REVIEW

doi:10.1038/nature19946

The coming of age of *de novo* protein design

Po-Ssu Huang[1,2*†], Scott E. Boyken[1,2,3*] & David Baker[1,2,3]

There are 20^{200} possible amino-acid sequences for a 200-residue protein, of which the natural evolutionary process has sampled only an infinitesimal subset. *De novo* protein design explores the full sequence space, guided by the physical principles that underlie protein folding. Computational methodology has advanced to the point that a wide range of structures can be designed from scratch with atomic-level accuracy. Almost all protein engineering so far has involved the modification of naturally occurring proteins; it should now be possible to design new functional proteins from the ground up to tackle current challenges in biomedicine and nanotechnology.

Proteins mediate the fundamental processes of life, and the beautiful and varied ways in which they do this have been the focus of much biomedical research for the past 50 years. Protein-based materials have the potential to solve a vast array of technical challenges. Functions that naturally occurring proteins mediate include: the use of solar energy to manufacture complex molecules; the ultrasensitive detection of small molecules (olfactory receptors[1]) and of light (rhodopsin[2]); the conversion of pH gradients into chemical bonds (ATP synthase[3]); and the transformation of chemical energy into work (actin and myosin[4]). Not only are these functions remarkable but they are encoded in sequences of amino acids with extreme economy. Such sequences specify the three-dimensional structure of the proteins, and the spontaneous folding of extended polypeptide chains into these structures is the simplest case of biological self-organization. Despite the advances in technology of the past 100 years, human-made machines cannot compete with the precision of function of proteins at the nanoscale and they cannot be produced by self-assembly. The properties of naturally occurring proteins are even more remarkable when considering that they are essentially accidents of evolution. Instead of a well-thought-out plan to develop a machine to use proton flow to convert ADP to ATP, selective pressure operated on randomly arising variants of primordial proteins, and there were also hundreds of millions of years in which to get it right.

In this Review, we propose that if the fundamentals of protein folding and protein biochemistry and biophysics can be understood, it should become possible to design from the ground up a vast world of customized proteins that could both inform basic knowledge of how proteins work and address many of the important challenges that society faces. We focus specifically on the problem of *de novo* protein design: the generation of new proteins on the basis of physical principles with sequences unrelated to those in nature. We describe the methodological advances that underlie progress in *de novo* protein design as well as provide an overview of the diversity of designed structures for which the high-resolution X-ray crystallography structure or nuclear magnetic resonance (NMR) structure is in atomic agreement with the design model. Almost all protein engineering so far has involved the modification of naturally occurring proteins to tune or alter their function using techniques such as directed evolution[5–7], which involves cycles of generating and selecting variation in the laboratory. Because these efforts have been extensively reviewed[8,9] and are essentially extensions of evolutionary processes, they will not be discussed here.

It is useful to begin by considering the fraction of protein sequence space that is occupied by naturally occurring proteins (Fig. 1a). The number of distinct sequences that are possible for a protein of typical length is 20^{200} sequences (because each of the protein's 200 residues can be one of 20 amino acids), and the number of distinct proteins that are produced by extant organisms is on the order of 10^{12}. Evidently, evolution has explored only a tiny region of the sequence space that is accessible to proteins. And because evolution proceeds by incremental mutation and selection, naturally occurring proteins are not spread uniformly across the full sequence space; instead, they are clustered tightly into families. The huge space that is unlikely to be sampled during evolution is the arena for *de novo* protein design. Consequently, evolutionary processes are not a good guide for its exploration — as discussed already, they proceed incrementally and at random. Functional folded proteins have been retrieved from random-sequence libraries[10–12] but this is a laborious (and non-systematic) process. Instead, it should be possible to generate new proteins from scratch on the basis of our understanding of the principles of protein biophysics.

Our approach is built on the hypothesis that proteins fold into the lowest energy states that are accessible to their amino-acid sequences, as originally proposed by Christian Anfinsen[13]. Given a suitably accurate method for computing the energy of a protein chain, as well as methods for sampling the space of possible protein structures and sequences, it should be possible to design sequences that fold into new structures. There are two challenges in implementing this approach: first, the energy of a system cannot be computed with perfect accuracy; and second, the space of possible structures and sequences is very large and therefore difficult to search comprehensively. In this Review, we describe the physical basis for the energy function used in the design calculations and the approaches that are used to overcome the sampling problem. The discussion is based on our experience of developing the Rosetta structure prediction and design methodology[14]; other *de novo* protein design software is described elsewhere[15–17].

Considerable recent progress in protein design is attributable not only to the advances in understanding and computational methods that are the focus of this Review, but also to advances in two other areas. The first is computing: *de novo* protein design is computationally expensive, and the steady increase in the availability of computing

[1]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA. [2]Institute for Protein Design, University of Washington, Seattle, Washington 98195, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. *These authors contributed equally to this work. †Present address: Department of Bioengineering, Stanford University, Stanford, California 94305, USA.

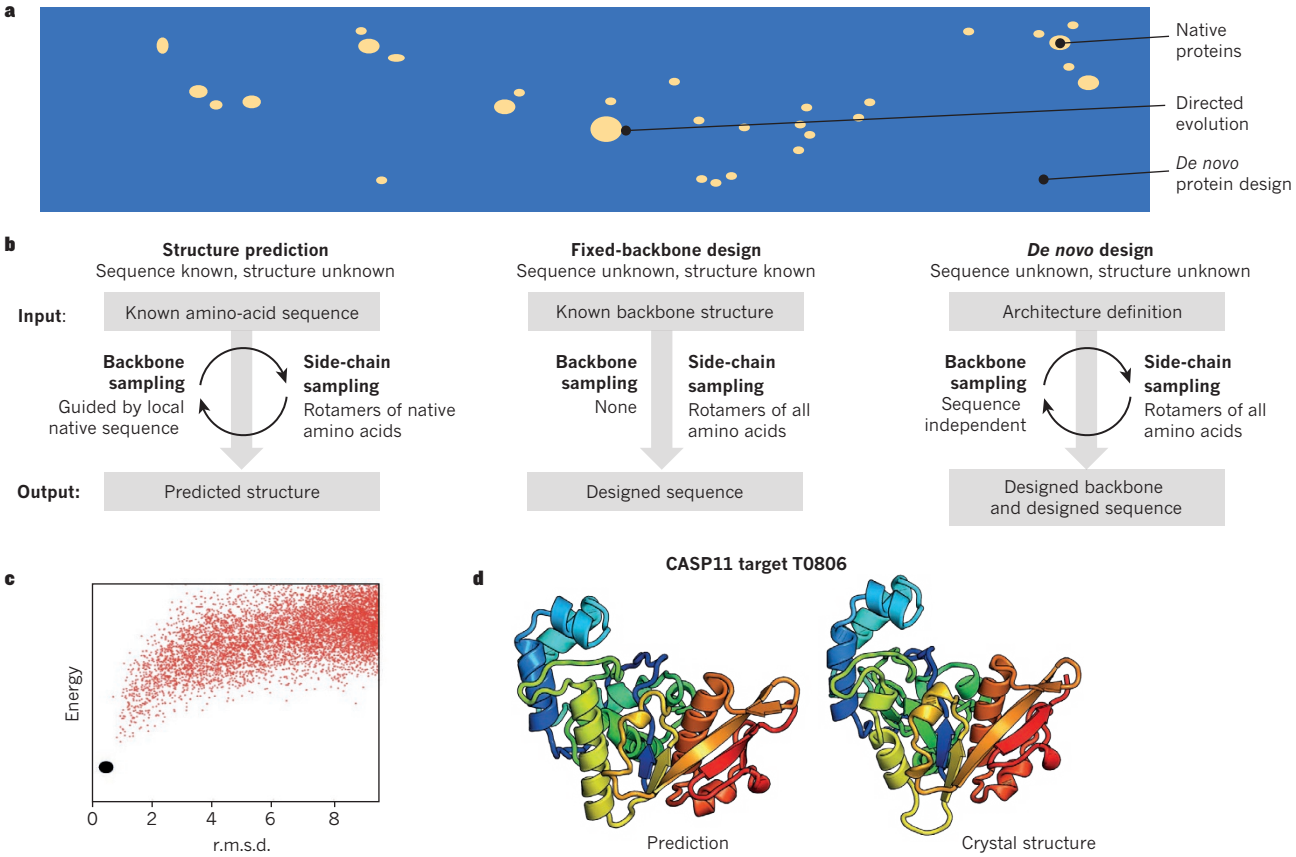

Figure 1 | Methods for *de novo* protein design. **a**, A schematic of the protein sequence space. Evolution has sampled only a tiny fraction of the total possible sequence space (blue), and the incremental nature of evolution results in tightly clustered families of native proteins (beige), which are analogous to archipelagoes in a vast sea of unexplored territory. Directed evolution is restricted to the region of sequence space that surrounds native proteins, whereas *de novo* protein design can explore the whole space. **b**, Structure prediction, fixed-backbone design and *de novo* protein design are global optimization problems with the same energy function but different degrees of freedom. In structure prediction, the sequence is fixed and the backbone structure is unknown; in fixed backbone protein design, the sequence is unknown but the structure is fixed; and in *de novo* protein design, neither is known. **c**, Example of an energy landscape generated from fixed-sequence protein-structure prediction calculations. The red dots represent lowest-energy structures from independent Monte Carlo trajectories, which are plotted according to their similarity to the target structure (black dot) along the *x* axis; structural similarity is measure by root-mean-square deviation (r.m.s.d.). In *de novo* design efforts, designed sequences for which the calculations converge on the target designed structure are selected for experimental characterization. **d**, Blind, *de novo* structure prediction (left) for the critical assessment of protein structure prediction (CASP)11 target T0806, which has no sequence similarity to any protein of known structure, using coevolution-derived contact constraints[27]. The crystal structure (Protein Data Bank accession code 5CJA) is shown for comparison (right). The ability to predict the structure of proteins with new folds with this level of accuracy enables large-scale structural genomics by means of computer calculation rather than experiment.

power has greatly enabled the work that we describe, much of which was completed using volunteer computing through the Rosetta@home project. The second advance is the synthetic manufacture of DNA. Because the proteins that are being designed do not exist in nature, genes that encode their amino-acid sequences also do not exist. To produce designed proteins in an organism such as *Escherichia coli*, synthetic genes that encode the designed amino-acid sequences must first be manufactured. Methods for DNA synthesis have improved dramatically in the past 10 years, greatly reducing the cost of synthesizing genes for *de novo* designed proteins and increasing the number of computational designs that can be tested experimentally.

Physical principles that underlie protein design

The driving force for protein folding is the burial of hydrophobic residues in the protein's core, away from the solvent. To minimize the size of the cavity that the protein occupies in water, and to maximize van der Waals forces, the side chains in the core must be packed closely but without energetically unfavourable atomic overlaps. Polar groups that interact with the solvent in the unfolded state that become buried upon protein folding must form intra-protein hydrogen bonds to compensate, otherwise the large energy cost of stripping water will disfavour folding[18]. The hallmark features of globular protein structures follow from these considerations: α-helical and β-sheet secondary structures, in which the polar carbonyl and amide groups of the polypeptide backbone can form hydrogen bonds, assemble in such a way that non-polar side chains fit together like the pieces of a jigsaw puzzle to form densely packed cores. Interactions of amino-acid side chains with neighbouring backbone atoms also contribute to the free energy of folding: these include hydrogen bonds at the termini of α-helices and steric and torsional effects that favour certain backbone geometries and disfavour others. For example, the amino acid proline has a rigid internal ring and is compatible with only a narrow range of backbones, whereas glycine, which lacks a side chain, enables tight bending of the backbone in loops between secondary structures.

This picture of protein folding is implemented in an energy function that captures the interactions of the atoms in proteins with each other and with the solvent. The main contributors to this energy function are van der Waals forces that favour close atomic packing, steric repulsion, electrostatic interactions and hydrogen bonds, solvation and the torsion energies of backbone and side-chain bonds. Predicting and designing protein structures using such an energy function requires methods for sampling alternative backbone and side-chain conformations to identify structures and sequences with very low energy. Different methods are used for backbone and side-chain sampling

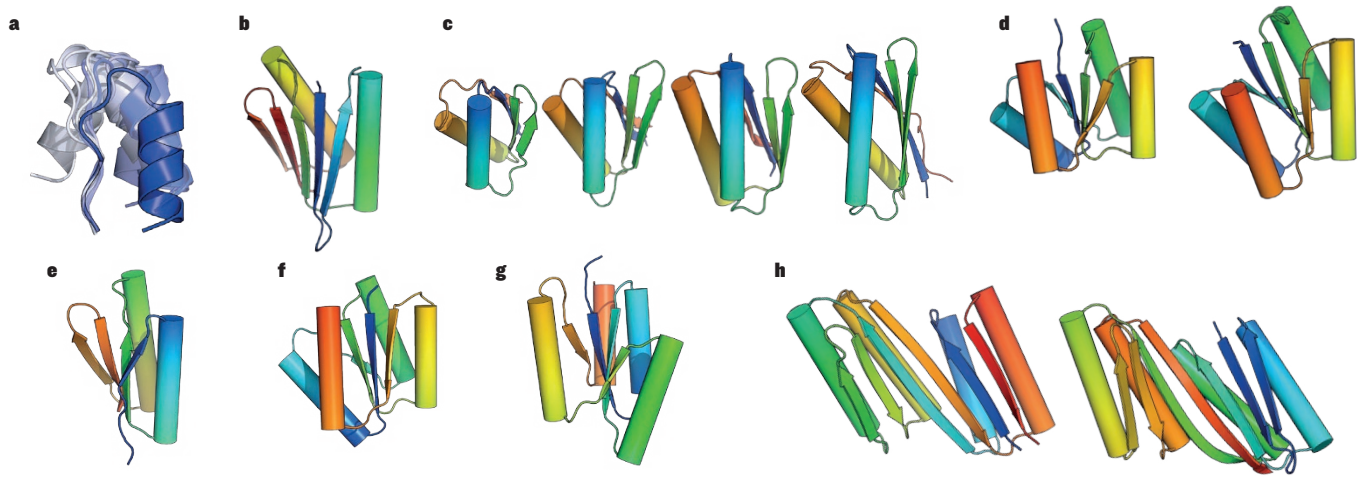

Figure 2 | Designing αβ proteins. **a**, Sampling alternative backbones for a β-strand-turn-α-helix blueprint through fragment assembly. **b–g**, *De novo* designed ideal αβ proteins with high-resolution NMR or X-ray structures that are in very close agreement with design models[28,36,37]. **b**, Top7. **c**, Ferredoxin folds of varying shapes and sizes. **d**, Rossmann 2×2 folds. **e**, IF3-like fold. **f**, P-loop 2×2 fold. **g**, Rossmann 3×1 fold. **h**, Larger, more complex structures that were generated from domains in **b** and **c**[38].

(Fig. 1b). In side-chain sampling, discrete combinatorial optimization is used to identify amino acids and side-chain conformations (known as rotamers) that lead to low-energy, closely packed protein cores[19–22]. If the amino-acid sequence is known in advance, such as in the protein structure prediction problem (predicting the structure of a protein from its amino-acid sequence), the amino-acid identities have already been fixed and the search covers the discrete rotameric states of each side chain. But if the sequence is unknown, such as in the protein design problem (finding a sequence that folds into a specified structure), both the amino-acid identities and the rotameric states are sampled. Backbone sampling often frames the initial stages of the search as a discrete optimization problem by taking advantage of biases in the local sequence towards a subset of possible local structures. In the later stages of refinement, continuous optimization methods such as quasi-Newton minimization are used to fine-tune the packing and the electrostatic interactions and hydrogen bonding of the structure.

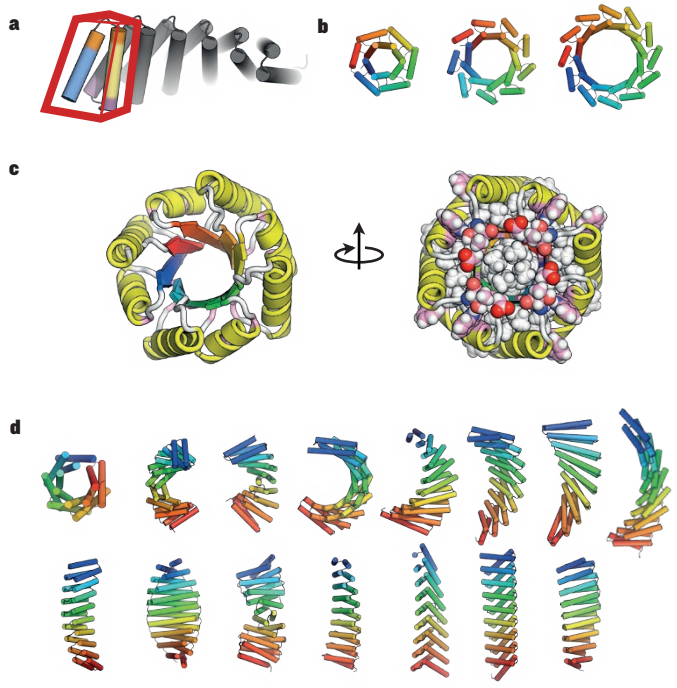

Figure 3 | Designing proteins with internal symmetry. **a**, The propagation of a single repeat unit generates a larger structure. **b–d**, *De novo* designed repeat proteins with high-resolution X-ray structures that are in very close agreement with design models. **b**, *De novo* α-helical toroids[41]. **c**, An ideal TIM barrel with four-fold symmetry. Packing features (white) and polar-fold determinants (pink spheres) are shown[42]. **d**, Tandem repeat proteins with a variety of twists and curvatures that go beyond the topologies that are observed in nature[43].

Protein–structure prediction

It is useful to first consider the *ab initio* structure prediction problem: finding the lowest energy structure for fixed amino-acid sequence in the absence of information about the structures of evolutionarily related proteins. Because the amino-acid sequence is fixed, side-chain combinatorial optimization covers only the various rotameric states and the backbone can be built from short fragments with similar local sequences[23]. An advantage of this approach is that sampling is very focused in regions where the local sequence strongly favours a particular local structure yet broad in regions where the local sequence is compatible with many conformations. It is still difficult to predict protein structures without homologues of known structure for all but the smallest proteins. The main challenge is the size of the backbone conformational space that must be sampled: the correct structure usually has a lower computed energy than all alternative structures, but it is very hard to find. However, if the sampling is guided by extra sources of information, such as co-evolution-based distance constraints[24,25], structure-prediction calculations can find the native-state energy minimum (Fig. 1c). In such cases, accurate, blind predictions of complex protein structures can be made[26,27] (Fig. 1d).

De novo protein design

Unlike in the structure-prediction and fixed-backbone design problems, in the general (*de novo*) protein design problem, both the sequence and the exact structure of the backbone are unknown (Fig. 1b). Given this, how do we effectively sample backbones from scratch? Because only a small proportion of backbone conformations can accommodate sequences with almost-perfect core packing and hydrogen bonding between the buried hydrogen-bond donors and acceptors, design calculations generally begin with a large set of (more than 10,000) alternative conformations. These initial backbones can be made either by assembling short peptide fragments[28,29] or by using algebraic equations to specify the geometry parametrically[30–35]. For each designed backbone conformation, combinatorial sequence-optimization calculations are used to identify the lowest-energy sequence for the structure. *Ab initio* structure-prediction calculations are then carried out to determine whether the designed structure is

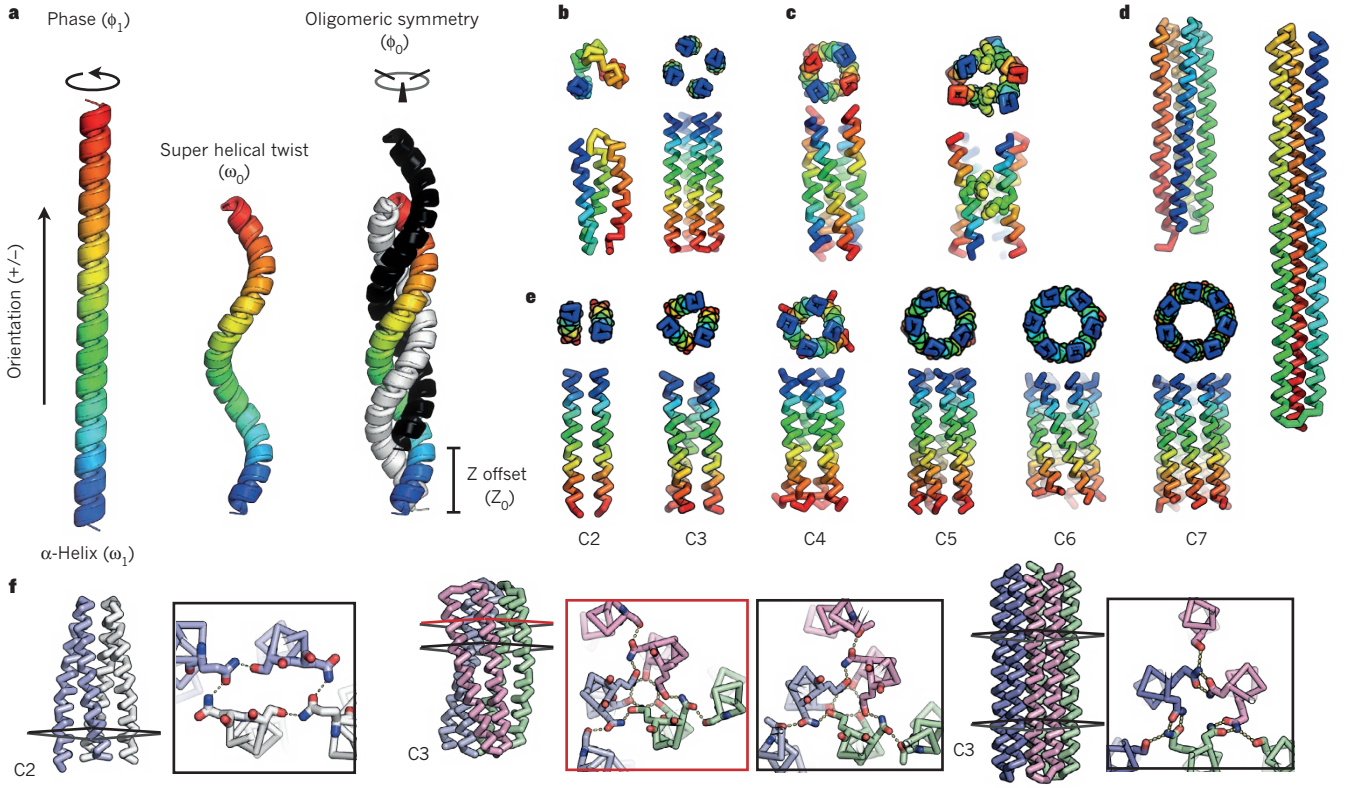

Figure 4 | *De novo* design using parametric backbone generation. **a**, Parameters that describe helical bundle geometry. **b**, The first *de novo* designed helical bundles to be structurally validated: α_3D (ref. 48) (left) and RH4 (ref. 30) (right), a right-handed coiled coil. **c**, Functional *de novo* helical bundles: a carbon nanotube-binding helix[53] (left), and a Zn^{2+} antiporter membrane protein (known as Rocker)[34]. **d**, Single-chain hyperstable helical bundles[33]: a right-handed four-helix bundle (left) and untwisted three-helix bundles (right). **e**, Homo-oligomeric single-ring helical bundles[31,33,51,52]. **f**, Homo-oligomeric *de novo* helical hairpins that form double-layered channels with hydrogen-bond network-mediated specificity[63]; the polar networks are shown as expanded cross-sections. C*n* indicates an *n*-fold cyclic symmetry operation: for example, C2 structures are homodimers and C3 structures are homotrimers.

the lowest-energy state of the designed sequence — this is an important *in silico* consistency check. *De novo* designs are usually experimentally characterized only if structure-prediction calculations that start from the designed sequence strongly converge on the designed structure (Fig. 1c).

Only a finite number of backbones can be sampled computationally. To tackle the important challenge of sequence-independent backbone construction, it is necessary to reduce the enormous space of possible backbone structures to those that are capable of being designed — that is, to those for which there is a reasonable probability that a sequence exists whose lowest-energy state is the structure. Progress towards this goal has required the investigation of sequence-independent constraints on backbone geometry. One such constraint comes from the connectivity of the polypeptide chain and the requirement that the polar atoms of the backbone either make hydrogen bonds within the chain in α-helices or β-sheets or come into contact with the solvent in exposed loops. This constraint immediately restricts the length of the secondary structures that are permitted for a given topology[36]. Another constraint comes from the limited flexibility of the polypeptide chain, which restricts the lengths of the loops that connect α-helices and β-sheets in various packing orientations[37]. Simulations and analyses of protein structures have revealed sequence-independent design principles that relate the lengths of helices, strands and loops when packed together that greatly facilitate the construction of topologies that consist of α-helices and β-sheets[36,37].

Even with these constraints, the space of possible backbones is still large. To meet the twin goals of bringing the principles that underlie protein folding and structure into sharp focus and generating robust and stable scaffolds for future functional design efforts, much *de novo* protein-design work has placed an emphasis on designing ideal protein structures with unkinked α-helices and β-strands and minimal loops. By contrast, most naturally occurring proteins contain irregular, non-canonical features that arise either from selection for function or from neutral drift. Such features complicate the structural analysis of proteins and reduce the free energy of folding. (During evolution, there was probably little pressure to optimize the free energy of folding beyond 8 kcal per mol, which corresponds to a folded-state population of more than 99.999%.)

Ideal αβ folds

A wide range of ideal αβ protein structures have been designed using the sequence-independent design principles[36,37] (Fig. 2). The design approach consists of several steps. First, an overall topology 'blueprint'[29] that is consistent with the backbone design principles is created to specify the lengths, packing arrangement and order of the constituent α-helices and β-strands, as well as the lengths of the connecting loops. Second, protein backbones that are compatible with the blueprint are assembled from protein structure fragments using a Monte Carlo approach (Fig. 2a). Third, combinatorial rotamer optimization is used to identify a low-energy amino-acid sequence for each backbone. Fourth, alternating cycles of backbone relaxation and sequence optimization are performed to achieve a sequence–structure pair with very low energy. Last, sequences that converge on the corresponding designed structure in structure prediction calculations are tested experimentally. This design approach was applied to the idealized backbones shown in Fig. 2. Synthetic genes encoding the new designed proteins were generated, and the proteins were produced in *E. coli*. The purified proteins were found to be extremely stable and had structures that were almost identical to those of the design models[28,36–38] (Fig. 2).

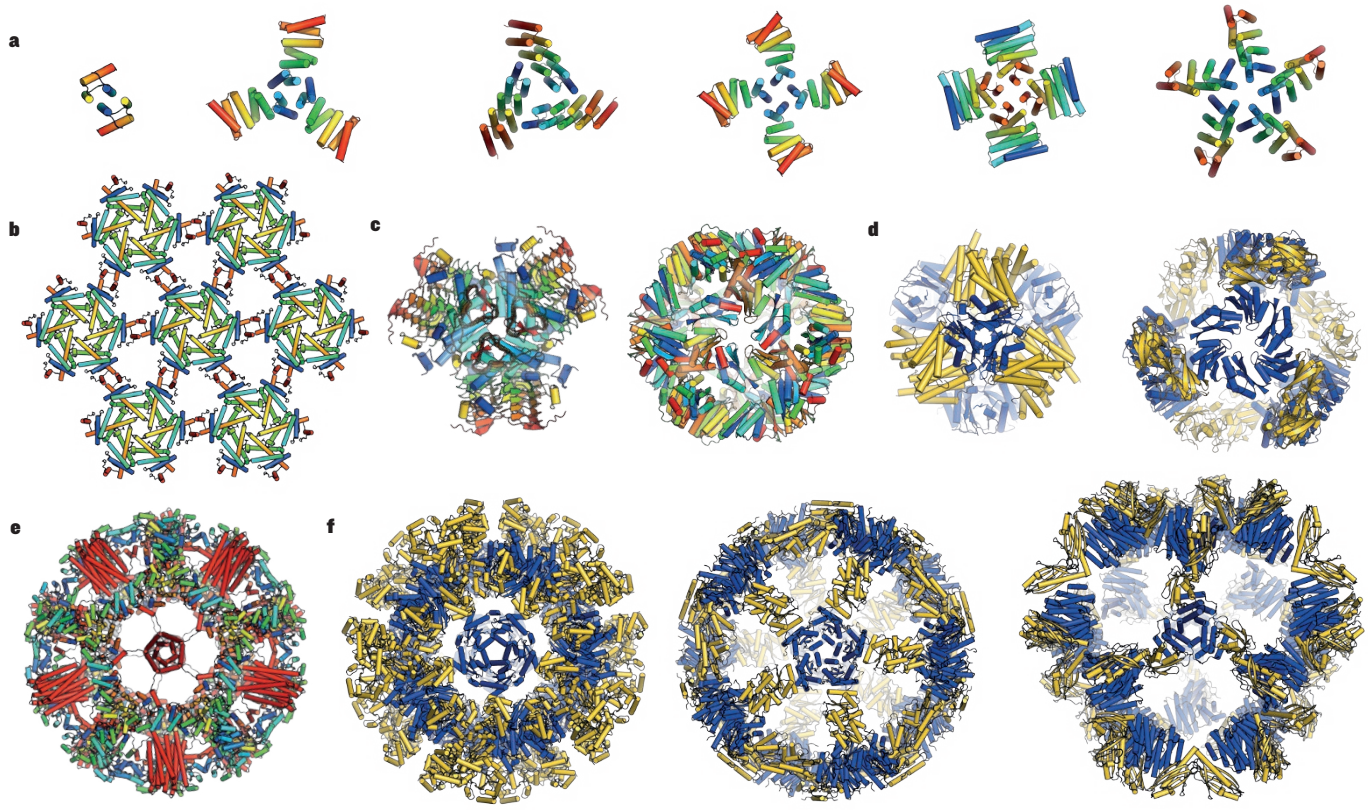

Figure 5 | Designing self-assembling nanomaterials. **a**, C2, C3, C4 and C5 symmetric homo-oligomers (ref. 78 and J. Fallas and G. Ueda, personal communication). **b**, Two-dimensional hexagonal lattice[81]. **c–f**, Self-assembling cages. **c**, A one-component tetrahedron (left) and a one-component octahedron[79] (right). **d**, Two-component tetrahedral nanoparticles[80]; the two asymmetric components are coloured in blue and yellow. **e**, A one-component hyperstable icosahedron with a *de novo* helical bundle (red helices) fused in the centre of the face[82]. **f**, Two-component megadalton-scale icosahedra[83]; the two components of each are coloured in blue and yellow.

Repeat proteins

The effort to construct *de novo* proteins with ideal backbone arrangements has led to the design of proteins with internal symmetry in which a single idealized unit is repeated numerous times[39–41] (Fig. 3). Internal symmetry reduces the size of the sequence space that must be searched and enables a relatively small unit with a known sequence–structure combination to be reused repeatedly to build larger proteins (Fig. 3a). The constraint of internal symmetry is particularly strong for closed structures in which the final repeat unit is juxtaposed with the first, such as in α-helical toroids[41] (Fig. 3b) and the TIM barrel[42] (Fig. 3c). In the TIM barrel, the backbone design principles, together with the geometry of closed β-sheets, makes four-fold symmetry the highest that can be attained and forces the two α-helices in each α–β–α–β unit to differ in length[42]. Both closed-repeat and open-repeat protein designs have been produced by introducing synthetic genes into *E. coli*, followed by experimental characterization of the purified proteins. High-resolution X-ray crystallography structures for the designs were found to be almost identical to the design models. The α-helical repeat structures have sequences and structures (Fig. 3d) that differ greatly from those found so far in nature, which suggests that naturally occurring proteins sample only a tiny fraction of the stable protein structures that can be realized[43]. These new repeated proteins are exceptionally stable; several of the open structures are denatured only by guanidine hydrochloride at concentrations of more than 6 M (D. Barrick, personal communication). By contrast, an approach to 'stitch' protein structures together from large helix-containing fragments of naturally occurring proteins generates structures with irregularities that are similar to those found in native structures[44] that present opportunities for the subsequent design of function. Contact information from native structures has also been used to guide the design of new backbone arrangements[45], including a scaffold that presents an epitope from respiratory syncytial virus to elicit a neutralizing immune response[46].

Parametric helical bundles

The use of parametric equations is a complementary approach to generating ideal backbone arrangements that provides considerable control over the global structure. Equations developed by Francis Crick enable the generation of idealized bundles of α-helices in parallel or antiparallel orientations in which the helices have arbitrary lengths, phasing, relative orientations and twists[47] (Fig. 4a). The helical bundles can be used directly in sequence-design calculations, yielding multiple-subunit oligomeric structures, or the helices can first be connected with loops to yield a single chain. Many helical bundles have been designed in this way[30,31,33,34,48–52] (Fig. 4), including a peptide that binds to carbon nanotubes[53], parallel self-assembling helical channels[31], an ion transporter[34], cages[54] and an α-helical barrel with installed hydrolytic activity[55]. The combination of parametric backbone generation with combinatorial side-chain optimization has enabled the design of larger, more diverse helical bundles[33]; like many *de novo* designed proteins, these parametrically designed proteins are extremely stable, remaining folded in 7 M guanidine hydrochloride at 95 °C.

Hydrogen–bond networks

The principles we have outlined for the *de novo* design of monomeric folds are necessary but not sufficient for controlling the specificity of protein interactions, which despite progress[56–60] remains a challenge[61]. Binding is driven by the balance between the burial of hydrophobic packing residues and peripheral polar interactions that help to solvate the monomeric state and provide structural specificity. In contrast to the double helix of DNA, in which regular arrays of central hydrogen bonds lead to the formation of a high-specificity heterodimer, the hydrogen

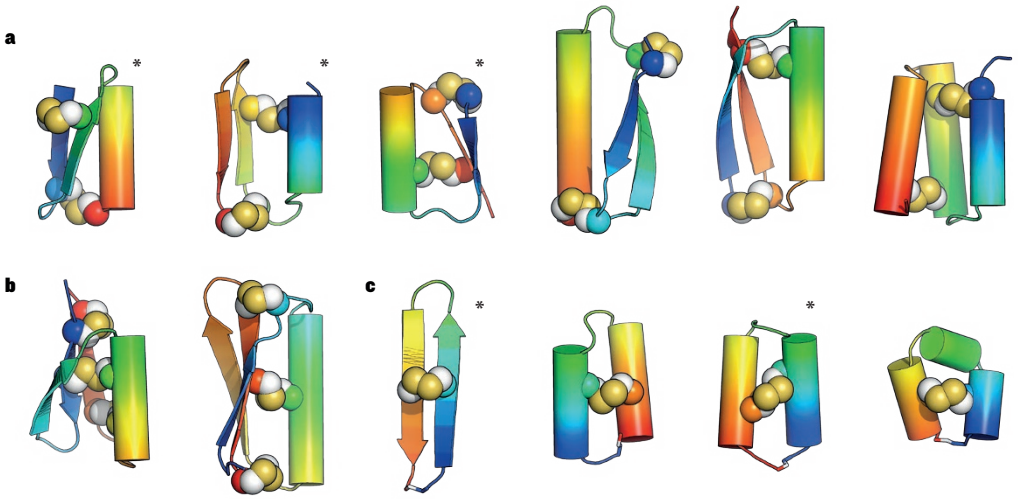

Figure 6 | Designing hyperstable *de novo* constrained peptides. **a**, **b**, Disulfide crosslinked miniproteins with two (**a**) or three (**b**) disulfide linkages (yellow spheres). **c**, Cyclic peptides with covalently linked N termini and C termini. An asterisk denotes a heterochiral design that contains a mixture of L-amino acids and D-amino acids.

bonds that form at the interfaces of naturally occurring proteins are placed irregularly and are very difficult to design[62].

A challenge when designing polar interactions is to ensure that all buried hydrogen-bond donors and acceptors form intraprotein hydrogen bonds. In the past year, it has become possible to design with atomic-level accuracy extensive networks of hydrogen bonds in which almost all of the donors and acceptors are satisfied[63]. This approach has enabled helical-bundle oligomers to be generated with a specificity that is determined by regular arrays of central hydrogen-bond networks, analogous to Watson–Crick base-pairing in DNA[64]. Identification of the rare backbones that can harbour more than one network of hydrogen bonds required the parametric generation of thousands of backbones. In the field of DNA nanotechnology[61], the limited set of Watson–Crick hydrogen bonds has been harnessed to build a wide range of shapes[65,66]; it should become possible to use similar 'digital' design principles to build structures from proteins using modular hydrogen-bond networks to encode specificity.

The design of new functions

The advances described in this Review, most of which were made in the past 3 years, demonstrate that a fundamental understanding of the principles of protein structure and protein folding has been achieved. This knowledge has enabled a wide variety of exceptionally stable protein structures and assemblies to be designed with atomic-level accuracy. (The high-resolution structures for all of the protein designs described in this Review, as determined by NMR, X-ray crystallography or electron microscopy, are in close agreement with the design models.) The potential for designing new functions on the basis of these scaffolds and the more general use of *de novo* backbone design methods is underscored by the achievements of computational protein-design efforts, in which scaffolds from naturally occurring proteins have been repurposed to carry out different functions. Such efforts have yielded enzymes that have attained high catalytic efficiencies through directed evolution[67–73], inhibitors of protein–protein interactions that can protect animals from viral infection[74] and small-molecule binding proteins that can be incorporated into *in vivo* biosensors[75–77]. The design of precise interfaces between protein subunits has enabled the creation of self-assembling, cyclic homo-oligomers (ref. 78 and J. Fallas and G. Ueda, personal communication), tetrahedra[79,80], octahedra[79] and open two-dimensional assemblies[81] (Fig. 5). Protein interface design methods have been used to create one- or two-component assemblies with icosohedral symmetry and 60 subunits[82] or 120 subunits[83], respectively. The high symmetry of these assemblies enables the multivalent presentation of antigens for vaccine applications, and the large volumes of their interior are well suited to packaging cargo for delivery to targets.

The design of constrained peptides

Because of the level of control that *de novo* protein design offers, the capabilities of the next generation of designed functional proteins could greatly exceed those of first-generation designed proteins based on native scaffolds. There is also the tremendous potential for *de novo* protein design to go beyond nature to discover new folds by incorporating new chemistries and unnatural amino acids. An example of this is the design of hyperstable peptides, which are constrained by disulfide crosslinks and cyclic peptide linkages that connect the N and C termini[84]. In this case, extensions to the design methodology enabled the use of L-amino acids and D-amino acids within the same protein design (Fig. 6). The structures of these peptides, determined experimentally through NMR and X-ray crystallography, are in close agreement to the design models, and despite the peptides being only 15–50 residues in length, most are extremely resistant to thermal and chemical denaturation.

Improving the robustness of *de novo* design

A limitation of *de novo* protein design is that only a fraction of protein designs adopt stable folded structures when produced in *E. coli*. The most frequent reasons for failure are insolubility and the formation of unintended oligomeric states (polydispersity) — experimentally determined high-resolution structures of soluble and monodisperse designs are almost always very similar to those of the design models. Insolubility and polydispersity probably arise from unanticipated intermolecular hydrophobic interactions. Increasing the robustness of designs will require improvements in the accuracy of the energy function that underlies the design process (for example, explicit modelling of the interactions of protein atoms with specific bound water molecules), more explicit negative design to disfavour alternative states and other advances in computational methodology. As the decreasing cost of synthesizing DNA enables the experimental characterization of larger numbers of protein designs, it should become increasingly possible to identify the features that differ between soluble and insoluble designs. Insight can be obtained by considering the success rate for each class of design that is described in this Review. The highest success rate from the work of our group was obtained for the cyclic and disulfide stapled peptides[84], for which seven of eight designs were soluble and monodisperse and had structures that were almost identical to the design models; the chemical staples limit alternative conformations of the designs in this class. These designs were also synthesized chemically — the lower success rate for proteins that are expressed recombinantly might be due in part to the toxicity of such proteins in *E. coli* or to other complexities of the bacterium's biology. The α-helical bundles that are mediated by networks of hydrogen bonds had a solubility of about 90%, and more than 60% of the bundles were monodisperse and in the designed oligomerization state[63]. Because a large energetic penalty is incurred if buried polar groups do not form hydrogen bonds, altered core-packing arrangements in which hydrogen bonds are not formed are disfavoured. Of the α-helical repeat designs[43], 90% were soluble and

64% were monodisperse. Almost all of the monodisperse designs had small-angle X-ray scattering data that were consistent with the design models[84]. Here, the sequence repetition probably favours structures with internal repeats over alternative structures.

Outlook and challenges

A fundamental problem encountered when redesigning naturally occurring proteins to deliver new functions such as catalytic sites is that the alteration of a large number of amino-acid residues to introduce the function will inevitably change aspects of the structure; this is demonstrated by crystal structures of designed enzymes that have unanticipated loop reconfigurations[85]. Native proteins are often marginally stable, and sequence changes can lead to unfolding or aggregation. The very high stability of *de novo* designed proteins should make them more robust starting points for creating new functions.

The next steps in protein design are not without challenges. The ideality of almost all of the *de novo* structures designed so far probably contributes to their stability, and the introduction of functional sites and binding interfaces will inevitably compromise this ideality. Proteins that bind to other proteins usually have hydrophobic residues on their surface and are therefore more prone to aggregation than the idealized polar surfaces of most of the proteins that have been described in this Review, and the active sites of enzymes have some mobility to enable substrates to enter and products to leave. Recessed cavities, which are not incorporated into most *de novo* designed proteins at present, will be required for ligand and substrate binding. Naturally occurring proteins provide numerous examples of the rich functionality, including allostery and signalling, that can emerge in protein systems with multiple low-energy states and moving parts that can be toggled by external stimuli. To achieve such capabilities, which could have widespread applications in the design of molecular machines to tackle problems ranging from tumour recognition to computing, will require proteins to be designed with multiple, distinct energy minima. (By contrast, the *de novo* designs in Figs 2–6 each have a single, deep energy minimum (Fig. 1c).) The creation of a zinc-transporting transmembrane protein that has two alternative states demonstrates that protein design can now start to achieve such complexity[34].

Overcoming these challenges in the years ahead is an exciting prospect. Success would signal a technological advance that is analogous to the transition from the Stone Age to the Iron Age. Instead of building new proteins from those that already exist in nature, protein designers can now strive to precisely craft new molecules to solve specific problems — just as modern technology does outside of the realm of biology. ■

Received 12 April; accepted 20 July 2016.

1. Firestein, S. How the olfactory system makes sense of scents. *Nature* **413,** 211–218 (2001).
2. Rosenbaum, D. M., Rasmussen, S. G. F. & Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **459,** 356–363 (2009).
3. Yoshida, M., Muneyuki, E. & Hisabori, T. ATP synthase — a marvellous rotary engine of the cell. *Nature Rev. Mol. Cell Biol.* **2,** 669–677 (2001).
4. Spudich, J. A. The myosin swinging cross-bridge model. *Nature Rev. Mol. Cell Biol.* **2,** 387–392 (2001).
5. Dougherty, M. J. & Arnold, F. H. Directed evolution: new parts and optimized function. *Curr. Opin. Biotechnol.* **20,** 486–491 (2009).
6. Arnold, F. H. The nature of chemical innovation: new enzymes by evolution. *Q. Rev. Biophys.* **48,** 404–410 (2015).
7. Goldsmith, M. & Tawfik, D. S. Directed enzyme evolution: beyond the low-hanging fruit. *Curr. Opin. Struct. Biol.* **22,** 406–412 (2012).
8. Khoury, G. A., Smadbeck, J., Kieslich, C. A. & Floudas, C. A. Protein folding and *de novo* protein design for biotechnological applications. *Trends Biotechnol.* **32,** 99–109 (2014).
9. Regan, L. *et al.* Protein design: past, present, and future. *Biopolymers* **104,** 334–350 (2015).
10. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410,** 715–718 (2001).
11. Fisher, M. A., McKinley, K. L., Bradley, L. H., Viola, S. R. & Hecht, M. H. *De novo* designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth. *PLoS ONE* **6,** e15364 (2011).
12. Murphy, G. S., Greisman, J. B. & Hecht, M. H. *De novo* proteins with life-sustaining functions are structurally dynamic. *J. Mol. Biol.* **428,** 399–411 (2016).
13. Epstein, C. J., Goldberger, R. F. & Anfinsen, C. B. The genetic control of tertiary protein structure: studies with model systems. *Cold Spring Harb. Simp. Quant. Biol.* **28,** 439-449 (1963).
14. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487,** 545–574 (2011).
15. Wood, C. W. *et al.* CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. *Bioinformatics* **30,** 3029–3035 (2014).
16. Negron, C. & Keating, A. E. Multistate protein design using CLEVER and CLASSY. *Methods Enzymol.* **523,** 171–190 (2013).
17. Smadbeck, J., Peterson, M. B., Khoury, G. A., Taylor, M. S. & Floudas, C. A. Protein WISDOM: a workbench for *in silico de novo* design of biomolecules. *J. Vis. Exp.* **77,** e50476 (2013).
18. Fleming, P. J. & Rose, G. D. Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci.* **14,** 1911–1917 (2005).
19. Ponder, J. W. & Richards, F. M. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193,** 775–791 (1987).
20. Dahiyat, B. I. & Mayo, S. L. Protein design automation. *Protein Sci.* **5,** 895–903 (1996).
21. Dahiyat, B. I. & Mayo, S. L. *De novo* protein design: fully automated sequence selection. *Science* **278,** 82–87 (1997).
22. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA* **97,** 10383–10388 (2000).
23. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268,** 209–225 (1997).
24. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79,** 1061–1078 (2011).
25. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc. Natl Acad. Sci. USA* **110,** 15674–15679 (2013).
26. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3,** e02030 (2014).
27. Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4,** e09248 (2015).
28. Kuhlman, B. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302,** 1364–1368 (2003).
Describes Top7, the first globular protein to be designed with a fold not observed in nature.
29. Huang, P.-S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS ONE* **6,** e24109 (2011).
30. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. High-resolution protein design with backbone freedom. *Science* **282,** 1462–1467 (1998).
Describes RH4, the first protein to be designed using flexible-backbone methods and parametric equations.
31. Thomson, A. R. *et al.* Computational design of water-soluble α-helical barrels. *Science* **346,** 485–488 (2014).
32. Grigoryan, G. & DeGrado, W. F. Probing designability via a generalized model of helical bundle geometry. *J. Mol. Biol.* **405,** 1079–1100 (2011).
33. Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346,** 481–485 (2014).
34. Joh, N. H. *et al. De novo* design of a transmembrane Zn^{2+}-transporting four-helix bundle. *Science* **346,** 1520–1524 (2014).
Presents the design of a functional *de novo* helical bundle (known as Rocker) that can transport Zn^{2+} and Co^{2+}, but not Ca^{2+}, across membranes.
35. Regan, L. & DeGrado, W. F. Characterization of a helical protein designed from first principles. *Science* **241,** 976–978 (1988).
36. Lin, Y.-R. *et al.* Control over overall shape and size in *de novo* designed proteins. *Proc. Natl Acad. Sci. USA* **112,** E5478–E5485 (2015).
37. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491,** 222–227 (2012).
Established sequence-independent design principles, which enabled the design of five αβ topologies.
38. King, I. C. *et al.* Precise assembly of complex beta sheet topologies from *de novo* designed building blocks. *eLife* **4,** e11012 (2015).
39. Rämisch, S., Weininger, U., Martinsson, J., Akke, M. & André, I. Computational design of a leucine-rich repeat protein with a predefined geometry. *Proc. Natl Acad. Sci. USA* **111,** 17875–17880 (2014).
40. Park, K. *et al.* Control of repeat-protein curvature by computational protein design. *Nature Struct. Mol. Biol.* **22,** 167–174 (2015).
41. Doyle, L. *et al.* Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **528,** 585–588 (2015).
42. Huang, P.-S. *et al. De novo* design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nature Chem. Biol.* **12,** 29–34 (2016).
The first structurally verified design of a TIM barrel.
43. Brunette, T. J. *et al.* Exploring the repeat protein universe through computational protein design. *Nature* **528,** 580–584 (2015).
44. Jacobs, T. M. *et al.* Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352,** 687–690 (2016).
45. Murphy, G. S. *et al.* Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core. *Structure* **20,** 1086–1096 (2012).
46. Correia, B. E. *et al.* Proof of principle for epitope-focused vaccine design. *Nature* **507,** 201–206 (2014).

47. Crick, F. H. C. The packing of α-helices: simple coiled-coils. *Acta Cryst.* **6,** 689–697 (1953).
48. Walsh, S. T., Cheng, H., Bryson, J. W., Roder, H. & DeGrado, W. F. Solution structure and dynamics of a *de novo* designed three-helix bundle protein. *Proc. Natl Acad. Sci. USA* **96,** 5486–5491 (1999).
49. Fletcher, J. M. *et al.* A basis set of *de novo* coiled-coil peptide oligomers for rational protein design and synthetic biology. *ACS Synth. Biol.* **1**, 240–250 (2012).
50. Zaccai, N. R. *et al.* A *de novo* peptide hexamer with a mutable channel. *Nature Chem. Biol.* **7,** 935–941 (2011).
51. Eisenberg, D. *et al.* The design, synthesis, and crystallization of an alpha-helical peptide. *Proteins* **1,** 16-22 (1986).
52. Keating, A. E., Malashkevich, V. N., Tidor, B. & Kim, P. S. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. *Proc. Natl Acad. Sci. USA* **98,** 14825-14830 (2001).
53. Grigoryan, G. *et al.* Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science* **332,** 1071–1076 (2011).
Describes the design of functional helical peptides that coat single-walled carbon nanotubes.
54. Fletcher, J. M. *et al.* Self-assembling cages from coiled-coil peptide modules. *Science* **340,** 595–599 (2013).
55. Burton, A. J., Thomson, A. R., Dawson, W. M., Brady, R. L. & Woolfson, D. N. Installing hydrolytic activity into a completely *de novo* protein framework. *Nature Chem.* http://dx.doi.org/10.1038/nchem.2555 (2016).
56. Grigoryan, G., Reinke, A. W. & Keating, A. E. Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* **458,** 859–864 (2009).
57. Reinke, A. W., Grant, R. A. & Keating, A. E. A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. *J. Am. Chem. Soc.* **132,** 6025–6031 (2010).
58. London, N. & Ambroggio, X. An accurate binding interaction model in *de novo* computational protein design of interactions: if you build it, they will bind. *J. Struct. Biol.* **185,** 136–146 (2014).
59. Gradišar, H. *et al.* Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nature Chem. Biol.* **9,** 362–366 (2013).
60. Gradišar, H. & Jerala, R. *De novo* design of orthogonal peptide pairs forming parallel coiled-coil heterodimers. *J. Pept. Sci.* **17,** 100–106 (2011).
61. Schreiber, G. & Keating, A. E. Protein binding specificity versus promiscuity. *Curr. Opin. Struct. Biol.* **21,** 50–61 (2011).
62. Stranges, P. B. & Kuhlman, B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci.* **22,** 74–82 (2013).
63. Boyken, S. E. *et al. De novo* design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352,** 680–687 (2016).
Describes the design of helical bundles with extensive buried hydrogen-bond networks that mediate interaction specificity in a manner analogous to DNA base pairing.
64. Seeman, N. C. DNA in a material world. *Nature* **421,** 427–431 (2003).
65. Linko, V. & Dietz, H. The enabled state of DNA nanotechnology. *Curr. Opin. Biotechnol.* **24,** 555–561 (2013).
66. Zhang, F., Nangreave, J., Liu, Y. & Yan, H. Structural DNA nanotechnology: state of the art and future perspective. *J. Am. Chem. Soc.* **136,** 11198–11211 (2014).
67. Hilvert, D. Design of protein catalysts. *Annu. Rev. Biochem.* **82,** 447–470 (2013).
68. Kries, H., Blomberg, R. & Hilvert, D. *De novo* enzymes by computational design. *Curr. Opin. Chem. Biol.* **17,** 221–228 (2013).
69. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational enzyme design. *Angew. Chem. Int. Edn Engl.* **52,** 5700–5725 (2013).
70. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. *De novo* enzyme design using Rosetta3. *PLoS ONE* **6,** e19230 (2011).
71. Baker, D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* **19,** 1817–1819 (2010).
72. Kiss, G., Röthlisberger, D., Baker, D. & Houk, K. N. Evaluation and ranking of enzyme designs. *Protein Sci.* **19,** 1760–1773 (2010).
73. Garrabou, X., Wicky, B. I. & Hilvert, D. Fast Knoevenagel condensations catalyzed by an artificial Schiff-base-forming enzyme. *J. Am. Chem. Soc.* **138,** 6972-6974. (2016).
74. Koday, M. T. *et al.* A computationally designed hemagglutinin stem-binding protein provides *in vivo* protection from influenza independent of a host immune response. *PLoS Pathog.* **12,** e1005409 (2016).
75. Tinberg, C. E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501,** 212–216 (2013).
76. Griss, R. *et al.* Bioluminescent sensor proteins for point-of-care therapeutic drug monitoring. *Nature Chem. Biol.* **10,** 598–603 (2014).
77. Feng, J. *et al.* A general strategy to construct small molecule biosensors in eukaryotes. *eLife* **4,** e10606 (2015).
78. Mou, Y., Huang, P.-S., Hsu, F.-C., Huang, S.-J. & Mayo, S. L. Computational design and experimental verification of a symmetric protein homodimer. *Proc. Natl Acad. Sci. USA* **112,** 10714–10719 (2015).
79. King, N. P. *et al.* Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336,** 1171–1174 (2012).
80. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510,** 103–108 (2014).
81. Gonen, S., Dimaio, F., Gonen, T. & Baker, D. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* **348,** 1365–1368 (2015).
82. Hsia, Y. *et al.* Design of a hyperstable 60-subunit protein icosahedron. *Nature* **535,** 136–139 (2016).
83. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353,** 389–394 (2016).
84. Bhardwaj, G., Mulligan, V. K., Bahl, C. D. & Baker, D. Accurate *de novo* design of hyperstable constrained peptides. *Nature* http://dx.doi.org/10.1038/nature19791 (2016).
The design of hyperstable constrained peptides that incorporate both L- and D-amino acids is described.
85. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nature Chem. Biol.* **9,** 494–498 (2013).

Acknowledgements We thank all members of the Baker laboratory and the Institute for Protein Design at the University of Washington, as well as the RosettaCommons community. We apologize to the researchers and protein designers whose work we were unable to acknowledge due to space and scope limitations. The authors are supported by the Howard Hughes Medical Institute (HHMI-027779).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/2bqnvlj. Correspondence should be addressed to D.B. (dabaker@uw.edu).