

# Advances in protein structure prediction and design

Brian Kuhlman <sup>1,2\*</sup> and Philip Bradley <sup>3,4\*</sup>

**Abstract** | The prediction of protein three-dimensional structure from amino acid sequence has been a grand challenge problem in computational biophysics for decades, owing to its intrinsic scientific interest and also to the many potential applications for robust protein structure prediction algorithms, from genome interpretation to protein function prediction. More recently, the inverse problem — designing an amino acid sequence that will fold into a specified three-dimensional structure — has attracted growing attention as a potential route to the rational engineering of proteins with functions useful in biotechnology and medicine. Methods for the prediction and design of protein structures have advanced dramatically in the past decade. Increases in computing power and the rapid growth in protein sequence and structure databases have fuelled the development of new data-intensive and computationally demanding approaches for structure prediction. New algorithms for designing protein folds and protein–protein interfaces have been used to engineer novel high-order assemblies and to design from scratch fluorescent proteins with novel or enhanced properties, as well as signalling proteins with therapeutic potential. In this Review, we describe current approaches for protein structure prediction and design and highlight a selection of the successful applications they have enabled.

The stunning diversity of molecular functions performed by naturally evolved proteins is made possible by their finely tuned three-dimensional structures, which are in turn determined by their genetically encoded amino acid sequences. A predictive understanding of the relationship between amino acid sequence and protein structure would therefore open up new avenues, both for the prediction of function from genome sequence data and also for the rational engineering of novel protein functions through the design of amino acid sequences with specific structures. The past decade has seen dramatic improvements in our ability to predict and design the three-dimensional structures of proteins, with potentially far-reaching implications for medicine and our understanding of biology. New machine-learning algorithms have been developed that analyse the patterns of correlated mutations in protein families, to predict structurally interacting residues from sequence information alone<sup>1,2</sup>. Improved protein energy functions<sup>3,4</sup> have for the first time made it possible to start with an approximate structure prediction model and move it closer to the experimentally determined structure by an energy-guided refinement process<sup>5,6</sup>. Advances in protein conformational sampling and sequence optimization have permitted the design of novel protein structures and complexes<sup>7,8</sup>, some of which show promise as therapeutics<sup>9</sup>.

These advances in protein structure prediction and design have been fuelled by technological breakthroughs as well as by a rapid growth in biological databases. Protein-modelling algorithms (BOX 1) are computationally demanding both to develop and to apply. The rapid increase in computing power available to researchers (both CPU-based and, increasingly, GPU-based computing power) facilitates rapid benchmarking of new algorithms and enables their application to larger molecules and molecular assemblies. At the same time, next-generation sequencing has fuelled a dramatic increase in protein sequence databases as genomic and metagenomic sequencing efforts have expanded<sup>10</sup>. Advances in software and automation have increased the pace of experimental structure determination, speeding the growth of the database of experimentally determined protein structures (the Protein Data Bank (PDB))<sup>11</sup>, which now contains close to 150,000 macromolecular structures. Deep-learning algorithms<sup>12</sup> that have revolutionized image processing and speech recognition are now being adopted by protein modellers seeking to take advantage of these expanded sequence and structural databases.

In this Review, we highlight a selection of recent breakthroughs that these technological advances have enabled. We describe current approaches to the prediction and design of protein structures, focusing primarily on template-free methods that do not require an

<sup>1</sup>Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC, USA.

<sup>2</sup>Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA.

<sup>3</sup>Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

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

\*e-mail: [bkuhlman@email.unc.edu](mailto:bkuhlman@email.unc.edu); [pbradley@fredhutch.org](mailto:pbradley@fredhutch.org)

<https://doi.org/10.1038/s41580-019-0163-x>

Box 1 | Navigating protein energy landscapes

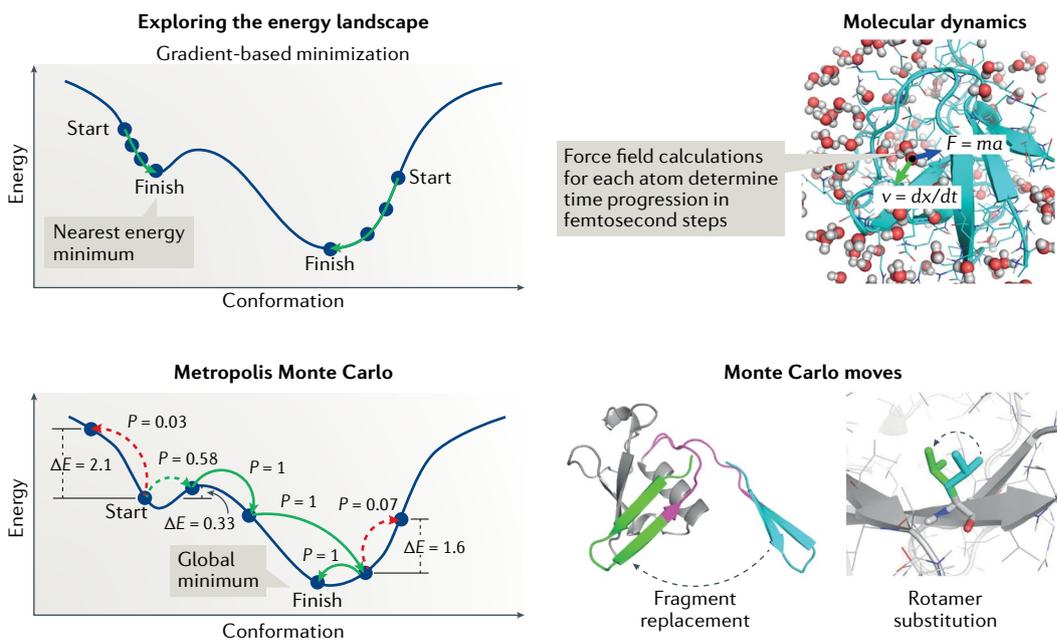
Protein conformational energy landscapes are complex, high-dimensional surfaces with many local minima. Navigating these landscapes in order to locate low-energy basins for prediction and design requires efficient sampling methods and accurate energy functions. In gradient-based optimization approaches (see the figure, upper left panel), the derivatives of the energy function with respect to the flexible degrees of freedom (e.g. the atomic coordinates or backbone torsion angles) are calculated in order to proceed in the direction in which the energy decreases most rapidly. Gradient-based optimization is effective at finding the nearest local minimum in the energy landscape, but it will not generally locate the global minimum. Monte Carlo sampling approaches employ randomly selected conformational moves and occasional uphill steps to escape local minima (see the figure, lower panels). In Metropolis Monte Carlo<sup>14</sup>, sampling moves are accepted (green arrows) or rejected (red arrows) on the basis of the change in energy: downhill moves that decrease the energy are accepted with probability 1, whereas uphill moves (dashed arrows) are accepted with a probability  $P$  that exponentially decreases as a function of the energy change. Examples of the move sets used for Monte Carlo simulations include fragment-replacement moves, in which a continuous backbone segment in the current conformation is replaced with an alternative conformation from a fragment library, and side-chain rotamer substitutions. A popular alternative to Monte Carlo sampling is molecular dynamics simulation (see the figure, upper right panel), in which the conformational sampling is dictated by Newton's laws of motion applied to the potential energy function of the molecular system. Given a starting set of atomic positions and velocities, the force acting on each atom is calculated by taking the gradient of the potential energy, and a resulting acceleration is derived from Newton's second law ( $F = ma$ ). A very small step forwards in time is taken (typically of the order of a few femtoseconds), and new positions and velocities are calculated on the basis of the size of the time step and the old positions, velocities and accelerations. With an accurate energy function and sufficiently small time steps, a long molecular-dynamics simulation provides broad sampling of the energy landscape and also gives a realistic picture of how individual molecules evolve over time. The challenge of modelling approaches based on molecular dynamics is that anywhere from millions to trillions of time steps must be conducted to reach biologically relevant time scales, requiring high-performance software and, in some cases, even special-purpose supercomputers<sup>170,171</sup>.

Protein energy functions

Functions that correspond to a mathematical model of the molecular forces that determine protein structures and interactions. The choice of an energy function defines a map from structures onto energy values, referred to as an energy landscape, which can guide structure prediction and design simulations. Typical protein energy functions are linear combinations of multiple terms, each term capturing a distinct energetic contribution (van der Waals interactions, electrostatics, desolvation), with the weights and atomic parameters for these terms chosen by a parameterization procedure that seeks to optimize the agreement between the quantities predicted from the energy function and the corresponding values derived from experiments or from quantum chemistry calculations on small chemical systems.

Deep learning

A form of machine learning that employs artificial neural networks with many internal-processing layers to recognize patterns in large and complex datasets, such as visual images and written and spoken language.



experimentally determined structure as a template. The strengths and weaknesses of these modelling approaches, as well as their current and potential applications, will be discussed. Finally, we comment on the broader practical implications of these developments for the fields of biology and medicine.

Protein-folding forces

Proteins possess the remarkable ability to fold spontaneously into precisely determined three-dimensional structures. Refolding experiments have established that the information required to specify a protein's folded conformation (its native state) is completely contained in its linear amino acid sequence<sup>13-15</sup>. According to Anfinsen's thermodynamic hypothesis, this information is encoded

in the shape of the energy landscape of the polypeptide: the native state is the one with the lowest free energy<sup>16,17</sup>. This hypothesis forms the basis for a general approach to protein structure prediction that combines sampling of alternative conformations with scoring to rank them by energy and identify the lowest energy state<sup>18-21</sup>. The chief obstacle to the success of this energy-guided approach, first identified by Cyrus Levinthal as a conceptual barrier to protein folding on biological timescales<sup>22</sup>, is the vast space of potential conformations: even supposing that each amino acid has only a limited, discrete set of possible backbone states, the total size of the conformational space that must be searched grows exponentially with chain length, and very quickly becomes astronomical. The solution to this dilemma lies in the recognition that

**Van der Waals interactions**

Inter-atomic or inter-molecular interactions that are individually weak (much weaker than covalent or ionic bonds) and relatively short-ranged.

**Rotamers**

A discrete set of conformations frequently adopted by amino acid side chains.

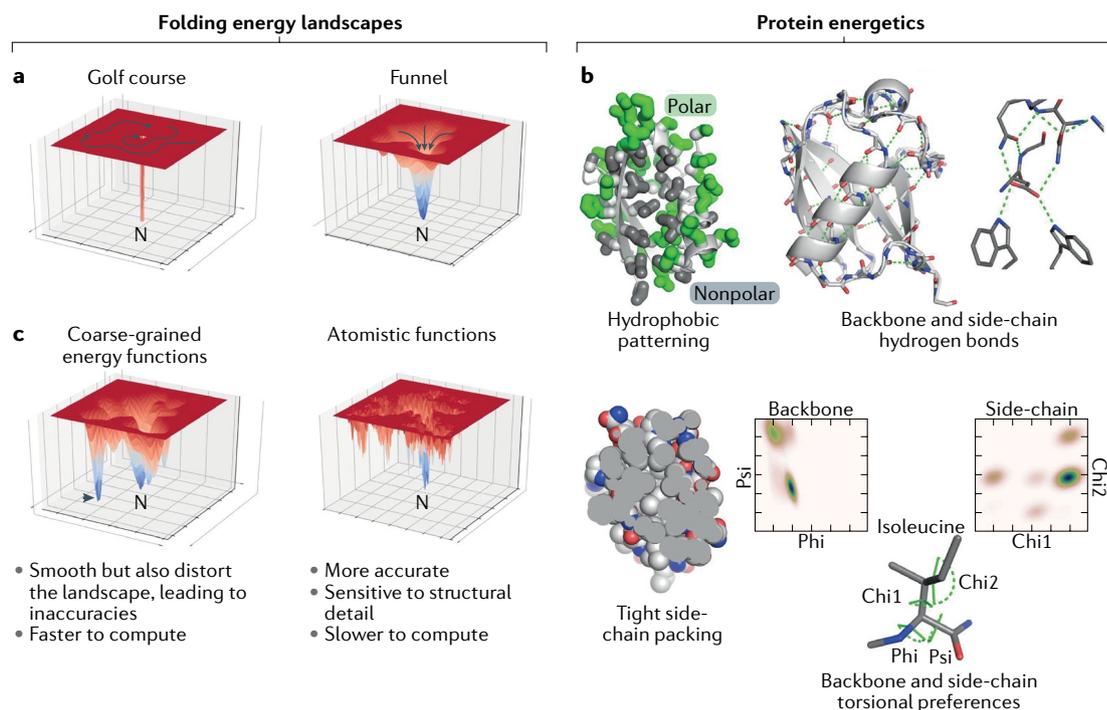
**Degrees of freedom**

The free parameters in a system that determine its structure and, hence, its energy. They can be continuous, such as a real-valued backbone torsion angle or atomic position, or discrete (permitting only a finite number of alternatives). Owing to strong torsional preferences, side-chain conformations can be successfully modelled using a discrete set of rotamers, identified by analysis of the structural database.

it is not necessary to explore the entire conformational space in order to identify the native state: the energy landscape is not a flat 'golf course' with a single native 'hole'; rather, directional cues impart an overall funnel shape to the landscape and guide sampling towards near-native conformations<sup>19,23</sup> (FIG. 1a). These directional cues can arise from sequence-local residue interactions that bias short stretches of the chain towards forming specific secondary structures, or from favourable long-range, non-local packing interactions that can be formed even before the global native fold is reached.

The driving force favouring the folding of water-soluble, globular proteins is thought to be the burial of hydrophobic amino acid side chains away from water<sup>24</sup>; folding is opposed by the loss of configurational entropy that accompanies the collapse of a flexible polypeptide chain into a defined 3D conformation. Tight packing of nonpolar side chains in the protein core enhances attractive van der Waals interactions and eliminates entropically unfavourable internal cavities (FIG. 1b). Moreover, this jigsaw puzzle-like packing is achieved while accommodating strong backbone and side-chain torsional preferences that restrict the observed torsion angle distributions

(lower panels in FIG. 1b), effectively reducing side-chain flexibility to the neighbourhood of a discrete set of rotamers at each position. Intra-protein hydrogen bonds and salt bridges largely compensate for the loss of interactions with water, as polar groups are buried during folding and hence these interactions contribute less to the stability of the native state than to its specificity (that is, they help discriminate the native state from other compact states). Whereas hydrophobic burial and backbone hydrogen bonding can be detected from low-resolution structural models, the tight core packing and absence of buried, unsatisfied polar groups that distinguish the native state require explicit modelling of the side-chain degrees of freedom. As a result, molecular modelling approaches for structure prediction and design often employ multiple levels of resolution: large-scale conformational sampling is performed with a computationally efficient coarse-grained energy function that captures hydrophobic burial, formation of secondary structure, and avoidance of atomic overlaps<sup>25–27</sup>; final protein model selection and refinement requires explicit modelling of the amino acid side chains using a more time-intensive, high-resolution atomistic energy function (FIG. 1c).



**Fig. 1 | Protein-folding landscapes and energies.** **a** | Simplified, two-dimensional representations of 'golf course' and 'funnel'-shaped energy landscapes. Identifying the native energy minimum ('N') in the landscape on the left requires exhaustive exploration, whereas a simple downhill search from most starting points will locate the native state in the landscape on the right. **b** | Energetic features that distinguish the protein native state include: hydrophobic patterning (shown here in a cutaway view of the small protein ubiquitin), with burial of nonpolar side chains in the protein core; backbone and side-chain hydrogen bonding (hydrogen bonds are shown as dotted green lines); tight side-chain packing (visible in a slice through a protein core); and restricted backbone and side-chain torsion angle distributions (evident in the highly focused two-dimensional probability distributions of backbone — phi angle versus psi angle — and side-chain — chi1 angle versus chi2 angle — torsion angles for the amino acid isoleucine). **c** | Computational models of protein energetics offer a trade-off between speed and accuracy. Coarse-grained models are computationally efficient and effectively smooth the energy landscape, permitting large-scale sampling; however, they also introduce inaccuracies such as false minima (for example, the blue basin to the left of the native minimum in this part, highlighted with an arrow). High-resolution, atomically detailed energy functions are more accurate, but also slower to evaluate and sensitive to structural detail, which introduces bumpiness (many local minima) into the landscape and makes them harder to navigate efficiently.

**Phi angle**

A torsion angle (or dihedral angle) that describes rotation about the bond that connects the backbone nitrogen and the backbone Ca carbon of an amino acid in a polypeptide chain. It is one of the two primary degrees of freedom (along with the psi angle) per amino acid residue that proteins use to adopt alternative conformations.

**Psi angle**

A torsion angle (or dihedral angle) that describes rotation about the bond that connects the backbone Ca carbon and backbone carbonyl carbon of an amino acid in a polypeptide chain. It is one of the two primary degrees of freedom (along with the phi angle) per amino acid residue that proteins use to adopt alternative conformations.

**Chi1 angle**

A torsion angle (or dihedral angle) in amino acid side chains that is numbered on the basis of the proximity in chemical connectivity of the bond to the protein backbone. Chi1 refers to rotation about the bond closest to the backbone, chi2 is the next closest position and so on. Some amino acids, such as alanine and glycine, have no rotatable bonds or torsion angles, while others, such as lysine and arginine, have up to four.

**Protein structure prediction**

There are two general approaches to predicting the structure of a protein of interest (the 'target'): template-based modelling, in which the previously determined structure of a related protein is used to model the unknown structure of the target; and template-free modelling, which does not rely on global similarity to a structure in the PDB and hence can be applied to proteins with novel folds. Historically, the methods applied in these two approaches have been quite distinct, with template-based modelling focusing on the detection of, and alignment to, a related protein of known structure, and template-free modelling relying on large-scale conformational sampling and the application of physics-based energy functions. Recently, however, the line between these approaches has begun to blur, as template-based methods have incorporated energy-guided model refinement, and template-free methods have employed machine learning and fragment-based sampling approaches to exploit the information in the structural database (although template-based methods still retain an increased accuracy for targets with detectable sequence similarity to the entries in the PDB). Here we provide a brief introduction to template-based modelling methods, and then turn to template-free modelling and describe recent developments in that area.

**Template-based modelling**

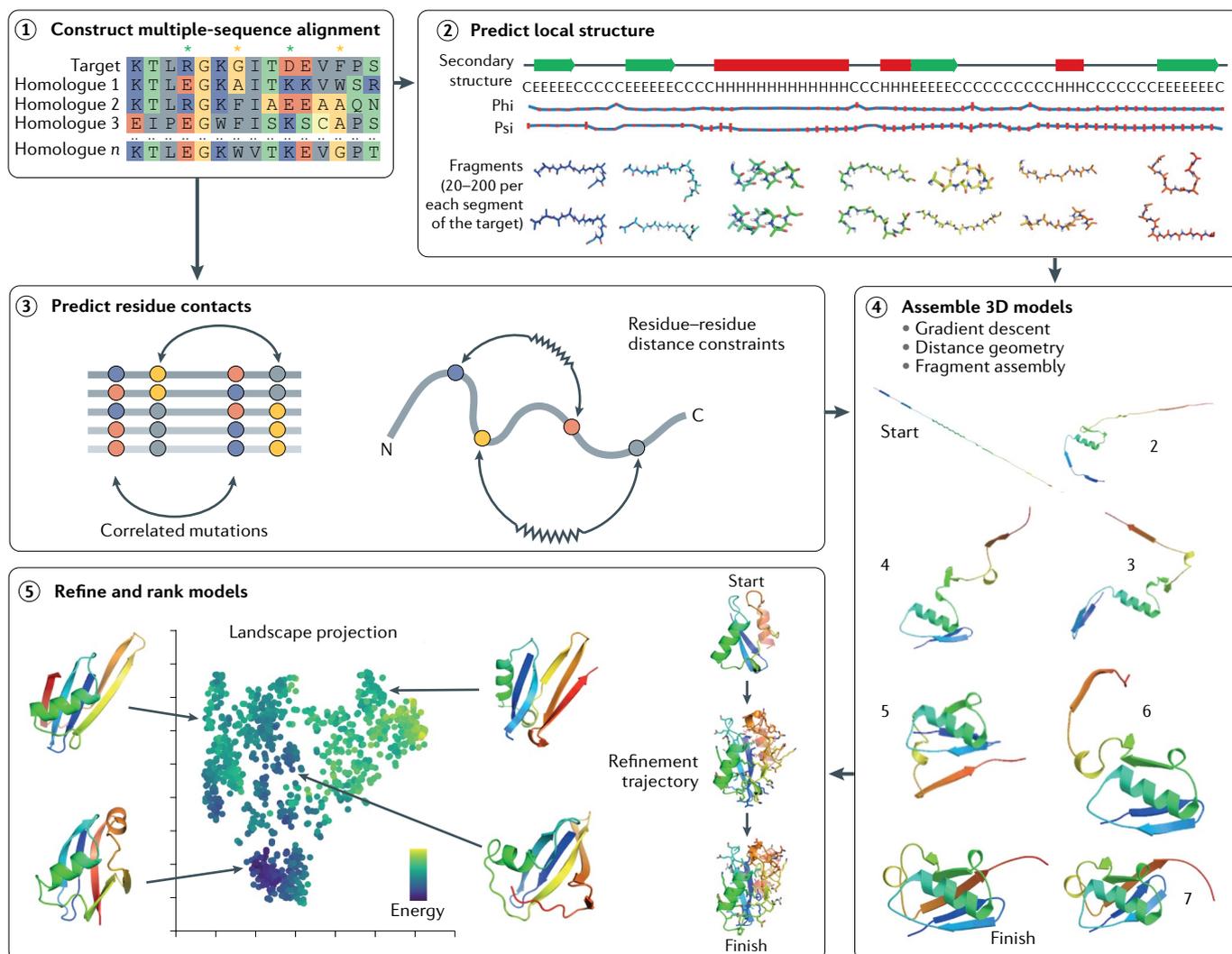
The steps in standard template-based modelling include selection of a suitable structural template; alignment of the target sequence to the template structure; and molecular modelling to account for mutations, insertions and deletions present in the target–template alignment. Closely related templates can be detected by using single-sequence search methods such as BLAST<sup>28</sup> to scan the PDB sequences. To detect more distantly related templates, a target sequence profile<sup>29,30</sup> built from a multiple-sequence alignment can be used to scan a database of sequence profiles for proteins of known structure by profile–profile comparison<sup>31,32</sup> or can be matched to a library of structural templates to assess sequence–structure compatibility<sup>33,34</sup>. Template selection methods return an initial target–template alignment that can be adjusted manually, often in an iterative manner after model building. Given an alignment to a template, established tools<sup>35–37</sup> can be used to quickly construct molecular models of the target sequence by performing side-chain optimization only at mutated positions and by rebuilding the backbone around insertions and deletions. For target protein sequences that are only distantly related to proteins of known structure, more sophisticated approaches that rely on multiple templates and perform aggressive backbone conformational sampling may be required<sup>37–39</sup>. Together with available crystal structures, template-based modelling approaches can provide structural information for roughly two-thirds of known protein families<sup>40</sup>.

**Template-free modelling**

Template-free modelling approaches can be applied to proteins without global structural similarity to a protein in the PDB. Lacking a structural template, these methods require a conformational sampling strategy for generating

candidate models, as well as a ranking criterion by which native-like conformations can be selected. The structure prediction process without a template (FIG. 2) typically begins with the construction of a multiple-sequence alignment of the target protein and related sequences. The sequences of the target and its homologues are then used to predict local structural features, such as secondary structure and backbone torsion angles, and non-local features, such as residue–residue contacts or inter-residue distances across the polypeptide chain. These predicted features guide the process of building 3D models of the target protein structure, which are then refined, ranked and compared with one another to select the final predictions.

**Fragment assembly.** One popular approach to conformational sampling is fragment assembly, in which models are built from short, contiguous backbone fragments (typically 3–15 residues in length) taken from proteins of known structure<sup>41–43</sup> (FIG. 2). Libraries of such fragments, typically 20–200 for each local sequence window of the target protein, are selected to provide a sampling of the possible local backbone structures. Fragment selection is typically guided by sequence similarity, as well as by predictions of local structural features such as secondary structure or backbone torsion angles. Building full-length 3D models from these fragments employs Monte Carlo simulations (BOX 1) that start from a random or fully extended conformation and proceed by repeatedly selecting a random window of the protein (e.g. residues 22–30) and inserting into that window the structure of a randomly selected fragment from the corresponding fragment library. The calculated energies of the protein model before and after the fragment insertion are then compared: if the energy is lower after the fragment insertion, the move is accepted, whereas if the energy is higher, the fragment insertion is accepted with a probability that decreases exponentially with the increase in energy (the Metropolis criterion<sup>44</sup>). To generate a population of hypothetical models, several thousand such simulations are conducted, each consisting of thousands of fragment insertion trials, leading to a final lowest-energy model. Fragment assembly simulations are typically conducted with a reduced representation (e.g. only backbone atoms and a single 'centroid' side-chain pseudo-atom are present) and a coarse-grained energy function that is fast to evaluate and defines a relatively smooth energy landscape appropriate for large-scale conformational sampling. Subsequent atomically detailed refinement simulations are then used to rank the candidate models and select the final predictions. Fragment assembly approaches have a number of advantages that contribute to their popularity: first, almost all protein structures in the PDB are locally similar to other, unrelated structures in the database<sup>45,46</sup>; second, the use of experimentally validated fragments ensures that the models will generally have protein-like local features; third, fragment libraries implicitly model the mapping between local sequence and structure without requiring an accurate energy model of the underlying interactions<sup>47</sup>; fourth, fragment assembly simulations have been proved empirically to be an efficient means of exploring the protein conformational space and are capable of sampling globally correct



**Fig. 2 | Key steps in template-free structure prediction.** An accurate multiple-sequence alignment between the target protein and its sequence homologues contains valuable information on the amino acid variation between the homologous sequences, including correlated patterns of sequence changes occurring at different positions (the green and yellow stars highlight pairs of alignment columns displaying amino acid charge and size swapping, respectively) (**step 1**). The target sequence and the multiple-sequence alignment form the basis for predictions of local backbone structure, including torsion angles (phi and psi predictions are shown, with red error bars indicating uncertainty) and secondary structure (**step 2**; PSIPRED<sup>67</sup> predictions are shown). Libraries of backbone fragments taken from proteins predicted to have similar local structures can also be assembled for use in model building. The multiple-sequence alignment can be used to predict residue pairs likely to be in spatial contact on the basis of observation of correlated mutations in pairs of alignment columns (**step 3**). These predictions of local structure and residue contacts guide 3D model building with techniques such as gradient-based optimization, distance geometry or fragment assembly (**step 4**; snapshots from a Rosetta<sup>42</sup> fragment assembly trajectory are shown). Initial 3D models are typically built with a reduced representation and a coarse-grained energy function; to better determine near-native predictions, these models are refined with an all-atom energy function and compared with one another to identify clusters of similar low-energy conformations, from which representative models are chosen as the final predictions (**step 5**; a 2D principal-component projection of the space of refined models is shown, in which each dot represents a single model).

#### Rosetta

A software package for the prediction and design of protein structures and interactions that implements a wide range of backbone and side-chain conformational sampling algorithms and sequence optimization methods.

fold models even if some local regions of the target protein are not well-represented in the fragment library<sup>48</sup>. In the absence of predicted or experimentally determined residue contact information, fragment assembly approaches generally work best on smaller  $\alpha$ -helical or mixed  $\alpha$ - $\beta$  protein domains, with all- $\beta$  and/or complex, non-local topologies presenting the greatest difficulties (since fragment insertion moves tend to perturb long-range contacts).

**Model refinement.** The high-resolution nature of the features that distinguish the native state (side-chain packing and hydrogen bonding, for example) means that the coarse-grained molecular representations and energy functions used in fragment assembly do not have the accuracy needed to reliably select near-native models, nor can they be used to fill in the atomistic details required by applications such as structure-based drug design. Thus, there is considerable interest in simulation

approaches that can take a crude starting model and move it closer to the native structure, while improving physical realism by eliminating features such as atomic overlaps or strained torsion angles. This process, termed model refinement, requires an accurate energy function as well as a strategy for exploring the conformational space nearby the starting model. One conformational sampling strategy that has been successfully applied to model refinement is molecular dynamics simulation (see BOX 1). In molecular dynamics-based refinement, the starting model is placed in a simulation box surrounded by water molecules, and the trajectory of the molecular system through many small steps forwards in time is simulated on the basis of Newton's laws of motion and a potential energy function. In one recent approach<sup>5</sup>, improved models sampled during this molecular dynamics simulation were identified using an all-atom energy function and were taken as starting points for subsequent molecular dynamics simulations to further enhance the conformational sampling. Atomically detailed Monte Carlo simulations (see BOX 1) that incorporate side-chain rotamer sampling and energy minimization with the Rosetta software package<sup>49</sup> have also been used successfully for model refinement<sup>6</sup>. The recent CASP protein structure prediction experiment<sup>50</sup> revealed progress in high-resolution model refinement, with several groups demonstrating substantial improvements in model quality for multiple prediction targets. These increases in performance were driven in part by enhancements in the accuracy of the underlying all-atom energy functions<sup>3,4</sup>, deriving from improved parameterization procedures.

**Contact predictions from residue covariation.** Analysis of multiple sequence alignments for proteins of known structure has revealed that pairs of alignment columns corresponding to residues in spatial contact often tend to show patterns of correlated mutations: when the amino acid in one column changes, the amino acid in the other column is also likely to change (see the two column pairs marked with an asterisk in the alignment in FIG. 2, step 1). This covariation between alignment columns has been attributed to the need to preserve favourable residue–residue interactions such as hydrogen bonds or tight packing. One key trend driving improvements in template-free modelling has been the increase in the accuracy with which these spatial contacts between residues can be predicted from the analysis of correlated mutations in multiple-sequence alignments. This increase in prediction accuracy has been fuelled by two factors: the growth in protein sequence databases driven by next-generation sequencing and metagenomics<sup>10</sup> (since larger databases mean deeper multiple-sequence alignments, and therefore more power to detect correlated mutations), and the development of global statistical methods that can separate direct residue coupling due to spatial proximity from indirect couplings. Early attempts to use covariation to predict spatial contacts employed 'local' measures, such as mutual information, that consider each pair of alignment columns independently<sup>51,52</sup>; these approaches suffered from false positives due, in part, to transitivity of the statistical

correlations: if position A is coupled to position B and position B is coupled to position C, then positions A and C will also tend to show significant coupling, even if they are not physically interacting. In 'global' methods<sup>53–56</sup>, a probabilistic graphical model of the entire multiple-sequence alignment is constructed by finding the set of direct interactions between alignment columns that most parsimoniously explains the observed sequence correlations. Putative interacting positions then correspond to the direct interactions with the highest weights in the model.

Predicted residue–residue contacts derived from covariation analysis have been shown to substantially improve the accuracy of template-free modelling for targets with many sequence homologues. Fragment assembly approaches have been extended to allow the inclusion of covariation-based distance constraints in the energy functions used for folding<sup>57–59</sup>. A recent study employing the Rosetta software package used protein sequence data from metagenomic projects to generate models for more than 600 protein families of unknown structure<sup>40</sup>. The EVfold<sup>60</sup> method uses software tools originally developed for structure determination from experimental distance constraints (e.g. from NMR or cross-linking experiments)<sup>61</sup> to build 3D models from covariation-derived contacts. This method has been successfully applied to the prediction of water-soluble<sup>60</sup> and transmembrane<sup>62</sup> protein structures, as well as to investigation of the structured states of intrinsically disordered proteins<sup>63</sup>. The major factor limiting the utility of covariation-based prediction approaches is their dependence on relatively deep multiple-sequence alignments (where deep has been defined, for example, as containing non-redundant sequences numbering at least 64 times the square root of the length of the target<sup>40</sup>). Although metagenomic projects have increased the available number of homologous sequences, their bias towards prokaryotic organisms means that the sequence coverage of many eukaryote-specific protein families remains too small for accurate predictions. New machine-learning approaches that can integrate diverse sources of sequence and structural information offer one promising avenue forwards in these cases.

#### **Machine learning in protein structure prediction.**

Machine-learning techniques have a long history of applications to protein structure analysis. Machine-learning models such as neural networks and support vector machines have been applied to prediction of 1D structural features such as backbone torsion angles, secondary structure and the solvent accessibility of residues<sup>64–68</sup>. Recently, the focus of machine-learning applications has shifted to 2D features, such as residue–residue contact maps and inter-residue distance matrices. Recognizing that contact maps are similar to 2D images — whose classification and interpretation have been among the striking successes of deep-learning approaches<sup>12</sup> — protein modellers have begun to apply deep learning to recognizing patterns in the sequences and structures of the proteins in the PDB. Convolutional neural networks (BOX 2 and Supplementary Fig. 1) have demonstrated excellent performance in image analysis tasks<sup>69</sup>, making them a natural choice for the

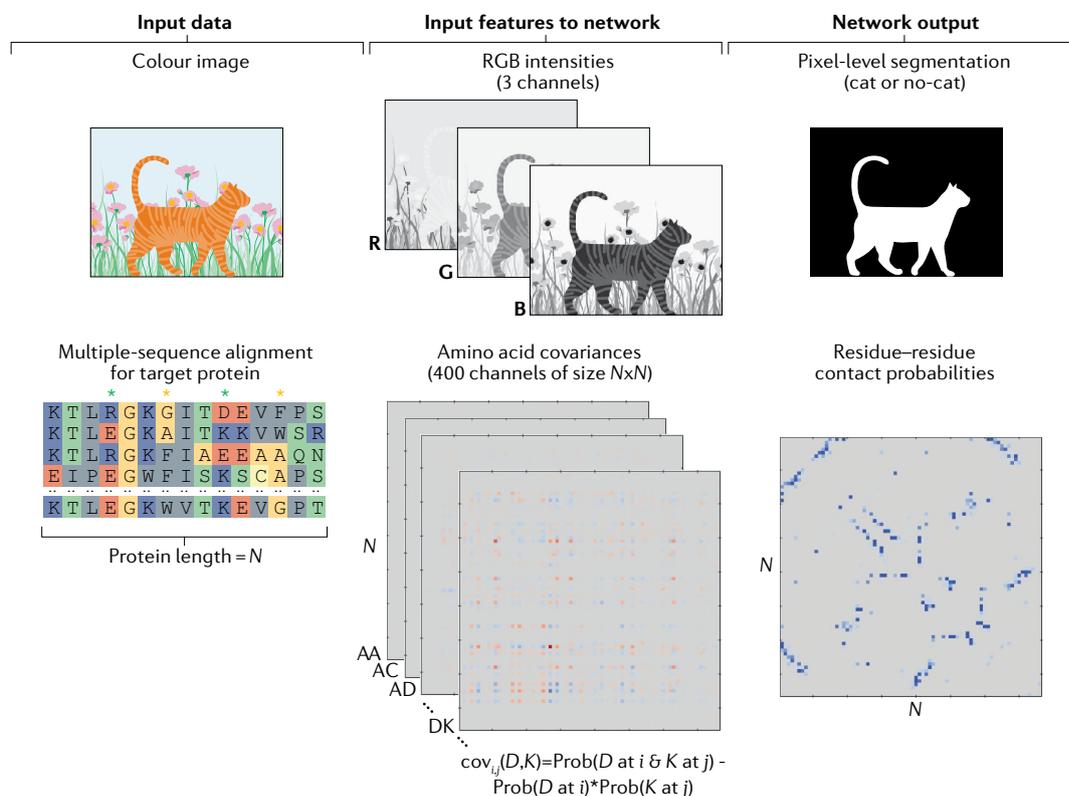
prediction of protein contact maps. The question of how best to encode information about the target protein for input to the neural network is an active research topic. Colour images, for example, are often encoded as three matrices of real numbers: the intensities of the red, green and blue colour channels for all the image pixels (BOX 2). Methods such as *DeepContact*<sup>70</sup> and *RaptorX-Contact*<sup>2</sup> use as input features the  $N \times N$  (where  $N$  is the number of

amino acids in the sequence of the target protein) residue–residue coupling matrices derived from covariation analyses for the target protein (augmented by predictions of local sequence features). In the *DeepCov*<sup>71</sup> and *TripletRes* (Y. Zhang, personal communication) approaches, more of the information in the target protein multiple-sequence alignment is provided to the network, in the form of 400 (the square of the number of standard amino acids)

### Box 2 | Deep convolutional neural networks in protein structural analysis

Artificial neural networks encompass one or more sequentially connected layers of processing units (the neurons), which transform input features to output predictions. Each individual processing unit of the network integrates weighted signals coming from connected units in the preceding layer into a single, nonlinear response and passes this response to downstream units in the next layer. The weights on the connections between processing units are fitted while training the network in order to maximize prediction accuracy. Deep neural networks have many internal layers of processing units (tens to thousands) that allow them to perform highly complex, nonlinear transformations of the input features. Convolutional neural networks are characterized by a very specific layered connectivity in which each unit receives inputs from a local window of units (corresponding to a small square of pixels in an input image, for example) in the preceding layer, via a matrix of weights called a convolutional filter (see Supplementary Fig. 1). The layers in convolutional neural networks are composed of multiple channels that correspond to different signals extracted from the inputs; each channel results from scanning a single convolutional filter over the entire preceding layer. In image classification, channels early in the network (closer to the inputs) might recognize specific local geometric features in the input image (edges or spots, for example), while later channels correspond to higher-order patterns such as composite features. The numbers of layers and channels in the network are typically optimized during the network training process.

A key challenge when applying neural networks to protein structures is the identification of appropriate encodings that convert information about the target protein into real-valued features suitable for input into the networks. For inter-residue contact prediction, where the desired output is a 2D matrix of contact predictions, one approach is to provide as input the 2D matrix of residue–residue covariation scores derived from a multiple-sequence alignment. By learning from many examples of known contact maps and covariation matrices fed into the network during training, the network is able to identify features of the protein contact maps (recurring secondary structure contacts, for example) that enable it to refine the sequence-based input predictions. Analysis of the early convolutional filters in such networks<sup>70</sup> has suggested that different filters are specialized to detect different modes of interaction ( $\beta$ -hairpins or  $\beta$ - $\alpha$ - $\beta$  motifs, for example). Recent work suggests that providing the neural network with even more information from the multiple-sequence alignment — 400 matrices of amino acid pair frequencies or covariances (one for each of the 400 ordered amino acid pairs, see the figure), rather than a single covariation matrix — can further improve amino acid contact and distance predictions<sup>71</sup>.



different  $N \times N$  feature matrices, each corresponding to a defined pair of amino acids, with the value at position  $(i, j)$  in a given matrix being either the pair frequency or the covariance for the given amino acid pair at alignment positions  $i$  and  $j$  (see also BOX 2 and Supplementary Fig. 1). The convolutional neural network is then tasked with integrating this massive set of features to identify spatial contacts, which it does by training on large sets of proteins of known structure and their associated contact maps and multiple-sequence alignments. The importance of incorporating machine learning in template-free modelling is highlighted by the top-performing CASP13 structure prediction methods, all of which rely on deep convolutional neural networks for, variously, predicting residue contacts or distances, predicting backbone torsion angles and ranking the final models.

### Protein design

Protein design is frequently referred to as the inverse protein-folding problem. Instead of searching for the lowest-energy conformation for a given protein sequence, the goal is to identify an amino acid sequence that will stabilize a desired protein conformation or binding interaction. Despite the reverse nature of the problem, the modelling tools needed for protein design are very similar to those needed for high-resolution structure prediction. It is critical to have an energy function to rank the relative favourability of different side-chain and backbone packing interactions, and it is necessary to have sampling protocols that can be used to search for low-energy sequences and protein conformations. Protein design efforts can be broadly divided into two categories. In template-based design, the sequence and structure of naturally evolved proteins are modified to achieve new functions. In *de novo* design, novel protein backbones and sequences are generated from scratch, guided by design requirements and physicochemical constraints. There are compelling reasons for pursuing each approach. *De novo* design provides a rigorous test of our understanding of protein structure and allows the creation of exceptionally stable proteins with sequences that are not restricted by evolutionary constraints that are irrelevant to the design goals<sup>7,72–74</sup>. Template-based design is well-suited to creating proteins with new functions, as the protein structure is predefined, and it is often possible to repurpose functional groups in the template<sup>75</sup>. Here we describe recent progress in both approaches with a focus on molecular-modelling strategies for generating novel protein backbones, complexes and functional sites.

### *De novo* protein design

In most projects in *de novo* protein design, the researcher begins by choosing a protein fold or topology to construct. In some cases, a particular fold is chosen because there is interest in developing better understanding of that fold; in other cases, a fold is chosen because it will be a good starting point for instilling a particular function in the protein. For instance,  $\beta$ -barrel proteins can form binding pockets for small molecules inside the barrel. Once the desired protein fold is chosen, a model of the polypeptide backbone adopting the target fold is constructed. While a protein can potentially populate

an enormous number of conformations for a given type of fold, only a small fraction of these conformations are consistent with forming a well-folded, thermodynamically stable protein. Much of the recent work in the field of *de novo* protein design has been focused on developing improved methods for constructing protein backbones that are physically realizable, in that they allow tight packing between the amino acid side chains, satisfy the hydrogen-bonding potential of the protein backbone (primarily through secondary structure formation) and have little strain in the backbone torsion angles<sup>72</sup>.

As with protein structure prediction, one common approach for ensuring the favourability of structural elements local in the primary sequence is to assemble the backbone from small fragments of naturally occurring proteins<sup>76</sup> (FIG. 3). In addition to providing well-formed secondary structure, these fragments can encode structural motifs that are energetically favourable at the beginning and end of secondary structures. To create a backbone that adopts a desired tertiary fold, fragment-based folding algorithms are frequently combined with user-defined distance constraints to specify how the secondary structural elements are positioned in 3D space<sup>77</sup>. This approach has been used to successfully design a variety of protein folds, including all helical proteins, repeat proteins, mixed  $\alpha$ - $\beta$  proteins and proteins consisting only of  $\beta$ -sheets and loops<sup>8,78–83</sup>. For some design projects, it may not be critical what tertiary fold the protein adopts, but instead, it may be important that the protein contain specific structural features that will allow it to perform a desired function. For example, it has been demonstrated that assembling proteins from naturally occurring helix–turn–helix fragments can be used to design large sets of alternative folds that contain features such as pockets and grooves that are frequently found in protein active sites<sup>84</sup>. Functional motifs such as metal-binding and protein interaction sites can also be explicitly incorporated into the assembly process<sup>85</sup>.

The process of constructing idealized folds from small protein fragments often reveals new information about the physical and structural constraints that dictate which conformations a protein can adopt. The first *de novo* design of an all- $\beta$ -sheet protein that adopts a jellyroll fold (double-stranded  $\beta$ -helices formed by eight antiparallel  $\beta$ -strands)<sup>82</sup> revealed strong couplings between loop geometry and loop length in the connections made between  $\beta$ -sheets. Similar rules have also been useful for designing the connections between secondary structural elements in mixed  $\alpha$ - $\beta$  proteins and  $\alpha$ -helical proteins<sup>86,87</sup>.

Another approach that is used to generate backbones for *de novo* protein design is to build a mathematical model that uses a small number of parameters to describe the structural variability typically observed for a type of protein fold. This approach has been particularly successful for the design of coiled-coil proteins, which consist of two or more  $\alpha$ -helices supercoiled around a central axis. Using a mathematical (parametric) model for the coiled coil first described by Francis Crick<sup>88</sup>, protein designers can rapidly generate large sets of protein backbones in which each model deviates from the next in a systematic manner. For instance, it is typical to scan through alternative values for the supercoil radius

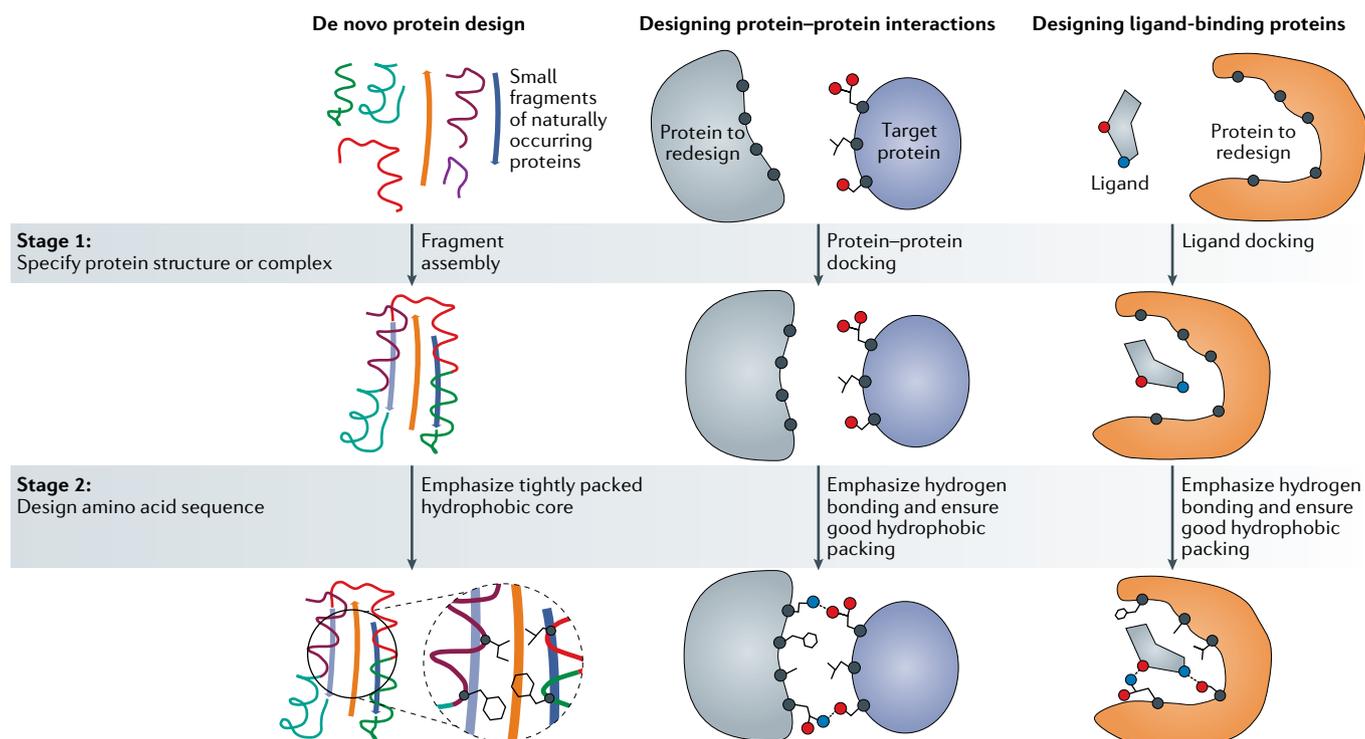


Fig. 3 | **Overview of the protein design process.** Projects in computational protein design can be distilled down to two key steps. First, a model of the desired structure and/or complex needs to be created. For de novo design, this can be accomplished by piecing together fragments of naturally occurring proteins (left column). Designing complexes requires moving the proteins (middle column) and/or ligands (right column) relative to each other (frequently referred to as docking) so that their surfaces are adjacent. After a model of the protein fold or complex is created, sequence optimization simulations are used to find sequences that stabilize the desired fold or complex (bottom row).

(which determines the distance between helices) and the supercoil twist (which determines how tightly the helices wrap around each other). These studies have revealed that certain regions of parameter space allow for stronger interactions between the helices, which can be used to create proteins with exceptional stability<sup>89</sup>. Recent applications of parametric design include the de novo design of transmembrane proteins<sup>7,90</sup> and  $\alpha$ -helical barrels<sup>91</sup>. In an extension of this approach, a parametric model was constructed to describe the backbone geometry of  $\beta$ -barrel proteins<sup>8</sup>. Interestingly, this approach did not succeed, because it was discovered that it was necessary to include non-idealized kinks in the backbone to reduce torsional strain. In the end, backbone assembly from protein fragments was better suited to this problem and allowed the first de novo design of a  $\beta$ -barrel<sup>8</sup>.

#### Template-based design

For many design goals, it may not be necessary to use a de novo-designed protein backbone and sequence; instead, high-resolution structures of naturally occurring proteins can serve as the starting point for the design process. This approach has frequently been used to design proteins that bind to other proteins, to design ligand-binding proteins and to design enzymes. Often it is not critical that a specific protein be used as the starting point for the design process, as the goal of the project is simply to create a protein that inhibits protein X or binds to ligand Y. In this case, it is advantageous to consider a

large set of naturally occurring proteins as templates for design<sup>92,93</sup>. Each template will have unique binding pockets and molecular surfaces that will make some templates more suitable than others for the design goal. Templates can be pre-screened to find proteins with binding pockets of an appropriate size for the target ligand, but then it is usually necessary to perform design simulations with multiple templates to find structures that can form a tight binding interaction with the target molecule. Potential templates are also often pre-screened in order to select proteins that are stable and can be produced easily. In some projects it is useful or necessary to start with a specific template because it has functionality needed for the design goals. Perhaps the starting protein is an enzyme that catalyses a target reaction or an antibody that activates an important cell surface receptor. The design goal in these cases is typically to improve the biophysical properties of the protein or to perturb activity in a specific way.

#### Optimizing the protein sequence

Once a model or set of models for the protein backbone has been generated, through either de novo or template-based methods, the next step in the design process is to identify an amino acid sequence that will stabilize the desired conformation or binding event. Sequence optimization software includes two key components: an energy function to evaluate the favourability of a particular sequence, and a protocol to search for more favourable sequences. Since the same physical properties need to be

**Dead-end elimination**

An algorithm for side-chain rotamer optimization that functions by eliminating rotamers that are not compatible with adopting the sequence with the lowest possible energy.

**Mean-field optimization**

A protocol for designing sequences that assigns a probability to observing each amino acid at each sequence position in the protein and calculates an average (mean-field) energy for the protein based on the assigned probabilities. The probabilities are then adjusted to lower the mean-field energy of the protein.

**Simulated annealing**

A probabilistic approach for identifying low-energy sequences that accepts or rejects sequence changes on the basis of the calculated change in the energy of the protein when a sequence change is made and the temperature of the modelled system. If a change lowers the energy of the protein, it is automatically accepted; if it raises the energy of the system, it is accepted with some probability that depends on how much the energy has increased (a bigger increase in energy is less likely to be accepted) and the current temperature (at higher temperatures it is more likely to accept changes that raise the energy of the system). The temperature is lowered as the simulation progresses, to identify low-energy sequences.

**Genetic algorithms**

A sequence optimization protocol that repeatedly modifies a population of sequences by applying rounds of energy-based selection. The energy of a sequence is calculated by modelling it in the desired protein conformation. Lower-energy sequences are more likely to progress to the next generation. Before each round of selection, the previous winning sequences are recombined with each other and small numbers of mutations are incorporated into the sequences, to search for lower-energy sequences.

optimized in high-resolution structure prediction and protein design (side-chain packing, hydrogen bonding, hydrophobic burial, backbone and side-chain strain), similar energy functions are frequently used for design and refinement<sup>94–96</sup>. The energy functions are typically parameterized using a variety of benchmarks that focus on reproducing the sequence and structural features of naturally occurring proteins. One exciting recent development is a demonstration that the energy functions for protein design and protein modelling can be improved by optimizing the energy function simultaneously with small-molecule thermodynamic data and high-resolution macromolecular structure data<sup>4</sup>. For instance, training the energy function to predict the experimentally measured free-energy changes associated with moving different chemical groups from the water phase to the vapour phase allows for more accurate modelling of the costs and benefits of burying hydrophilic and hydrophobic groups in a protein. This approach led to improved results in traditional protein design benchmarks such as native-sequence recovery (an approach that aims to produce native-like sequences when redesigning naturally occurring proteins) and prediction of free-energy changes by mutation, as well as improvements in structure prediction.

The second component of a design simulation is the search for lower-energy sequences and side-chain conformations. A variety of methods have been developed for this problem, including deterministic approaches such as dead-end elimination and mean-field optimization, as well as stochastic approaches such as simulated annealing and genetic algorithms<sup>97</sup>. Most methods simplify the side-chain optimization problem by restricting side-chain motion to a set of commonly observed conformations (or rotamers) observed in high-resolution structures from the PDB<sup>98</sup>. The design software Rosetta uses Monte Carlo sampling with simulated annealing to identify low-energy sequences and rotamers<sup>99</sup>. Starting from a random sequence, the Rosetta protocol accepts or rejects single amino acid mutations or rotamer substitutions on the basis of the Metropolis criterion (meaning that mutations that lower the energy of the system are accepted, and mutations that raise the energy of the system are accepted at some probability dictated by the energy change and the temperature of the simulation; see also BOX 1). Despite the simplicity of this approach, independent simulations that start with different random sequences converge on similar sequences, and one strength of the method is that it is rapid (typically <10 minutes on a single processor for proteins of less than 200 residues). One limitation of stochastic approaches like simulated annealing is that they do not guarantee that the lowest-energy sequence will be identified. For this reason, research teams have also developed approaches that can identify the lowest-energy sequence for a given protein backbone and rotamer library<sup>97,100</sup>. In addition to finding the lowest-energy sequence, the protein design package OSPREY has methods for rank ordering the lowest-energy sequences<sup>101</sup>.

One challenge in computational protein design is that the optimal amino acid sequence for a protein can be very sensitive to the precise 3D structure of the protein. In many cases, a small change in backbone conformation can

lead to dramatic drops in protein energy and large changes to the most favourable sequence. These large changes in energy reflect the stiffness of chemical interactions — very small displacements (<1 Å) in atomic coordinates can lead to strong steric repulsion or loss of a favourable interaction such as a hydrogen bond (FIG. 1b). To account for this rugged energy landscape, protein designers have developed a variety of methods for performing backbone sampling along with sequence optimization<sup>102</sup>. One approach that has worked well when designing de novo proteins is to iterate between rotamer-based sequence optimization and gradient-based minimization of torsion angles (backbone and side chain)<sup>72,77</sup> (see also BOX 1). With this approach, it is possible to simultaneously optimize angles throughout the entire structure. An alternative approach is to couple discrete changes in the conformation of the protein backbone with side-chain rotamer substitutions. These coupled moves allow the backbone to adjust to an amino acid substitution before scoring the favourability of the change<sup>103,104</sup>.

Because energy calculations based on all-atom models of proteins are hypersensitive to small changes in backbone conformation, and because the energy functions used for scoring protein conformations are empirical models that do not fully capture all the phenomena that contribute to the energy of a protein, knowledge-based approaches for sequence design that do not rely on explicit modelling of the amino acid side chains have also been developed<sup>105,106</sup>. In one approach, the protein is divided into sets of spatially adjacent residues and then the PDB is searched for residues that are in a similar structural environment (called tertiary structural motifs, or TERMS)<sup>105</sup>. The sequence preferences from the PDB for each residue position in a TERM are then used to derive an energy function that is used for sequence optimization<sup>106</sup>. Despite using naturally occurring proteins to derive the sequence preferences, this approach will likely also work for de novo protein design, in that a limited number of TERMS (hundreds) are able to describe most of the local structural environments observed in natural proteins<sup>105</sup>.

Many design scenarios benefit from specialized sequence optimization schemes. For instance, using step-wise rotamer optimization for designing buried hydrogen bond networks — which are particularly important for protein–ligand and protein–protein interactions — is a difficult task, as the network is not energetically favourable until all polar groups have a hydrogen bonding partner. In other words, obtaining a sequence with a fully connected network from a sequence with no hydrogen bond network requires that the sequence optimization simulation pass through intermediate sequences with high energies. To address this issue, a side-chain sampling protocol that uses a graphical representation of potential hydrogen bond partners was developed to enumerate all possible side-chain hydrogen bond networks for an input backbone structure<sup>107–109</sup>. This method allowed the design of helical oligomers with large numbers of buried polar amino acids and high binding specificities between the protein chains. Creating networks like these will be essential for designing new enzymes, because active site residues are frequently supported by extensive hydrogen bonding.

### Multi-state design algorithms

An approach to designing sequences that satisfy multiple constraints simultaneously. For instance, such algorithms can be used to find protein sequences that are predicted to bind ligand X but not ligand Y. Alternatively, they can be repurposed to find sequences that are simultaneously good at binding both ligand X and ligand Y.

### Yeast surface display

An experimental approach for probing a large library of protein sequences (up to tens of millions) for binding to another molecule. In the final yeast library, each yeast cell contains the DNA for one member of the protein library, and this protein is expressed as a fusion protein that presents the protein on the outside of the cell. The cells are mixed with the target molecule, which has been labelled with a fluorescent dye, and then fluorescence-activated cell sorting is used to identify the cells that contain a designed protein that binds the target protein. DNA sequencing is used to identify the designs that passed selection.

### BCL-2 protein family

A family of structurally related proteins that interact with each other to induce or repress apoptosis.

Another frequent design problem is the desire to favour one binding interaction (referred to as positive design) over another binding interaction (referred to as negative design). To perform simultaneous positive and negative design, multi-state design algorithms have been developed that evaluate amino acid sequences on alternative conformations and explicitly search for sequences that will increase the energy gap between the two states<sup>110–114</sup>.

### Validating computational predictions

The most exciting and nerve-racking moment in a protein design project is when experimental validation is performed. Various experiments are typically used to characterize designed proteins, including stability analysis with denaturation experiments, measurement of oligomerization state, high-resolution structure determination, and functional assays, when appropriate. For many design goals, it is necessary to experimentally characterize a set of alternative designs to identify a small number of successful designs. The number of designs that need to be characterized can vary widely, depending on the design goal. Methods for stabilizing proteins (see also below) have progressed to the point that often only a handful of designs need to be characterized<sup>115</sup>, while more challenging problems, such as redesigning the surface of a protein to bind another protein, may require screening of thousands of alternative sequences. It is common in a protein design project to progress through multiple generations of designs, modifying the design and/or computational strategy with each generation. In this way, it is possible to learn about the minimal structural features needed to accomplish the design goal. In one excellent example of this process, a high-throughput screening strategy based on yeast surface display was combined with next-generation sequencing to simultaneously evaluate thousands of de novo-designed proteins<sup>116</sup>. Iterative rounds of adjusting the design protocol and screening resulted in a pipeline that more robustly produced folded proteins.

### Applications of protein design

One of the great promises of computational protein design is that it will allow the creation of new proteins that have valuable applications in medicine, industry and research. Over the past 10 years, this promise has started to be realized<sup>117</sup>. This progress is being enabled by improved computational methods as well as by advances in DNA synthesis and sequencing<sup>118</sup>. It is now affordable to order large sets of computationally designed sequences, allowing protein designers to rapidly explore multiple solutions to a problem. To give a flavour of the types of problems that computational protein design can be applied to, we highlight a subset of protein-engineering projects from the past few years that have made use of molecular modelling as an important step in the design process.

### Stabilizing proteins

There is a long-standing interest in using computer-based methods to identify mutations that will increase the thermodynamic stability of proteins<sup>119–122</sup>, as this often results in higher expression levels of recombinant protein

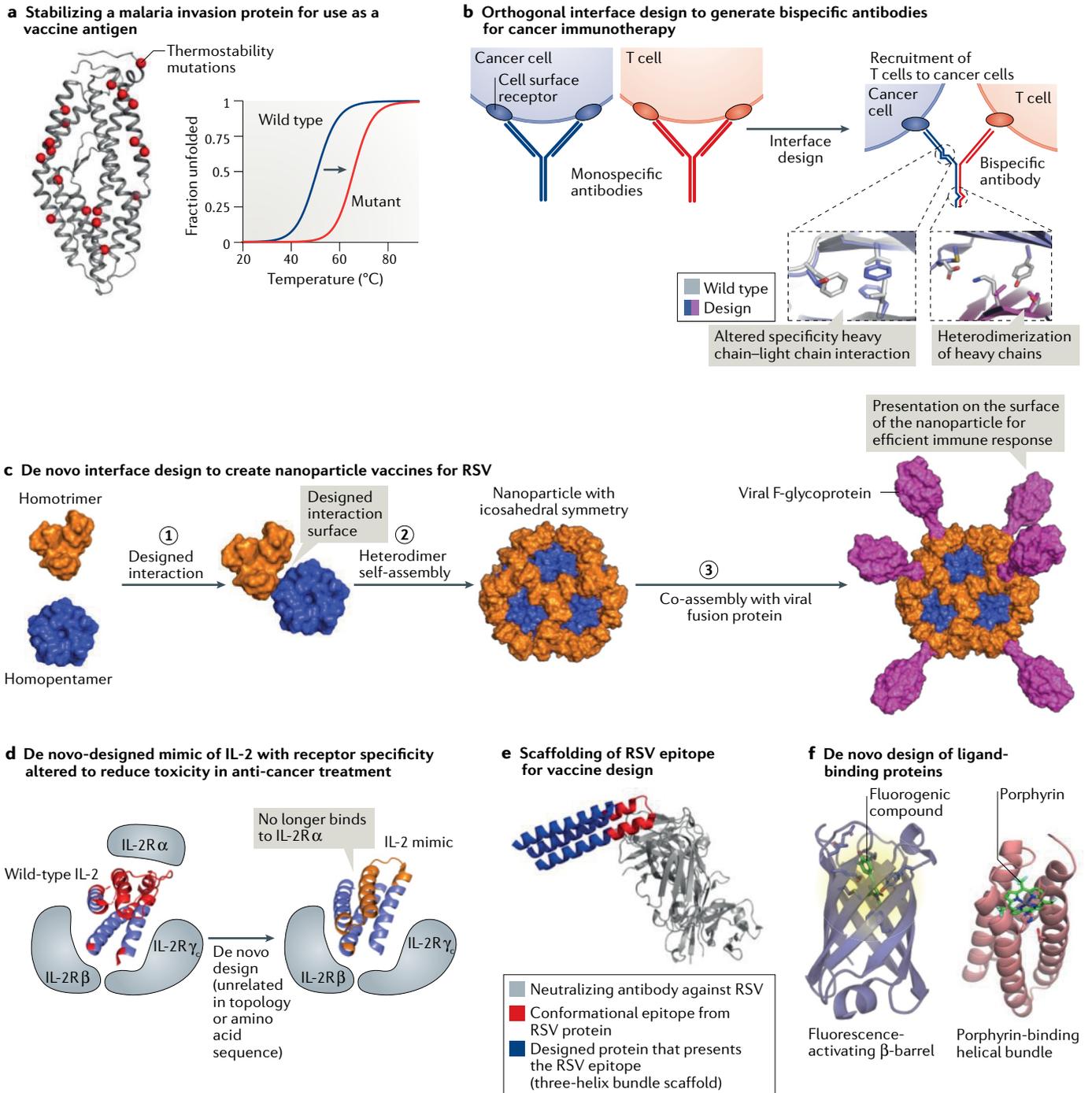
and can reduce the propensity to aggregate. It has been apparent for many years that naturally occurring proteins are often not optimized for stability, and that a global redesign (a simulation in which all residues in the protein are allowed to mutate) of a protein can raise the thermal unfolding temperature dramatically, by over 30 °C in some cases<sup>123,124</sup>. In most real-world applications, however, it is preferred to raise stability by making a more selective set of mutations. One approach that is proving to be well suited to this task is to combine information from computational design simulations with sequence preferences in a multiple-sequence alignment generated by identifying homologues of the protein of interest in the NCBI non-redundant protein database<sup>125</sup>. Many groups have demonstrated that replacing an amino acid that is poorly represented in a multiple-sequence alignment with the most preferred amino acid at that position can often raise protein stability<sup>120,126</sup>. To reduce false positives in this type of prediction and identify mutations that are most likely to dramatically improve stability, it is possible to use information from multiple-sequence alignments along with Rosetta design simulations<sup>115</sup>. In a striking demonstration of this approach, a one-step design process was sufficient to identify 18 mutations that raise the thermal tolerance of the malaria invasion protein RH5 by over 15 °C, while retaining the ligand-binding and immunogenic properties needed for vaccine development<sup>127</sup> (FIG. 4a).

### Controlling binding specificities

Modulating the specificity of protein interactions is a powerful approach for studying and redirecting cell signalling pathways<sup>128,129</sup>. Altered interaction interfaces can also be used to assemble novel molecular assemblies. In one demonstration of this capability, multi-state design simulations were used to redesign contacts between the heavy and light chains of antibodies to create antibody variants that would no longer interact with their wild-type counterparts<sup>130–132</sup>. These redesigns allowed the creation of bispecific antibodies in which one arm of an IgG-like antibody recognizes one antigen and the other arm recognizes a separate antigen. Antibodies of this type are useful therapeutically — for example, as a form of anti-cancer immunotherapy, when bispecific antibodies can recruit a patient's T cells to their cancer cells<sup>133</sup> (FIG. 4b). Another approach to manipulate the specificity of an interaction is to expand the interaction interface to design new contact sites for regions of the target protein that are not conserved between other potential binding partners. This approach was used to design novel, specific inhibitors for each of the six members of the pro-survival BCL-2 protein family. This allowed the investigators to probe the role of each family member in pro-survival signalling in human cancer cell lines<sup>134</sup>.

### De novo interface design

To design a novel interaction between two proteins, a model must first be created with the proteins brought near each other so that their surfaces are adjacent. The computational process of bringing the two proteins near each other is frequently referred to as docking. Once the proteins are adjacent, sequence optimization simulations



**Fig. 4 | Using computational design to create proteins that have valuable applications in research and medicine. a** | Increasing protein stability. Energy calculations from protein design simulations were combined with sequence conservation information to identify 18 mutations that raise the thermostability of a malaria invasion protein by more than 15 °C, thereby improving its recombinant production for use as a vaccine immunogen<sup>127</sup>. **b** | Manipulating binding specificity. Redesign of interactions between antibody chains allowed the self-assembly of two unique light chains and two unique heavy chains into bispecific antibodies that can simultaneously bind two different antigens. These antibodies can be used for a variety of applications, including the recruitment of T cells to cancer cells as a form of immunotherapy<sup>130–132</sup>. **c** | Design of interaction interfaces. The design of an interaction between two homo-oligomers (orange trimer and blue pentamer; **step 1**) induced self-assembly of a large protein cage (**step 2**) and allowed for multivalent display of an antigen from respiratory syncytial

virus (RSV), thereby establishing a nanoparticle vaccine candidate<sup>138,142</sup> (**step 3**). This nanoparticle with the viral antigen on the surface induced neutralizing antibody responses that were ~10-fold higher than when the antigen was provided alone. **d** | De novo design of an interleukin mimic that binds to a subset of interleukin receptors, allowing the protein to maintain anti-cancer activity while reducing toxicity<sup>9</sup>. The designed protein maintains selected binding surfaces (shown in blue) in their naturally occurring orientations while embedding them in a new protein scaffold. **e** | De novo design of a protein scaffold that presents a conformational epitope from RSV. In vivo, this epitope-focused immunogen elicited antibodies that neutralize the virus, and currently is being tested as a vaccine<sup>172</sup>. **f** | De novo design strategies can also be used to design proteins optimized to bind certain ligands. For example, two custom-built backbones with different protein folds — a  $\beta$ -barrel and a helical bundle — were generated, respectively, to bind and activate a fluorescent ligand and to bind porphyrin<sup>8,158</sup>.

can be used to search for amino acids that will stabilize the interaction. Since many docked conformations may not be designable — for example, because the residues are not appropriately positioned to allow a high density of favourable interactions — it is important to consider large sets of alternatively docked complexes. One type of *de novo* interface design that has been particularly successful is the creation of symmetrical homo-oligomers. There are several inherent advantages in this design approach: favourable interactions are replicated through symmetry; the degrees of freedom to sample during docking are reduced; and obligate homo-oligomers can be stabilized primarily with hydrophobic interactions at the interface, as the proteins do not need to be soluble in the unbound form. In general, computational design methods have been more successful at designing nonpolar than polar interactions<sup>135</sup>. A striking example of symmetrical oligomer design is a generation of self-assembling nanocages<sup>136–138</sup> (FIG. 4c). These designs encompass complexes of homodimers or homotrimers with designed interfaces between the complexes to stabilize formation of the nanocages. Importantly, only a single new interface needs to be designed to induce self-assembly into a cage. Exciting applications of this technology include *de novo* cages that package RNA, similar to viral capsids<sup>139</sup>, ordered presentation of small proteins for structure determination by cryo-EM<sup>140,141</sup> and display of viral antigens in nanoparticle vaccines<sup>142</sup> (FIG. 4c). Protein cages have also been designed using approaches that do not require the creation of novel protein–protein interfaces. Similar to how DNA base pairing can be used to assemble large macromolecular complexes (DNA origami), helices that form coiled-coil dimers can be used to drive the assembly of protein cages<sup>143</sup>, and engineering rigid linkers between naturally occurring homo-oligomers can be used to favour the formation of symmetrical cages<sup>144</sup>. The design of symmetrical interactions has also been used to engineer ordered, crystal-like lattices and proteins that polymerize into 2D sheets and filaments<sup>145–148</sup>.

One-sided interface design (in which only one of the protein partners is mutated) can be used to engineer proteins that bind and regulate signalling proteins involved in disease. However, to date, computer-based one-sided interface design has proved more difficult than two-sided design. One-sided design often requires the design of more polar contacts, as the surface of the protein that is being targeted cannot be mutated and is likely to contain some polar amino acids. Successful one-sided designs have generally bound to surface-exposed hydrophobic patches on the target protein, as was the case in the design of a protein that binds and neutralizes the influenza virus<sup>149</sup>. Because one-sided interface design is so challenging, investigators have taken advantage of recent advances in DNA synthesis and yeast surface display to experimentally screen large numbers (tens of thousands) of computationally designed sequences to find tight, stable binders<sup>150</sup>.

One type of one-sided interface design that has been particularly elusive is the computer-based design of antibodies that bind tightly to defined surfaces on binding partners. This is a challenging problem because, in

addition to designing interactions across the antibody–antigen interface, it is necessary to stabilize the antibody variable loops in a conformation that is favourable for binding; these loops can access a broad range of conformations, but during naturally occurring affinity maturation in antibodies, they typically evolve to adopt unique conformations<sup>151,152</sup>. There has been encouraging recent progress in using libraries of antibody loop conformations from the PDB<sup>153</sup> to create stable antibodies with defined loop structure<sup>154</sup>, but there are no reports of *de novo* interfaces between a designed antibody and a binding partner that have been validated with a high-resolution structure.

### Scaffolding protein binding sites

For some projects involving protein–ligand or protein–protein interactions, a structural and sequence motif that is favourable for binding is already known, but it would be useful to scaffold the binding motif within a well-folded protein. Scaffolding a motif can increase binding affinity by pre-ordering it in a binding-competent conformation and can improve the stability and solubility of the protein. The design of BCL-2 protein inhibitors described above is an example of this. In another example, mimics of the natural cytokines IL-2 and IL-15 were created that display binding surfaces for interacting with only a subset of interleukin receptors<sup>9</sup> (FIG. 4d). Selected helices from IL-2 that are important for binding the desired subset of interleukin receptors were incorporated into a *de novo*-designed protein in a way that kept the helices appropriately positioned to bind their receptors. Like naturally occurring IL-2, the designed protein has anti-cancer activity when used in mouse models of melanoma and colon cancer, but is less toxic to the mice. Motif grafting has also been used to display conformational epitopes for vaccine development. This approach has shown promise in eliciting response from naive B cells towards scaffolded antigens derived from HIV and respiratory syncytial virus (RSV)<sup>155</sup> (FIG. 4e). Of note, besides the increased immunogenicity desired for the generation of vaccines, protein design can also be used to make proteins less immunogenic. As an example, design simulations have been used to search for protein mutations that will maintain stability and activity but remove sequence motifs known to be good substrates for major histocompatibility complexes (MHCs)<sup>156</sup>.

### Designed ligand binding and catalysis

Introducing new ligand and substrate binding sites into existing or *de novo*-designed proteins is of great interest, as this can be used to create new imaging reagents, catalysts and sensors (FIG. 4f). In this type of design, amino acid side chains surrounding a pocket in the protein are varied so as to form favourable contacts with the desired ligand. In many cases a particular protein structure may not have residues appropriately positioned to form strong interactions with the desired ligand. To allow tighter contacts, the binding pocket can be adjusted with small changes to the protein backbone, or alternative proteins can be considered as the starting point for design<sup>92,157</sup>. *De novo* design methods offer one strategy for creating a set of protein structures that can be computationally

#### DNA origami

A term describing approaches that use the high sequence specificity of DNA interactions to design DNA sequences that will fold into complex and predictable two- and three-dimensional shapes.

#### Major histocompatibility complexes

(MHCs). A set of cell surface proteins that bind to antigens from foreign pathogens and present them for recognition by other proteins and cells from the immune system. They are a key component of the acquired immune system.

screened for favourable binding pockets<sup>81,158</sup>. In one recent study, a set of de novo-designed  $\beta$ -barrels were used to bind a fluorescent molecule in a conformation that enhances fluorescence emission and can be used for imaging in cells<sup>8</sup>. Binding-site design can also be used to predict mutations that will lead to resistance against therapeutics<sup>159</sup>.

The de novo design of catalysts is especially challenging, as the enzyme must not only bind the substrate but also stabilize the transition state and release the final product. Computational methods have been developed for rapidly scanning a set of protein structures to find constellations of residues properly positioned to build an enzyme active site<sup>160</sup>. Once the key catalytic residues are in place, further side-chain design is performed on residues surrounding the catalytic residues (second shell) to stabilize the catalytic residues in the desired conformations. This approach has achieved modest success, but rationally designed enzymes have so far not achieved catalytic proficiencies comparable to those of naturally occurring enzymes<sup>161</sup>. It is not established whether this poor performance occurs because the current methods cannot place and/or stabilize active-site residues with enough precision (perhaps they need to be designed with sub-ångstrom accuracy, whereas many computational designs are only accurate to within 1 or 2 Å) or because efficient enzymes incorporate many other features, such as mechanisms for the substrate and product to rapidly enter and exit the binding site.

#### **Design of protein switches**

One important property of proteins is that they frequently switch between alternative conformations. This switching is integral to a variety of protein functions, including signal transduction, molecular motors and catalysis. Multi-state design simulations can be used to search for sequences that have low energies when adopting alternative backbone conformations. This approach has been used to design a small protein that can switch between alternative folds, and to design a transmembrane protein that rocks between alternative conformations to transport metals across a membrane<sup>162,163</sup>. In an extension of this strategy, multi-state design was used to introduce conformational transitions occurring at functionally relevant timescales (milliseconds) into an existing bacterial protein<sup>164</sup>. Molecular modelling can also be used to aid the design of chimeric proteins that combine naturally occurring conformational switches with functional motifs. This strategy has been used to create photoactivatable proteins for controlling oligomerization and catalysis in living cells<sup>165–167</sup>. Protein design software can also be used to inform the placement of photoactivatable chemical groups in proteins to control conformational switching<sup>168,169</sup>.

#### **Conclusions and perspective**

Tools for protein structure prediction and design have advanced considerably in the past decade, but many challenges remain. The energy functions that guide prediction and design — necessarily approximate, for reasons of computational efficiency — still struggle to

accurately balance polar and nonpolar interactions and solvation effects, particularly at interfaces. As a result, the success rates for interface-modelling applications, such as protein docking with backbone flexibility, one-sided interface design and enzyme design, remain low. Hybrid approaches that explicitly model a subset of ‘structurally important’ ordered water molecules making key interactions with the modelled protein surface represent one promising avenue forwards, but these remain challenging to parameterize due to their computational cost and the need to balance interactions with explicit waters and with bulk water being modelled implicitly. Loop-mediated interactions such as those between T cell receptors and peptide–MHC or antibodies and antigens also remain difficult to predict and engineer. Here, the challenges of interface energetics are compounded by the need to accurately model the conformational preferences of irregular polypeptide segments that may sample an ensemble of structures in the unbound state. More broadly, new approaches are needed to robustly predict and design the protein conformational flexibility and motion that are often critical for protein function. Approaches that combine molecular-dynamics trajectories with analysis of energy landscapes may be required to capture the dynamic aspects of these flexible systems.

In spite of these challenges, we believe that protein prediction and design approaches, as they continue to mature, are poised to have an increasingly important role in biology and medicine. One remarkable feature of naturally occurring proteins is that, to fulfil their role in the cell, they are often involved in multiple activities and binding reactions. For instance, Rho family GTPases catalyse the hydrolysis of GTP, undergo well-defined conformational changes, bind to multiple downstream signalling proteins and are regulated by guanine exchange factors and guanine-activating proteins. We anticipate that by combining directed-evolution approaches with advances in molecular modelling, the design of proteins with similar complexity will soon be in reach and will enable a variety of exciting applications. For instance, de novo-designed protein cages are likely to be useful in vaccine development as scaffolds for computationally designed immunogens<sup>142</sup>, and in cancer therapy as ‘smart’ drug delivery vehicles capable of integrating multiple targeting cues (for example, aberrant cell-surface protein expression or MHC-presented neo-epitopes). Engineering control mechanisms into proteins (ligand binding, light activation, enzyme activation, etc.) should allow the design of therapeutics that are active only in a particular environment and that reduce toxic side effects.

As the protein structure databases continue to grow, the availability of new sets of protein backbones and side-chain packing arrangements will increase, opening up possibilities to repurpose them to create novel binding sites and functions. It will also be exciting to see whether machine learning and pattern recognition can advance the field of protein design, just as they are doing for structure prediction.

Published online 15 August 2019

1. Jones, D. T., Singh, T., Kosciolk, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999–1006 (2015).
2. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).  
**This paper presents an accurate deep learning method that predicts residue–residue contacts by integrating 1D sequence features with 2D residue covariation and pairwise interaction features.**
3. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–75 (2017).
4. Park, H. et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
5. Heo, L. & Feig, M. Experimental accuracy in protein structure refinement via molecular dynamics simulations. *Proc. Natl Acad. Sci. USA* **115**, 13276–13281 (2018).
6. Park, H., Ovchinnikov, S., Kim, D. E., DiMaio, F. & Baker, D. Protein homology model refinement by large-scale energy optimization. *Proc. Natl Acad. Sci. USA* **115**, 3054–3059 (2018).  
**Heo et al. and Park et al. report substantial progress in refinement of protein structure models by physics-based simulations.**
7. Mravic, M. et al. Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science* **363**, 1418–1423 (2019).  
**This study reports on the design of helical membrane proteins with only apolar interactions between side chains, which demonstrates that hydrogen bonding between helices is not required for the folding and stability of membrane proteins.**
8. Dou, J. et al. De novo design of a fluorescence-activating  $\beta$ -barrel. *Nature* **561**, 485–491 (2018).  
**The first de novo design of a functional  $\beta$ -barrel protein, which reveals that symmetry breaking within the barrel is required to eliminate backbone strain and maximize hydrogen bonding between  $\beta$ -strands.**
9. Silva, D.-A. et al. De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019).
10. Chen, I.-M. A. et al. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).
11. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
12. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
13. Anson, M. L. & Mirsky, A. E. Protein coagulation and its reversal: the preparation of insoluble globin, soluble globin and heme. *J. Gen. Physiol.* **13**, 469–476 (1930).
14. Lumry, R. & Eyring, H. Conformation changes of proteins. *J. Phys. Chem.* **58**, 110–120 (1954).
15. Anfinsen, C. B., Haber, E., Sela, M. & White, F. H. Jr The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl Acad. Sci. USA* **47**, 1309–1314 (1961).
16. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
17. Anfinsen, C. B. & Scheraga, H. A. Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.* **29**, 205–300 (1975).
18. Lazaridis, T. & Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10**, 139–145 (2000).
19. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316 (2008).
20. Karplus, M. The Levinthal paradox: yesterday and today. *Fold. Des.* **2**, S69–75 (1997).
21. Levitt, M. & Warshel, A. Computer simulation of protein folding. *Nature* **253**, 694–698 (1975).
22. Levinthal, C. How to fold graciously. *Mossbauer Spectrosc. Biol. Syst.* **67**, 22–24 (1969).
23. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195 (1995).
24. Dill, K. A. Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).
25. Monticelli, L. et al. The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* **4**, 819–834 (2008).
26. Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **15**, 144–150 (2005).
27. Mairuradze, G. G., Senet, P., Czaplewski, C., Liwo, A. & Scheraga, H. A. Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field. *J. Phys. Chem. A* **114**, 4471–4485 (2010).
28. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
29. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
30. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
31. Sadreyev, R. & Grishin, N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **326**, 317–336 (2003).
32. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
33. Bowie, J. U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170 (1991).
34. Jones, D. T., Taylor, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* **358**, 86–89 (1992).
35. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
36. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Struct. Funct. Bioinf.* **77**, 778–795 (2009).
37. Webb, B. & Sali, A. Protein structure modeling with MODELLER. *Methods Mol. Biol.* **1654**, 39–54 (2017).
38. Yang, J. et al. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
39. Song, Y. et al. High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).
40. Ovchinnikov, S. et al. Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017).  
**This study shows that inclusion of sequence data from metagenomics triples the number of protein families for which accurate structural models can be built using folding simulations that incorporate covariation-derived residue–residue contact predictions.**
41. Jones, D. T. & McGuffin, L. J. Assembling novel protein folds from super-secondary structural fragments. *Proteins* **53**, 480–485 (2003).
42. 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).
43. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715–1735 (2012).
44. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
45. Jones, T. A. & Thirup, S. Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822 (1986).
46. Baeten, L. et al. Reconstruction of protein backbones from the BriX collection of canonical protein fragments. *PLoS Comput. Biol.* **4**, e1000083 (2008).
47. Byströf, C., Simons, K. T., Han, K. F. & Baker, D. Local sequence–structure correlations in proteins. *Curr. Opin. Biotechnol.* **7**, 417–421 (1996).
48. Bujnicki, J. M. Protein-structure prediction by recombination of fragments. *ChemBiochem.* **7**, 19–27 (2006).
49. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecular complexes. *Methods Enzymol.* **487**, 545–574 (2011).
50. Moutl, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Struct. Funct. Bioinf.* **23**, ii–iv (1995).
51. Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W. & Dress, A. W. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.* **17**, 164–178 (2000).
52. Fodor, A. A. & Aldrich, R. W. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**, 211–221 (2004).
53. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
54. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–301 (2011).
55. 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).
56. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
57. Nugent, T. & Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl Acad. Sci. USA* **109**, E1540–E1547 (2012).
58. Ovchinnikov, S. et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248 (2015).
59. Zhang, C., Mortuza, S. M., He, B., Wang, Y. & Zhang, Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins* **86**, 136–151 (2018).
60. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
61. Brünger, A. T. et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
62. Hopf, T. A. et al. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
63. Toth-Petroczy, A. et al. Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170.e12 (2016).
64. Cheng, J., Randall, A. Z., Sweredoski, M. J. & Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* **33**, W72–W76 (2005).
65. 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).
66. Faraggi, E., Zhang, T., Yang, Y., Kurgan, L. & Zhou, Y. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J. Comput. Chem.* **33**, 259–267 (2012).
67. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
68. Karplus, K. SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res.* **37**, W492–W497 (2009).
69. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
70. Liu, Y., Palmedo, P., Ye, Q., Berger, B. & Peng, J. Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst.* **6**, 65–74.e3 (2018).
71. Jones, D. T. & Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **34**, 3308–3315 (2018).
72. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
73. 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).
74. Woolfson, D. N. et al. De novo protein design: how do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.* **33**, 16–26 (2015).
75. Coluzza, I. Computational protein design: a review. *J. Phys. Condens. Matter* **29**, 143001 (2017).

76. Mackenzie, C. O. & Grigoryan, G. Protein structural motifs in prediction and design. *Curr. Opin. Struct. Biol.* **44**, 161–167 (2017).
77. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
78. Brunette, T. J. et al. Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).
79. Huang, P.-S. et al. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
80. Doyle, L. et al. Rational design of  $\alpha$ -helical tandem repeat proteins with closed architectures. *Nature* **528**, 585–588 (2015).
- First de novo design of repeat proteins that adopt 'doughnut'-like structures with the N and C termini adjacent in three-dimensional space.**
81. Marcos, E. et al. Principles for designing proteins with cavities formed by curved  $\beta$  sheets. *Science* **355**, 201–206 (2017).
82. Marcos, E. et al. De novo design of a non-local  $\beta$ -sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25**, 1028–1034 (2018).
83. Murphy, G. S. et al. Computational de novo design of a four-helix bundle protein—DND\_4HB. *Protein Sci.* **24**, 434–445 (2015).
84. Jacobs, T. M. et al. Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).
85. Guffy, S. L., Teets, F. D., Langlois, M. I. & Kuhlman, B. Protocols for requirement-driven protein design in the Rosetta modeling program. *J. Chem. Inf. Model.* **58**, 895–901 (2018).
86. Lin, Y.-R. et al. Control over overall shape and size in de novo designed proteins. *Proc. Natl Acad. Sci. USA* **112**, E5478–85 (2015).
87. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
88. Crick, F. H. C. The Fourier transform of a coiled-coil. *Acta Crystallogr.* **6**, 685–689 (1953).
89. Huang, P.-S. et al. High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014).
90. Lu, P. et al. Accurate computational design of multipass transmembrane proteins. *Science* **359**, 1042–1046 (2018).
91. Thomson, A. R. et al. Computational design of water-soluble  $\alpha$ -helical barrels. *Science* **346**, 485–488 (2014).
92. Tinberg, C. E. et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013).
93. Röthlisberger, D. et al. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
94. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
95. Boas, F. E. & Harbury, P. B. Potential energy functions for protein design. *Curr. Opin. Struct. Biol.* **17**, 199–204 (2007).
96. O'Meara, M. J. et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J. Chem. Theory Comput.* **11**, 609–622 (2015).
97. Gainza, P., Nisonoff, H. M. & Donald, B. R. Algorithms for protein design. *Curr. Opin. Struct. Biol.* **39**, 16–26 (2016).
98. Dunbrack, R. L. Jr Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
99. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA* **97**, 10383–10388 (2000).
100. Traoré, S. et al. Fast search algorithms for computational protein design. *J. Comput. Chem.* **37**, 1048–1058 (2016).
101. Hallen, M. A. et al. OSPREY 3.0: open-source protein redesign for you, with powerful new features. *J. Comput. Chem.* **39**, 2494–2507 (2018).
102. Lapidoth, G. et al. Highly active enzymes by automated combinatorial backbone assembly and sequence design. *Nat. Commun.* **9**, 2780 (2018).
103. Ollikainen, N., de Jong, R. M. & Kortemme, T. Coupling protein side-chain and backbone flexibility improves the re-design of protein-ligand specificity. *PLoS Comput. Biol.* **11**, e1004335 (2015).
104. Hallen, M. A. & Donald, B. R. CATS (coordinates of atoms by Taylor series): protein design with backbone flexibility in all locally feasible directions. *Bioinformatics* **33**, i5–i12 (2017).
105. Mackenzie, C. O., Zhou, J. & Grigoryan, G. Tertiary alphabet for the observable protein structural universe. *Proc. Natl Acad. Sci. USA* **113**, E7438–E7447 (2016).
106. Frappier, V., Jensen, J. M., Zhou, J., Grigoryan, G. & Keating, A. E. Tertiary structural motif sequence statistics enable facile prediction and design of peptides that bind anti-apoptotic Bfl-1 and Mcl-1. *Structure* **27**, 606–617 (2019).
- Instead of using an all-atom model of the complex to calculate interaction energies, Frappier et al. employed a knowledge-based approach with sequence preferences from structural motifs similar to the designed interface to predict binding energies.**
107. Boyken, S. E. et al. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352**, 680–687 (2016).
108. Chen, Z. et al. Programmable design of orthogonal protein heterodimers. *Nature* **565**, 106–111 (2019).
109. Maguire, J. B., Boyken, S. E., Baker, D. & Kuhlman, B. Rapid sampling of hydrogen bond networks for computational protein design. *J. Chem. Theory Comput.* **14**, 2751–2760 (2018).
110. 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).
111. Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A generic program for multistate protein design. *PLoS ONE* **6**, e20937 (2011).
112. Negron, C. & Keating, A. E. Multistate protein design using CLEVER and CLASSY. *Methods Enzymol.* **523**, 171–190 (2013).
113. Allen, B. D. & Mayo, S. L. An efficient algorithm for multistate protein design based on FASTER. *J. Comput. Chem.* **31**, 904–916 (2010).
114. Löffler, P., Schmitz, S., Hupfeld, E., Sterner, R. & Merkl, R. Rosetta:MSF: a modular framework for multi-state computational protein design. *PLoS Comput. Biol.* **13**, e1005600 (2017).
115. Goldenzweig, A. et al. Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell* **63**, 337–346 (2016).
- This study uses protein design simulations coupled with sequence conservation information to create an effective protocol for identifying sets of mutations that increase protein thermostability and expression.**
116. Rocklin, G. J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
117. Gainza-Ciraucui, P. & Correira, B. E. Computational protein design — the next generation tool to expand synthetic biology applications. *Curr. Opin. Biotechnol.* **52**, 145–152 (2018).
118. Wrenbeck, E. E., Faber, M. S. & Whitehead, T. A. Deep sequencing methods for protein engineering and design. *Curr. Opin. Struct. Biol.* **45**, 36–44 (2017).
119. Malakauskas, S. M. & Mayo, S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**, 470–475 (1998).
120. Magliery, T. J. Protein stability: computation, sequence statistics, and new experimental methods. *Curr. Opin. Struct. Biol.* **33**, 161–168 (2015).
121. Goldenzweig, A. & Fleishman, S. J. Principles of protein stability and their application in computational design. *Annu. Rev. Biochem.* **87**, 105–129 (2018).
122. Borgo, B. & Havranek, J. J. Automated selection of stabilizing mutations in designed and natural proteins. *Proc. Natl Acad. Sci. USA* **109**, 1494–1499 (2012).
123. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460 (2003).
124. 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).
125. Bednar, D. et al. FireProt: energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.* **11**, e1004556 (2015).
126. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta* **1543**, 408–415 (2000).
127. Campeotto, I. et al. One-step design of a stable variant of the malaria invasion protein RH5 for use as a vaccine immunogen. *Proc. Natl Acad. Sci. USA* **114**, 998–1002 (2017).
128. Kapp, G. T. et al. Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. *Proc. Natl Acad. Sci. USA* **109**, 5277–5282 (2012).
129. Jensen, J. M., Ryan, J. A., Grant, R. A., Letai, A. & Keating, A. E. Epistatic mutations in PUMA BH3 drive an alternate binding mode to potentially and selectively inhibit anti-apoptotic Bfl-1. *eLife* **6**, e25541 (2017).
130. Froning, K. J. et al. Computational design of a specific heavy chain/ $\kappa$  light chain interface for expressing fully IgG bispecific antibodies. *Protein Sci.* **26**, 2021–2038 (2017).
131. Leaver-Fay, A. et al. Computationally designed bispecific antibodies using negative state repertoires. *Structure* **24**, 641–651 (2016).
132. Lewis, S. M. et al. Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface. *Nat. Biotechnol.* **32**, 191–198 (2014).
- In this study, multi-state design simulations are used to create altered specificity interactions between antibody constant domains, allowing the proper assembly of IgG antibodies that recognize two separate antigens simultaneously.**
133. Krishnamurthy, A. & Jimeno, A. Bispecific antibodies for cancer therapy: a review. *Pharmacol. Ther.* **185**, 122–134 (2018).
134. Berger, S. et al. Computationally designed high specificity inhibitors delineate the roles of BCL2 family proteins in cancer. *eLife* **5**, e20352 (2016).
135. 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).
136. King, N. P. et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171–1174 (2012).
137. King, N. P. et al. Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–108 (2014).
138. Bale, J. B. et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389–394 (2016).
- One of several papers in which this team demonstrate that protein interface design combined with modelling of higher-order symmetries can be used to create large, multi-component protein cages.**
139. Butterfield, G. L. et al. Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).
140. Liu, Y., Gonen, S., Gonen, T. & Yeates, T. O. Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. *Proc. Natl Acad. Sci. USA* **115**, 3362–3367 (2018).
141. Liu, Y., Huynh, D. T. & Yeates, T. O. A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold. *Nat. Commun.* **10**, 1864 (2019).
142. Marcandalli, J. et al. Induction of potent neutralizing antibody responses by a designed protein nanoparticle vaccine for respiratory syncytial virus. *Cell* **176**, 1420–1431 (2019).
143. Ljubetic, A. et al. Design of coiled-coil protein-origami cages that self-assemble in vitro and in vivo. *Nat. Biotechnol.* **35**, 1094–1101 (2017).
144. Lai, Y.-T., Cascio, D. & Yeates, T. O. Structure of a 16-nm cage designed by using protein oligomers. *Science* **336**, 1129 (2012).
145. Shen, H. et al. De novo design of self-assembling helical protein filaments. *Science* **362**, 705–709 (2018).
146. 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).
147. Zhang, H. V. et al. Computationally designed peptides for self-assembly of nanostructured lattices. *Sci. Adv.* **2**, e1600307 (2016).
148. Tian, Y. et al. Nanotubes, plates, and needles: pathway-dependent self-assembly of computationally designed peptides. *Biomacromolecules* **19**, 4286–4298 (2018).
149. Fleishman, S. J. et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**, 816–821 (2011).
150. Chevalier, A. et al. Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
151. Adolf-Bryfogle, J. et al. RosettaAntibodyDesign (RABD): a general framework for computational antibody design. *PLoS Comput. Biol.* **14**, e1006112 (2018).

152. Kundert, K. & Kortemme, T. Computational design of structured loops for new protein functions. *Biol. Chem.* **400**, 275–288 (2019).
153. Adolf-Bryfogle, J., Xu, Q., North, B., Lehmann, A. & Dunbrack, R. L. Jr PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Res.* **43**, D432–D438 (2015).
154. Baran, D. et al. Principles for computational design of binding antibodies. *Proc. Natl Acad. Sci. USA* **114**, 10900–10905 (2017).
155. Kulp, D. W. & Schief, W. R. Advances in structure-based vaccine design. *Curr. Opin. Virol.* **3**, 322–331 (2013).
156. Salvat, R. S. et al. Computationally optimized deimmunization libraries yield highly mutated enzymes with low immunogenicity and enhanced activity. *Proc. Natl Acad. Sci. USA* **114**, E5085–E5093 (2017).
157. Bick, M. J. et al. Computational design of environmental sensors for the potent opioid fentanyl. *eLife* **6**, e28909 (2017).
158. Polizzi, N. F. et al. De novo design of a hyperstable non-natural protein-ligand complex with sub-Å accuracy. *Nat. Chem.* **9**, 1157–1164 (2017).
159. Reeve, S. M. et al. Protein design algorithms predict viable resistance to an experimental antifolate. *Proc. Natl Acad. Sci. USA* **112**, 749–754 (2015).
160. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational enzyme design. *Angew. Chem. Int. Ed. Engl.* **52**, 5700–5725 (2013).
161. Baker, D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* **19**, 1817–1819 (2010).
162. Ambroggio, X. I. & Kuhlman, B. Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J. Am. Chem. Soc.* **128**, 1154–1161 (2006).
163. Joh, N. H. et al. De novo design of a transmembrane Zn<sup>2+</sup>-transporting four-helix bundle. *Science* **346**, 1520–1524 (2014).
164. Davey, J. A., Damry, A. M., Goto, N. K. & Chica, R. A. Rational design of proteins that exchange on functional timescales. *Nat. Chem. Biol.* **13**, 1280–1285 (2017).
165. Guntas, G. et al. Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins. *Proc. Natl Acad. Sci. USA* **112**, 112–117 (2015).
166. Dagliyan, O. et al. Engineering extrinsic disorder to control protein activity in living cells. *Science* **354**, 1441–1444 (2016).
167. Dagliyan, O. et al. Computational design of chemogenetic and optogenetic split proteins. *Nat. Commun.* **9**, 4042 (2018).
168. Blacklock, K. M., Yachnin, B. J., Woolley, G. A. & Khare, S. D. Computational design of a photocontrolled cytosine deaminase. *J. Am. Chem. Soc.* **140**, 14–17 (2017).
169. Hoersch, D., Roh, S.-H., Chiu, W. & Kortemme, T. Reprogramming an ATP-driven protein machine into a light-gated nanocage. *Nat. Nanotechnol.* **8**, 928–932 (2013).
170. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
171. Dror, R. O. et al. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* **348**, 1361–1365 (2015).
172. Correia, B. E. et al. Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201–206 (2014).

**Correia et al. used a de novo protein design to generate a small protein that mimics a conformational epitope from RSV and elicits neutralizing antibodies in animal studies.**

### Acknowledgements

The authors apologize to the scientists whose important work could not be cited in this Review owing to space constraints. This work was supported by NIH grants R01GM117968 and R55GM131923 to B.K. and R01GM121487 and R01GM123378 to P.B.

### Author contributions

Both authors contributed equally to all aspects of the article.

### Competing interests

The authors declare no competing interests.

### Peer review information

*Nature Reviews Molecular Cell Biology* thanks W. DeGrado, T. O. Yeates and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

### Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Supplementary information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41580-019-0163-x>.

### RELATED LINKS

CASP protein structure prediction experiment: <http://predictioncenter.org/casp13>  
 DeepContact: <https://github.com/largelymfs/deepcontact>  
 DeepCov: <https://github.com/psipred/DeepCov>  
 EVfold: <http://evfold.org/evfold-web/evfold.do>  
 RaptorX-Contact: <http://raptorex.uchicago.edu/ContactMap/>  
 TripletRes: <https://zhanglab.cccb.med.umich.edu/TripletRes/>  
 OSPREY: <https://www2.cs.duke.edu/donaldlab/osprey.php>  
 PSIPRED: <http://bioinf.cs.ucl.ac.uk/psipred/>  
 Rosetta: <https://www.rosettacommons.org/>