Accurate *de novo* design of hyperstable constrained peptides

Gaurav Bhardwaj[1,2,*], Vikram Khipple Mulligan[1,2,*], Christopher D. Bahl[1,2,*], Jason M. Gilmore[1,2], Peta J. Harvey[3], Olivier Cheneval[3], Garry W. Buchko[4], Surya V.S.R.K. Pulavarti[5], Quentin Kaas[3], Alexander Eletsky[5], Po-Ssu Huang[1,2], William A. Johnsen[6], Per Greisen[1,2,7], Gabriel J. Rocklin[1,2], Yifan Song[1,2,8], Thomas W. Linsky[1,2], Andrew Watkins[9], Stephen A. Rettie[2], Xianzhong Xu[5], Lauren P. Carter[2], Richard Bonneau[10,11], James M. Olson[6], Evangelos Coutsias[12], Colin E. Correnti[6], Thomas Szyperski[5], David J. Craik[3], and David Baker[1,2,13]

*These authors contributed equally to this work.

Affiliations:

[1]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA

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

[3]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072 Australia

[4]Seattle Structural Genomics Center for Infectious Diseases and Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA

[5]Department of Chemistry, State University of New York at Buffalo, Buffalo, New York 14260, USA

[6]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA

[7]Global Research, Novo Nordisk A/S, DK-2760 Måløv, Denmark

[8]Cyrus Biotechnology, Seattle, Washington 98109, USA

[9]Department of Chemistry, New York University, New York, NY 10003, USA

[10]Department of Biology, New York University, New York, NY 10003, USA

[11]Center for Computational Biology , Simons Foundation , NY, NY 10010

[12]Applied Mathematics and Statistics and Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, USA

[13]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA

Corresponding Author: David Baker (**dabaker@uw.edu**)

Summary

Naturally occurring, pharmacologically active peptides constrained with covalent cross-links generally have shapes evolved to fit precisely into binding pockets on their targets and can have excellent biopharmaceutical properties, combining the stability and tissue penetration of small molecule drugs with the specificity of much larger protein therapeutics. The ability to design constrained peptides with precisely specified structures would enable the design of shape complementary inhibitors for a wide range of targets. Here we describe the development of computational methods for *de novo* design of conformationally-restricted peptides, and the use of these methods to design 15-50 residue disulfide-crosslinked and heterochiral N-C backbone-cyclized peptides. These peptides are exceptionally stable to temperature and chemical denaturation, and twelve experimentally determined X-ray and NMR structures are nearly identical to the computational models. The computational design methods and stable scaffolds provide the basis for development of a new generation of peptide-based drugs.

Main Text

The vast majority of drugs currently approved for use in humans are either proteins or small molecules. Lying between the two in size, constrained peptides are an underexplored frontier for drug discovery combining advantages of both classes[1,2]. Naturally-occurring peptides rigidified by disulfide bonds or backbone cyclization, such as conotoxins, chlorotoxin, knottins, and cyclotides, play critical roles in signaling, virulence and immunity and are among the most potent pharmacologically active compounds known[3]. These peptides have pre-stabilized

binding-competent conformations that precisely complement their targets. Inspired by the potency of naturally occurring compounds, there have been considerable efforts to generate new bioactive molecules by re-engineering existing constrained peptides using loop grafting and sequence randomization followed by selection[4]. Although powerful, these approaches are hindered by the limited variety of naturally-occurring constrained peptide structures and the inability to achieve global shape complementarity with targets. There is a clear need for a method of creating constrained peptides with new structures and functions that provides precise control over the size and shape of the designed molecules. A method with sufficient generality to incorporate non-canonical backbones and unnatural amino acids would enable access to broad regions of peptide structure and function space not explored by evolution.

Although there have been recent advances in protein design methodology[5,6], the computational design of covalently-constrained peptides with new structures and non-canonical backbones presents new challenges. In previous protein design work, backbones have been generated either using parametric equations or by assembling short fragments of known protein structure, neither of which are compatible with non-canonical components. Likewise, most structure prediction methods rely on properties derived from the Protein Data Bank and cannot handle non-canonical backbones. This limitation complicates the use of structure prediction calculations as an *in silico* consistency check (the lowest energy structure found for a designed sequence in structure prediction calculations should be the design model). Hence, both backbone generation and design validation require new backbone sampling methods. Furthermore, methods are needed for incorporation of multiple geometric constraints (auxiliary covalent bonds) without introduction of conformational strain. Finally, energy calculations must correctly model amino acid chirality.

Here we describe the development of new computational methods that meet these challenges, opening this exciting frontier to computational design. We demonstrate the power of the methods by designing a structurally diverse array of 15-50 residue peptides spanning two broad categories: (i) genetically encodable disulfide-rich peptides, and (ii) heterochiral peptides with non-canonical architectures and sequences. To explore the folds accessible to genetically encoded constrained peptides under 50 amino acids, we selected nine topologies: αα, ααα, βαβ, ββα, αβββ, βαββ, ββαβ, βββα, and βββββββ (**Fig. 1**; we define a "topology" as the ordered sequence of secondary structure elements in the folded peptide). To explore the expanded design space revealed by inclusion of non-canonical amino acids and backbone cyclization, we sought to cover all topologies containing two to three canonical secondary structure elements: αα, ααα, ββα, βαβ, αββ, and ββ. To explore the expanded conformational space of heterochiral peptides further, we also designed a cyclic topology with right- and left-handed helices: $\alpha_L\alpha_R$. This broad array of structures provides a range of starting points for structure-based drug design, and the computational methods complement peptide drug development efforts that use high-throughput biological libraries (*e.g.* yeast and phage display) and/or synthetic peptide libraries (*e.g.* split-and-pool solid phase methods).

All of the design calculations described in this paper were carried out with the Rosetta software suite[7] and follow the same basic approach. Large numbers of peptide backbones are stochastically generated as described in the following sections, combinatorial sequence design calculations are carried out to identify sequences (including disulfide cross-links) stabilizing each backbone conformation, and the designed sequence-structure pairs are assessed by determining the energy gap between the designed structure and the lowest energy alternative structures found in large scale structure prediction calculations starting from the designed sequence. A subset of the designs in deep energy minima are then produced in the laboratory, and their stabilities and structures are experimentally determined.

Genetically encodable disulfide-constrained peptides

We first sought to create genetically encodable peptides with folds stabilized by disulfide bonds. The advantages of genetic encodability for downstream applications are that powerful selection methods that couple phenotype to genotype, such as phage display, ribosome display, and yeast display, can readily be applied to optimize binding affinity, and such peptides can readily be produced *in vivo*.

To design new genetically encodable peptides, for each topology we created a "blueprint" specifying the lengths of each secondary structure and connecting loop (see **Methods**). Ensembles of backbone conformations were generated for each blueprint by Monte Carlo-based assembly of short protein fragments[8], or in case of αα and ααα topologies, by varying the parameters in parametric generating equations[9]. The backbones were scanned for sites capable of hosting near-ideal geometry disulfide bonds (see **Methods**), and 1 to 3 disulfide bonds were incorporated. Low-energy amino acid sequences were designed for each disulfide-crosslinked backbone using iterative rounds of Monte Carlo-based combinatorial sequence optimization while allowing the backbone and disulfide linkages to relax in the Rosetta all-atom force field (see **Methods**). Except for the βαββ topology, we performed no manual amino acid sequence optimization. Rosetta *ab initio* structure prediction calculations were carried for each designed sequence, and synthetic genes were obtained for a diverse set of 130 for which the target structure was in a deep global free energy minimum (**Fig. 2a**,**b**).

Disulfide bond-containing peptides are unlikely to fold in the reducing environment of the cytoplasm, so we developed a new *Escherichia coli* expression and secretion system (see **Methods**; expression screening was also performed for a subset of designs using the

mammalian cell culture-based Daedalus expression system[10]). Following pilot-scale purification, disulfide formation was assessed by a gel-shift assay under reducing and nonreducing conditions, and secondary structure was assessed by circular dichroism (CD) spectroscopy. Twenty nine of the designs exhibited redox-sensitive HPLC migration and/or a CD spectrum consistent with the designed topology (see **Supplementary Document 4**). All twenty nine contain at least one non-alanine hydrophobic residue on each secondary structure element contributing van Der Waal interactions in the core; these hydrophobic core residues likely are important for proper peptide folding. We chose one representative design from each topology for large-scale, high-purity production and detailed biochemical characterization. As eight of the nine topologies contained four or more cysteine residues, we used multiple-stage mass spectrometry to investigate the disulfide connectivity, and in all cases the data were consistent with the designed connectivity (see **Supplementary Document 3**).

The stability of the designs to thermal and chemical denaturation was assessed by CD spectroscopy. Samples were heated to 95°C (**Fig. 2d**), or incubated with increasing concentrations of guanidinium hydrochloride (GdnHCl) (**Fig. 2e**). The contribution of disulfide bonds to protein folding was assessed by incubating samples with a ~100-fold molar excess of the reductant tris(2-carboxyethyl)phosphine (TCEP). Design gHH_44 consists of two α-helices with a single disulfide bond connecting the termini, and partial helical structure was retained following reduction with TCEP. Design gHEEE_02, from topology αβββ, was both thermostable and completely resistant to chemical denaturation, even at saturating concentrations (6 M) of GdnHCl. This design contains three disulfide bonds, with each secondary structure element participating in at least one disulfide bond, and no two secondary structure elements sharing more than one disulfide bond. Design gEEEH_04, representing topology βββα, has two (of the three total) disulfide bonds linking the N-terminal β-strand to the C-terminal α-helix; this peptide also exhibited robust thermal and chemical stability. The βββββββ design, gEEEEEE_02, was

also both thermally and chemically stable; it consists of two antiparallel β-sheets packing against one another in a sandwich-like arrangement, with each β-sheet stabilized by a disulfide bond linking a mainchain terminus to its adjacent β-strand.

We obtained crystals for design gEHEE_06, and determined the structure to a resolution of 2.09 Å (**Fig. 3**, **Supplementary Information Table S2-2**). The crystals had three-fold non-crystallographic symmetry, and each protomer aligns to the design model with a mean all-atom RMSD of 1.12 Å. All three of the designed disulfide bonds are well-defined by electron density (**Extended Data Fig. 1**), and rotamers of core residues exhibited excellent agreement with the design model. The protein was thermostable and completely resistant to chemical denaturation (**Fig. 2d**,**e**). While gEHEE_06 shares the short-chain scorpion toxin topology, the length of secondary structure elements and loops and the position of the disulfide bonds are entirely divergent from known natural peptides.

As crystallization efforts for other designs were unsuccessful (with phase-separation rather than protein precipitation in crystal drops), we sought to determine structures by nuclear magnetic resonance (NMR) spectroscopy[11]. The designed proteins were expressed in *E. coli* with isotope labels, and structures were solved using standard protocols[12] (see **Methods**). Upfield chemical shifts of the cysteine β-carbons[13] further confirmed the formation of designed disulfide bonds. Design gEEHE_02, from topology ββαβ with one disulfide bond connecting the termini within the β-sheet and two between the α-helix and β-sheet, aligns to the NMR ensemble with a mean all-atom RMSD of 1.44 Å. This design is impervious to both thermal and chemical denaturation, and it remains partially folded in the presence of TCEP. The final three designs are each composed of three secondary structure elements, with termini located at opposite ends of the molecule and two disulfide bonds which connect each terminus to the middle structural element or adjacent loop. The ββα topology design, gEEH_04, has an antiparallel β-sheet and is less

stable than the others, but it has a structure nearly identical to the design model (mean all-atom RMSD of 1.29 Å). Design gEHE_06, from topology βαβ, contains a parallel β-sheet and aligns to the NMR ensemble with an all-atom mean RMSD of 1.95 Å; it is thermally and chemically stable and remains folded in the presence of TCEP. Design gHHH_06, from topology ααα, partially unfolds upon heating to 95°C and fully returns to the folded state upon cooling. The gHHH_06 design model aligns to the NMR ensemble with a mean all-atom RMSD of 1.74 Å**.** Taken together, the X-ray crystallographic and NMR structures demonstrate that our computational approach enables accurate design of protein mainchain conformation, multiple disulfide bonds, and core residue rotamers with atomic-level accuracy.

Synthetic heterochiral disulfide-constrained peptides

We next sought to design shorter, chemically synthesizable, disulfide-constrained peptides incorporating both L- and D-amino acids. Chemical synthesis methods enable the production of peptides containing non-canonical amino acid residues, providing access to an enormously expanded but sparsely explored sequence and conformational space. Since chemical synthesis is labour-intensive, we prioritized the development of automated computational screening techniques, limiting the amount of experimental screening needed to obtain structured designs. For additional confidence in the *in silico* selections, we supplemented Rosetta *ab initio* screening with molecular dynamics (MD) based evaluation.

Previously-established, Rosetta-based protein design tools are compatible with peptides composed of canonical amino acids, but those incorporating non-canonical components require significant extensions to the computational methodology. We addressed these challenges by fully generalizing the Rosetta energy function to support D-amino acids, inverting the torsional potentials used for the equivalent L-amino acids (see **Methods** and **Supplementary**

Information). The Rosetta sequence design algorithms were also extended to enable mixed-chirality design.

Large numbers of disulfide-constrained backbones for topologies αββ, βαβ, and ββα were generated by fragment assembly as described above for genetically encodable peptides. Sequences were designed (permitting D-amino acids at positive-phi positions), and the resultant low-energy designs were evaluated using MD and *ab initio* structure prediction (in the latter, D-amino acid positions were replaced with glycine so the Rosetta fragment-based approach could be used) (see workflow flowchart in **Extended Data Fig. 2**). A single, low-energy design which underwent only small (< 0.5 Å RMSD) fluctuations in the MD simulations (**Extended Data Fig. 3**) and had a significant energy gap in the structure prediction calculations was selected for each topology (selected designs shown in **Extended Data Fig. 4**), and the peptides were chemically synthesized and structurally characterized by NMR. In all three cases, the NMR spectra had well-dispersed, sharp peaks and secondary alpha-proton (αH) chemical shifts consistent with the secondary structure of the design model (**Supplementary Fig. S2-5**).

High-resolution NMR solution structures were determined for each of the designs (**Supplementary Information Table S2-3**). The αββ design NC_HEE_D1 is a 27-residue peptide with a D-proline, L-proline turn at the β-β junction. (Here, Rosetta identified a motif known previously to stabilize type I' or II' turns[14,15]). The NMR structure closely matches the design model: the C_α RMSD is 0.99 Å between the designed structure and the lowest-energy NMR model (**Fig. 4, top row**). The βαβ design NC_EHE_D1 is a 26-residue peptide stapled using two disulfide bonds (C1-C21, C2-C24). The design algorithm placed a D-arginine residue in the β-α loop and a D-asparagine residue as the C-terminal capping residue for the α-helix. The design model has a 1.9 Å C_α RMSD to the lowest-energy NMR ensemble member, and 0.68 Å to the closest member of the ensemble over all the C_α atoms (**Fig. 4, middle row**; the

last two residues at C-terminal vary considerably in the ensemble). Isolated pairs of solvent-exposed parallel β-strands, as designed in this topology, are found very rarely in natural protein structures[16]. NMR characterization of an initial design for the ββα topology showed an unwound C-terminal α-helix adopting an extended conformation, differing from the design model (design NC_EEH_D1, **Extended Data Fig. 5**). We hypothesized that substantial strain was introduced by the angle between the helix and the preceding strand, and the placements of disulfide bonds at both ends of the helix. A second design for this topology, NC_EEH_D2, has a Type I' turn at the β-β connection and a different placement of disulfide bonds (C2-C11, C5-C26). The NMR ensemble for NC_EEH_D2 is very close to the design model (0.86 Å C_α RMSD to the lowest-energy NMR model; **Fig. 4, bottom row**).

The designs were created *de novo* without sequence information from natural proteins. Searches for similar sequences in the PDB and non-redundant database using PSI-BLAST found a significant alignment (e-value < 0.01) only for NC_EHE_D1. This sequence has weak similarity (e-value of 2×10^{-4}) to the zinc-finger domain of lysine-specific demethylase (PDB ID: 2MA5), but the aligned regions adopt different structures (**Extended Data Fig. 6**).

We explored the stability of the designed peptides using CD to monitor thermal and chemical denaturation. All three peptides are very thermostable; there is no loss in secondary structure for NC_HEE_D1 and NC_EEH_D2 at 95 °C, and only a small decrease for NC_EHE_D1 (**Fig. 4f**). Quite remarkably, NC_HEE_D1 does not denature at 6 M GdnHCl (**Fig. 4g, top row**). Treatment with TCEP causes unfolding of all three designs, highlighting the importance of disulfide bonds.

Backbone-cyclized peptides

Next, we explored the design of backbone-cyclized, heterochiral peptides. Backbone cyclization can increase stability and improve pharmacokinetic properties in peptides by protecting against exopeptidases. To generate such backbones without dependence on fragments of known structures, we implemented a generalized kinematic loop closure[17,18] method (named "GenKIC") to sample peptide bonds, disulfide bonds, or other covalent linkages capable of connecting the termini. Each GenKIC chain-closure attempt begins by perturbing multiple mainchain degrees of freedom, then analytically solving kinematic equations to enforce loop closure with ideal peptide bond geometry in the case of N-C cyclic peptides (see **Methods**, **Supplementary Information**, and **Extended Data Fig. 7**). Residues intended to be in helical or strand conformations were initialized to ideal mainchain dihedral values, then perturbed by small degrees prior to analytical closure; initial mainchain dihedral angles for loop residues were drawn randomly from a Ramachandran-biased distribution. Sequence design, backbone relaxation, and *in silico* structure validation using MD simulation and Rosetta *ab initio* structure prediction were carried out with bond geometry constraints between the termini (**Extended Data Fig. 2**).

Cyclic peptides were chemically synthesized for three topologies (ββ, αα, and ααα) and their structures determined by NMR spectroscopy. The 18 residue ββ design (NC_cEE_D1) has a single disulfide bond connecting the strands in addition to the terminal peptide bond, with D-proline, L-proline Type II' turns at both ends. This design has a similar overall fold to natural theta-defensins, but has just one (rather than three) disulfide bonds and different turn types connecting the two strands[19]. This design showed a somewhat broader minimum in Rosetta *ab initio* structure prediction, and bigger fluctuations during MD simulations, than designs for the other two cyclic topologies (**Fig. 5e, top row**). The lowest-energy NMR model has a C_α RMSD of 1.26 Å to the designed structure. The variability in the curvature of the sheets across the NMR ensemble is similar to the variability observed by the structure calculations (**Fig. 5, top**

row). The 26 residue NC_cHH_D1 design, which has one disulfide bond between C9 and C22, has a 1.03 Å C_α RMSD from the lowest-energy NMR structure (**Fig. 5, second row**). The 22-residue NC_cHHH_D1 design has three short regions of α-helical structure and a disulfide bond between C5-C18. The NMR structure of the design was again very close to the design model (**Fig. 5, third row**), with a C_α RMSD of 1.06 Å to the lowest-energy NMR structure.

All three cyclic topologies were found to be extremely stable in thermal denaturation experiments, retaining secondary structure when heated to 95 °C (**Fig. 5f**). NC_cHHH_D1 showed loss of secondary structure in 6M GdHCl, however, NC_cHH_D1 and NC_cEE_D1 showed no change in secondary structure in the presence of 6 M GdnHCl (**Fig. 5g**). After reducing the disulfide bonds with 2.5 mM TCEP, both NC_cHH_D1 and NC_cHHH_D1 lost secondary structure, but the CD spectrum of NC_cEE_D1 was not changed by reduction of the central disulfide bond (**Fig. 5g, top row**). Overall, the cyclic designs showed exceptional thermodynamic stability.

Beyond natural secondary and tertiary structure

As a final test of the generality of the new design methodology, we designed a heterochiral, backbone-cyclized, two-helix topology with one right-handed helix and one left-handed helix ($\alpha_L\alpha_R$) assembling into a tertiary structure not observed in natural proteins. As before, we validated designs by MD; however, for validation by *ab initio* structure prediction it was necessary to develop a new protocol (see **Extended Data Fig. 8**, **Methods**, and **Supplementary Information**) since the standard Rosetta *ab initio* structure prediction method utilizes fragments of native proteins, which typically do not contain left-handed helices. The selected design for the cyclic $\alpha_L\alpha_R$ topology, NC_cHh_DL_D1, is a 26-residue peptide with one D-cysteine, L-cysteine disulfide bond connecting the right-handed and left-handed α-helices. There is an excellent match between the NMR structure ensemble and design model (C_α

RMSD: 0.79 Å) (**Fig. 6**). As expected for the nearly achiral topology, the CD signal is very small (as observed for a previously studied two chain, four helix mixed L D system[20]), and no change was observed on heating up to 95 °C. The secondary alpha proton chemical shifts also show no significant change on heating to 75 °C (**Fig. 6g, Supplementary Fig. 2-6**), indicating that the peptide is thermostable. Successful design of this topology demonstrates that our computational methods are sufficiently versatile and robust to design in a conformational space not explored by nature.

Conclusions

The considerable progress in computational design of new globular protein folds[6,8,9,21,22] had prior to this study not been matched with methods for rationally designing the covalently-stapled peptides essential for unlocking the pharmacological potential of peptide-based therapeutics. The key advances in computational design presented here — notably the methods for designing constrained peptide backbones spanning a broad range of topologies and incorporating natural and non-natural building-blocks — enable high-accuracy design of new peptides with exceptional thermostability and resistance to chemical denaturation. All 12 experimentally-determined structures are in close agreement with the design models, including one with helices of different chirality. Unlike the natural constrained peptide families, designed peptides are not limited to particular shapes, sizes, nucleating motifs, or disulfide connectivities; indeed, the sequences of these *de novo* peptides are quite different from anything found in nature thus far. The automated methods can access broad regions of shape space well beyond natural secondary and tertiary structures. In this paper, we have focused on extending sampling and scoring methods to permit design with D-amino acid residues and cyclic backbones, but the new tools are fully generalizable to peptides containing more exotic building-blocks, such as amino acids with noncanonical sidechains[23] or noncanonical backbones[24].

The hyperstable molecules presented in this study provide robust starting scaffolds for generating peptides that bind targets of interest. Both the computational predictions and experimental results suggest high mutational tolerance, and as described in the **Supplementary Information**, solvent-exposed hydrophobic residues can be introduced on the surface without impairing folding and solubility (**Extended Data Figs. 9** and **10**, **Supplementary Fig. S2-6**), allowing the introduction of target-binding residues to construct binders, agonists, or inhibitors. There has been considerable effort in both academia and industry to develop small naturally occurring proteins as alternatives to antibody scaffolds for library selection based affinity reagent generation. Our genetically encoded designs offer considerable advantages as starting points for such approaches because of their very high stability, small size, and diverse shapes. Furthermore, having been designed exclusively to be robust and stable, they lack the often destabilizing structural idiosyncrasies that arise in naturally occurring proteins from evolutionary selective pressure for a particular function. Similarly, the heterochiral designs described here provide starting points for split-pool and other selection strategies compatible with non-canonical amino acids.

Going beyond the adaptation of the sequences of our hyperstable designs to bind targets of interest, the methods developed in this paper can be used to specifically design new backbones to fit into target binding pockets. The advantage of such “on-demand” target specific scaffold generation is that the shape complementarity is likely to be considerably higher than that of scaffolds generated without knowledge of the target. High affinity binding peptides could be obtained from such shape complementary starting points by optimizing the binding interface using library selection methods as described in the previous paragraph, computational protein-protein interface design[25], or perhaps most effectively, by computation followed by experimental optimization.

Acknowledgements

We thank the Hyak supercomputing network at the University of Washington, and the Argonne Leadership Computational Facility, for providing computing and data storage resources. We thank the volunteer participants of the Rosetta@Home project on BOINC for providing additional computing resources. We are grateful for access to the facility of the Queensland NMR Network. We thank Darwin Alonso, Yuan Liu, Sanjay Srivatsan, Miriam Williamson for their help and advice. We also thank Ratika Krishnamurty, Parisa Hosseinzadeh, and Anastassia Vorobieva for critical comments and suggestions on the manuscript. This work was supported by NIH grant P50 AG005136 supporting the Alzheimer's Disease Research Center program, philanthropic gifts from the Three Dreamers and Washington Research Foundation, and funding from the Howard Hughes Medical Institute. DJC is funded by the Australian Research Council (ARC) as an Australian Laureate Fellow (FL150100146). TS acknowledges NIH support (GM094597), and SVSRKP, AE and XX were supported with NESG funds. EC is funded by NIGMS GM090205. The authors thank Peter Rupert and Roland K. Strong at the Fred Hutchinson Cancer Research Center for aiding in the crystallographic data collection and refinement of the gEHEE_06 structure. Additionally, we thank Lance Stewart, Binchen Mao, Victor Ovchinnikov, Nicholas Hasle, Damon May, Damian Ekiert, Gira Bhabha, Nobu Koga, Alexander Ford, Brandon Keir, and James Bardwell for helpful discussions, advice, and assistance. GWB was funded by the National Institute of Allergy and Infectious Diseases, National Institute of Health, Department of Health and Human Services, under Federal Contract number HHSN272201200025C. Some of this research was performed at the W.R. Wiley Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility located at Pacific Northwest National Laboratory (PNNL) and sponsored by U.S. Department of Energy's Office of Biological and Environmental Research (BER) program. Battelle operates PNNL for the U.S. Department of Energy.

Author Contribution

CDB, GB, VKM and DB designed the study. Algorithms were developed by VKM with help from AW, EC, YS, GB, RB, CDB, GJR, and TWL. CDB and JMG designed the canonical peptides with help from DB, GJR, and TWL. GB designed the non-canonical heterochiral and backbone cyclized peptides with help from VKM, DB, PG, and PSH. CDB expressed and characterized the designed canonical peptides from *E. coli* with help from JMG and SAR, and JMG performed MS analysis. WAG and CEC purified canonical peptides *via* Daedalus and determined the X-ray crystal structure. GWB, SVSRKP, AE, and TS determined the NMR solution structures of canonical peptides, purified with isotopic labeling by CDB. OC and GB synthesized, purified and characterized the designed non-canonical peptides. PJH and DJC determined the NMR solution structures of non-canonical peptides. PJH, QK and DJC analysed the data from structure determination of non-canonical peptides. CDB, GB, VKM, and DB wrote the manuscript with help from all the authors.

Author Information

NMR solution structures are deposited to RCSB Protein Data Bank with accession codes 5JG9, 2ND2, 2ND3, 3JHI, 5JI4. Reprints and permissions information is available at www.nature.com/reprints. Authors declare no competing financial interests. Correspondence and requests for materials should be addressed to David Baker (dabaker@uw.edu)

References

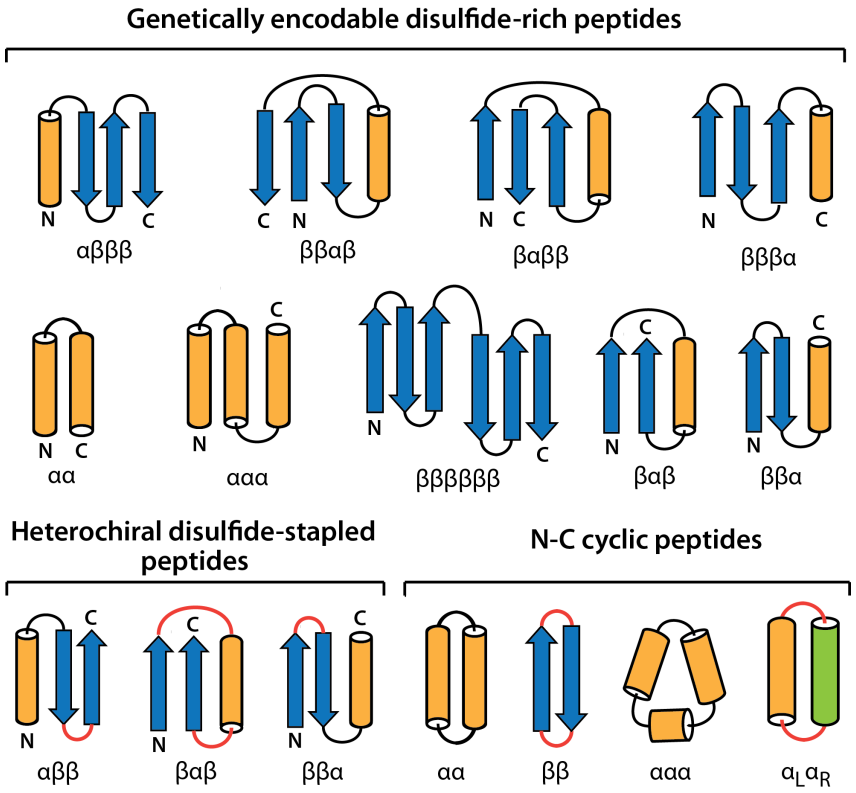
1. Conibear, A. C. *et al.* Approaches to the stabilization of bioactive epitopes by grafting and peptide cyclization. *Biopolymers* **106,** 89–100 (2016).

2. Craik, D. J., Fairlie, D. P., Liras, S. & Price, D. The future of peptide-based drugs. *Chem. Biol. Drug Des.* **81,** 136–147 (2013).
3. Góngora-Benítez, M., Tulla-Puche, J. & Albericio, F. Multifaceted roles of disulfide bonds. Peptides as therapeutics. *Chem. Rev.* **114,** 901–926 (2014).
4. Kimura, R. H., Levin, A. M., Cochran, F. V. & Cochran, J. R. Engineered cystine knot peptides that bind alphavbeta3, alphavbeta5, and alpha5beta1 integrins with low-nanomolar affinity. *Proteins* **77,** 359–369 (2009).
5. Boyken, S. E. *et al.* De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352,** 680–687 (2016).
6. Brunette, T. J. *et al.* Exploring the repeat protein universe through computational protein design. *Nature* **528,** 580–584 (2015).
7. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487,** 545–574 (2011).
8. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491,** 222–227 (2012).
9. Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346,** 481–485 (2014).
10. Bandaranayake, A. D. *et al.* Daedalus: a robust, turnkey platform for rapid production of decigram quantities of active recombinant proteins in human cell lines using novel lentiviral vectors. *Nucleic Acids Res.* **39,** e143 (2011).
11. Sagaram, U. S. *et al.* Structural and functional studies of a phosphatidic acid-binding antifungal plant defensin MtDef4: identification of an RGFRRR motif governing fungal cell entry. *PLoS One* **8,** e82485 (2013).
12. Liu, G. *et al.* NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 10487–10492 (2005).
13. Sharma, D. & Rajarathnam, K. 13C NMR chemical shifts can predict disulfide bond

formation. *J. Biomol. NMR* **18,** 165–171 (2000).

14. Syud, F. A., Stanger, H. E. & Gellman, S. H. Interstrand side chain--side chain interactions in a designed beta-hairpin: significance of both lateral and diagonal pairings. *J. Am. Chem. Soc.* **123,** 8667–8677 (2001).
15. Lai, J. R., Huck, B. R., Weisblum, B. & Gellman, S. H. Design of non-cysteine-containing antimicrobial beta-hairpins: structure-activity relationship studies with linear protegrin-1 analogues. *Biochemistry* **41,** 12835–12842 (2002).
16. Richardson, J. S. beta-Sheet topology and the relatedness of proteins. *Nature* **268,** 495–500 (1977).
17. Coutsias, E. A., Seok, C., Jacobson, M. P. & Dill, K. A. A kinematic view of loop closure. *J. Comput. Chem.* **25,** 510–528 (2004).
18. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6,** 551–552 (2009).
19. Trabi, M., Schirra, H. J. & Craik, D. J. Three-dimensional structure of RTD-1, a cyclic antimicrobial defensin from Rhesus macaque leukocytes. *Biochemistry* **40,** 4211–4221 (2001).
20. Sia, S. K. & Kim, P. S. A designed protein with packing between left-handed and right-handed helices. *Biochemistry* **40,** 8981–8989 (2001).
21. Lin, Y.-R. *et al.* Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5478–85 (2015).
22. Doyle, L. *et al.* Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **528,** 585–588 (2015).
23. Renfrew, P. D., Douglas Renfrew, P., Choi, E. J., Richard, B. & Brian, K. Incorporation of Noncanonical Amino Acids into Rosetta and Use in Computational Protein-Peptide Interface Design. *PLoS One* **7,** e32637 (2012).

24. Drew, K. *et al.* Adding diverse noncanonical backbones to rosetta: enabling peptidomimetic design. *PLoS One* **8,** e67051 (2013).
25. Fleishman, S. J. *et al.* Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332,** 816–821 (2011).



480

481

482

483

484 **Figure 1: Designed peptide topologies**

485 The designed secondary structure architectures for each of the three classes of constrained
486 peptides (genetically encodable disulfide stapled, heterochiral disulfide-stapled, and cyclic) span
487 most of the topologies that can be formed with four or fewer secondary structure elements.
488 Arrows: β-strands, orange cylinders: right-handed α-helices, green cylinder: left-handed α-helix;
489 red: loop segments containing D-amino acid residues.

490

491

492

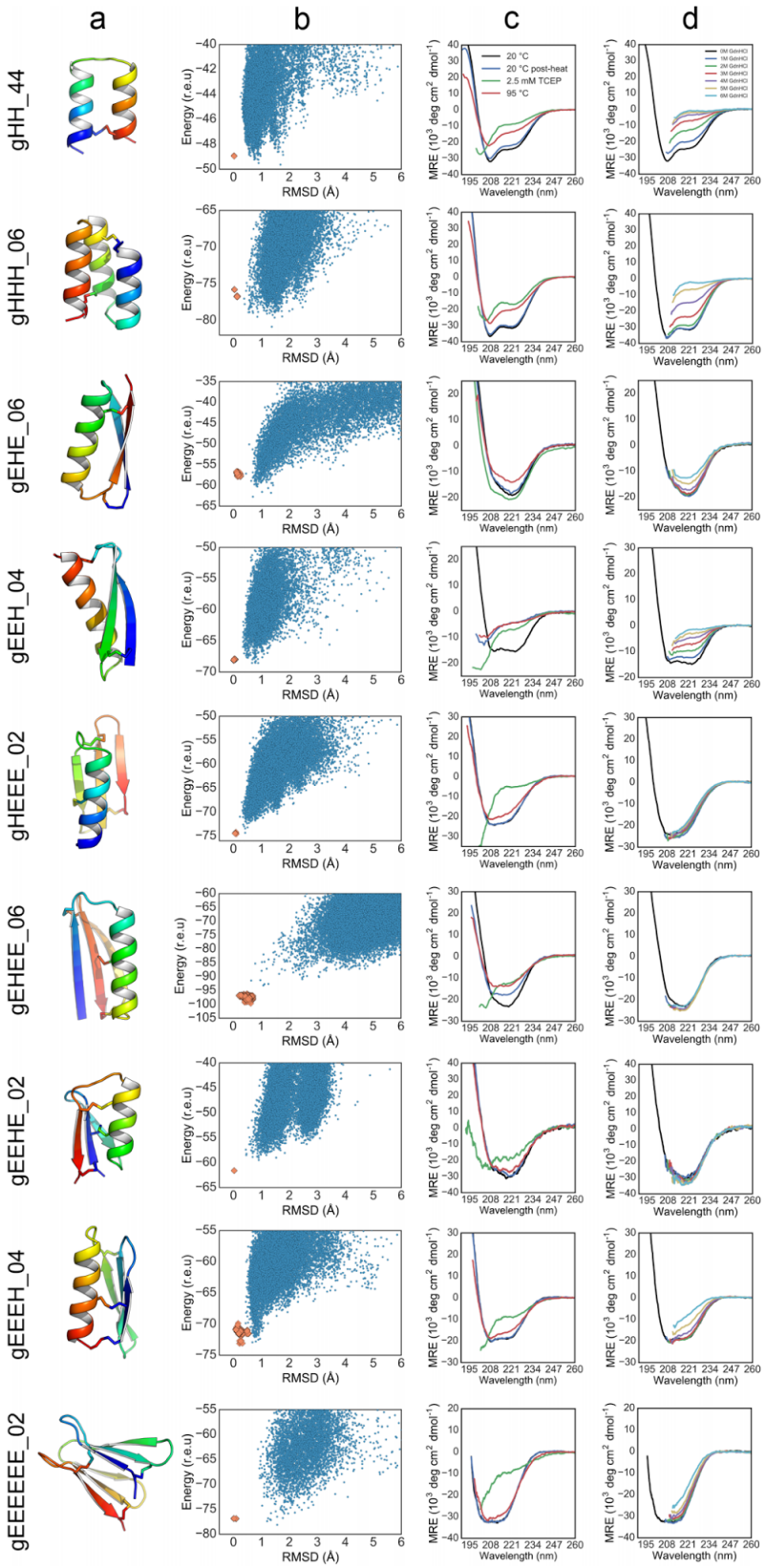


Figure 2: Computational design and biophysical characterization of genetically encodable disulfide-rich peptides. Genetically encodable designed peptides are named using the following convention: a lowercase prefix “g” to indicate genetic encodability, the topology (H for α-helix, E for β-sheet), and a number to differentiate designs that share a common topology. (column a) Cartoon renderings of the design representing each topology are shown with rainbow colouring from the N-terminus (blue) to the C-terminus (red), and disulfide bonds are shown as sticks. (column b) The energy landscape of each designed sequence was assessed by Rosetta structure prediction calculations starting from an extended chain (blue dots) or from

the design model (yellow dots); lower energy structures were in some cases sampled in the former because disulfide constraints were only present in the latter. (column c) CD steady-state wavelength spectra at 20°C (blue line), after heating to 95°C (red line), and upon cooling back to 20°C (green line). The contribution of disulfide bonds to protein folding was assessed by reduction with 2.5 mM TCEP (purple line). (column d) CD steady-state wavelength spectra at different concentrations of the chemical denaturant GdnHCl.

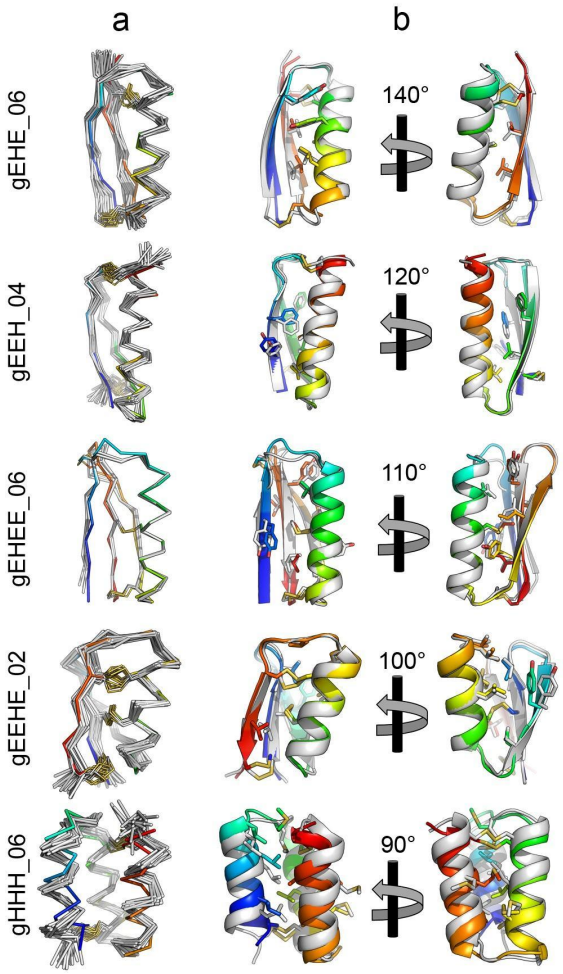


Figure 3: X-ray crystal structures and NMR solution structures of designed peptides are very close to design models. Structures for gEHE_06, gEEH_04, gEEHE_02, and gHHH_06 were determined by NMR spectroscopy, and the structure for gEHEE_06 was determined by X-ray crystallography. (column a) Cα traces of NMR ensembles, or superimposed members of the asymmetric unit, (gray) are aligned against the design model (rainbow). Disulfide bonds are shown with sidechain atoms rendered as sticks with sulfur atoms coloured yellow. (column b) A cartoon representation of the lowest energy conformer of each NMR ensemble or crystallographic asymmetric unit (gray) is shown aligned to the design model (rainbow). Sidechain atoms of hydrophobic core residues are rendered as sticks.

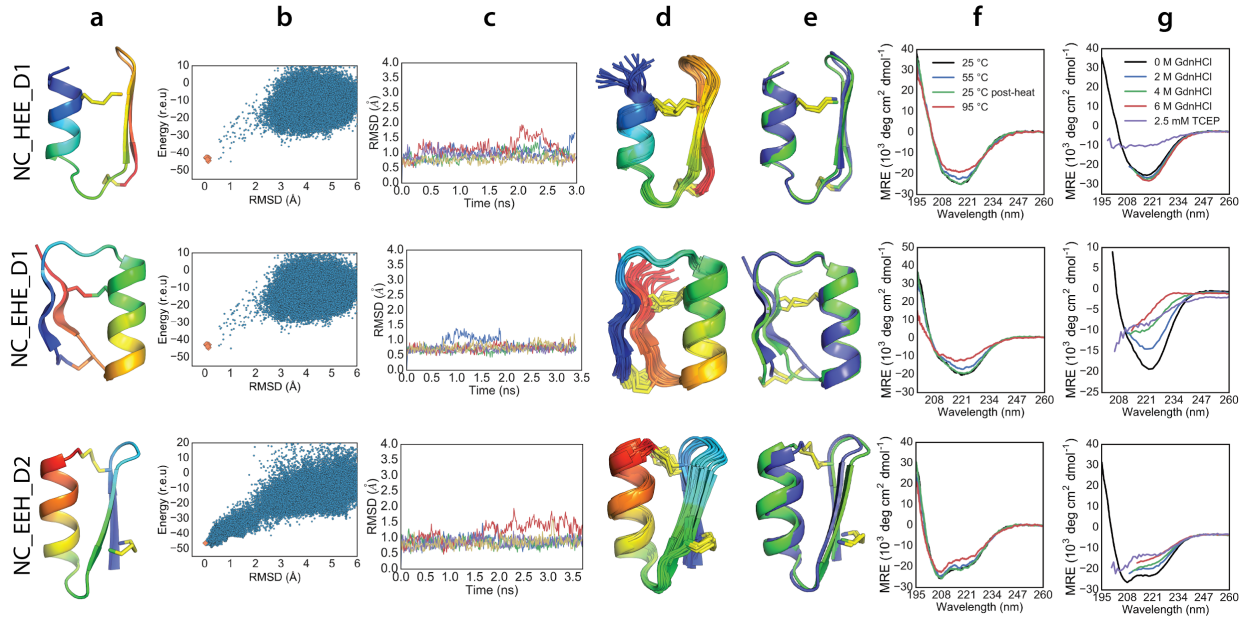


Figure 4: Design and characterization of heterochiral disulfide-constrained peptides

Non-canonical designed peptides are named using the following convention: prefix “NC” denotes non-canonical sequence or backbone architecture, the topology (H for α-helix, E for β-sheet), and a number to differentiate designs that share a common topology. *Column a*: Cartoon representations of design models with the N-terminus in blue and C-terminus in red. *Column b:* Folding energy landscapes from Rosetta@home Rosetta *ab initio* structure prediction calculations. Blue dots indicate lowest-energy structures identified in independent Monte Carlo trajectories. Orange dots are from trajectories starting with the design model. All three design models are in deep energy minima. (r.e.u: Rosetta Energy Units, RMSD: root mean square distance to the designed topology). *Column c:* Five representative trajectories from a total of 50 independent molecular dynamics simulations starting from the design model with different initial velocities. *Column d:* NMR-determined structure ensembles. Cartoon representations coloured and oriented as in column a. *Column e*: Superposition of the designed structure with the lowest-

energy NMR structure. Green: NMR model, Blue: Design model. *Column f:* Thermal denaturation of designed peptides. CD steady-state wavelength spectra between 195 nm and 260 nm recorded at 25 °C (black), 55 °C (blue), 95 °C (red), and after cooling back to 25 °C (green). *Column g:* Chemical denaturation studies conducted in presence of GdnHCl or TCEP. CD spectra recorded at 0 M GdnHCl (black), 2 M GdnHCl (blue), 4 M GdnHCl (green), 6 M GdnHCl (red), and 2.5 mM TCEP / 0 M GdnHCl (purple). Data are truncated in the far-UV region for spectra acquired in the presence of high GdnHCl concentrations (due to GdnHCl absorbance).

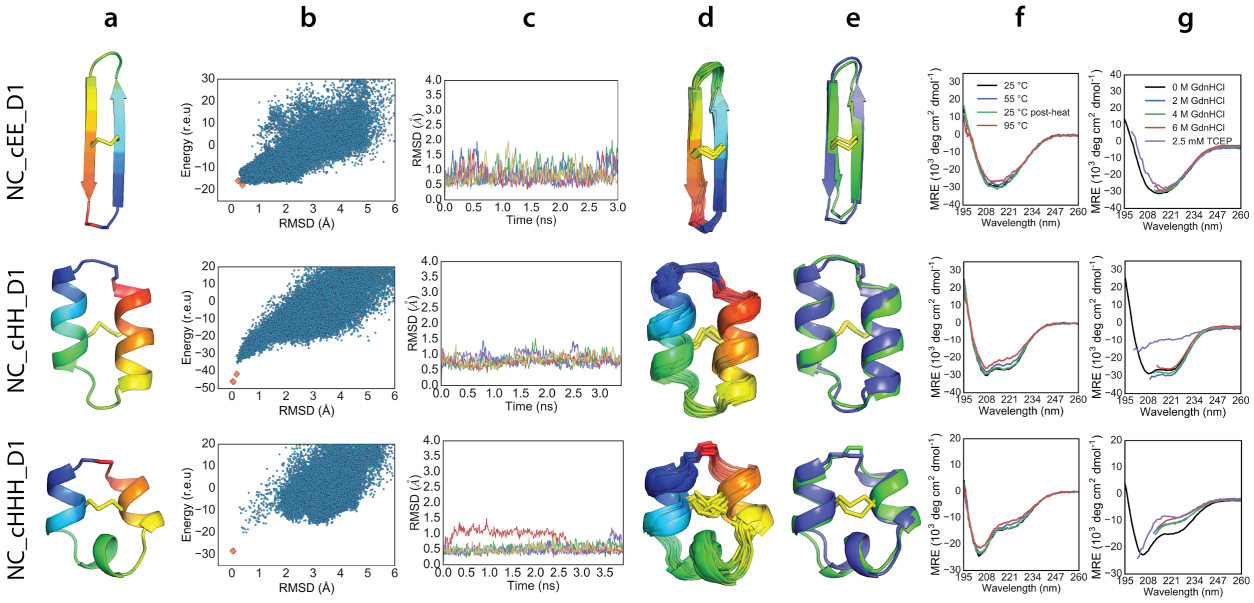


Figure 5: Design and characterization of N-C backbone cyclic peptides

Columns are as indicated in Figure 4 legend. A lowercase “c” in the peptide name indicates N-C cyclic backbone.

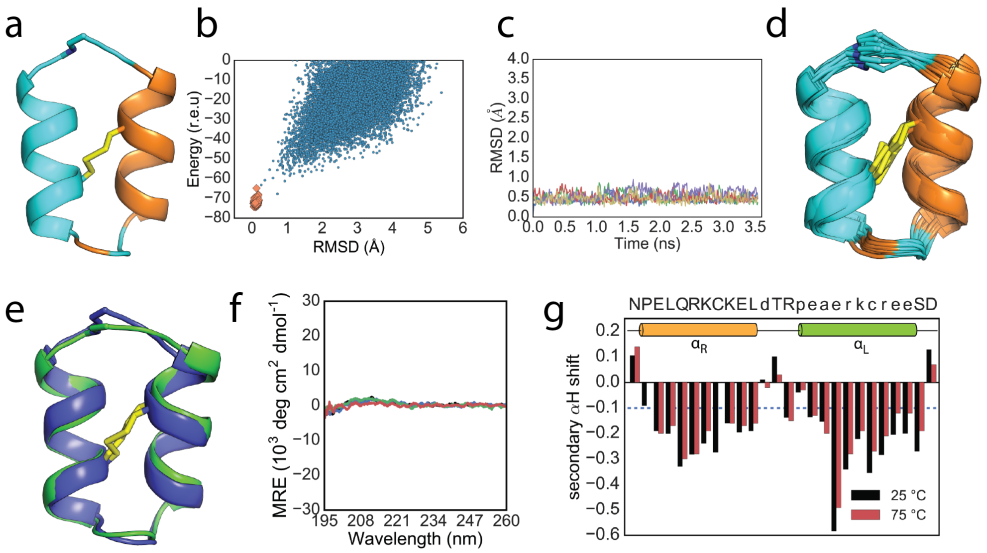


Figure 6: Design and characterization of cyclic 2-helix topology with non-canonical secondary and tertiary structure.

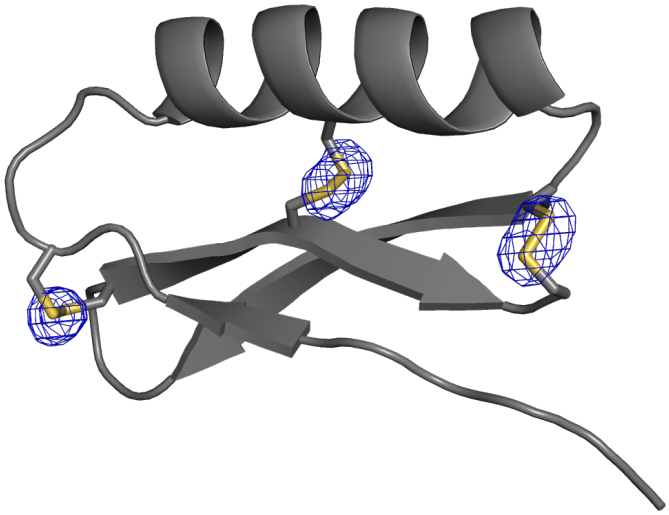
a) Cartoon representation of NC_cHh_DL_D1 design with L-amino acid residues coloured cyan and D-amino acid residues coloured orange. **b)** Folding energy landscape generated using a new structure prediction algorithm compatible with noncanonical secondary structures; see methods and **Supplementary Information** for details. **c)** Five representative trajectories from a total of 50 independent molecular dynamics simulations starting from the design model with different initial velocities. **d)** NMR-determined structure ensembles. Cartoon representations coloured and oriented as in first panel **e)** Superposition of the designed structure with the lowest-energy NMR structure. Green: NMR model, Blue: Design model. **f)** Thermal denaturation of designed peptides. CD spectra between 195 nm and 260 nm recorded at 25 °C (black), 55 °C (blue), 95 °C (red), and after cooling back to 25 °C (green). The CD steady-state wavelength spectrum of the cHh_DL_D1 design exhibits only very weak signals because the L- and D-

596 helical signals largely cancel. **g)** Secondary alpha proton chemical shifts (in ppm) calculated at

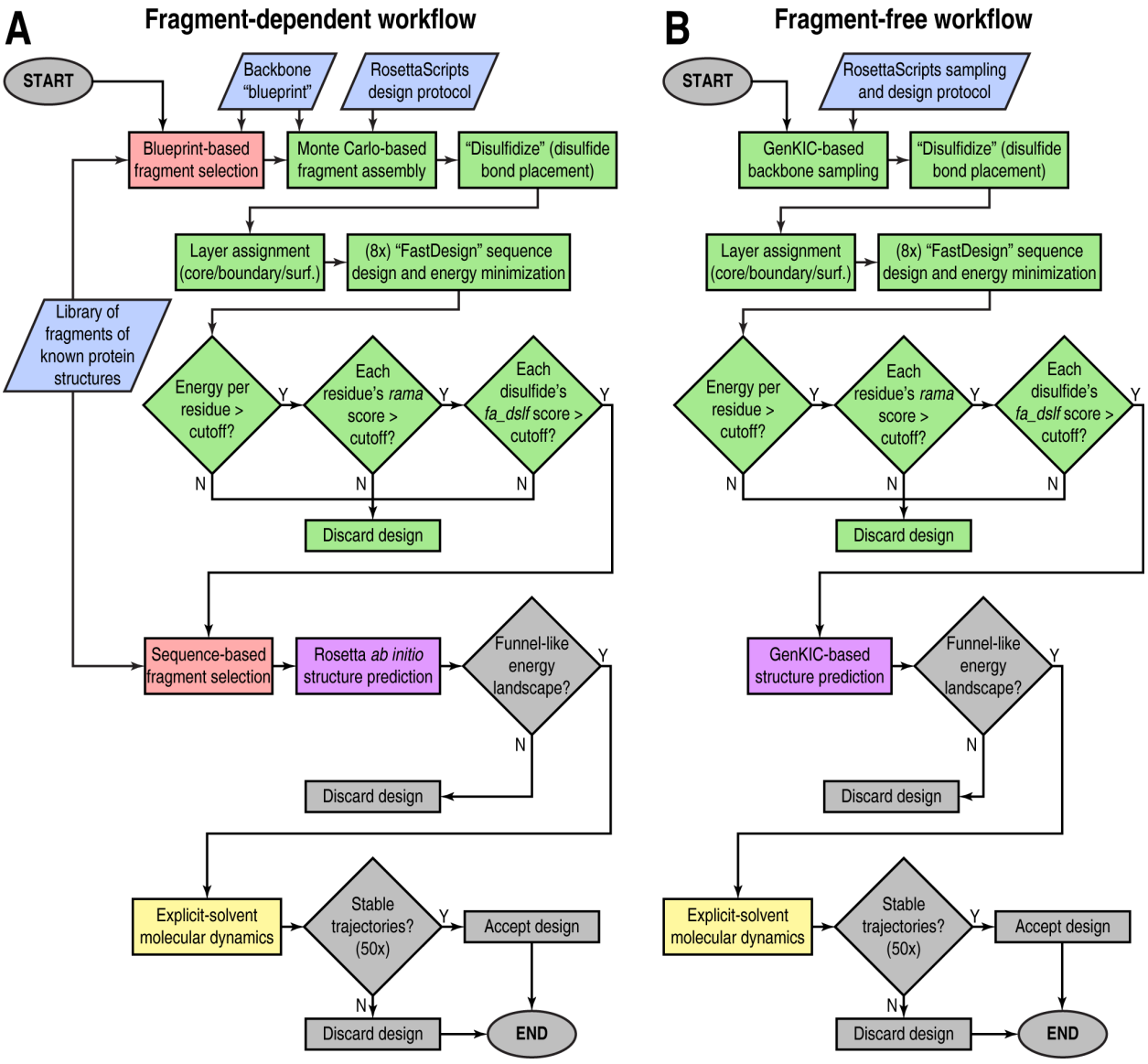
597 25 °C (black) and 75 °C (red) show no change, indicating that the peptide is thermally stable.

598

599



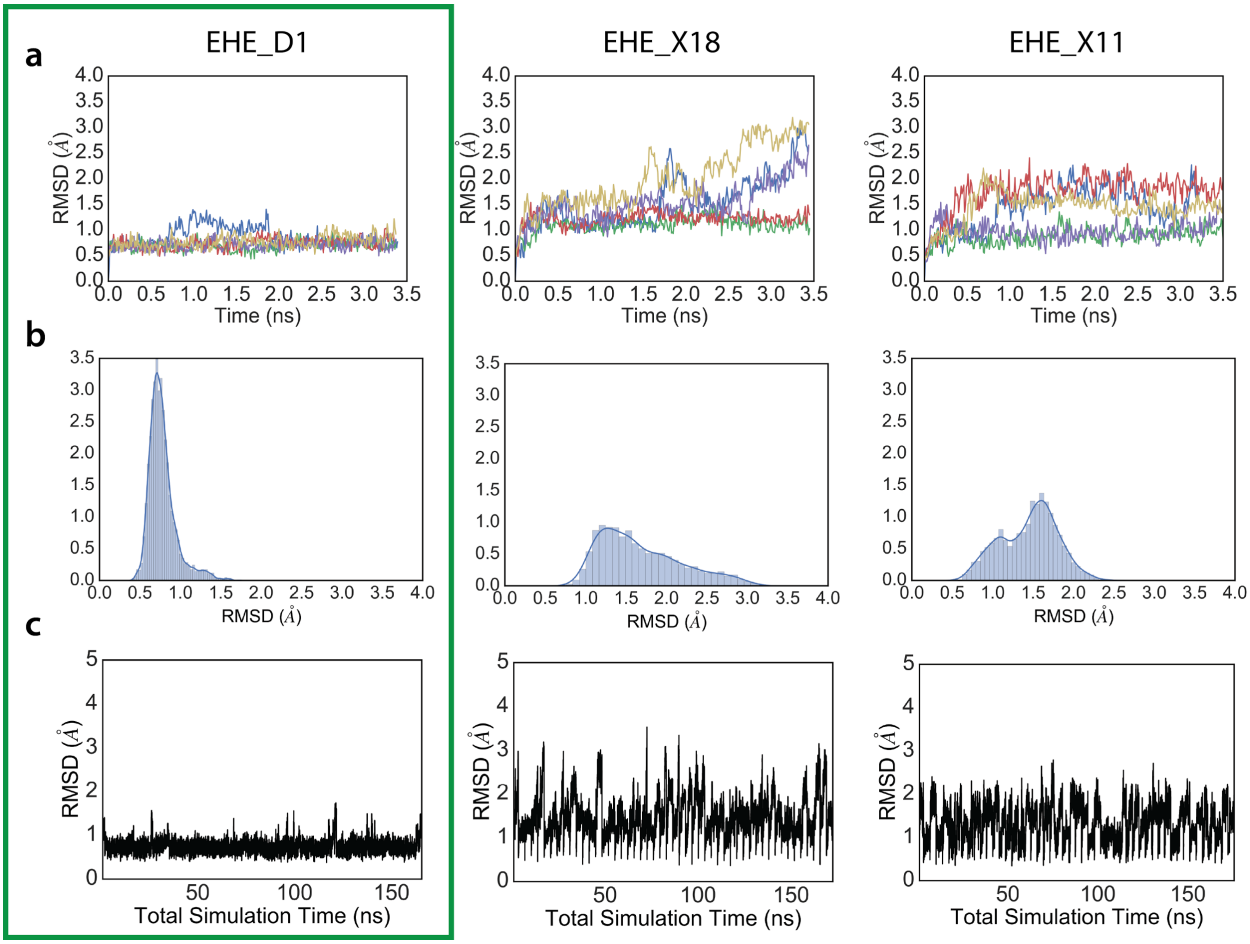
Extended Data Figure 1: Disulfide bonds are well defined by X-ray crystallography. An $F_o - F_c$ omit-map is shown contoured at 4 σ for design gEHEE_06. Disulfide sulfur atoms were removed, and the omit-map was calculated following real-space refinement.



Extended Data Figure 2: Flowchart of pipelines for designing non-canonical cyclic peptides

Inputs are shown in blue, RosettaScripts-automated parts of the pipeline are in green, parts carried out by Rosetta standalone applications are shown in pink (the fragment picker application) and purple (the various structure prediction applications), parts performed with molecular dynamics software are coloured yellow, and manual steps are shown in grey. a) Fragment assembly-based design pipeline. Final computational validation was carried out using

MD simulations and fragment-based Rosetta *ab initio* structure prediction. For peptides containing isolated D-amino acids, these residues were mutated to glycine for Rosetta *ab initio* structure prediction. b) Fragment-free, GenKIC-based design pipeline. This approach permits design of noncanonical topologies like the mixed $\alpha_L\alpha_R$ topology, which occurs in no known natural protein. The GenKIC-based structure prediction algorithm is described in **Extended Data Figure 7** and in the **Supplementary Information**.



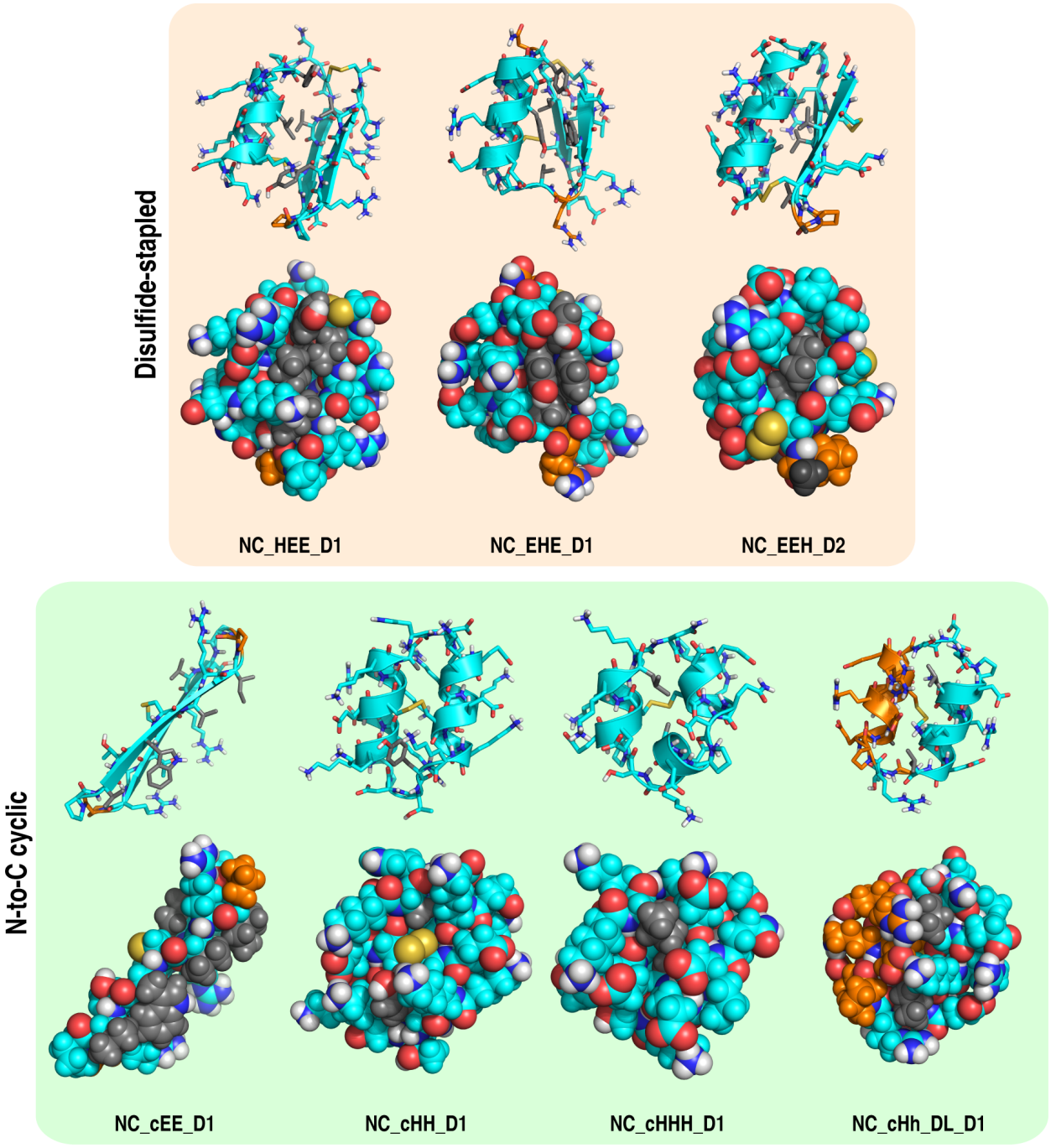
Extended Data Figure 3: Molecular dynamics screening of designed peptides

Fifty independent molecular dynamics (MD) simulations in explicit solvent conditions, and all starting from the designed peptide, were used for discriminating good (*e.g.* EHE_D1) designs from non-optimal designs of the same topology (*e.g.* EHE_X18 and EHE_X11). a) Five representative trajectories from MD simulation runs. Designs that showed good convergence, and smaller fluctuations were selected for further experimental characterization. b) RMSD distribution from all 50 trajectories. Only the last one-third of the trajectory was used for this analysis. Designs with narrower distributions were picked for further testing. c) Concatenated

639 trajectory of all 50 independent runs show lower fluctuations for the more optimal designs.

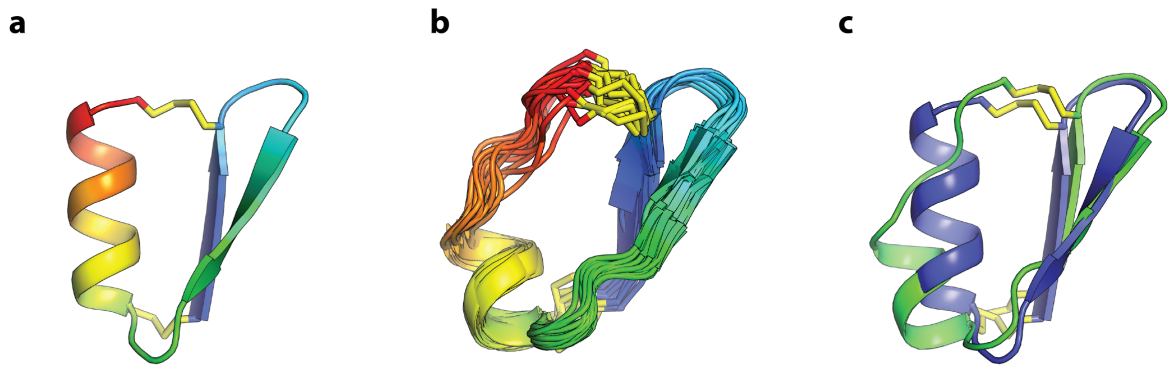
640

641



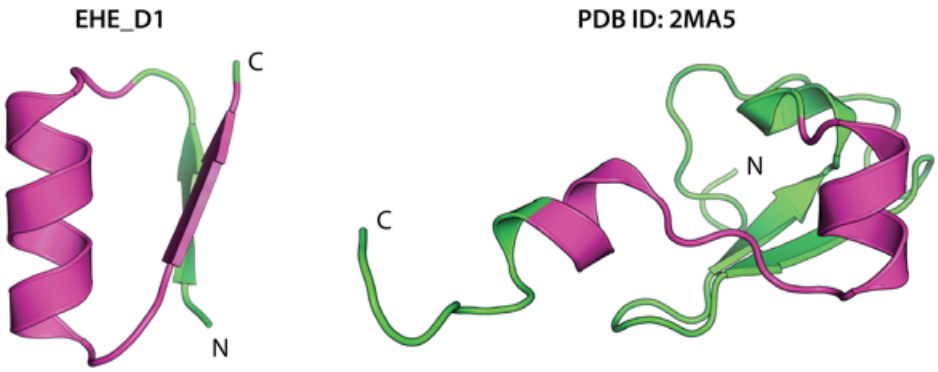
Extended Data Figure 4: Sidechain placement in non-canonical peptide designs chosen for experimental characterization

Designs are shown as cartoon and stick representations (top row in each box) and as van der Waals spheres showing sidechain packing (bottom row in each box). L-amino acid residues are shown in cyan, and D-amino acid residues are shown in orange. Sidechains of D- or L-variants of alanine, phenylalanine, isoleucine, leucine, valine, tryptophan, and tyrosine are coloured grey to aid visualization of hydrophobic packing interactions.



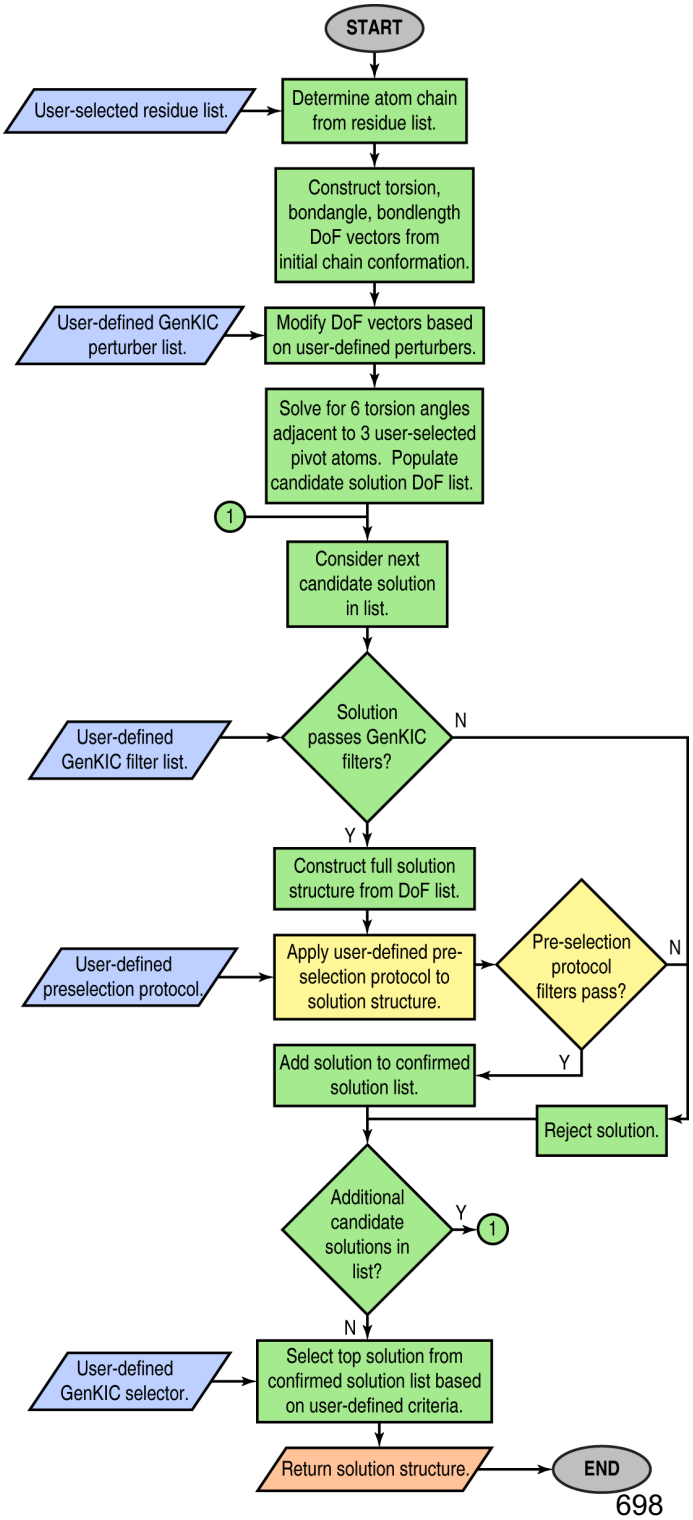
Extended Data Figure 5: Structural characterization of EEH_D1

NMR structure of EEH_D1 does not match the designed topology. a) Rosetta-designed model for EEH_D1. b) Ensemble of conformers representing the NMR solution structure. c) Superposition of the designed model (blue) with a representative NMR conformer (green).



Extended Data Figure 6: Structural mapping of sequence aligned region between EHE_D1 and 2MA5

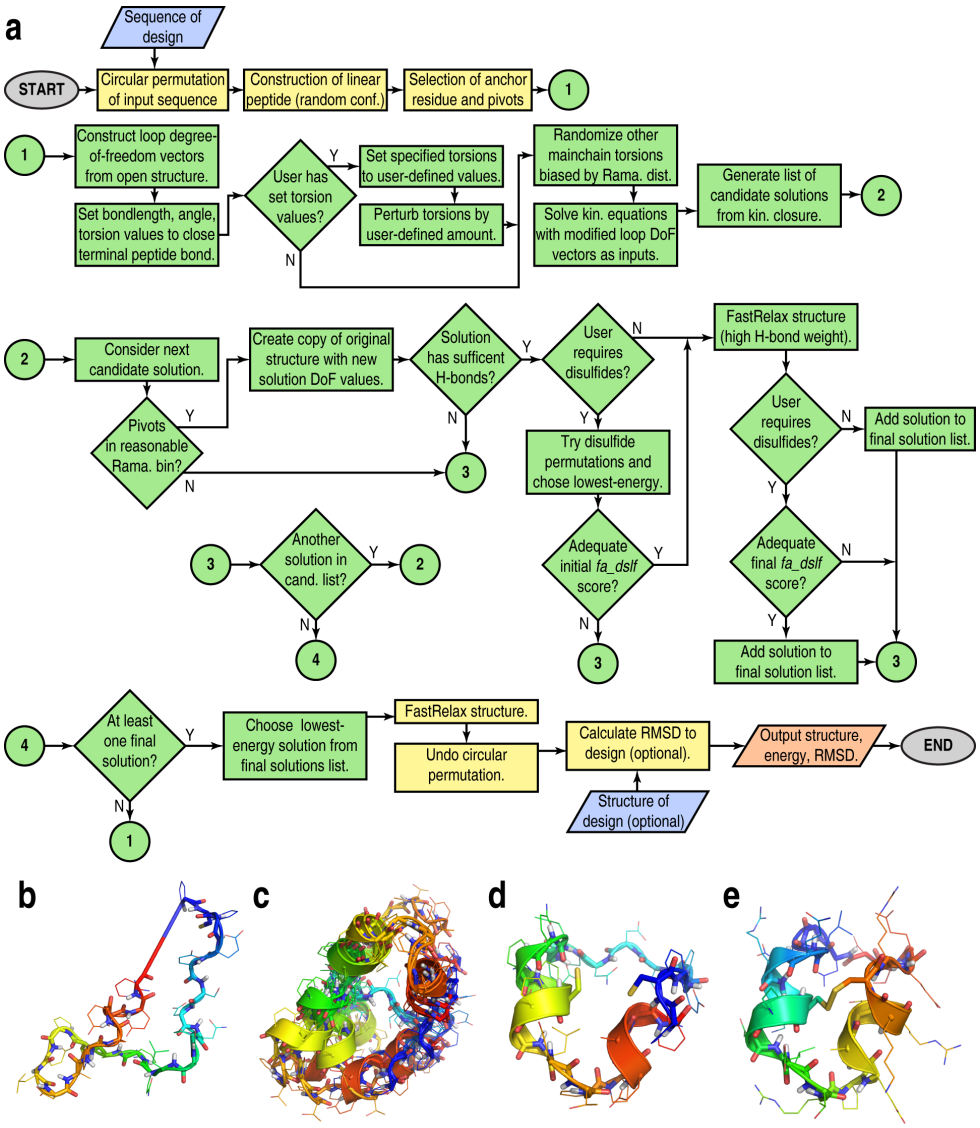
Design NC_EHE_D1 and PDB entry 2MA5 show weak but significant (e-value: 2×10^{-4}) sequence alignment, which is highlighted in purple. The aligned region folds into very different structures in the different contexts of peptide and protein.



Extended Data Figure 7: Generalized kinematic closure algorithm flowchart

GenKIC permits the sampling of closed conformations of arbitrary chains of atoms. These chains can pass through canonical or noncanonical backbone or sidechain linkages. Bond length, bond angle, and torsional degrees of freedom in the chain can be fixed, perturbed from a starting value by small amounts, set to user-defined values, or sampled randomly, as the user sees fit. The algorithm then solves for six torsion angles adjacent to three user-defined pivot atoms in order to enforce closure of the loop. The many solutions from the closure are then filtered internally, and each can be subjected to arbitrary user-defined Rosetta protocols and filtration in order to further prune the solution list. A single solution is selected from those passing filters by user-defined selection criteria. This flowchart shows the steps in a single invocation of the algorithm; for sampling, a user may specify that the algorithm be applied any number of times. User inputs are shown in

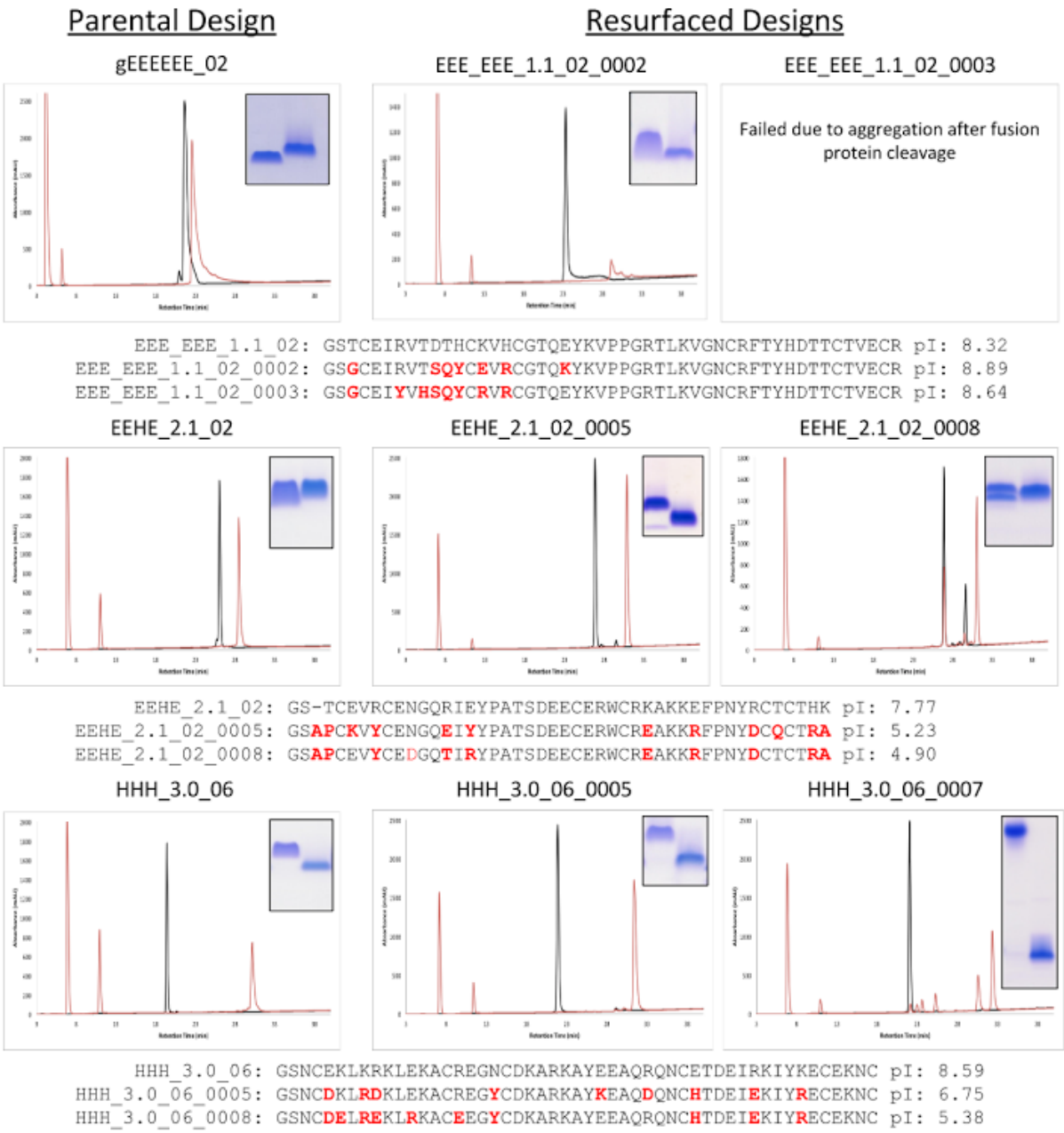
blue, steps carried out by the GenKIC algorithm itself are in green, steps carried out by Rosetta code external to the GenKIC algorithm are shown in yellow, and outputs are shown in salmon.



Extended Data Figure 8: A new fragment-free structure prediction algorithm

a) Flowchart diagramming the steps to generate a single sampled conformation. In typical usage, this process would be repeated tens of thousands of times to produce many samples. Inputs (the peptide sequence and an optional PDB file for the design structure) are shown in blue, and outputs (the sampled structure, its energy, and its RMSD from the design structure) are shown in salmon. Steps performed by the Generalized Kinematic Closure algorithm are shaded green, and setup and completion steps performed by the **simple_cycpep_predict**

application are shown in yellow. Further details of this algorithm are discussed in the **Supplementary Information** available online. b) The initial, random peptide conformation with bad terminal peptide bond geometry. c) Ensemble of closed conformations found for a single closure attempt. In this example, residue 7 (cyan) is the fixed anchor residue. Certain regions of the peptide have been set to left- or right-handed helical conformations prior to solving closure equations. d) A single closed solution with relative cysteine sidechain orientations that pass the initial, low-stringency filter for disulfide (*fa_dslf*) conformational energy. e) The resulting structure, following sidechain repacking, energy-minimization, and cyclic de-permutation.



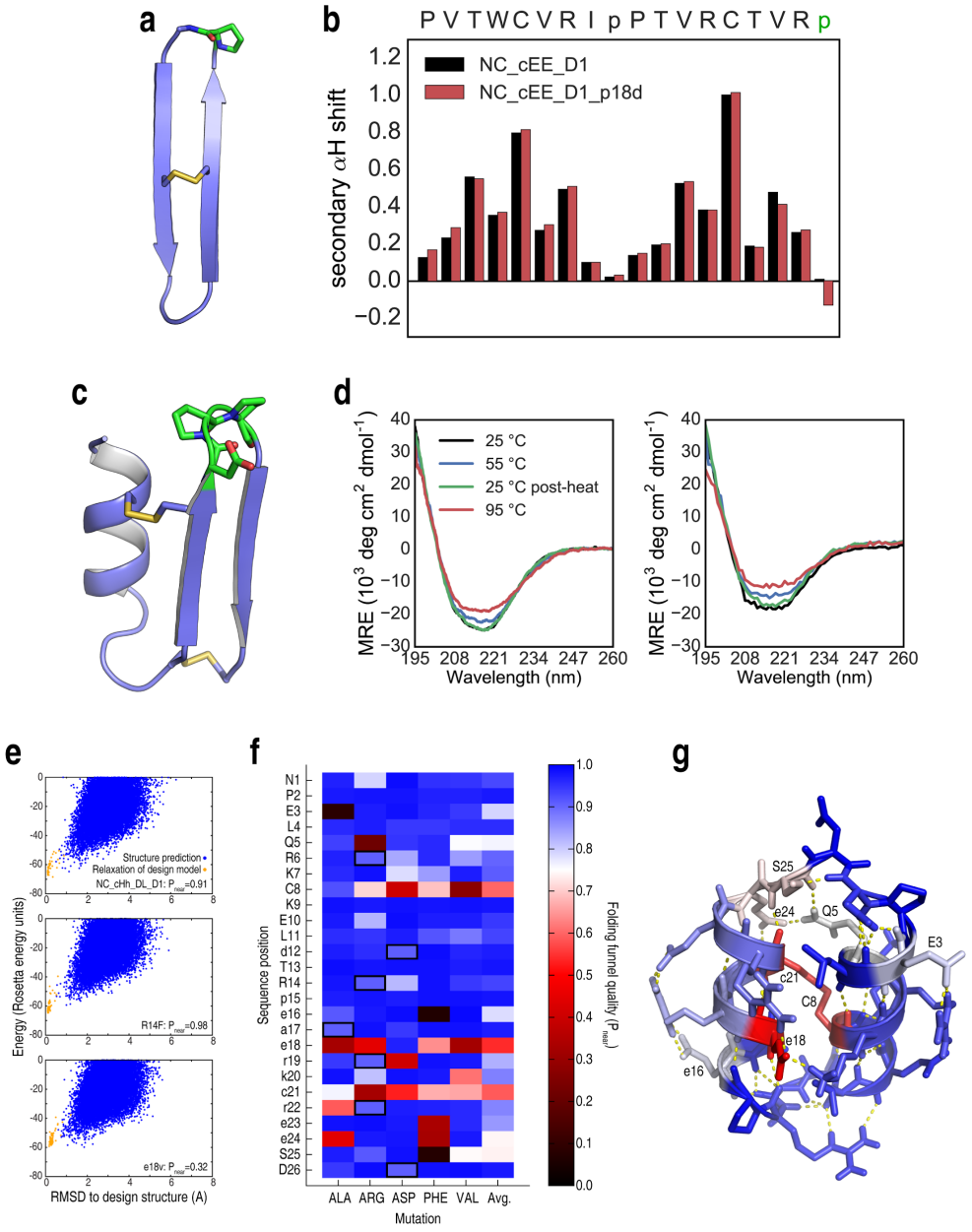
Extended Data Figure 9: Mutational tolerance of selected genetically encodable designs

RP-HPLC traces for the parental designs are shown next to the redesigned variants where applicable. Proteins run under oxidized conditions are shown in black while proteins run following reduction with 10mM DTT are shown in red. Insets within each panel are shown only to highlight the SDS-PAGE mobility of each purified protein under oxidizing (left band) and reducing conditions (right band). Sequence alignments are shown with the mutated positions

730 highlight in red, along with theoretical isoelectric points as calculated by ProtParam.

731

732



Extended Data Figure 10: Mutational tolerance of selected NC designs

a-b) Mutational tolerance of D-proline, L-proline loop of design NC_cEE_D1 (green in panel a), assessed by NMR chemical shift indices for the design sequence (black bars in panel b) and the p18d loop mutation (red bars). Eliminating this key proline residue does not result in loss of β-strand signal. c-d) Mutational tolerance of loop region of design NC_HEE_D1 (green in panel c),

as assessed by CD spectroscopy for the design sequence (left plot, panel d) and for the D19T, p20q, P21D triple mutant (right plot, panel d). Both proline residues may be mutated without loss of secondary structure or major change in the thermal stability. e-g) Computationally predicted mutational tolerance of design NC_cHh_DL_1, across the entire sequence. Each position was successively mutated *in silico* to D- or L-alanine, arginine, aspartate, phenylalanine, or valine (preserving the position's chirality), and full folding simulations were carried out with the Rosetta simple_cycpep_predict application. Folding funnel quality was evaluated using the P_{near} metric described in the methods. e) Representative plots of energy *vs.* RMSD to design for the design sequence (top), for the non-disruptive R14F mutation (middle), and for the e18v mutation (bottom). Results from structure prediction runs are shown in blue, and relaxation runs, in orange. Note that the bottom case shows many sampled states far from the design state with energy equal to or less than the design state energy. f) Mutational tolerance by position (vertical axis) and mutation (horizontal axis). Blue rectangles represent well-tolerated mutations, and red to black rectangles represent disruptive mutations, based on P_{near} evaluation of the folding funnel. Black borders indicate the design sequence. g) Mutational tolerance mapped onto the NC_cHh_DL_1 structure, with colours as in the previous panel. Most positions tolerate mutation well, with only the disulfide bridge (C8-c21) and the salt bridges formed by e18 being highly sensitive. The hydrogen bond networks formed by residues Q5, e24, and s25 show some moderate sensitivity to mutation, as do residues E3 and e16.

Methods

Computational design

De novo design of stapled peptides can be divided into three main steps: backbone assembly and sequence design. Practically, our peptide design pipeline has been optimized to permit these two steps to be performed in immediate succession with a single set of inputs, with no need for export or manual curation of generated backbones prior to the sequence design. (A third and final validation step is typically performed separately.)

For backbone assembly, we used two different approaches in this report: disulfide-constrained topologies were sampled using a fragment assembly method, while backbone-cyclized peptide topologies were sampled using a fragment-independent kinematic closure-driven approach. Example scripts and command lines for each step in the design workflow are available in the **Supplementary Information**.

Disulfide positioning

To design disulfide bonds, we first evaluated all residue pairs with C_β atoms ≤ 5 Å apart for geometry suitable to disulfide bond formation[26], selected backbones that could harbor disulfide bonds with near-ideal geometry, and incorporated one to three disulfide bonds. To select an ideal disulfide configuration from the set of all sterically possible combinations of disulfide bonds for a given backbone, we ranked disulfide configurations according to their effect on the unfolded state configurational entropy. The reduction in unfolded state entropy due to a set of multiple cross-links was computed according to a random flight model using Eq. 6 in Harrison *et al.*[27], with $\Delta V = 29.65$ Å^3 and $b = 3.8$ Å^3. This method has been implemented in the Rosetta software suite as the Disulfidize Mover and DisulfideEntropy Filter, both of which are accessible to the RosettaScripts scripting language.

Backbone design using fragment assembly

In the case of disulfide-stapled designs, the topology was defined using a "blueprint" that specifies secondary structure and torsion bins for each amino acid residue, the latter defined using the *ABEGO* alphabet system described previously[8,21]. The *ABEGO* nomenclature assigns a letter to each of five regions, or bins, in Ramachandran space. These correspond to the α-helical region (*A*), the β-sheet region (*B*), the region with positive phi values typically accessed by glycine (*G*), and the remainder of the Ramachandran space (*E*). (The fifth bin, *O*, represents residues with *cis*-peptide bonds, and was not used here.) The blueprint is the input for a Rosetta Monte Carlo-based fragment assembly protocol[7,8,21,26] that generates backbone conformations matching the blueprint architecture. Briefly, the fragment assembly protocol uses the defined blueprint to pick backbone fragments from a database of non-redundant high-resolution crystal structures. The insertion of fragments serves as the moves in a Monte Carlo search of backbone conformation space. For searches of the ββα topology, loop types were limited to *ABEGO* bins *EA* and *GG* for the ββ connection, and *BAB* and *GBB* for the αβ connection. For sampling of the βαβ topology, βα connections were limited to *GBB*, *BAB*, and *AB*, while αβ connections were limited to *GB*, *GBA*, and *AGB*. For sampling αββ topology, αβ connections were limited to *BAAB*, *GB*, *GBA*, and *AGB*, while ββ connections were limited to *EA* and *GG*.

Backbone design using generalized kinematic closure

While the fragment-based approaches described above are powerful, they are limited to conformations favored by peptides composed primarily of L-amino acids. For N-C cyclic designs — NC_cHHH_D1, NC_cHH_D1, NC_cEE_D1, NC_cHh_DL_D1 — we chose to focus on fragment-independent methods that are better suited to explore conformations that are only

accessible to mixed D/L peptides. We therefore turned to generalized kinematic closure (GenKIC).

GenKIC-based sampling works by treating a peptide as a single loop, or series of loops, to be "closed". The torsion values of an initial, "anchor" residue are randomly selected; this residue is then fixed, and the rest of the peptide is treated as a loop closure problem. The particular covalent linkages serve as a set of geometric constraints for loop closure. The GenKIC algorithm performs a series of user-controlled perturbations to the torsion angles of the peptide chain, which inevitably disrupt the geometry of the closure points. GenKIC then mathematically solves for the value of six "pivot" torsion angles that restore the geometry of the closure points and permit the loop to remain closed[17,18,28]. Since the algorithm can return up to sixteen solutions per closure attempt, a series of filters are applied to eliminate solutions with amino acid residues in energetically unfavorable regions of Ramachandran space or with other geometric problems, such as clashes with other residues. The "best" solution is then chosen based on the Rosetta score function[7].

During the sampling steps, regions in the designed topology that were intended to form helices or sheets were initialized to ideal phi/psi values, and were either kept fixed or perturbed by only small amounts (<20 degrees). In loop regions, the perturbation was carried out by drawing torsion values randomly, biased by the Ramachandran preferences of the amino acid residue. The allowed torsion value range either covered the entire Ramachandran space, or, in cases in which known loop *ABEGO* patterns could connect secondary structure elements, the mainchain torsion values were were limited to those *ABEGO* bins. For example, during the design of the cEE topology, connection types were limited to the 'GG' and 'EA' torsion bins for the 2-residue loops.

Modifications to Rosetta to permit design of cyclic backbones and mixed D/L peptides

D-amino acid residues allow access to regions of conformational space normally only accessed by glycine. When placed correctly, they can provide greater rigidity than glycine, stabilizing glycine-dependent structural motifs and, thereby, the overall fold[29]. Because the Rosetta software suite has primarily been used for designing proteins consisting of the 19 canonical L-amino acids and glycine, a number of modifications were necessary in order to permit robust design of peptides containing mixtures of D- and L-amino acids. First, Rosetta's default scoring function, called *talaris2013*, was updated to permit D-amino acids to be scored with mirror symmetry relative their L-counterparts. Terms in the score function that are based on mainchain or sidechain torsion values were modified to invert D-amino acid torsion values before applying the equivalent L-amino acid potentials. Those score function terms that are based on interatomic distances required minimal changes. To permit energy minimization, score function derivatives were also modified to invert torsion derivative values for D-amino acids. Rosetta's rotameric search algorithm, the *packer*, was modified to use L-amino acid rotamers with sidechain chi torsion values inverted for D-amino acid rotamer packing. Finally, we added an option to symmetrize the energy tables for the mainchain torsion preferences of glycine, which are asymmetric by default because they are based on statistics taken from the Protein Data Bank. (Glycine, in the context of L-amino acids only, occurs disproportionately in the positive-phi region of Ramachandran space, but should have no asymmetric preferences in a mixed D/L context.) Details of these modifications are described in the **Supplementary Information**.

Because Rosetta has traditionally been used to build linear polymers, a number of core Rosetta libraries had to be modified to permit N-C cyclic geometry to be sampled and scored properly. The assumption that residue i is connected to residues i+1 and i-1, which is invalid for cyclic peptides, has been removed and replaced with proper lookups of connected residue indices.

Note that, as of 11 March 2016, the default Rosetta score function has been changed to *talaris2014*, which re-weights a number of score terms and introduces one new term. The talaris2014 score function has also been made fully compatible with D-amino acids and cyclic geometry. A newer, experimental score function, currently called *beta_nov15*, has also been made fully compatible with D-amino acids and cyclic geometry.

Sequence design and filtering

Backbone assembly using Fragment Assembly or GenKIC was followed by a sequence design step. Sequence design involved eight rounds each of alternating sidechain rotamer optimization (during which sidechain identities were permitted to change) and gradient descent-based energy minimization. Each amino acid position was sorted into a layer ("core", "boundary", or "surface") based on burial, and the layer dictated the possible amino acid types allowed at that position. Hydrophobic amino acid residues, for example, were only permitted at core positions. To favor more proline residues during sequence design, the reference weight for proline in the Rosetta score function was reduced by 0.5 units. Backbones were allowed to move during the relaxation steps. For each topology ~80,000 structures were generated, and filtered based on the overall energy per residue, score terms related to backbone quality, and score terms related to the disulfide geometry.

Rosetta-based computational validation

Typically, the number of designs that can be created *in silico* exceeds the number that can be produced and examined experimentally. We therefore used Rosetta to prune the list of designs, by one of two methods. For design consisting of canonical amino acids, Rosetta's fragment-based *ab initio* algorithm[30] was utilized. Disulfide bonds were not allowed to form during these simulations; the designed disulfide bonds are intended to stabilize the folded conformation

rather than direct protein folding. Designs which incorporate short stretches of D-amino acids were also validated using Rosetta's fragment-based *ab initio* algorithm; the amino acid sequences of designs, with all D-amino acids mutated to glycine, were provided as input, and we allowed Rosetta to generate on the order of 30,000 predicted structures as output. Unlike the standard *ab initio* protocol, we did not use secondary structure predictions in fragment picking. Additionally, the length of small and large fragments was set to 4 and 6 amino acid residues, instead of the default 3 and 9; we found that this produced better sampling for peptides. After conformational sampling, the D-amino acid positions were changed to their original identities, and rescored. A small modification to the *ab initio* algorithm permitted it to build a terminal peptide bond for the N-C cyclic designs during the full-atom refinement stages of the structure prediction. Those designs that showed no sampling near the design conformation, or for which the design conformation was not the unique, lowest-energy conformation, were discarded.

Since fragment-based methods are poorly suited to the prediction of structures with large amounts of D-amino acid content, such as NC_cHh_DL_D1, we developed a new, fragment-free algorithm for validation of these topologies. This algorithm, which we call "simple_cycpep_predict", uses the same GenKIC-based sampling approach used to build backbones for design, with additional steps of filtering solutions based on disulfide geometry, optimizing sidechain rotamers, and gradient-descent energy minimization. Because the search space is vast, even with the constraints imposed by the N-C cyclic geometry and the disulfide bond(s), we further biased the search by setting mainchain torsion values for residues in the middle of the helices to helical values (a Gaussian distribution centred on phi=-61°, psi=-41° for the α_R helix and on phi=+61°, psi=+41° for the α_L helix).

Molecular dynamics-based computational validation

We carried out further molecular dynamics-based validation on those designs for which the *ab initio* or simple_cycpep_predict algorithms predicted high-quality energy landscapes. Similar to strategies described previously[31,32], we used multiple short and independent trajectories, starting with different initial velocities to analyze the conformational flexibility and kinetic stability of designed peptides. MD simulations were performed in explicit solvent conditions using the AMBER12 package and Amber ff12sb force field[33]. A rectangular water box with 10 Å buffer of TIP3P water[34] in each direction from the peptide was used for simulations. Sodium and chloride counterions were added to neutralize the system. The solvated system was minimized in two steps: solvent was first minimized for 20,000 cycles while keeping restraints on the peptide, followed by minimization of the whole system for another 20,000 cycles. At the start of simulations, the system was slowly heated from 0 K to 300 K under constant volume with positional restraints on the peptide of 10 kcal/(mol·Å) for 0.1 ns. For each selected peptide, 50 independent simulations starting with different initial velocities were performed. Each simulation started with the energy-minimized designed model, and was carried out for ~3.5 ns. Periodic boundary conditions were used with a constant temperature of 300 K using the Langevin thermostat[35] and a pressure of 1 atm with isotropic molecule-based scaling. A cutoff of 10 Å was used for the Lennard-Jones potential and the Particle Mesh Ewald method[36] to calculate long-range electrostatic interactions. The SHAKE algorithm[37] was applied to all bonds involving H atoms and an integration step of 2 fs was used for the simulations with amber12 PMEMD in the NPT ensemble. At the conclusion of the simulations, all the trajectories were analysed using the Amber12 package, and VMD[38], for fluctuations in RMSD, and the convergence (or the lack thereof) to the designed structure among all the trajectories. Distribution of RMSD values at the end of all trajectories was also analyzed, although the beginning two-thirds of each trajectory was discarded as a burn-in period. An example of using MD-based screening for three designs of the same topology is shown in Extended Data Figure 3.

Prediction of mutational tolerance

Since the designed peptides presented in this study are intended to be used as starting points for designing binders to targets of therapeutic interest, we sought to examine the extent to which the designs can tolerate mutations (such as those that must be introduced to create a binding surface). Due to the computational expense of the mutational analysis, we focused on the NC_cHh_DL_1 design, mutating each position in sequence to each of alanine, arginine, aspartate, phenylalanine and carrying out a full structure prediction simulation for each. These mutations covered each class of mutation (elimination of the sidechain, introduction of a positive or negative charge, introduction of a bulky aromatic sidechain, or introduction of a small aliphatic sidechain). Mutations preserved chirality (*i.e.* only D-amino acid to D-amino acid or L-amino acid to L-amino acid mutatinos were considered). Simulation runs were carried out on the Argonne Leadership Computing Facility's Blue Gene/Q supercomputer ("Mira") using a version of the Rosetta simple_cycpep_predict application parallelized using the Message Passing Interface (MPI). A typical prediction run for a single mutation occupied 512 16-core nodes for 2.5 hours (approx. 20,000 CPU-hours per run), and produced on the order of 25,000 sampled, closed conformations with good disulfide geometry. For each mutation considered, 50 trajectories were also carried out in which the mainchain was perturbed slightly and relaxed. The resulting collection of samples (from structure prediction and relaxation) was then used to calculate a goodness-of-funnel metric, termed P_{near}, by the following expression:

$$(1) \qquad P_{near} = \frac{\sum_{i=1}^{N} e^{-RMSD_i^2/\lambda^2} e^{-E_i/(k_B T)}}{\sum_{j=1}^{N} e^{-E_j/(k_B T)}}$$

The value of P_{near} ranges from 0 (a poor funnel with low-energy alternative conformations or poor sampling close to the design conformation) to 1 (a funnel with a unique low-energy conformation very close to the design conformation). N is the number of samples, and E_i and

$RMSD_i$ represent the Rosetta score and RMSD from the design structure of the i^{th} sample, respectively. The parameter controls how close a state must be to the design if it is to be considered native-like. This was set to 1 Å. Similarly, the parameter k_BT governs the extent to which the shallowness or depth of the folding funnel affects the score. This was assigned a value of 1 Rosetta energy unit. The P_{near} metric provided a basis for comparison for the mutations considered.

Code availability

All the methods described in this report were implemented in the Rosetta software suite (www.rosettacommons.org). Rosetta software is available free to academic and non-commercial users. Commercial licenses for the suite are available *via* the University of Washington Technology Transfer Office. Design protocols were implemented using the RosettaScripts interface available within the Rosetta software suite. Input files and command line arguments for each step in our peptide design pipeline are available in the **Supplementary Information**.

Protein purification of genetically encodable disulfide-rich peptides

Genes of designed disulfide-rich peptides were cloned into the vector pCDB180 (which we have made available *via* Addgene) using Gibson Assembly[39]. Protein expression from *E. coli* was carried out using a large N-terminal fusion domain consisting of: the native *E. coli* protein OsmY to direct periplasmic and extracellular localization[40], a deca-histidine tag for protein purification, and the SUMO protein Smt3 from *Saccharomyces cerevisiae* to chaperone folding and provide a mechanism for scarless cleavage of the fusion from the designed protein[41]. Designed proteins were expressed from BL21*(DE3) *E. coli* (Invitrogen), and expression cultures were grown overnight with incubation at 37 °C and shaking at 225 RPM. Following expression *via* Studier autoinduction[42], a periplasmic extract[43] was prepared by washing cells with: 20% sucrose, 30 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 1 mg/mL lysozyme. Protein was purified

from the bacterial-conditioned medium and/or the periplasmic extract by immobilized metal-affinity chromatography (IMAC). During screening, fusion protein was purified from the bacterial-conditioned medium of 50 mL cultures, which typically yielded 9 ± 4 mg of protein (prior to removal of the fusion protein). Protein expression from mammalian cells was carried out using the Daedalus[10] system, as previously described in detail. With both purification systems, purified fusion proteins were cleaved by a site-specific proteins, SUMO protease for *E. coli* and TEV protease for Daedalus, followed by a secondary IMAC step. The final designs were purified to homogeneity by reverse-phase high-performance liquid chromatography on an Agilent 1260 HPLC equipped with a C-18 Zorbax SB-C18 4.6 x 150mm column. Solvent A (Water + 0.1%TFA) and solvent B (Acetonitrile + 0.1%TFA) were run using the following gradient: 0-5% solvent B (5 minutes), 5-45% solvent B(40 minutes).

Synthesis and purification of non-canonical peptides

Linear and cyclic peptides were synthesized as previously described[44]. Briefly, peptides were synthesized using automated solid phase peptide synthesis with Fmoc (9-fluorenylmethyloxycarbonyl) strategy. Cyclic reduced peptides were obtained after cleavage of the sidechain-protected peptides from the resin, ligation of both termini and the cleavage of sidechain protecting groups. Linear reduced peptides were collected by cleaving the sidechain protecting groups and resin from the peptides simultaneously. All linear or cyclic reduced peptides were oxidized at room temperature in a buffer containing 0.1 M NH_4HCO_3, where the peptide concentration was 0.25 mg/mL. After 48 h, the mixture was acidified with trifluoroacetic acid, loaded onto a semi-preparative column and purified by RP-HPLC.

Mass spectrometry

Intact samples for each genetically encodable peptide were diluted in loading buffer with 0.1% formic acid and analyzed on a Thermo Scientific Orbitrap Fusion Tribrid Mass Spectrometer via

data-dependent acquisition. Liquid chromatography consisted of a 60 minute gradient across a 15 cm column (75 µm internal diameter) packed with C_{18} resin with a 3cm kasil frit trap (150 µm internal diameter) packed with C_{12} resin. For disulfide connectivity analysis, peptides were digested with sequencing grade modified trypsin (Promega) at 1:50, enzyme to substrate, concentration for 1 hour at 37°C then desalted via mixed-mode cationic exchange (MCX). Peptide samples were dried under vacuum and resuspended in 0.1% formic acid. Digested samples were analyzed using both data-dependent acquisition and targeted methods.

Thermal and chemical denaturation experiments

Circular dichroism (CD) wavelength and temperature scans were recorded on AVIV model 420 or Jasco J-1500 CD spectrometer. For thermal denaturation, peptides samples were prepared at 0.07-0.2 mg/ml final concentration in 10 mM sodium phosphate buffer (pH 7.0). Wavelength scans from 195 nm to 260 nm were recorded at 25 °C, 55 °C, 95 °C, and again after cooling back to 25 °C. For chemical denaturation experiments, samples for each peptide were prepared in the presence of 0 M to 6 M GdnHCl concentrations. The concentration of GdnHCl was measured by refractometry[45]. Peptide samples were also prepared in the presence of 2.5 mM TCEP (TCEP was pre-equilibrated to pH 7.0 prior to addition), and incubated for 3 hours. Peptide concentrations were the same across all samples. Wavelength scans from 190 nm to 260 nm were recorded for each sample in 0.1 cm cuvette.

NMR analysis and structure determination of genetically encodable disulfide-rich peptides

Agilent NMR spectrometers operating at ^{1}H resonance frequencies between 500 to 750 MHz equipped with $^{1}H\{^{15}N,^{13}C\}$ probes were used to acquire NMR data for gEHE_06, gEEHE_02, gEEH_04, and gHHH_06. The peptides were all uniformly ^{15}N-labeled with gEEH_04 and

gHHH_06 also ~10% labeled with ^{13}C. The peptides were suspended in 50 mM sodium chloride, 20 mM sodium acetate, pH 4.8 (gEHE_06 and gEEHE_02) or 50 mM sodium phosphate, 4 μM 4,4-dimethyl-4-silapentane-1-sulfonic acid, 0.02% sodium azide, pH 6.0 (gEEH_04 and gHHH_06) at concentrations between 1.5 and 0.5 mM. The ^{1}H, ^{13}C, and ^{15}N chemical shifts of the backbone and sidechain resonances were assigned by analysis of two-dimensional [^{15}N,^{1}H] HSQC, [^{13}C,^{1}H] HSQC (aliphatic and aromatic), [^{1}H,^{1}H] TOCSY, and [^{1}H,^{1}H] NOESY spectra, and three-dimensional (3D) ^{15}N-resolved [^{1}H,^{1}H] TOCSY, ^{15}N-resolved [^{1}H,^{1}H] NOESY, HNCA, HNCO, and HNHA spectra acquired at 20 °C (for gEHE_06 and gEEHE_02) and 25 °C (gEEH_04 and gHHH_06), respectively. Mixing times of 90 ms (gEHE_06 and gEEHE_02) and 200 ms (gEEH_04 and gHHH_06) were used for 2D and 3D NOESY, respectively. Slowly exchanging amides were identified for gEHE_06 and gEEHE_02 by lyophilizing a ^{15}N-labeled protein, re-dissolving in D_2O, and collecting a 2D [^{15}N,^{1}H] HSQC spectrum ~10 minutes after re-dissolving the protein. The resulting D_2O sample was subsequently used to collect additional 2D [^{1}H-^{1}H] TOCSY and [^{1}H-^{1}H] NOESY data. Stereospecific assignments for the Val and Leu methyl groups were obtained for gEEH_04 for the 10% fractionally ^{13}C-labelled sample[46,47]. Because it was not economical to prepare uniformly ^{13}C-labelled peptides by autoinduction, established triple-resonance NMR backbone assignment protocols could not be used. Instead, the carbon resonances were assigned by analyzing the 2D [^{1}H,^{1}H] TOCSY spectra along with [^{13}C,^{1}H] HSQC spectra (collected at natural ^{13}C abundance for gHHH_06, gEHE_06 and gEEHE_02). For gEEH_04, which was 10% fractional ^{13}C-labeled,the assignments were complemented with HNCA spectra. NMR data were processed using the Felix2007 (MSI, San Diego, CA) and PROSA (v6.4) programs and were analyzed using the programs Sparky (v3.115), XEASY, or CARA. Proton chemical shifts were referenced to internal DSS, while ^{13}C and ^{15}N chemical shifts were referenced indirectly via gyromagnetic ratios. Chemical shifts, NOESY peak lists and time domain NMR data were deposited in the BioMagResBank (for accession numbers see **Supplemental Table 2-1**).

Isotropic overall rotational correlation times of 1.6 - 1.3 ns were inferred from averaged backbone ^{15}N spin relaxation times (www.nmr2.buffalo.edu/nesg.wiki), indicating that all peptides are monomeric in solution. The ^{1}H, ^{13}C, and ^{15}N chemical shift assignments and NOESY peak lists were used for iterative structure calculations using the program CYANA (v 2.1 and 3.97). Chemical shifts were used to derive dihedral Ψ and Φ angle constraints using the program TALOS+[48] for residues located in well-defined regular secondary structure elements. For the final structure calculation, hydrogen bond restraints[11] were also introduced for gEHE_06 and gEEHE_02, for slowly exchanging amide protons. The resulting ensemble of 20 CYANA conformers was refined by restrained molecular dynamics in an 'explicit water bath' using the program CNS (v1.3)[49]. Structural quality was assessed using the online Protein Structure Validation Suite (PSVS, v1.5)[50]. The structural statistics are summarized in **Supplemental Table 2-1**. The coordinates for the 20 conformers representing the solution structures were deposited in the PDB (for accession numbers see **Supplemental Table 2-1**).

NMR analysis and structure determination of non-canonical peptides

Each non-canonical peptide (1 mg) was dissolved in 500 mL of 10% D_2O/90% H_2O or 100% D_2O (~pH 4). NMR spectra were recorded at 298K on a Bruker Avance-600 spectrometer. Two-dimensional NMR experiments included TOCSY with an 80 s MLEV-17 spin lock, NOESY (200 ms mixing time), ECOSY, as well as natural-abundance ^{13}C and ^{15}N HSQC. Solvent suppression was achieved using excitation sculpting. Spectra were processed using Topspin 2.1 then analysed using CcpNmr Analysis[51]. Chemical shifts were referenced to internal 2,2-dimethyl-2-silapentane-5-sulfonate (DSS).

Initial structures were generated using CYANA and were based upon distance restraints derived from NOESY spectra recorded in both 10% and 100% D_2O. The following restraints were also

included: disulfide bonds, hydrogen bonds as indicated by slow D_2O exchange and sensitivity of amide proton chemical shift to temperature, chi1 restraints from ECOSY and NOESY data, and backbone phi and psi dihedral angles generated using the program TALOS-N[52]. The final set of structures was generated within CNS[53] using torsion angle dynamics, refinement and energy minimization in explicit solvent and protocols as developed for the RECOORD database[54]. Final structures were assessed for stereochemical quality using MolProbity[55].

X-ray crystallography

The gEHEE_06 peptide was purified by size exclusion chromatography on an AKTA Pure using a GE HiLoad 16/600 Superdex 75 pg column, concentrated to 50mg/ml and crystallized by vapor diffusion over well solutions of 100mM citrate (pH 3.5), and 25% PEG3350. Selected crystal was transferred to a cryo-solution of 100mM citrate (pH 3.5), 20% PEG3350, with 15% glycerol, and diffraction data were collected on a Rigaku Micromax-007HF with a Saturn944+ CCD detector and integrated and scaled with HKL-2000. Initial phases were determined by molecular replacement using Phaser[56] as implemented in the CCP4 software suite with coordinates derived from a Rosetta model for the scaffold. Molecular replacement found 2 molecules per asymmetric unit (ASU). This solution was iteratively refined with the program Refmac followed by model building with COOT, yielding a crystallographic R-values (Rcryst = 39.9%, Rfree = 42.5%). Based on the Matthews' coefficient, the crystals should have contained 3 molecules per ASU in order to have a reasonable solvent content of 45%. At this point positive electron density appeared that allowed for the manual positioning of a third molecule in the ASU and improving the R-values (R^{cryst} = 32.0%, R_{free} = 34.9%). The model was further improved by including solvent molecules and TLS refinement. The quality of the final model was assessed using ProCheck and Molprobity (overall score: 100th percentile). The final model has been deposited in the PDB with accession code 5JG9. Crystallographic statistics are reported in **Supplementary Table S2-2**.

Surface redesign

In attempt to reduce solubility and enhance crystallization, we performed a redesign solvent-exposed residues of designs representing each major topological category (mixed α/β, all β-sheet, all α-helical). Two re-surfaced variants were selected for each design bearing between one to two solvent-exposed tyrosine residues. We then expressed and purified these resurfaced designs using Daedalus, all of which expressed solubly and exhibited a redox-sensitive migration time by reverse-phase HPLC. We were only able to obtain diffracting protein crystals for re-design gEEHE_2.1_02_0008, from topology ββαβ, which diffracted to 2.90 Å resolution (**Supplemental Table 2-2**). However, Matthews calculations predicted non-crystallographic symmetry with approximately nineteen copies in the asymmetric unit, and attempts to phase the crystal by molecular replacement were unsuccessful, as were attempts at reproducing the crystal outside of the initial screen.

Additional References

26. Huang, P.-S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One* **6,** e24109 (2011).

27. Harrison, P. M. & Sternberg, M. J. Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage. *J. Mol. Biol.* **244,** 448–463 (1994).

28. Lee, J., Lee, D., Park, H., Coutsias, E. A. & Seok, C. Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins* **78,** 3428–3436 (2010).

29. Rodriguez-Granillo, A., Annavarapu, S., Zhang, L., Koder, R. L. & Nanda, V. Computational design of thermostabilizing D-amino acid substitutions. *J. Am. Chem. Soc.* **133,** 18750–18759 (2011).

30. Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309,** 1868–1871 (2005).

31. Caves, L. S., Evanseck, J. D. & Karplus, M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.* **7,** 649–666 (1998).

32. Wijma, H. J. *et al.* Computationally designed libraries for rapid enzyme stabilization. *Protein Eng. Des. Sel.* **27,** 49–58 (2014).

33. D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Götz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P.A. Kollman. *AMBER 12*. (University of California, San Francisco, 2012).

34. Jorgensen, W. L. & Corky, J. Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density. *J. Comput. Chem.* **19,** 1179–1186 (1998).

35. Loncharich, R. J., Brooks, B. R. & Pastor, R. W. Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **32,** 523–535 (1992).

36. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N · log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089 (1993).

37. Ryckaert, J.-P., Giovanni, C. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23,** 327–341 (1977).

38. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol.*

Graph. **14,** 33–8, 27–8 (1996).

39. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).

40. Kotzsch, A. *et al.* A secretory system for bacterial production of high-profile protein targets. *Protein Sci.* **20,** 597–609 (2011).

41. Marblestone, J. G. *et al.* Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Sci.* **15,** 182–189 (2006).

42. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41,** 207–234 (2005).

43. Neu, H. C. & Heppel, L. A. The release of enzymes from Escherichia coli by osmotic shock and during the formation of spheroplasts. *J. Biol. Chem.* **240,** 3685–3692 (1965).

44. Cheneval, O. *et al.* Fmoc-based synthesis of disulfide-rich cyclic peptides. *J. Org. Chem.* **79,** 5538–5544 (2014).

45. Pace, C. N. Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* **131,** 266–280 (1986).

46. Neri, D. *et al.* Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional carbon-13 labeling. *Biochemistry* **28,** 7510–7516 (1989).

47. Herve du Penhoat C, E. al. The NMR solution structure of the 30S ribosomal protein S27e encoded in gene RS27_ARCFU of Archaeoglobus fulgidis reveals a novel protein fold. - PubMed - NCBI. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15096641. (Accessed: 13th July 2016)

48. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44,**

213–223 (2009).

49. Linge, J. P., Williams, M. A., Spronk, C. A. E. M., Alexandre M J & Michael, N. Refinement of protein structures in explicit solvent. *Proteins: Struct. Funct. Bioinf.* **50,** 496–506 (2003).

50. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66,** 778–795 (2007).

51. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins: Struct. Funct. Bioinf.* **59,** 687–696 (2005).

52. Shen, Y. & Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **56,** 227–241 (2013).

53. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2,** 2728–2733 (2007).

54. Nederveen, A. J. *et al.* RECOORD: A recalculated coordinate database of 500 proteins from the PDB using restraints from the BioMagResBank. *Proteins: Struct. Funct. Bioinf.* **59,** 662–672 (2005).

55. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 12–21 (2010).

56. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).