
DLRR_A	Ncap-L22 <sup>6</sup>	6 (6)	1	1	1	1	1	73	1.4 (0.8, 2.0)	0.4 (0.5, 1.0)
DLRR_B	Ncap-L24 <sup>7</sup>	7 (7)	1	1	1	1	1	78	1.7 (1.5, 2.9)	0.3 (0.6, 0.4)
DLRR_C	{L28→L29} <sup>5</sup>	5 (10)	5	5	1	#		71		
DLRR_D	Ncap-L22 <sup>4</sup> →L24 <sup>5</sup>	9 (9)	1	1	1	1		87		
DLRR_E	Ncap-L24 <sup>5</sup> →L22 <sup>5</sup>	10 (10)	1	1	1	1	1	77	2.1 (1.4, 2.0)	0.4 (0.7, 0.7)
DLRR_F	Ncap-L22 <sup>4</sup> -JN <sub>L22</sub> →L28→L29→{L28→L29} <sup>3</sup>	9 (13)	6	6	6	6		77		
DLRR_G	Ncap-L24 <sup>5</sup> -JN <sub>L24</sub> →L28→L29→{L28→L29} <sup>3</sup>	10 (14)	6	6	6	6	1	81	2.6 (3.1, 3.8)	0.8 (0.8, 2.2)
DLRR_H	Ncap-L24 <sup>2</sup> -JN <sub>L24</sub> →L32→L24-L24 <sup>2</sup>	5 (7)	4	4	4	2	1	65	0.9 (0.5, 1.0)	0.8 (0.7, 1.2)
DLRR_I	Ncap-L24 <sup>2</sup> -JN <sub>L24</sub> →L32→L24-JN <sub>L24</sub> →L32→L24-L24 <sup>2</sup>	6 (10)	1	1	1	1	1	53	1.7 (1.2, 2.3)	0.5 (0.5, 0.7)
DLRR_J	Ncap-L22 <sup>4</sup> →L24 <sup>2</sup> -JN <sub>L24</sub> →L28→L29→{L28→L29} <sup>2</sup>	10 (13)	1	1	1	1		82		
DLRR_K	Ncap-L24 <sup>2</sup> -JN <sub>L24</sub> →L32→L24-L24 <sup>3</sup> - JN <sub>L24</sub> →L28→L29→{L28→L29} <sup>2</sup>	10 (15)	1	1	1	1	1	75	†	1.1 (1.2, 3.9)
DLRR_L	Ncap-L22 <sup>3</sup> →L24 <sup>3</sup> -JN <sub>L24</sub> →L32→L24-L24 <sup>3</sup> - JN <sub>L24</sub> →L28→L29→{L28→L29} <sup>2</sup>	14 (19)	1	1	1	1		83		

T<sub>m</sub> is estimated by calculating the inflection point of the melting curve at 218 nm, and the highest T<sub>m</sub> value is represented when multiple designs exist.

RMSD<sup>1</sup>: C $\alpha$  RMSD (Å) between crystal structure and model generated from design models of building blocks and junction modules.

RMSD<sup>2</sup>: C $\alpha$  RMSD (Å) between crystal structure and model generated using crystal structures of building block and junction modules (Supplementary Fig. 5c)

RMSD values for the first and the last unit in global structure alignment are provided in parenthesis.

<sup>§</sup>The superscripts represent the number of repeat units

<sup>¶</sup>Alternatively occurring two-unit {L28→L29} is considered as one module.

Experimental data of DLRR\_B is from Parmeggiani et al.<sup>22</sup>

HHMI Author Manuscript

HHMI Author Manuscript

HHMI Author Manuscript

#DLRR\_C forms a dimer

† Model of DLRR\_K is generated by module assembly without an initial design model.